Fine Tuning Text-to-Image Diffusion Models with Style Transfer

Apoorv Walia
Department of Computer Science
Rice University
Apoorv.Walia@rice.edu

Benson Thomas
Department of Computer Science
Rice University
Benson.Thomas@rice.edu

Shraiysh Vaishay
Department of Computer Science
Rice University
Shariysh.Vaishay@rice.edu

Abstract

Large-scale text-to-image synthesis models have significantly advanced the field of artificial intelligence by enabling the high-quality and diverse generation of images from textual descriptions. DreamBooth[1], a particular model, has demonstrated impressive results for subject-driven fine-tuning. Our research aims to build upon this progress by training a model not only on the subject (human model) but also on the object, specifically a piece of clothing, to create automated photo shoots for clothing companies without the need for human models.

Video — Presentation — Code

1. Introduction

In recent years, text-to-image synthesis models have seen significant improvements, with large models like DreamBooth[1] achieving remarkable results for subject-driven fine-tuning. DreamBooth is fine-tuned on top of Imagen[2]. Imagen is a text-to-image diffusion model developed by Google Research Brain Team that has an unprecedented degree of photorealism and a deep level of language understanding. It builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation1. Imagen uses a large frozen T5-XXL encoder to encode the input text into embeddings and a conditional diffusion model maps the text embedding into a 64×64 image1. In this project, we propose to take this a step further by creating model-esque photoshoots using only a few images of a human model and a few images of the item of clothing. Our approach aims to not only reduce costs but also give companies a competitive edge by reducing the time it takes to get the item to market.

To achieve our goal, we plan to use DreamBooth, a state-of-the-art text-to-image synthesis model. We will fine-tune the model on both the subject (human model) and object (piece of clothing) to generate images that look like a real photoshoot. This approach will allow us to generate images of the item of clothing on a human model without the need for an actual photoshoot.

One of the major advantages of our approach is that it allows for different lighting and settings to be used in the photoshoot. This is because the model is not limited by the constraints of a physical photoshoot. Instead, we can generate images with different lighting conditions and settings, which can help to showcase the item of clothing in different ways.

Furthermore, our approach has the potential to be used for a wide range of items of clothing. This is because the same human model can be used for multiple items of clothing, reducing the cost and time needed for each individual photoshoot. In addition, our approach can be used to create images of the same item of clothing in different colors and styles, giving companies a wider range of options to showcase their products.

2. Model

DreamBooth is a technique to fine-tune text-to-image diffusion models for subject-driven generation. The model architecture consists of the following components:

- A text encoder that maps text to a sequence of embeddings using a pretrained transformer language model such as T5-XXL, BERT or CLIP.

- A base 64x64 text-to-image diffusion model that maps the text embeddings to a low-resolution image using a U-Net architecture with cross attention over the text embeddings at multiple resolutions. The diffusion model can be any pretrained model such as Imagen or Stable Diffusion.
- Two super-resolution diffusion models that upsample the 64x64 image to 256x256 and then to 1024x1024 using noise conditioning augmentation and text cross attention.
- A rare-token identifier that is used to bind a specific subject with a unique word that has a weak prior in both the language model and the diffusion model. The identifier is constructed by finding rare tokens in the vocabulary and decoding them into text space.
- A class-specific prior preservation loss that leverages the semantic prior of the diffusion model on the class of the subject and encourages it to generate diverse instances of the same class using the class name in a text prompt.

3. Current Approach

Our approach of combining subject and object training has several potential advantages over traditional methods of clothing photography. By training a model on a specific piece of clothing, we can generate images that are consistent in lighting, pose, and other factors, allowing for easy comparison between products. Additionally, automated photoshoots would eliminate the need for hiring human models, saving time and money for clothing companies.

To accomplish this task, we first implemented the DreamBooth model and ran several experiments. While our results were promising, we encountered some issues with accuracy in facial features and exact details of the clothing.

During the fine-tuning process, we introduced a dummy keyword that is linked to the dataset of images of the model and a different dummy keyword linked to the piece of clothing, enabling the model to learn how to incorporate the model style as well as the piece of clothing into the image synthesis process.

After the fine-tuning process, when generating new images, we provide the model with a prompt that includes both the subject (human model) and the product's keyword, such as ”dummy-keyword-subject wearing dummy-keyword-product.” We train our model on a dataset of a shirt containing 16 images of the shirt in different conditions as well 12 images of a human model. We also fine-tune the model's architecture and hyperparameters to optimize its output.

4. Experiments and Result

The following are the details of our experiments:

- Optimizer: AdamW optimizer
  - Learning Rate = 5×10^{-6}
  - Betas = (0.9, 0.999)
  - Weight decay = 10^{-2}
  - Epsilon = 10^{-8}
- Tokenizer: CLIPTokenizer
- Text encoder: CLIPTextModel
- Variational autoencoder: AutoencoderKL
- Noise Scheduler: DDPMScheduler
- UNet: UNet2DConditionModel
- Image size: 512 pixels
- Epochs on human model images: 20 epochs
- Epochs on product images: 40 epochs
- Human Model image prompt for training: ”a photo of sks person”
- Product image prompt for training: ”a photo of tft shirt”
- Prompt for prediction: ”sks person wearing tft shirt”

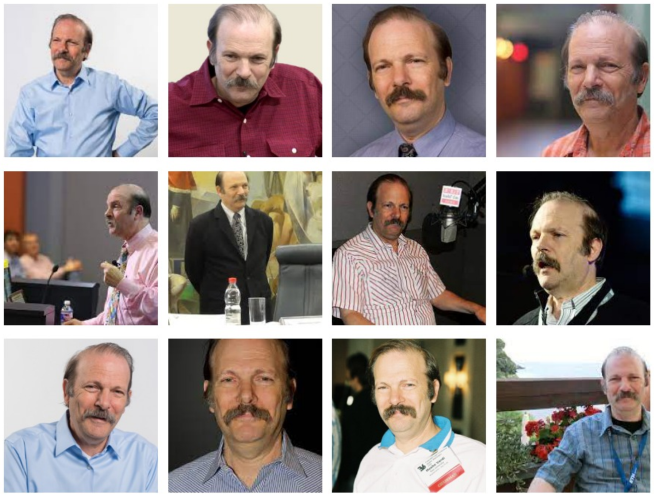

Figure 1. Human model images, trained on the phrase ”a photo of sks person”

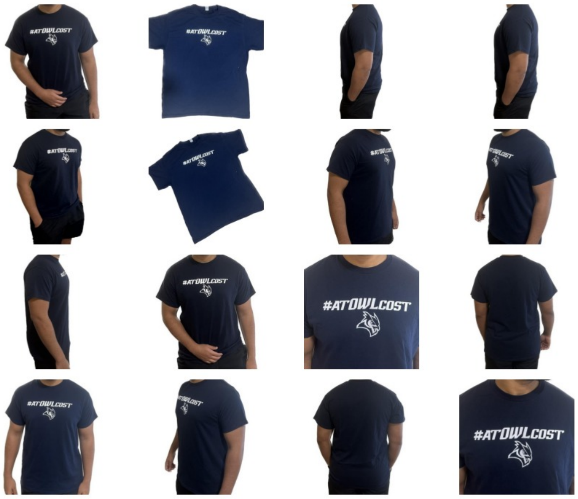

Figure 2. Product Images, trained on the phrase ”a photo of tft shirt”

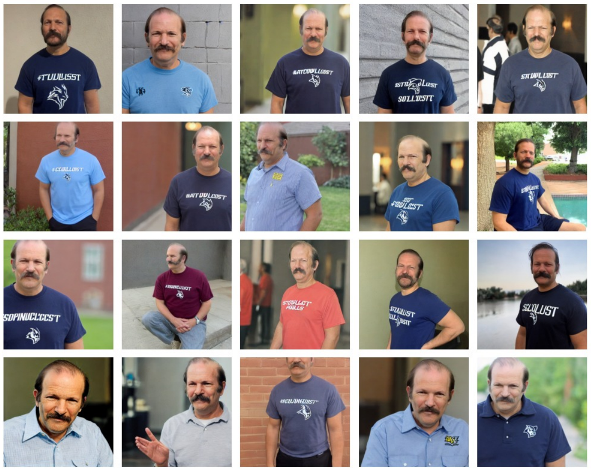

Figure 3. Generated images by the model for the phrase ”sks person wearing tft shirt”

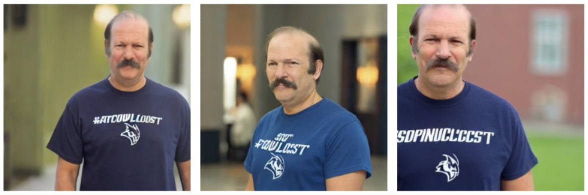

Figure 4. Good images generated by the model

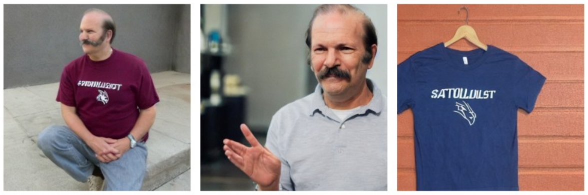

Figure 5. Erroneous images generated by the model

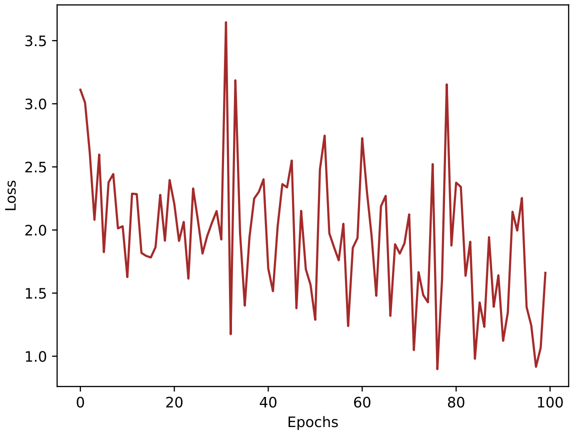

Figure 6. Our loss graph. Due to compute constraints we have only trained for 100 epochs.

On average, we were able to achieve good outputs on 12 images out of a total 30 samples generated. Below we provide examples of “sks” person and “tft” t shirt that the model was trained on as well as some positive and negative results. We never observed the wrong human model in the pictures generated. There are two kinds of negative results:

- Images of human model wearing a different shirt.
- Images of only shirt without a human model.

5. Future Work

Future work on this problem will involve additional experimentation and longer training periods to improve the model’s performance. We can also improve our approach by exploring various avenues such as class-based training, text metadata, and prompt engineering.

Class-based training involves training the model on images of same classes that do not belong to our dummy token, such as other dresses, tops, pants, and so on. This approach to training the model by providing negative examples could help the model learn better representations for each class, resulting in more accurate and realistic output images.

Text metadata is additional descriptive information that can be used to improve the quality of the generated images. This information can include color, texture, style, and other relevant details that can guide the model to create more appropriate images. Incorporating this information into the training process can help the model to generate images that better match the text prompts.

Prompt engineering involves modifying the text prompts to be more effective at guiding the model to create appropriate images. This can be achieved by adding more descriptive words, emphasizing certain aspects of the image, or rephrasing the prompt to make it clearer and more specific.

Another direction for this work would to generate new styles and colours for the human model and shirt. For example, training the model on blue shirt and a human model, and then asking it to generate images of the human model in a pink variant of the same shirt. We tried this on our model, and it could generate ∼5-10% good images even though it wasn't trained for this purpose.

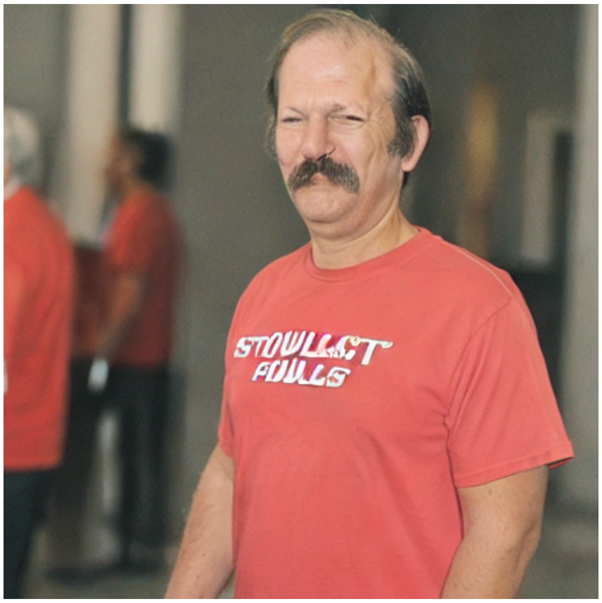

Figure 7. Generated image for the prompt "sks person wearing pink tft shirt."

6. Conclusion and Evaluation Metrics

For evaluation, we sent out a survey to a group of 10 students at Rice University. We explained the problem in the survey, and asked them to rate the generated images on a scale of one to five. A score of five means that the generated image seems like a real picture, where the exact input human model is wearing the input shirt. They are free to decrease the score if there is a problem with the human model or with the shirt in the generated image. The average score for generated images was 3.72.

Based on these scores, we classified the images into "good" and "bad" images. The good images had an average rating of 4.62/5.0 with a standard deviation of 0.18 averaged over all the respondents and over all the images. On an average, there were 8 such images out of every 20 images. The bad images had an average rating of 3.13 with a standard deviation of 0.43 calculated in a similar fashion.

Based on these results, we can say that our model is expected to give ∼40% good images.

In conclusion, our research aims to leverage the recent advances in large-scale text-to-image synthesis to automate clothing photoshoots. By training a model on both the subject and object, we hope to generate high-quality and consistent images of clothing items for clothing companies, without the need for human models. Our future work will focus on improving the model's accuracy and detail, through additional experimentation and longer training periods, using state-of-the-art machine learning techniques.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.