

DREAMBOOTH: FINE TUNING TEXT-TO-IMAGE DIFFUSION MODELS FOR SUBJECT-DRIVEN GENERATION [1]

Grant Rinehimer¹, Adam Faridi¹, Armaan Tewary¹, Ignazio Perez Romero¹, Elliot Kim¹

¹Cornell University

Introduction/Motivation

Recent text-to-image models like Stable Diffusion and DALL-E 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 (Figure 1) in different contexts.

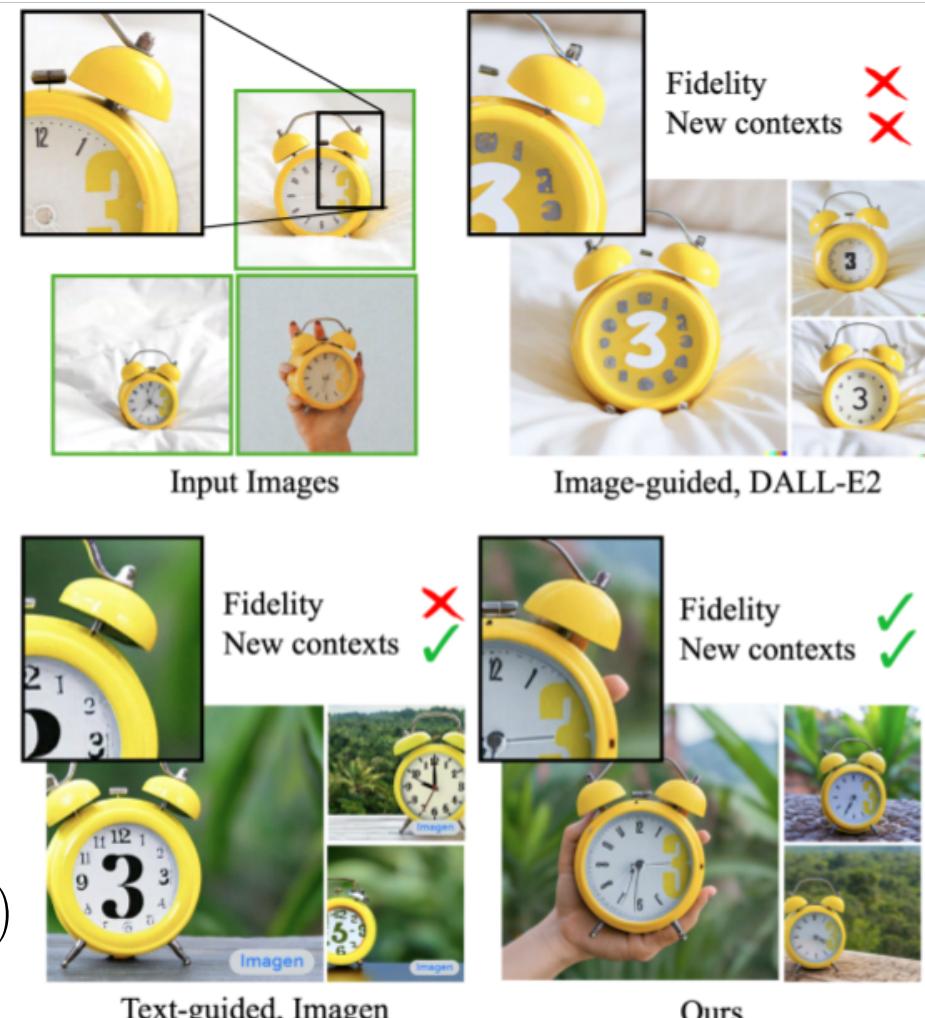


Figure 1: Comparison of subject-driven generation methods [1]

GOALS

- Re-implement DreamBooth to personalize pre-trained T2I models.
- Enable consistent generation of a subject in diverse contexts.
- Evaluate with fidelity metrics: **DINO**, **CLIP-I**, and **CLIP-T**.
- Analyze impact of **Prior Preservation Loss (PPL)** on performance.

Methodology

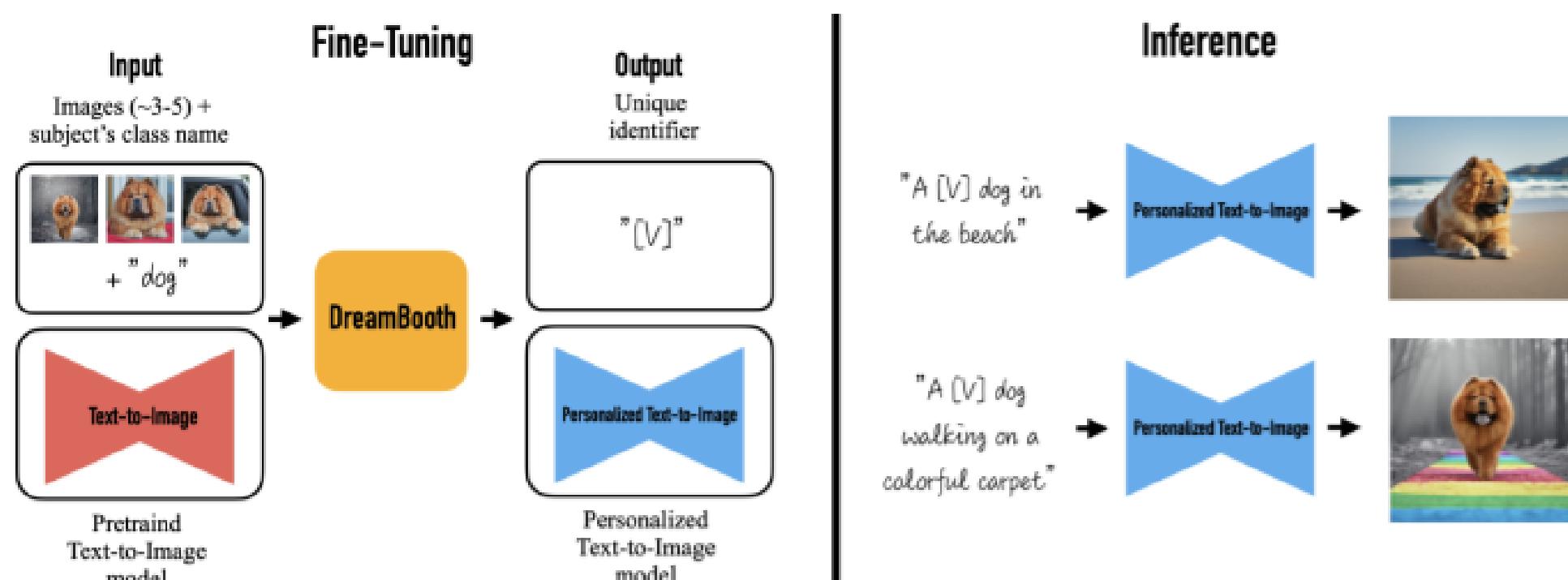


Figure 2: Overview of Dreambooth fine-tuning and inference process [1]

We use an existing dataset of 30 subjects (21 objects, 9 live subjects/pets) and 8 prompts. For the evaluation suite, we generate two images per subject and per prompt, totaling 480 images.

Fine Tuning, Loss and Architecture

Fine Tuning

1. Use Prior Preservation Loss
⇒ Mitigates loss of class diversity while preserving subject!
2. Use low and high resolution crops of subject images
⇒ Enhances model's ability to generalize with quality detail.

$$\text{Preservation Term} \\ \mathbb{E}_{x,c,\epsilon,t} \left[w_t \| \hat{x}_\theta(\alpha_t x + \sigma_t \epsilon, c) - x \|_2^2 \right] + \\ \lambda \mathbb{E}_{x_{pr},c_{pr},\epsilon',t'} \left[w_{t'} \| \hat{x}_\theta(\alpha_{t'} x_{pr} + \sigma_{t'} \epsilon', c_{pr}) - x_{pr} \|_2^2 \right]$$

Diversification Term

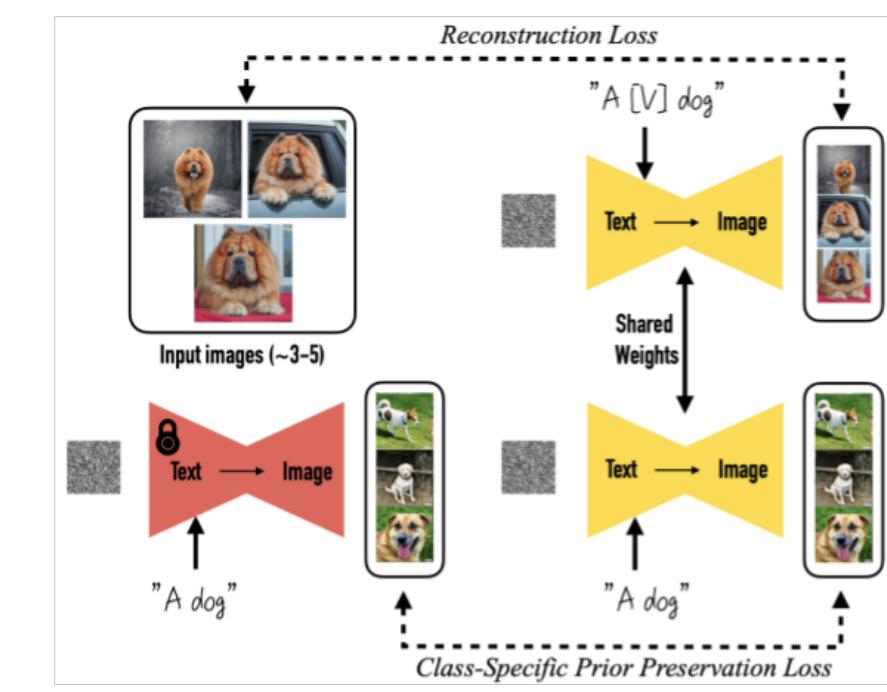


Figure 3: Dreambooth Architecture [1]

As in the paper, for the CLIP Tokenizer, we use **rarer tokens** located at the upper indices of the Tokenizer, which have weak priors.
⇒ Avoids confusing model with tokens like "dog" with strong semantics.

Results

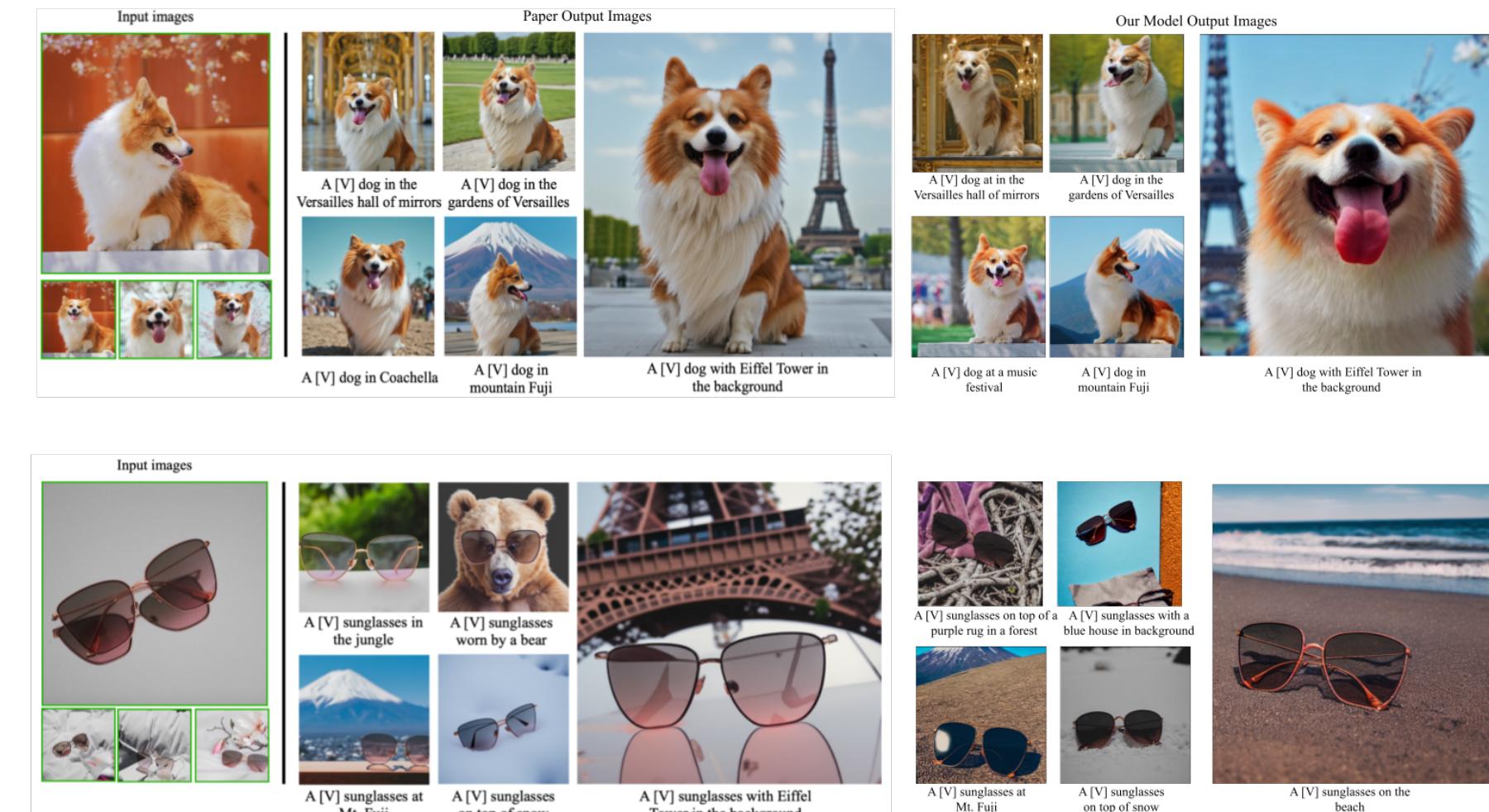


Figure 4: Comparison of recontextualization samples of a dog and sunglasses subject instances between paper output and our model output. [1]

We observe that our model (fine-tune Stable Diffusion v1-5) produces similar results with the paper model (fine-tune Imagen), with slight differences in fidelity.

Results cont.

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. Comparison of subject-driven generation methods.

Metric	Meaning
PRES	measures collapse of the "prior class onto my subject"
DIV	measures avg. cosine sim. of gen. images of same subject & prompt.
DINO	measures fidelity to the "subject I asked for"
CLIP-I	measures visual sim. of subject input and gen. images
CLIP-T	measures semantic alignment of the gen. image and prompt

Additional Results



Figure 5: Dog instance in the style of Vincent Van Gogh, pop art, and statue
Input images Generated images



Figure 6: Instances where Eiffel Tower was generated on backpack, not background

Conclusion & Future Work

- Still observed similar trends like PPL in fine-tuning increases diversity and prior class preservation. Also, PPL needs **hella training**.
⇒ Results produced with PPL adhere to prompt better, at the cost of image quality/detail.
- As in original paper, PPL did not increase subject preservation.
- In the future, improve PPL model and try new subjects, like Kilian.

References

- [1] 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, 2023.