Image Synthesis based on Prompts utilizing Multiple Subjects

Rithik Bhandary, Trived Katragadda, Thiru Satya Surya Mahaveer Bonagiri
Boston University
{rithik, trivedk, mahaveer}@bu.edu

ABSTRACT

A text-to-image model takes as input a natural language description and produces an image matching that description. Text-to-image models combine a language model with a generative image model. In this work, we present a new approach for personalization of text-to-image diffusion models. Given as input just a few images of multiple subjects, we fine-tune a pre-trained text-to-image model such that it learns to bind a unique identifier with those specific subjects. Once the subjects are embedded in the output domain of the model, the unique identifier can then be used to synthesize fully novel photo-realistic images of the multiple subjects contextualized in different scenes. By leveraging the semantic prior embedded in the model, along with a new autogenous class-specific prior preservation loss, this technique enables synthesizing the subject in diverse scenes, poses, views, and lighting conditions that do not appear in the reference input images of the subjects.

INTRODUCTION

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. Although, a task that is difficult to do with these models is the rendering of images such that we can synthesise instances of specific subjects in new contexts such that they naturally and seamlessly blend into the scene.

"DreamBooth[7]: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Ruiz et al., 2022 " represents a given subject class with rare token identifiers and fine-tunes a pre-trained, diffusion-based text-to-image framework. The goal is to expand the language-vision dictionary of the model such that it binds new words with specific subjects that the user wants to generate. Once the new dictionary is embedded in the model, it can use these words to synthesize novel photo-realistic images of the subject, contextualized in different scenes, while preserving their key identifying features.

The initial objective is to implant the subject into the output domain of the model such that it can be synthesized with a unique identifier. We represent a given subject with rare token identifiers and fine-tune a pre-trained, diffusion- based text-to-image framework. Taking a few images as the input, defined as the subject and providing a prompt as context we synthesise the subject in the context without losing the key identifying features of the subject. The paper makes use of Prior-Preservation Loss to help prevent overfitting of the output to the few images given as input and to prevent language drift.

Goal

We propose to enhance the aforementioned Dreambooth model, which only uses one subject synthesized in a particular context, by trying to upgrade the model for multiple subjects. We will assign a special rare token identifier for each of the subject classes. The user provides multiple images of a subject as input along with a rare token identifier to distinguish the particular class. The different classes along with their rare token identifier is trained and the model can perform synthesis of blending the multiple subject classes into a context based on the user entered prompt. The prompt should contain a distinct rare token identifier for each of the subject classes so as to distinguish them as different from the global class dictionary.

Dataset

The original stable diffusion text-image model[6] was trained on pairs of images and captions taken from LAION-5B dataset. The publicly licensed weights of pre-trained model have been downloaded from Hugging Face, which we use to fine-tune our model. The generative model will synthesise novel renditions of the subject within the context, based on few images of a subject provided as input by the user, along with a text prompt with a unique identifier which behaves as the context.

RELATED WORK AND REFERENCES

Synthesizing a given subject in different contexts has been accomplished in which image composition techniques aim to clone a given subject into a new background such that the subject melds seamlessly into the scene. However, these techniques [[3],[9]] are quite limited in their ability to adapt to correct lighting from the background, cast proper shadows, or fit the background content in a semantically aware manner. The challenge is even greater when the requirement is to synthesize the subject in novel articulations and views that were not seen in the few given images of the

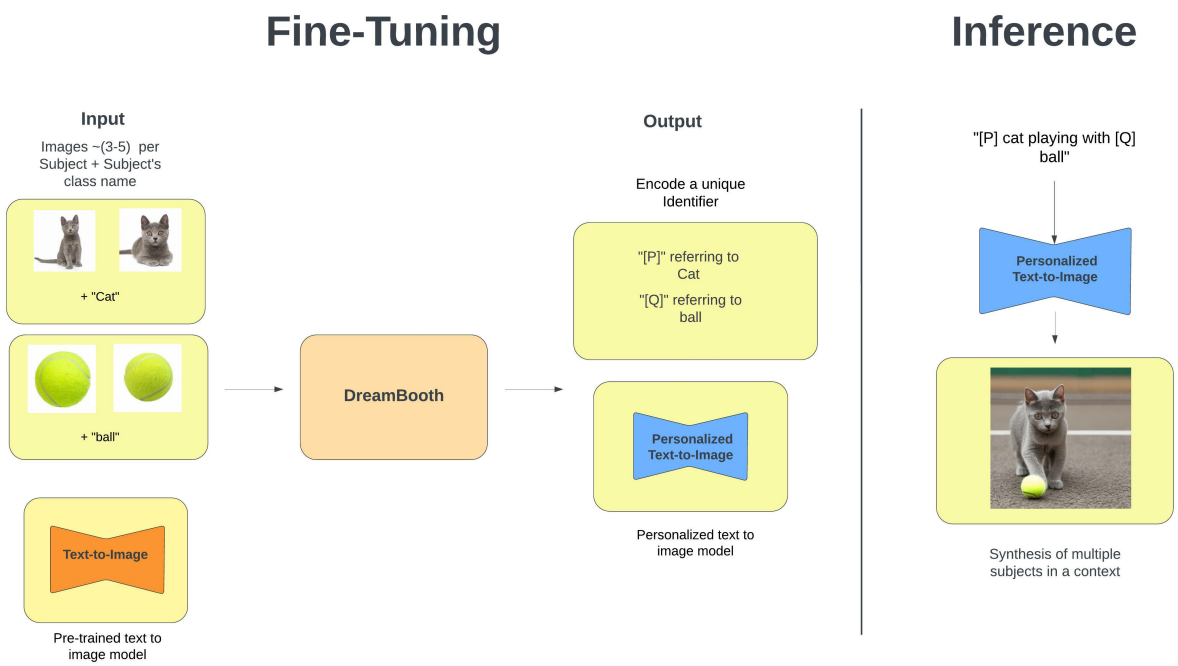

Fig. 1: **Architecture and design of the fine-tuning process**

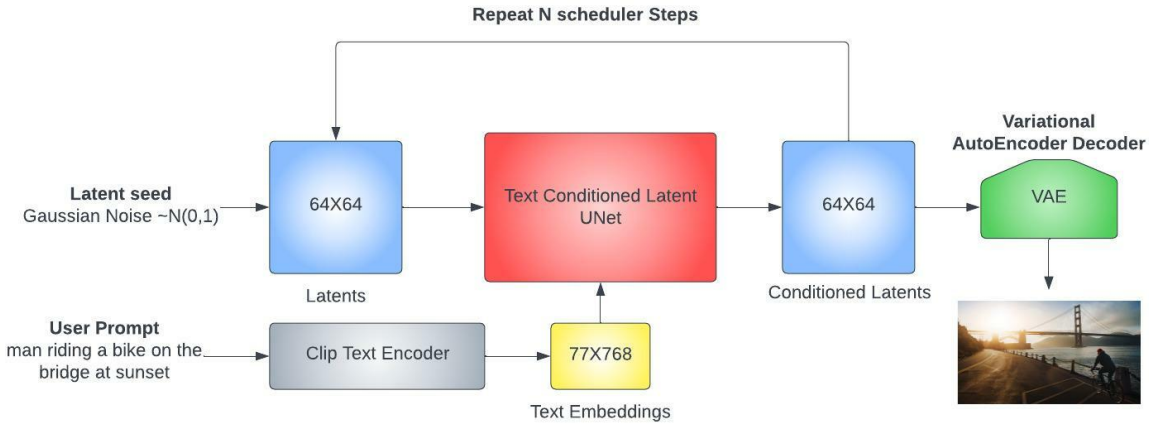

Fig. 2: **Model Backbone, Stable Diffusion**

subject. One way of preserving object appearance while changing context is image editing. Text-driven image manipulation has recently achieved significant progress using GANs, but some of the techniques [[1]] require the user to provide a spatial mask so as to direct the location of the edit. However, while GAN-based image editing approaches succeed on highly-curated, e.g., human faces, they struggle over diverse datasets with many subject types. Most of these editing approaches [[4], [1], [2]] allow modification of global properties or local editing of a given image, none of them enables generating novel renditions of a given subject in new contexts.

Recent large text-to-image models have great semantic power such as Imagen[[8]], DALL-E2[[5]], but these models fail to preserve the identity of a subject consistently across different images, since modifying the context in the prompt also modifies the appearance of the subject.

PROBLEM FORMULATION

We need to synthesize novel renditions of multiple subjects using a few images of the different subjects as input for training, and with the guidance of a text prompt as a context for those subjects. A key problem is that fine-tuning on a small set of image of our subject is prone

to overfitting on the given images. Each of the subjects will be assigned to a unique rare token identifier which we use so as to prevent the new subject from overfitting into the global class of the subject, which is known as language drift. For example, if we use the prompt " a photo of ['v'] dog ", the model should not overfit the global class 'dog' to the newly created subject [v] dog, thus enabling the model to embed a new language dictionary that maintain its previous properties. Another problem is most such models lack the ability to mimic the appearance of subjects in a given reference set, and synthesize novel renditions of those same subjects in different contexts by taking care of the essential features of the subject, carrying them forward and not losing them.

METHODOLOGY

The back bone of our model is the Stable diffusion model which is used to generate the image given a text prompt via a diffusion process as the Imagen used by the original Dream-booth is not open-source. The general working of the backbone can be seen in Fig.2 as it takes gaussian noise which is the random seed along with the text embeddings which are the unique identifiers for different subjects, other than these there is also CLIP text encoder, UNet and Variational Auto Encoder(VAE) to process the user prompts and generate images according them.

The key idea is to embed the subjects instance in the output domain of a text-to-image diffusion model by binding the subjects to a unique identifier called the rare token identifier so as to avoid the increase in computational and training time because if the general terms that are present in the vocabulary are used as indicators then our model will have to re-align its learning which can be avoided using rare token identifier. The diffusion is done in two steps:

- First, fine-tune the low-resolution text-to-image model with the input images and text prompts containing a unique identifier followed by the class name of the subject
- Fine-tune the super-resolution component with pairs of low-resolution and high-resolution versions of the input images. This allows the model to maintain high fidelity to small details of the subject.

Model

Given only a few images of multiple subjects, without any textual description, our objective is to generate new images of the subjects with high detail fidelity and by following the text prompt which serves as the context for these subjects. The first task is to implant the subject instance into the output domain of the model and to bind the subject with a unique identifier. There is an autogenous class-specific prior preservation loss, where we alleviate overfitting and prevent language drift by encouraging the diffusion model to keep generating diverse instances of the same class as our subject. To enhance the details preservation, the super-resolution components of the model should also be fine-tuned. A detailed sketch of our proposed training procedure is shown in Fig[1].

- **Representing the Subject with a Rare-token Identifier:** We need to implant a new (key, value) pair into the diffusion model's dictionary such that, given the key for our subject, we are able to generate fully novel images of this specific subject with meaningful semantic modifications guided by a text prompt. 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.). We specifically use a class descriptor in the sentence in order to tether the prior of the class to our unique subject. A problem with using existing English words as identifiers is that they tend to have a stronger prior due to occurrence in the training set of text-to-image diffusion models. We generally find increased training time and decreased performance when using such generic words to index our subject, since the model has to both learn to disentangle them from their original meaning and to re-entangle them to reference our subject.
  In a nutshell, our approach is to find relatively rare tokens in the vocabulary, and then invert these rare tokens into text space. In order to do this, we first perform a rare-token lookup in the vocabulary and obtain a sequence of rare token identifiers f($\hat{V}$). Then, by inverting the vocabulary using the detokenizer on f($\hat{V}$) we obtain a sequence of characters that define our unique identifier $\hat{V}$.

- **Class-specific Prior Preservation Loss:** The input image set is quite small, and fine-tuning large image generation models with these small sets can overfit to both the context and the appearance of the subject in the given input images. A language model that is pre-trained on a large text corpus and later fine-tuned for a specific task progressively loses syntactic and semantic knowledge of the language as it learns to improve in the target task, is known as language drift. An autogenous class-specific prior-preserving loss counters both the overfitting and language drift issues. In essence, the method is to supervise the model with its own generated samples, in order for it to retain the prior once the few-shot fine-tuning begins. Specifically, we generate data $x_{pr} = \hat{x}(z_{t1}, c_{pr})$ by using the ancestral sampler on the frozen pre-trained diffusion model with random initial noise z_{t1} N(0,I) and conditioning vector $c_{pr} := \Gamma(f("a\ [class\ noun]"))$. The loss becomes:
  $$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+\lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2\right]$$
  where lambda controls for the relative weight of the prior-preservation term.

RESULTS

Training input images

The generative model will synthesize novel renditions of the subject within the context, based on a few images of a subject provided as input by the user. The user also inputs a text prompt with rare token unique identifier which behaves as the context for the subject. Here, we took (4-7) 512 X 512 resized images of each of the six classes as the input. These input subject images are assigned different rare token identifiers in the form of “a [identifier] [class noun]”. The input images are shown in Fig. 3.

Fig. 3: The Stable diffusion model dictionary of text-image generation is fine-tuned with the following identifiers and subjects- pbt catwoman, rbp bike, mtr dog, bmt man, ktp cat and lmw ball

Intermediate checkpoint outputs

Fig. 4: Training checkpoint for every 1000 steps

During our training we created a checkpoint for every 1000 iterations for which we assigned a variable, save sample prompt as "photo of bmt man walking a mtr dog photo", and the model performs a text-to-image generation of 4 samples at each of these checkpoints with the weights it has learned at that checkpoint. As you can see in Fig. 4, initially the model just outputs images of the man walking random dogs and slowly as the model learns, improving with the number of iterations, it

combines the two new subjects 'bmt man' and 'mtr dog'. After training the model for 8,000 iterations and using the learned weights.

Fig. 5: The prompt 'photo of bmt man holding mtr dog' which generated the following result.

Fig. 6: For the prompt 'Photo of laughing bmt man and ktp cat in the park' the generated image is as shown. As it can be seen, Stable diffusion sometimes does badly with faces.

Fig. 7: We can also see that the model doesn't overfit to our new 'bmt man' class by providing a prompt as 'Photo of laughing man looking at ktp cat in the park' which generates the image as shown. Here, we provide man from the general class and not the man to which we have attached a unique rare token identifier, thus showing that our model is embedding a new language dictionary without causing the issue of language drift.

Fig. 8: By setting the right parameters for the seed, the number of inference steps and the right guidance scale value, we were successful in generating an image that synthesizes 3 subjects, mtr dog, ktp cat and lmw ball, following the prompt 'photo of lmw ball next to ktp cat and mtr dog'.

Although generating an image which clubs 3 subjects is difficult, the model does get it right if we fine-tune the parameters just right, thus the model with further improvement does have scope to be implemented for 3 or more subjects. We also have included some results with prompts as shown below.

Fig. 9: Photo of mtr dog playing with lmw ball.

Fig. 10: Photo of pbt catwoman riding rbp bike.

Fig. 11: Photo bmt man and mtr dog by the ocean looking at sunset.

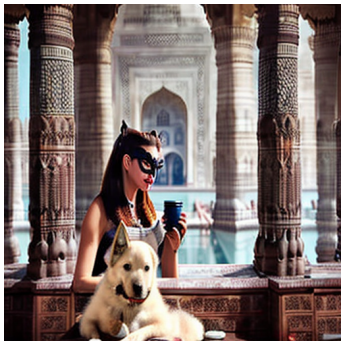

Fig. 12: Photo of pbt catwoman drinking tea at the taj mahal and looking at mtr dog.

Adversarial Results:

Fig. 13: Photo of bmt man and pbt catwoman in an amusement park.

Apart from the results that we generated above, we also observed certain adversarial effects. The prompt 'Photo of bmt man and pbt catwoman in an amusement park' generated an image in which the 'bmt man' gets blended into the attire of 'pbt catwoman'. The prompt 'Photo of laughing man looking at ktp cat in the park' generated an image in which the man also turned into a cat. The prompt 'photo of mtr dog looking at rbp bike' synthesized an image in which the 'mtr dog' blended into 'rbp bike', the following results are shown in Fig. 13, 14, 15.

Fig. 14: Photo of laughing man looking at ktp cat in the park.

Fig. 15: Photo of mtr dog looking at rbp bike.

CODE REFERENCES

- github.com/mahaveer220/MS-Dreambooth
  (Our Github repo)
- CompVis/stable-diffusion-v1-4
  (Used to fine-tune the weights)
- github.com/ShivamShrirao/Dreambooth-Stable-Diffusion

CONCLUSION

- We found that number of training steps are critical in producing good results. Also, we observed that 3-4 images do not suffice in reproducing the subject with high fidelity, and that we at least need 5-7 image in each case.
- We have observed that results are more accurate and consistent with 2 custom subjects provided in the prompt and it performed a little less consistent on prompts with >=3 subjects with missing details of the subject or entirely different subjects.
- Upon performing training with varied parameter values, we observed that we achieved the best results with 175-200 generalized images of each class. Further, we reckoned that setting the number of training steps to follow this equation '(sum of no of sample images in each custom subject)*200' generated good results, with the idea of making sure each sample

image is seen 200 times while training. Finally, the learning rate is set to 10^5.

REFERENCES

[1] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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. 2022.

[8] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[9] 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.