Improving Few-Shot Image Generation using Multi-Subject Decomposition in DreamBooth

Shane Davis, Joe Fioresi, Mitchell Klingler, Nyle Siddiqui
Center for Research in Computer Vision, University of Central Florida, Orlando, USA
{sdavis175, joefioresi718, mitchell.klingler99, nylesiddiqui}@knights.ucf.edu

Abstract

Text-to-image generation has recently been dominated by the advent of denoising diffusion probabilistic models (DDPM). Given a text prompt, or any conditioning criterion, these models are capable of generating novel, photo-realistic, and high-fidelity images starting from complete Gaussian noise. DreamBooth [6] introduced a novel method of fine-tuning DDPMs for personal use, wherein a pre-trained diffusion model can be fine-tuned on a new subject and generate images of the subject with different backgrounds, poses, appearances, etc. However, this method is limited in the sense that a single model cannot be fine-tuned on multiple subjects; each new subject would require its own model. Furthermore, DreamBooth requires a sequential process of fine-tuning such that even if it was capable of fine-tuning on multiple subjects, it would be impossible to do simultaneously. Therefore, we propose an improvement on the DreamBooth technique that not only allows a single model to be fine-tuned on multiple subjects, but is able to do so simultaneously. We show through various qualitative examples that our method is not only capable of fine-tuning on multiple subjects at the same time, but retains the semantic structure of each subject and is even able to generate new subjects in the same image together. We achieve these results by proposing two new losses to the DreamBooth process in addition to improving upon the training process.

1. Introduction

Text to image synthesis has recently been becoming popular with state-of-the-art results by using diffusion models such as Stable Diffusion [19] and Imagen [21]. However, they only handle generic classes and cannot generate images of custom or novel classes that did not appear during training. To solve this problem, fine-tuning methods such as DreamBooth [20] have introduced a way to train a pre-existing diffusion model on novel subjects, who can then be generated in an image with the flexibility of a DDPM.

Figure 1. Our model is capable of learning multiple subjects simultaneously, maintaining the ability to realistically generate them in variable text-guided contexts.

They achieve this by associating a unique token identifier with a novel subject, which is covered more in Section 2.

The resulting fine-tuned model can then be used to generate photorealistic images using the unique token identifier to directly reference a fine-tuned subject in the text prompt. However, DreamBooth's fine-tuning process is only capable of training and fine-tuning on a single subject that the user provides. If a user wishes to generate images of other unique subjects, each subject requires their own separate model to be fine-tuned for image generation

We expand on the DreamBooth process by enabling a single model to fine-tune on and generate novel images of multiple subjects. We attain this by assigning and training a unique token identifier for each custom subject the user provides. The naive approach would be take a diffusion model $\hat{\mathbf{x}}_\theta$ with parameters θ that generates images $\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\theta(\epsilon, \mathbf{c}, \mathbf{t})$ given some Gaussian noise, ϵ, conditioning text prompt $\mathbf{c}$, and timestep $\mathbf{t}$. Then, the model would be fine-tuned using the DreamBooth process on a custom token $[V_1]$ and custom subject s_1. The same process would need to be repeated for another custom subject s_2 with a new corresponding custom token $[V_2]$. This fine-tuning technique necessitates a sequential training process resulting in an inefficient linear time scale cost, since each individual subject requires a repetition of the fine-tuning process. More importantly, the results in Figure 2 show that this naive approach struggles to generate a realistic image of two fine-tuned subjects next to each other. Furthermore, the results in Figure 3 show that the model's ability to generate an individual subject by itself degrades when fine-tuned on multiple subjects. Thus, the model forgets the individual subject's semantic structure during fine-tuning of other subjects.

To solve this, we first implemented a custom training scheme (covered more in Section 3.1) which allows the model to fine-tune on all subjects simultaneously. We build upon HuggingFace's pre-existing implementation of DreamBooth [6] and our code can be found here: GitHub link. In addition to changing the training scheme, we implemented two custom losses: a Text Space Loss designed to disentangle the unique identifier tokens for each subject (covered more in Section 3.3), and an Image Space Loss designed to enforce image-wise similarity between different instances of the same subject (covered more in Section 3.4). Moreover, DreamBooth has been noted to have problems with over-fitting on the user's images which requires sensitive fine-tuning of the hyperparameters for each specific subject. Since we are training on multiple subjects that may require different amounts of training steps, we propose an individual stopping method (covered more in Section 3.2) to allow for early-stopping during training on subjects that tend to overfit faster than others. Much better quality images are generated with our novel approach compared to the original DreamBooth technique, while only requiring a single generalizable model. Our results are shown in Section 4, with some limitations detailed in Section 5.

Figure 2. An example output from the standard DreamBooth fine-tuning technique after training on both the dog and the rubbery duck toy as new subjects. Given the prompt "a sks dog next to a zwx toy", we see that instead of separating the two learned subjects, the model blends them into a single object.

(a) A sks dog (b) A zwx cat

Figure 3. Two noisy sampled images after applying the naive training multi-subject training method.

2. Related Work

2.1. Text Guided Diffusion

Recently, many DDPMs have been proposed that improve upon the diffusion process to generate higher quality images driven by some kind of textual input [3, 10, 18]. These models take in a text describing a scene, and are able to output imagery that is remarkably realistic and descriptive of the input text. Imagen [21] has been one such of these models that is able to produce photo-realistic imagery in any general context. Text-guided diffusion was largely motivated by the recent development of combining textual encodings with imagery, such as CLIP [17]. Another such proposal is GLIDE [16]. This model utilizes classifier free guidance [12] in an attempt to produce more photorealistic imagery. Classifier guidance works by training a single model to guide a DDPM via a parameterized score estimator. This allows the model to learn the latent space much like it would in CLIP. This combined latent representation of text and image paired with denoising diffusion proba-

bilistic models have allowed for the programmatic production of novel, realistic imagery, as well as adaptation to new and unseen scenarios. DDPMs have also expanded diffusion to a plethora of other domains, such as action generation [25], person synthesis [4], video generation [9], object detection [2, 5], and more [1, 11, 15, 22].

2.2. Fine-Tuning DDPMs

Single Subject Despite the aforementioned promising results via DDPMs, there are still a large gaps in the problem domain of image generation, such as generating/fine-tuning on human faces individuals in any novel environment or situation. DreamBooth [20] has taken this idea and developed a fine-tuning process that maintains the model's knowledge of previously learned classes and their structures, but at the same time learns the fine details of this instance we want to recreate. The proposal with this methodology is to use a rare token-identifier that is unique enough such that it becomes associated with the novel subject, while still maintaining knowledge of prior classes.

Other research has also been conducted into this single subject image generation modality. Unified Multi-Modal Diffusion [14] has also emerged to produce photo-realistic imagery of a subject instance via embedding this new instance into the latent space, and reprojecting the instance into the image space. This technique has some novelty, however the model will lose all prior knowledge of the subject class and can degrade when producing unexplored features in the output domain, untrained on the instance in the dataset.

Multiple Subjects Tangentially being developed alongside of our work, there have also been some techniques attempting to approach multiple subject instances produced from a single model, usually two. These techniques all have their own merits, use cases, and methodologies of production. Some of which have utilized modifying the latent representations of the text to achieve this, and others have modified the input imagery. Looking first at textual modifications Training-Free Structured Diffusion [7] has seen some success generating multiple classes into photorealistic imagery. They do this modifying the latent space to better break apart subjects via cross attention layers. However this does not combat the issue of specific subject instances, only production of multiple generics. Building on this work, SVDiff [8] utilizes the same idea of cross-attention maps, but extends the DreamBooth [20] training process to contain some input image modifications called cut-mix-unmix. This process entails passing in split imagery of subjects with different token identifiers to be learned in the textual latent space, and the DDPM. This is more so along the lines of what we desire, but is not extendable to n subjects. Expanding on this idea, concurrent work to this is Custom Diffusion [13], which trains instances subjects in parallel, or even combines already trained models on instances.

3. Methodology

We devise a comprehensive, flexible training framework that allows for a user to personalize a pretrained text-to-image diffusion model, $\hat{\mathbf{x}}_\theta$, by injecting any arbitrary number of novel subjects into the model's learned representations. Given only a few images of each subject, we train the model to generate novel text-guided images of each subject with high fidelity. In some cases, multiple learned subjects can be generated in the same image. Figure 4 shows a schematic overview of our fine-tuning process. Section 3.1 explains how we prepare the training data to fine-tune on multiple subjects simultaneously. Section 3.2 shows how we mitigate individual subject overfitting. Finally, Section 3.3 and Section 3.4 describe the novel loss functions we propose to guide the multi-subject fine-tuning process.

Fine-Tuning Text-to-Image Diffusion Models Given a pre-trained text-to-image diffusion model $\hat{\mathbf{x}}_\theta$, we can fine tune on a ground truth subject image $\mathbf{x}$ (randomly selected from a set of 3-5 images of the same subject). To achieve this, we can use the standard diffusion loss equation:

$$\mathcal{L}_{simple} = \omega_t ||\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x}||_2^2 \qquad (1)$$

where $\mathbf{c}$ is a text prompt conditioning vector, ϵ is the sampled noise at a timestep t, and $\alpha_t, \sigma_t, \omega_t$ are additional noise scheduling and sample quality functions [20]. Using this equation for a sufficient number of iterations yields a fine-tuned text-to-image diffusion model $\hat{\mathbf{x}}'_\theta$ that is capable of generating images featuring the subject of the ground truth images.

Personalized Diffusion Objective A fine-tuned model using the above loss as the main training objective tends to exhibit language drift on the training examples, forgetting how to generate other examples belonging to the same class. The original DreamBooth [20] paper introduced an additional loss, denoted as prior preservation loss (Eq. 2), to the fine-tuning process that prevents this undesired language drift. Specifically, this loss is identical to Eq. 1, except that instead of comparing $\mathbf{x}_{\text{gen}}$ to $\mathbf{x}$, we generate coarse class images from the fine-tuned model using text prompt $\mathbf{c}_{\text{pr}}$ and compare them to the output from a frozen pretrained model $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}_\theta(\epsilon, \mathbf{c}_{\text{pr}}, \mathbf{t})$ with the same prompt. To ensure that fine-tuning preserves prior class knowledge, this loss function is combined with Eq. 1.

$$\mathcal{L}_{ppr} = w_{t'} ||\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{pr} + \sigma_{t'} \epsilon^{'}, \mathbf{c}_{pr}) - \mathbf{x}_{pr}||_2^2 \qquad (2)$$

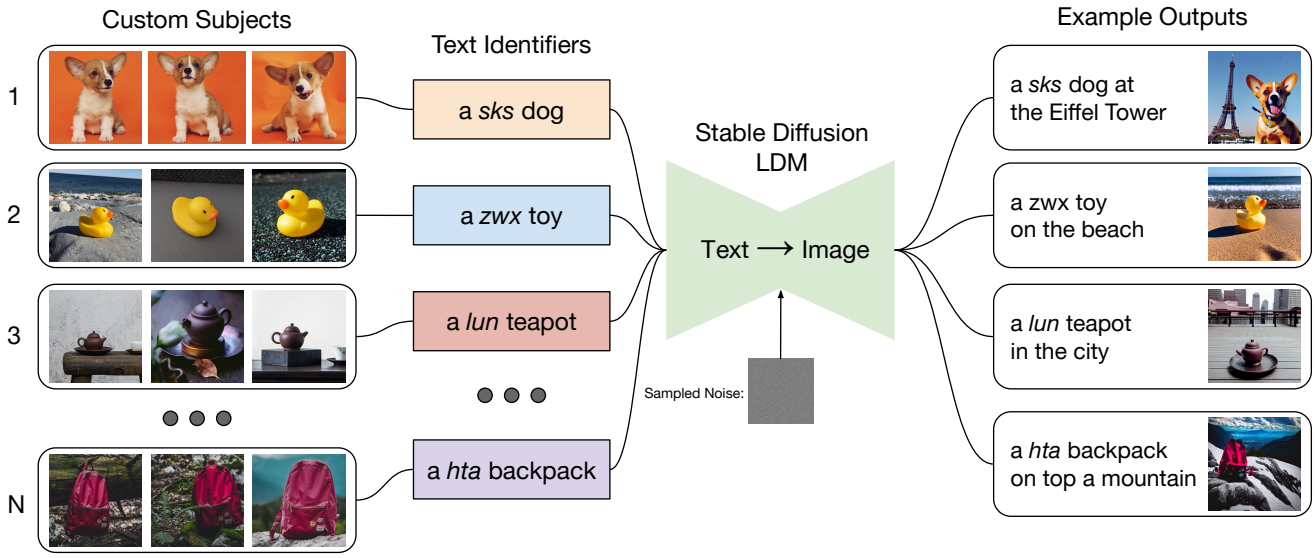


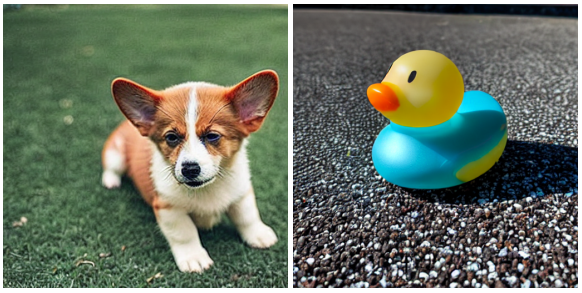
Figure 4. Given any number of custom subjects and unique text identifiers, our fine-tuned model is able to learn each one simultaneously, maintaining object fidelity when generating them in novel settings.

3.1. Dataset Loader

To enable the simultaneous training of multiple subjects, as opposed to sequentially in [13], we needed to define a directory structure and data collection paradigm. Each instance would have its own prompt, the associated imagery, and ample data for prior preservation (Eq. 2). It becomes trivial to associate all of the data and pass them with dependence to the model. Batching can now be done such that we sample a few of the novel subjects with their text prompts and pair them with each of their prior preservation class examples. With choosing a subject arbitrarily, we can ensure that good distribution of the subjects will be chosen and sampled enough to achieve a low enough loss when taking the cosine similarity between it and another randomly chosen subject. Because the loss is cosine, if the two subjects are identical, the loss will not be contributed and it will simply move on.

3.2. Individual Stopping

With DreamBooth, a common problem is that the fine-tuning process tends to over-fit on certain subjects if there are too many training steps or the learning rate is too high. This is due to the few training images that are used in the fine-tuning process. The amount of training steps and learning rate is heavily dependent on both the class of the subject and the number of training images. Current guidelines [24] suggest to use 800-1200 training steps with a learning rate of 1e-6, starting from a lower training step and gradually increasing. It is noted that generic objects typically take less training steps than more complex and detailed classes, such as humans. However, with our multi-subject approach, the class and complexity of subjects can vary, impacting the number of training steps required per subject. Therefore, we risk over-fitting on some individual subjects to continue training on other subjects, which can be seen in Figure 5.



(a) A photo of sks dog (b) A photo of zwx toy

Figure 5. This is a sample of two validation images taken from training on two subjects, a dog and a toy, at 600/1000 training steps. On the left, the model has fully learned our given dog images and is able to generate realistic outputs. However on the right, the model has still not learned the toy and is not able to generate realistic outputs.

To mitigate this effect, we implement an early-stopping method dependent on the class and complexity of each subject. The step at which the subjects will be stopped training on must be defined by the user. Once the specified step is reached, the subject will no longer be accessible to be sampled by the data loader, therefore stopping the training on the specific subject.

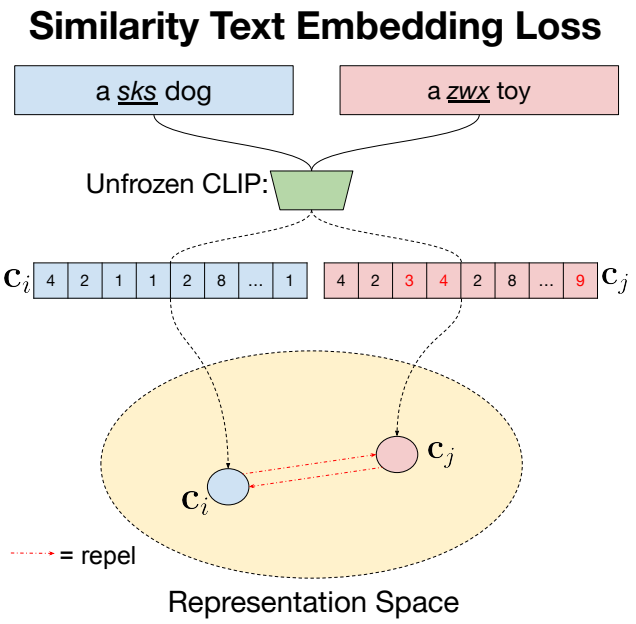


Figure 6. The unique tokens are passed through the CLIP text encoder, producing latent representations $\mathbf{c}_i$ and $\mathbf{c}_j$, which are repelled from each other during training.

3.3. Text Space Loss

In our initial experiments with multi-subject image generation, we observed that generating images with multiple objects often resulted in features of the learned objects blending together, producing undesirable outputs such as Figure 2. This is likely caused by the model learning both subjects simultaneously without properly learning how to discriminate between subjects. Other works have shown similar findings, noticing that the CLIP text encoder does not perform well at multi-object decomposition [7, 14, 18, 21]. While the model performs well on each individual object, at test time the diffusion model may receive the CLIP text embeddings of both unique subjects simultaneously. This presents an unseen situation, and by default the model is not well equipped to handle this.

To address this issue, we first unfreeze the CLIP text encoder. This adds computational complexity, but allows for the model to better learn these new unique prompts. Additionally, we introduce a cosine similarity loss shown in Figure 6 that compares the token embeddings of the unique identifiers. Specifically, given token embeddings $\mathbf{c}_1$ and $\mathbf{c}_2$, we compute the cosine similarity (Eq. 3), then take the complement to finally compute the dissimilarity between the embeddings in Eq. 4.

$$cos(\mathbf{c}_1, \mathbf{c}_2) = \frac{\mathbf{c}_1 \cdot \mathbf{c}_2}{||\mathbf{c}_1||_2 \cdot ||\mathbf{c}_2||_2} \tag{3}$$

$$\mathcal{L}_{text} = \frac{1}{N} \sum_{i=1, i \neq j}^{N} 1 - cos(\mathbf{c}_i, \mathbf{c}_j) \tag{4}$$

where N is the number of subject instances in a single batch.

The loss value ranges from 0 to 1, where a value of 0 indicates identical embeddings and a value of 1 indicates maximum dissimilarity. By subtracting this loss from the overall loss objective, we encourage the model to push the embeddings of unique subject identifier tokens apart, facilitating better discrimination between the different subjects.

3.4. Image Space Loss

In addition to the losses enforced in the text space, we hypothesize that similarly implementing losses in the image space will further encourage the model to learn and retain the visual structure for multiple subjects. Specifically, we apply a loss directly on the latent space produced by the stable diffusion model [19]. It is noted in [19] that one of the many benefits of latent diffusion is the encoding of only the high-level information in the latent space. Diffusing on every pixel in the image space is redundant and computationally expensive, thus diffusing on a latent space in which imperceptible details are not retained is more efficient. More importantly, the losses we impose on this latent space will then only affect the high-level structure of the generated image, as opposed to impacting the entire image at the pixel-level.

The loss we implement in the latent space is meant to aid the model in retaining its ability to generate the previous subjects it has been fine-tuned on, while still learning new subjects. As discussed in Section 3.1, each training batch contains a random selection of subjects to fine tune on, with each subject having one or more instances in a single batch. To this end, we enforce the latent features produced by the model to be similar when different instances of the same fine-tuned subject are being produced. The latent features produced for one instance of a subject, denoted f_i, and the latent features produced for another instance of the same subject, denoted f_j, have their distance minimized using an MSE loss:

$$\mathcal{L}_{image} = \frac{1}{N} \sum_{i=1, i \neq j}^{N} (f_i - f_j)^2 \tag{5}$$

where N is the number of instances of that subject in a single batch. Therefore, the high-level structure of a specific fine-tuned subject is learned to be the same across different instances, assisting our model in learning and retaining better representations. A similar loss formulation enforcing dissimilarity across different subjects was also explored, but yielded poor results and is discussed in Section 5.

Our overall training loss objective is described as follows:

$$\mathcal{L} = \mathcal{L}_{simple} + \lambda_1 \mathcal{L}_{ppr} - \lambda_2 \mathcal{L}_{text} + \lambda_3 \mathcal{L}_{image} \quad (6)$$

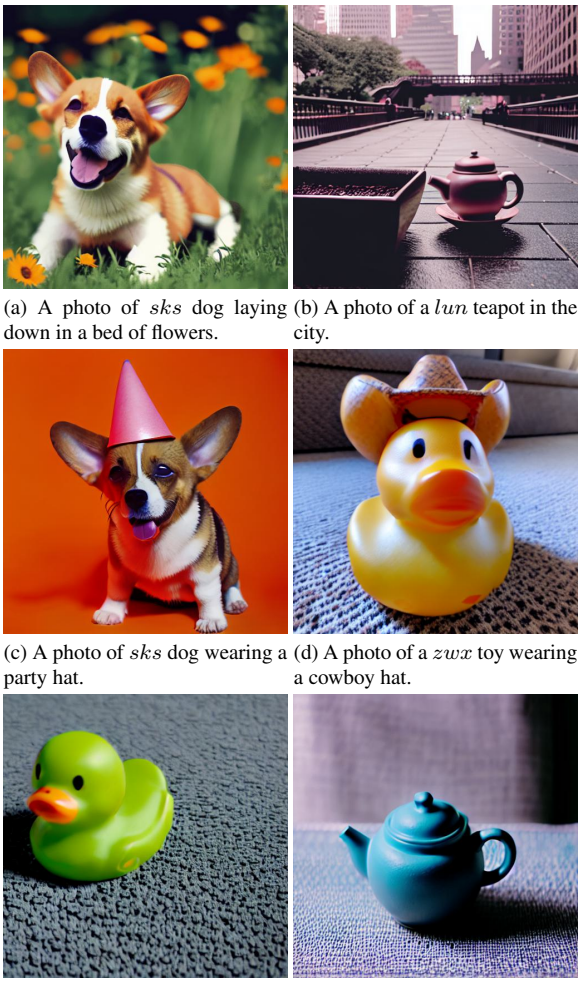
where $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.05$ are regularization hyperparameters to control the strength of each auxiliary loss.

4. Results

Each Subject in Individual Images A text-to-image diffusion model fine-tuned using our multi-subject methodology maintains the base DreamBooth [20] ability to generate each novel images of each subject individually. It does not experience any performance degradation while learning multiple different subjects. The following figures demonstrate the ability of the model to synthesize different types of imagery with a high degree of object fidelity and prompt accuracy given simple prompts. Figures 7a, 7b show the input subjects placed into contexts unseen during the tuning. Figures 7c, 7d highlight the capability of the model to accessorize the subjects. Figures 7e, 7f prove that the model is able to reliably modify properties of the subjects while still maintaining object fidelity.

Subjects in Combined Image Diffusion models excel at producing images that could potentially come from the distribution in which it is trained on. When fine-tuning on unique subjects, the model is able to effortlessly blend them into novel concepts aligned with the trained distribution. However, an additional challenge is met when we prompt the model to generate multiple subjects in the same image. While it may seem like a simple task, we find that it is actually quite difficult. Given multiple subjects and unique tokens, the model has not quite learned how to parse them properly, resulting in blended images such as the one seen in Figure 2. Figure 8 demonstrates an example in which our method is able to alleviate this issue by producing both objects in the same space without any feature blending.

Prior Preservation Ability Following the original DreamBooth method [20] the model should maintain its ability to generate examples from the fine-tuned subject class. For example, it should still be able to produce other dogs after tuning on a specific dog. Figure 9 shows images output from our 4 subject tuned model, where two of the subjects are the orange dog and the brown teapot. Given prompts with no unique identifiers, the model preserves its ability to generate examples of the tuned-subject class priors.



(a) A photo of sks dog laying down in a bed of flowers.

(b) A photo of a lun teapot in the city.

(c) A photo of sks dog wearing a party hat.

(d) A photo of a zwx toy wearing a cowboy hat.

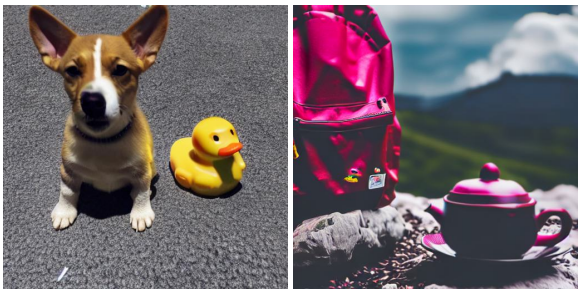
(e) A photo of green zwx toy.

(f) A photo of a blue lun teapot.

Figure 7. Example of multiple subjects being properly learned. Every image was prompted using using the same fine-tuned model. (a), (b) show the objects being seamlessly re-contextualized. (c), (d) show the subjects being accessorized by wearing different hats. (e), (f) are examples of property modification, where the color of the objects are changed.

5. Limitations

While our method is more competent in producing multiple subjects in the same image, we find that this ability has limits. The more unusual subjects we try to combine into one image, the more difficult it gets. Even though the model is perfectly capable of generating each subject individually, an image where both of these subjects are present is unlikely to have occurred anywhere in the training distribution, so the model struggles to coherently generate them together. This indicates that diffusion models are still quite sensitive to their training datasets and can benefit from more generalizable training, if possible. Figure 10 shows the base Stable Diffusion model trying to generate an unusual com-



(a) A sks dog next to a zwx toy (b) A hta backpack next to a lun teapot

Figure 8. Given prompts with multiple subjects in the same image, the personalized model is able to generate each subject with high fidelity and no feature blending.

(a) A photo of a dog (b) A photo of a teapot

Figure 9. Images produced by our tuned model given generic prompts without the unique subject identifier. The synthesized subjects match the quality of the model prior to fine-tuning.

(a) A teapot, a dog, and a rubber duck (b) A teapot, a rubber duck, and a backpack

Figure 10. Base Stable Diffusion v1.5 failure cases produced by trying to synthesize the same unusual combinations of subjects.

(a) A lun teapot, a sks dog, and a zwx toy (b) A lun teapot, a zwx toy, and a hta backpack

Figure 11. Failure cases produced by our model trying to synthesize the same unusual combination of trained subjects as Figure 10.

bination of subjects. Figure 11 shows the same examples on one of our fine-tuned models trying to generate our input subjects in the same unusual combinations. From this, we can see that it is not necessarily our new methodology that causes these failures, but rather stems from the original pre-trained diffusion model itself.

Moreover, we found that the choice of mean parameterization for the posterior distribution in the diffusion process affects the latent-based loss, and furthermore limited our implementation of any variants. It has been established in [10, 23] that the learned estimated mean of the desired distribution in the reverse process can take one of three forms: estimating the source image, x_0, at each timestep, estimating the previous timestep, x_{t-1}, at timestep t, and estimating the inputted noise, ϵ, at each timestep. Since [6] modeled the noise at each timestep, we adopted the same approach in this work. With the success of enforcing similar latents for the same subject, we attempted to enforce dissimilar latents between different subjects to further improve the quality of generated images. However, this dissimilarity loss resulted in the nonsensical images as seen in Figure 12. Estimating the inputted noise at each timestep has already been accepted as an computationally efficient and effective parameterization of the reverse diffusion process, but since it discards/ignores the overall semantic information of the current image at a given timestep by only focusing on the imparted noise, the dissimilarity latent loss is not able to properly distinguish what high-level features differ between different subjects in the image space (i.e, the structural difference between a dog and a cat).

In other words, directly estimating the noise at each timestep works for the latent similarity loss since each instance image of a subject is relatively similar in color, background, etc. such that reversing the noising process between two subject instances is quite similar. Two instances from different subjects differ greatly, however, and the imparted noise for each image and its impact at each timestep will also differ greatly. Thus, the dissimilarity latent loss struggles in capturing what semantically is different between the two subjects, while still learning a meaningful reverse process for image generation. This confusion leads to the images we see in Figure 12, and we believe further testing where our approach is trained with estimating the source image, or the image at the previous timestep, as the parameterization could yield better results.

(a) A photo of a zwx cat (b) A photo of a sks dog

Figure 12. Sample output with given prompts of different subjects when trained with dissimilarity loss. The poor results could be due to an unsuitable mean parameterization choice or an insufficient representation of the high-level structure in the latent space.

6. Conclusion

In this work we propose an update to the DreamBooth [20] specific subject fine-tuning method that allows for any number of subjects to be trained simultaneously. To achieve this, we assign unique token identifiers to each subject, enabling subject-specific text prompting. To mitigate potential problems that may arise in a mulit-subject setting, we propose several techniques, including subject-dependent early stopping, a text encoder loss, and an image space loss. We demonstrate that our proposed model preserves all the capabilities of existing text-to-image diffusion models while learning multiple subjects.

References

[1] Korbinian Abstreiter, Sarthak Mittal, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion-based representation learning. *arXiv preprint arXiv:2105.14257*, 2021. 3

[2] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *arXiv preprint arXiv:2211.12500*, 2022. 3

[5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 3

[6] Hugging Face. Dreambooth. https://huggingface.co/docs/diffusers/training/dreambooth. 1, 2, 7

[7] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis, Feb. 2023. arXiv:2212.05032 [cs]. 3, 5

[8] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning, Mar. 2023. arXiv:2303.11305 [cs]. 3

[9] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 7

[11] 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):1–33, 2022. 3

[12] Jonathan Ho Ho and Tim Salimans. Classifier-Free Diffusion Guidance, July 2022. arXiv:2207.12598 [cs]. 2

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion, Dec. 2022. arXiv:2212.04488 [cs]. 3, 4

[14] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified Multi-Modal Latent Diffusion for Joint Subject and Text Conditional Image Generation, Mar. 2023. arXiv:2303.09319 [cs]. 3, 5

[15] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[16] Nichol 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, Mar. 2022. arXiv:2112.10741 [cs]. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. arXiv:2103.00020 [cs]. 2

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, Apr. 2022. arXiv:2204.06125 [cs]. 2, 5

[19] 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. 1, 5

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Mar. 2023. arXiv:2208.12242 [cs]. 1, 3, 6, 8

[21] 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, May 2022. arXiv:2205.11487 [cs]. 1, 2, 5

[22] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[23] 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. 7

[24] Valentine Kozin Suraj Patil, Pedro Cuenca. Training stable diffusion with dreambooth using diffusers. https://huggingface.co/blog/dreambooth. 4

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3