DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Authors: Grant Rinehimer, Adam Faridi, Armaan Tewary, Ignazio Perez Romero, Elliot Kim

https://github.com/grantrinehimer/dreambooth_diffusion

I. Introduction

Recent text-to-image (T2I) models have shown incredible performance in generating high-quality images given a text prompt. However, these models, given a reference set of subjects, lack the ability to generate accurate and consistent renditions in different contexts. Dreambooth personalizes pre-trained T2I models by fine-tuning them on just 3–5 images of a subject, using a rare token identifier and class-specific prompts to synthesize subjects. The paper [1] contributes a class-specific Prior Preservation Loss (PPL) that allows T2I models to generate a diverse and contextually appropriate image of a subject while effectively preserving its identity.

II. Chosen Result

We extended Figure 12 and Table 3 from the original paper [1] to qualitatively compare our reimplementation [Figure 1] and quantitatively assess the effect of PPL on subject fidelity and output diversity—key factors in evaluating the success of subject-driven generation.

III. Methodology

For our reimplementation, we finetuned a pretrained Stable Diffusion v1.5 T2I model. The original paper used Imagen (proprietary) and Stable Diffusion (the version was not disclosed). We used the paper's original dataset [1], consisting of 30 subjects and 15 classes. The original paper used 25 prompts and generated 4 images per subject and prompt. We used a subset of 8 prompts and generated 2 images per subject and a prompt for computational feasibility.

We utilized a T4 GPU with 16 GB of GPU memory hosted on Google Cloud. We finetuned two models for each subject, one using PPL and one without. For PPL, we generated 100 class samples before finetuning, rather than the 1000 used in the paper (although the paper noted that less could be used). We also trained each model over 400 training steps rather than 1000. On testing, we found this did not change the quality of results. We also used a variety of techniques to solve memory issues. We used an FP16 quantized pretrained model, FP16 mixed precision training, an 8-bit Adam optimizer, and gradient checkpointing. The extent to which these methods are used in the paper isn't clear.

We made use of a Huggingface dreambooth CLIP training script [2] for our implementation and adapted it for use in our script to train models and run inference on all 30 subjects. Our script allows config files to be passed to generate results on any number of subjects in a batched format. We wrote scripts to compute CLIP and DINO embeddings on our results to recreate the main table result of the paper. We also used our own prompts to recreate the various subject adaptation images in the paper.

IV. Evaluation

We computed four complementary metrics to assess the performance of our fine-tuned models under both the no-PPL (baseline) and PPL conditions. Our calculated scores for these metrics were compared with the original paper's scores [1] present in their Table 3. We measured subject fidelity to quantify how faithfully our fine-tuned models reproduced the unique visual identity of the reference subject via two metrics: the first is DINO, which found the average cosine similarity between ViT-S/8 DINO embeddings of real and generated images. This differs from Dreambooth's [1] calculation of DINO (and all other metrics using ViTs pretrained with the DINO self-supervised learning method [3]), as Ruiz et al. utilize ViT-S/16 DINO instead. The second subject fidelity metric is CLIP-I, standing for the average cosine similarity between CLIP embeddings of real and generated images In addition, we measure prompt fidelity to see how well a generated image matches its conditioning text prompt, done through CLIP-T: the average cosine similarity between CLIP prompt text embeddings and generated image embeddings from that same prompt.

We evaluated prior collapse through the PRES (preservation) metric, found by the average

cosine similarity of ViT-S/8 DINO embeddings between real images of a subject and generated images of other subjects in the same class. The last metric computed was DIV, representing the diversity between generated samples of the same prompt. We quantify DIV by computing the mean LPIPS distance between all pairs of generated images sharing the same prompt and then averaging those per-prompt scores across the full prompt set. Concretely, for each subject, we took its few real reference images and its generated outputs, extracted normalized embeddings, formed all pairwise comparisons, and averaged to obtain per-subject scores.

V. Results and Analysis

Our reimplementation showed good results qualitatively for subjects with a strong prior. We can observe from Figure 1 that our model produces outputs for the dog that are more consistent with those from the original paper than it does for the sunglasses. This discrepancy is likely due to dogs having a stronger prior in the pre-trained diffusion model: dogs appear more frequently and with more diverse contexts in the model's training data, enabling better subject recognition and preservation. In contrast, sunglasses are less common and more context-dependent, making it harder for the model to generate high-fidelity, contextually appropriate images for that class. Furthermore, Figure 2 shows that our model is able to preserve subject identity even under strong stylistic transformations.

Across all metrics, our reimplementation reproduces the paper's trends, with slightly lower absolute values owing to our reduced prompt set (8 vs. 25) and shorter fine-tuning (400 vs. 1000 steps). This is highlighted below [Table 1]; the same trends persist between the Dreambooth metrics [1] and ours. Adding PPL sharply reduces prior collapse (lower PRES) in both cases, meaning the model no longer "hallucinates" the fine-tuned subject when generating random class samples. We interpreted this as a lack of overfitting to the original subject; PPL aids in understanding the key components of what features make up the class without recreating the original subject, which is evident when prompts contain another subject from the same class.

Moreover, PPL boosts sample diversity under both pipelines, as generated images vary more in pose, background, and articulation. Nevertheless, our significant difference in DIV scores in PPL and no-PPL of 0.245 and 0.207, respectively, is indicative of our reduced amount of output images per prompt compared to the Dreambooth [1] output (2 vs. 4): having more output images mitigates average variance. When you only have two images, that one distance completely determines the mean; with four images, you average over six distances, which smooths out outliers/reduces overall DIV value.

Importantly, even with 2 outputs per prompt and 400 training steps, PPL maintained its benefits: relative PRES reductions and DIV gains closely match those reported by Ruiz et al. [1]. This robustness suggests that class-specific prior preservation can be deployed under constrained compute budgets while preserving subject identity and encouraging diverse generations.

VI. Reflections

Our group learned multiple lessons related to fine-tuning diffusion models. First, dealing with heavy compute can make it challenging to obtain quality large-scale results. Just on reproducing a small subset of results from the paper, we consumed our available compute credits. Second, a related issue was that setting up VMs correctly, actually using them (and their GPUs), and debugging to avoid OOM errors during training was *more* than half the challenge. We realized that a lot of working with diffusion models is setting up the compute architecture to sustain the model, as well as adapting code across our different computers to work with the specs of each machine. This presented a pain point for our team and not everyone was able to help with running code for training and/or inferencing. Lastly, fine-tuning can cause some seemingly magical things to happen, like the Eiffel Towel being morphed into the smile of the dog backpack (see Figure 3). It's challenging to understand why these weird things happen, let

alone fine-tune correctly. This resulted in trial and error making up a large portion of our work. As an aside, we learned that the aforementioned behavior in Figure 3 reflects how, intuitively, PPL causes generation better reflecting the prompt, but as reflected in the original paper, sometimes at the expense of the quality of the original subject. This was an exaggerated case, but all of these oddities of fine-tuning diffusion models combined with the difficulties of actually training/running the model presented an excellent learning opportunity for our group.

VII. References

[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” 2023.

[2] P. von Platen et al., Diffusers: State-of-the-art diffusion models. GitHub, 2022. [Online]. Available: https://github.com/huggingface/diffusers

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve´ Je´gou, 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.

Figures

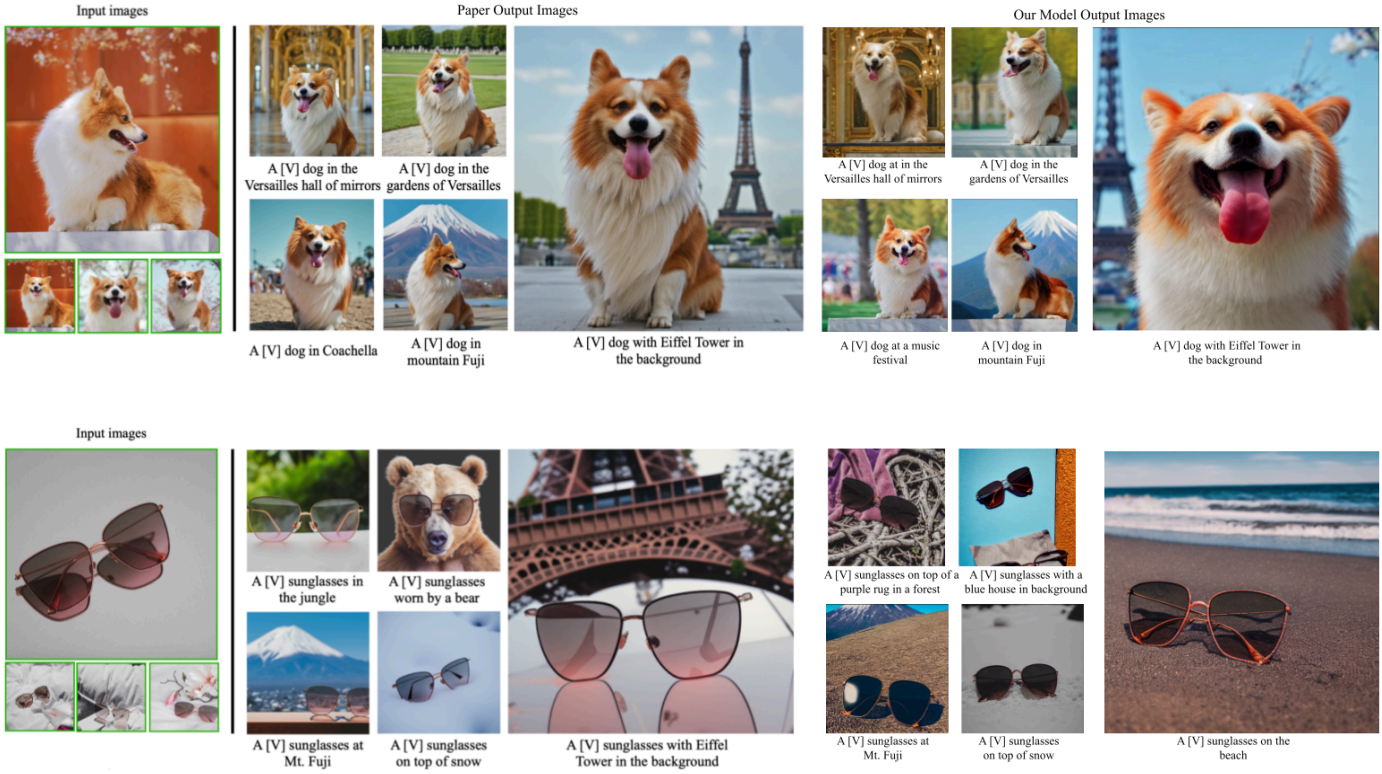

Figure 1: An extended version of Figure 12 from the original paper [1], showing a side-by-side comparison between outputs from the original Imagen-based model and our reimplementation using Stable Diffusion v1.5. For the sunglasses subject, we included alternative prompts in cases where the original ones produced unsatisfactory results.

Figure 2: Inspired by Figure 13 from the original paper [1], we generated our own artistic renditions of another dog instance. The first two images adopt the style of Vincent Van Gogh, the third mimics a bold pop art aesthetic, and the fourth reimagines the dog as a public statue.

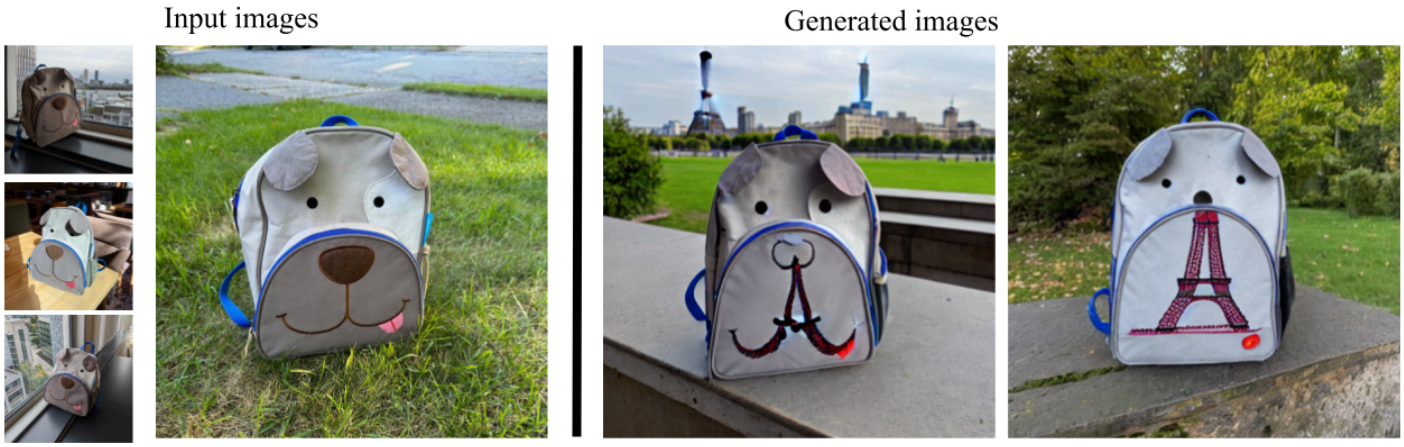

Figure 3: A failure case from our reimplementation, where the model misinterpreted the prompt and rendered the Eiffel Tower as part of the backpack design rather than placing the subject in the intended background.

Tables

Method	PRES ↓	DIV ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
DB (Imagen, PPL)	**0.493**	**0.391**	0.684	0.815	**0.308**
DB (Imagen)	0.664	0.371	**0.712**	**0.828**	0.306
DB (Stable Diff., Ours, PPL)	**0.288**	**0.636**	0.556	0.752	0.201
DB (Stable Diff., Ours)	0.316	0.578	**0.693**	**0.804**	0.201

Table 1: An extended version of Table 3 from the original paper [1], comparing results from our Stable Diffusion reimplementation with the original DreamBooth models using Imagen.