Fine Tuning Text-to-Image Diffusion Models

Using

DreamBooth

Project report in partial fulfillment of the requirement for the award of the degree of

Bachelor of Technology

In

Computer Science And Technology

Submitted By

Yash Shankar Tiwary	University Roll No.12019009022091
Sushant Kumar Suman	University Roll No. 12019009022213
Malyaban Ganguly	University Roll No. 12019009022051
Subimal Sahu	University Roll No. 12019009001190

Under the guidance of

Prof. Nilanjan Chatterjee

&

Prof. Arnab Chatterjee

Department of Computer Science and Technology

UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA

University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.

CERTIFICATE

This is to certify that the project titled **Fine Tuning Text-to-Image Diffusion Models Using DreamBooth**

submitted by **Yash Shankar Tiwary(University Roll No. 12019009022091), Sushant Kumar Suman (University Roll No. 12019009022213)**, **Malyaban Ganguly(University Roll No. 12019009022051) and Subimal Sahu**(**University Roll No.** 12019009001190**)** students of UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA, in partial fulfillment of requirement for the degree of Bachelor of Computer Science and Technology, is a bonafide work carried out by them under the supervision and guidance of Prof. Nilanjan Chatterjee & Prof. Arnab Chatterjee during 8th Semester of academic session of 2022 - 2023. The content of this report has not been submitted to any other university or institute. I am glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

Signature of Guide

Signature of Guide

Signature of Head of the Department

ACKNOWLEDGEMENT

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Nilanjan Chatterjee and Prof. Arnab Chatterjee of the Department of Computer Science, UEM, Kolkata, for his wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

Yash Shankar Tiwary

Sushant Kumar Suman

Malyaban Ganguly

Subimal Sahu

TABLE OF CONTENTS

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/

CHAPTER – 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. arXiv:2208.12242v2 [cs.CV] 15 Mar 2023 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

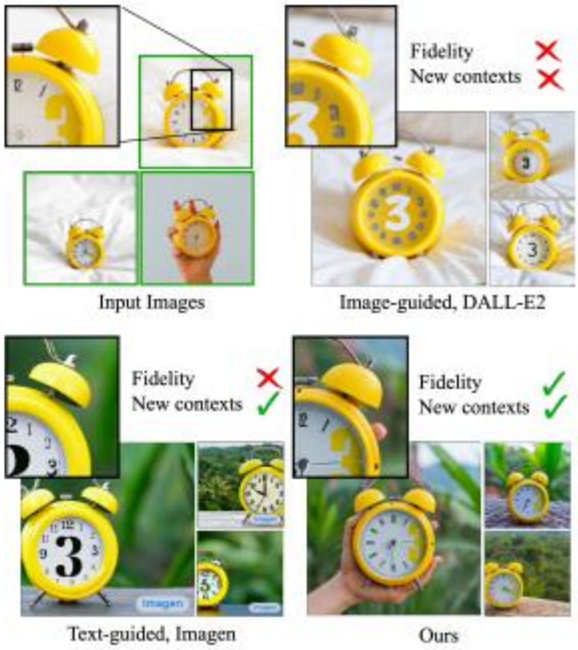

Figure 2. **Subject-driven generation.** Given a particular clock (left), it is hard to generate it while maintaining high fidelity to its key visual features (second and third columns showing DALL-E2 [54] image-guided generation and Imagen [61] text-guided generation; text prompt used for Imagen: "*retro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

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.

CHAPTER – 2: LITERATURE SURVEY

2.1 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.

2.2 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

2.3 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 identitypreserving 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.

CHAPTER – 3: PROBLEM STATEMENT

Image generation is an exciting field of research that has seen tremendous progress in recent years, thanks to advances in deep learning and computer vision. One of the most popular approaches to generating images is through the use of generative adversarial networks (GANs), which have shown impressive results in generating realistic and diverse images.

DreamBooth is a project that aims to make the process of generating images using GANs more accessible to the general public. It is a web-based application that allows users to upload an image and generate new images based on that input. The user can then select the images they like and continue the generation process, creating a virtually endless stream of unique and creative images.

The problem with image generation using GANs is that it can be computationally intensive, requiring powerful hardware and specialized software. DreamBooth aims to overcome this challenge by providing a user-friendly interface that abstracts away the technical details, making it easier for anyone to create high-quality images.

Another challenge with image generation using GANs is that the generated images may not always be of high quality or may not match the user's expectations. DreamBooth addresses this by allowing users to fine-tune the generated images using sliders that control various parameters, such as brightness, contrast, and saturation.

In summary, the problem statement for image generation using DreamBooth is to provide a user-friendly and accessible platform that allows anyone to generate high-quality and creative images using GANs. This involves overcoming the technical challenges associated with GANs and providing tools that allow users to fine-tune the generated images to match their preferences. With DreamBooth, we hope to democratize image generation and inspire creativity among people from all walks of life.

CHAPTER – 4: PROPOSED SOLUTION

Given only a few 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 textto-image diffusion models, then present our finetuning technique to bind a unique identifier with a subject described in a few images , and finally propose a class-specific prior-preservation loss that enables us to overcome language drift in our fine-tuned model.

4.1. 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-toimage diffusion model xˆθ that, given an initial noise map e∼ N (0, I) and a conditioning vector c = Γ(P) generated using a text encoder Γ and a text prompt P, generates an image xgen = xˆθ(e, c). They are trained using a squared error loss to denoise a variably-noised image or latent code zt := αtx + σte as follows:

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

where x is the ground-truth image, c is a conditioning vector (e.g., obtained from a text prompt), and αt, σt, wt are terms that control the noise schedule and sample quality, and are functions of the diffusion process time t ∼ U([0, 1]). A more detailed description is given in the supplementary material.

4.2. 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 finetuning 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 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.

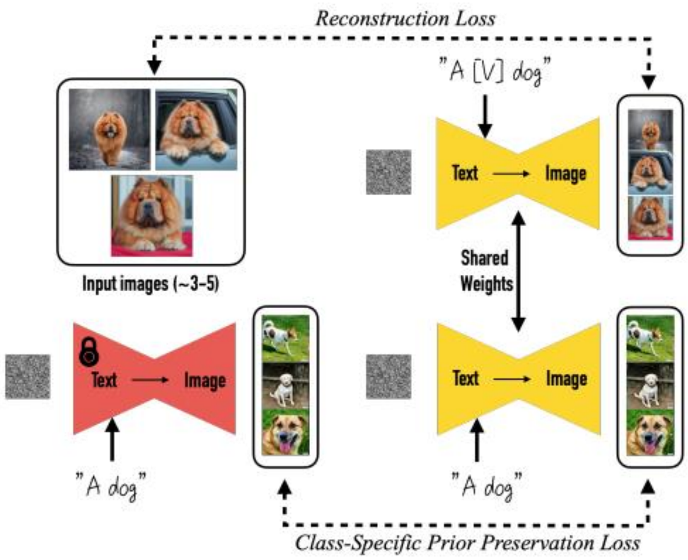

Figure 3. **Fine-tuning.** Given ∼ 3 − 5 images of a subject we finetune a text-to-image diffusion model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., "A [V] dog"), in parallel, we apply a class-specific prior preservation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

4.3 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.

CHAPTER – 5 : EXPERIMENTAL SETUP AND RESULT ANALYSIS

5.1 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.

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.

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. We also collected 25 prompts: 20 recontextualization prompts and 5 property modification prompts for objects; 10 recontextualization, 10

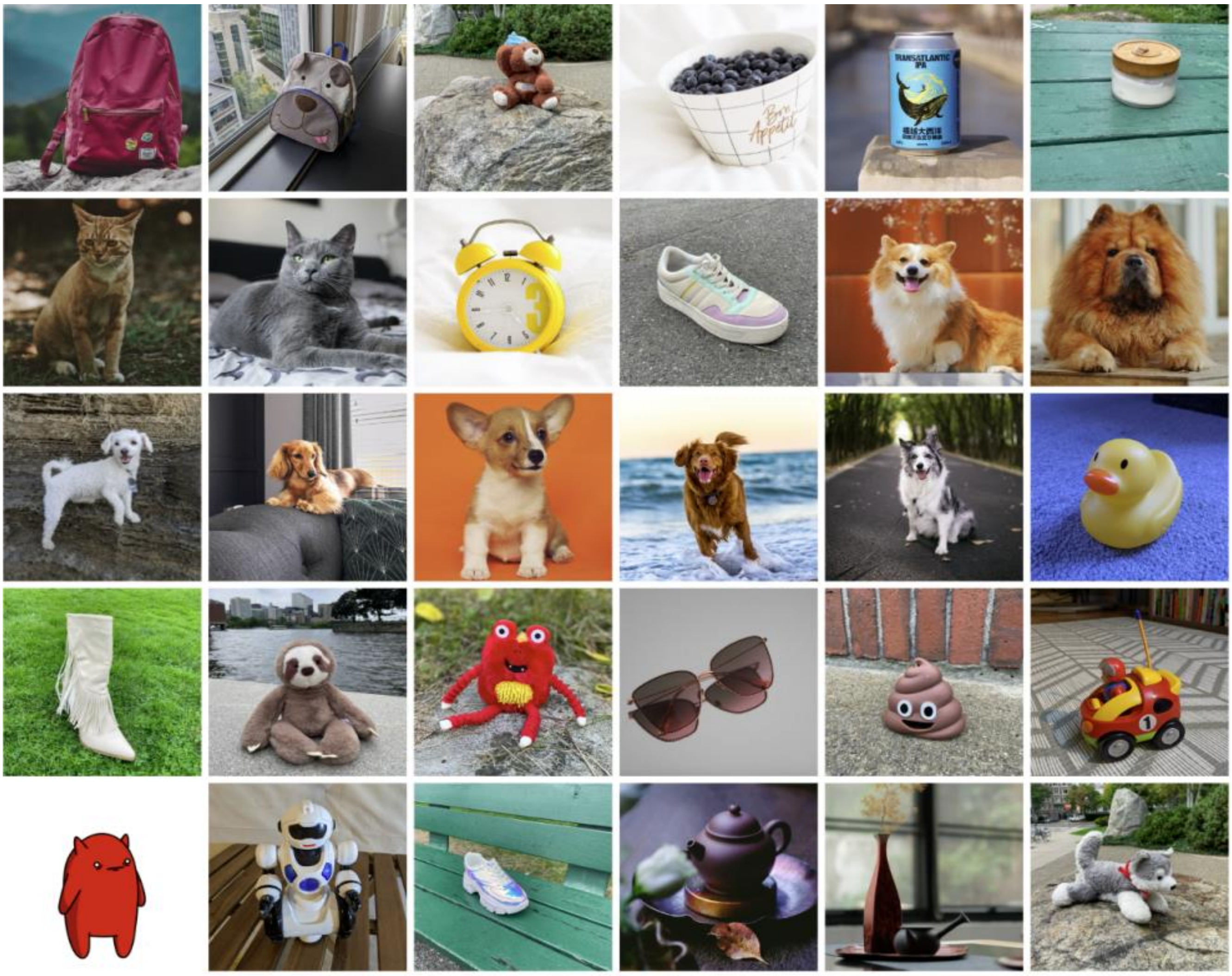

Figure 5. **Dataset**. Example images for each subject in our proposed dataset.

5.2 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 ViTS/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.

CHAPTER – 6 : CONCLUSION & FUTURE SCOPE

6.1 FUTURE SCOPE

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.

6.2 CONCLUSION

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.

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, Herve J ´ egou, ´ 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. Instanceconditioned 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: Scenebased 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 CohenOr. An image is worth one word: Personalizing text-toimage 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 Hark ¨ onen, ¨ 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