Customizing and Enhancing Stable Diffusion for Style LoRA and Subject-Specific Generation

Katherine Chen

Department of Computer Science, Stanford University

This project involves 1) fine-tuning **Stable Diffusion** using **LoRA** on a **dataset of Carla Cordelia's art style**, and 2) exploring **personalized text-to-image generation** using **Textual Inversion** and **Dream-Booth** on a **dataset of a stuffed animal cat**.

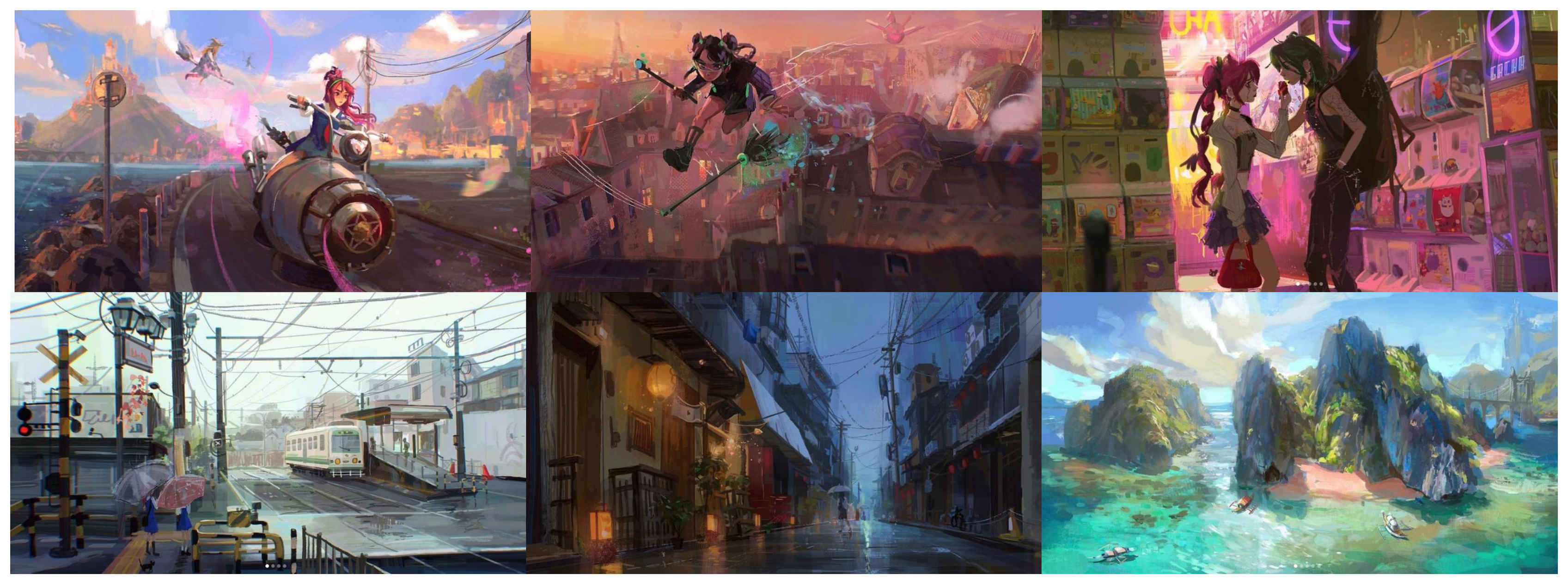


Figure 1. Images from the dataset featuring artist Carla Cordelia's digital paintings.

Figure 2. Images from the subject-specific dataset of my stuffed animal cat, Erica.

Problem & Background

- Default project with no extensions
- **Stable Diffusion** (text → image) is powerful but not enough for custom use cases
- Can we have the model output **images of a certain artist's style**?
- Can we have the model output **images of a certain subject in diverse contexts**?

Dataset

Dataset	Number of images	Dimensions	Used for
Cordelia's art	122	890 x 490 pixels	**LoRA**
Stuffed animal cat (Erica)	10	3024 x 4032 pixels	**Textual Inversion DreamBooth**

Methods

Fine-tuning with LoRA (**Lo**w-**R**ank **A**daption)

- **Goal**: Capture Cordelia's unique **art style**
- **Motivation**: Downstream fine-tunings have low intrinsic dimension
- Train weight matrix to be in the form of **W + AB^T**
  - **W** is the **pretrained weights** (kept frozen)
  - **AB^T** is the rank-r **residual matrix** (fine-tuned), where **r << min(d, k)**

Subject-specific fine-tuning: Textual Inversion

- **Goal**: Generate images of my stuffed animal cat based on textual descriptions
- **String with <placeholder> → tokens → embeddings → text transformer**
- Embedding vector optimized with reconstruction objective against <placeholder>

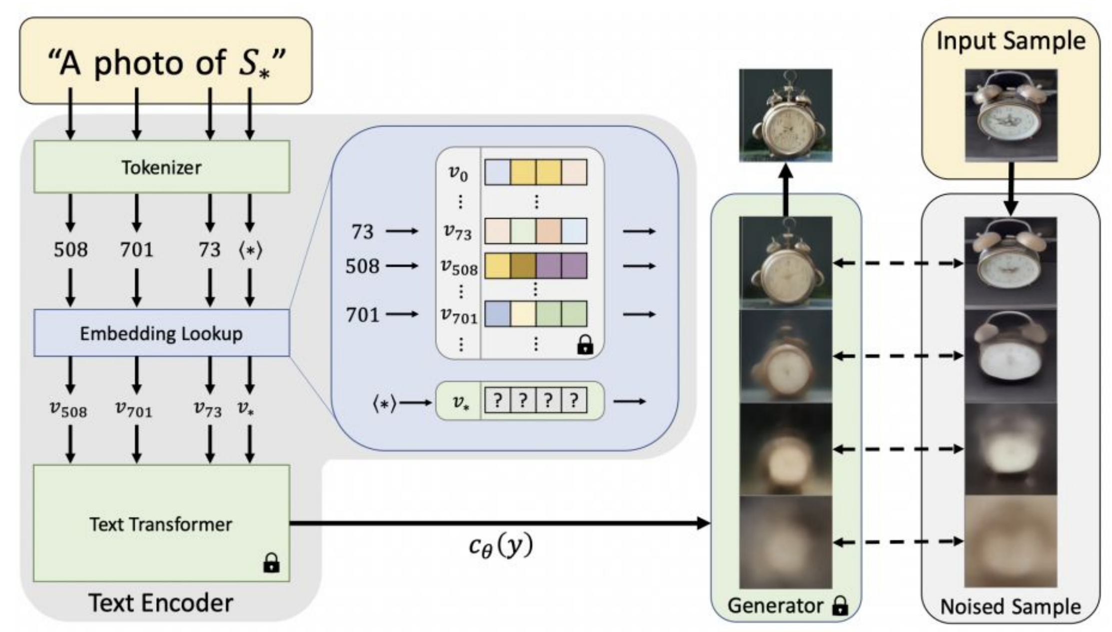


Figure 3. Textual Inversion model architecture.

Subject-specific fine-tuning: DreamBooth

- **Goal**: Same as **Textual Inversion**
- Model leverages existing semantic knowledge on the subject's **class**

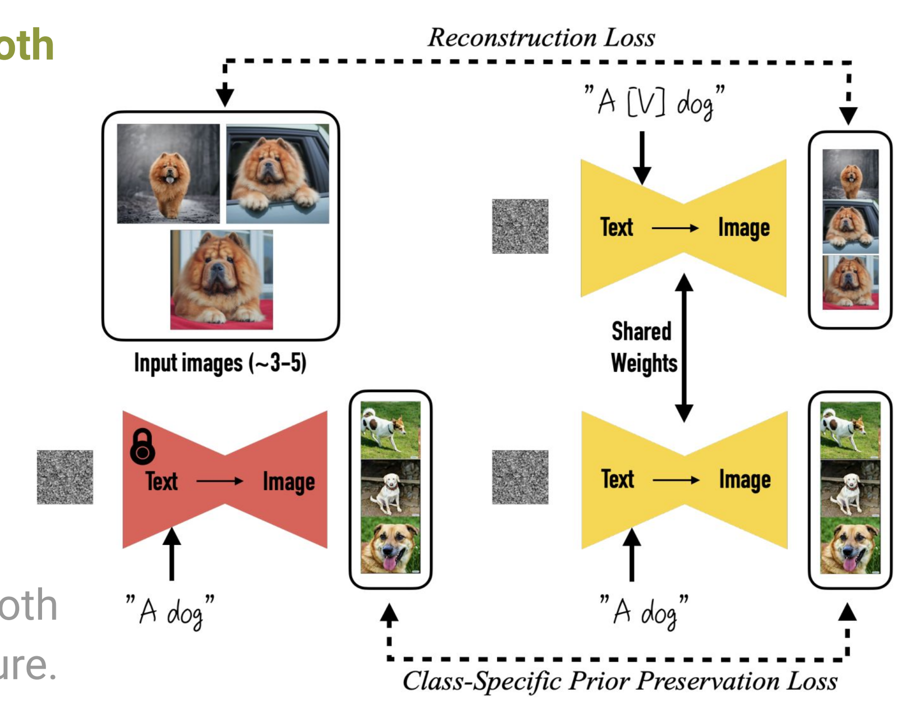


Figure 4. DreamBooth model architecture.

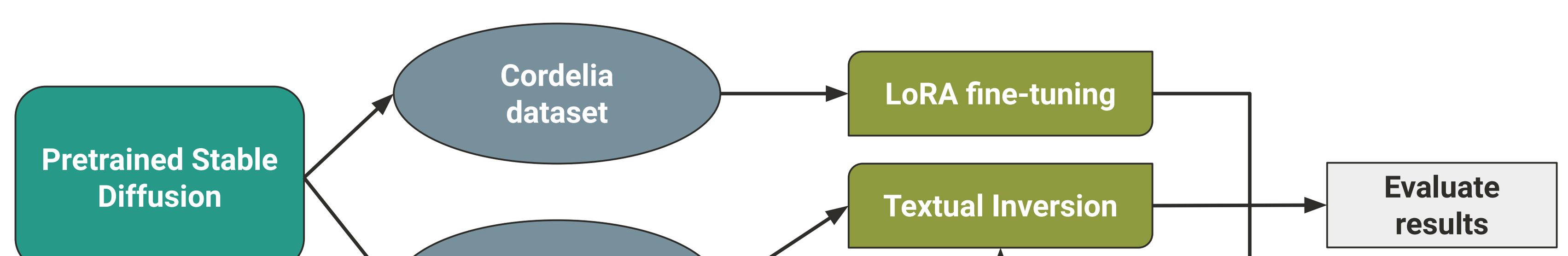


Figure 5. Experimental procedure.

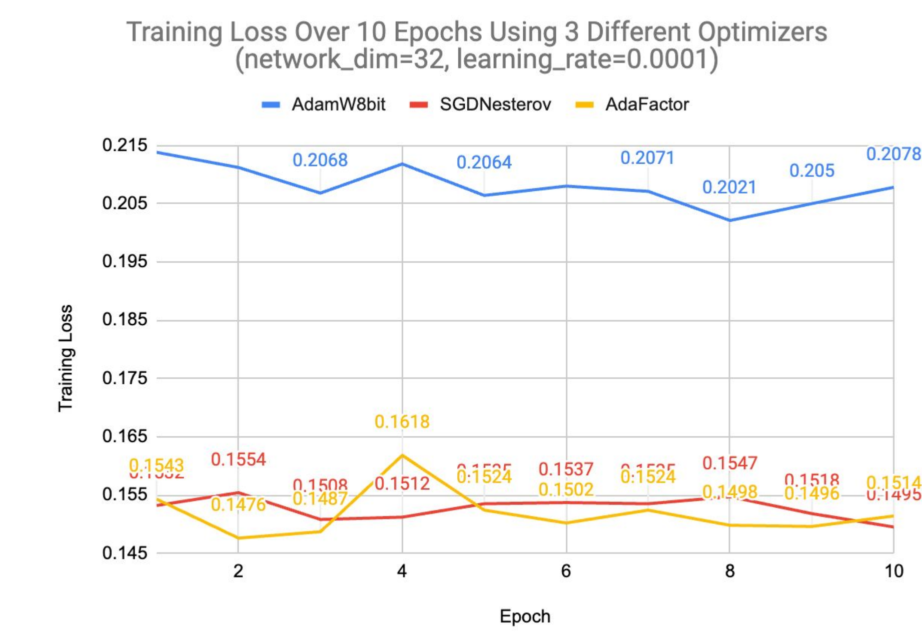
Experiments & Results

"Girl with pink braids in a fantasy city scenery" | "Cat in city streets under neon lights" | "Train with falling sakura over a river"

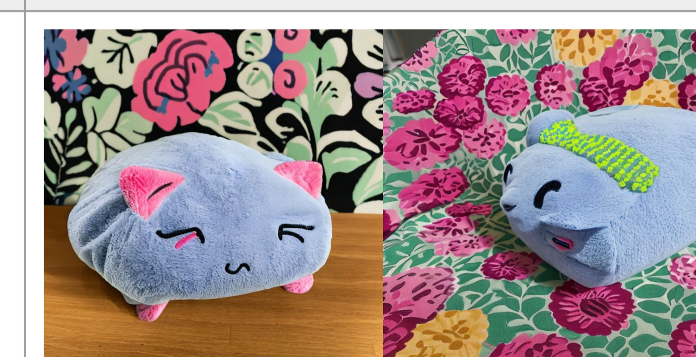
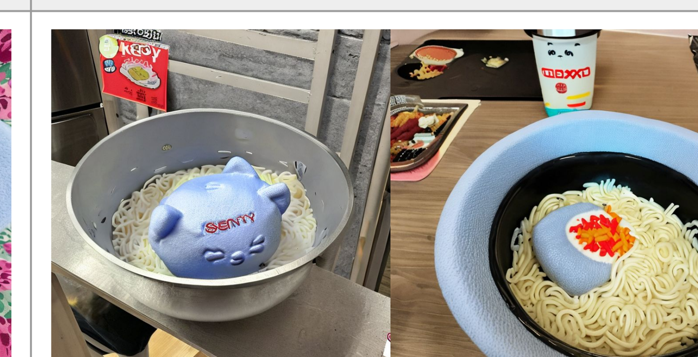
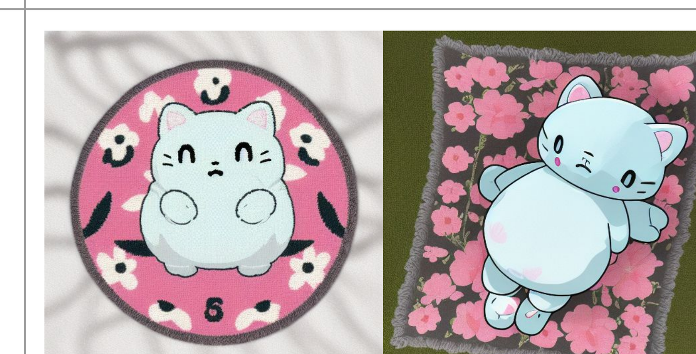
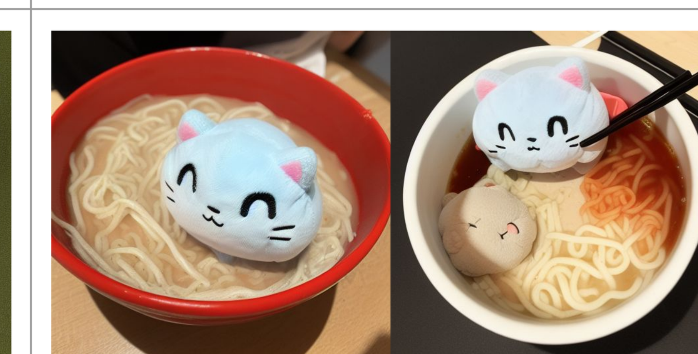
Figure 6. Images generated by LoRA after fine-tuning on the dataset of Cordelia's art.

- **LoRA** → qualitatively did well in replicating Cordelia's art style
- **AdaFactor** & **SGDNesterov** were better optimizers than **AdamW**
- **Hyperparameter tuning** on **learning rate** (with AdaFactor)

Learning rate	5e-5	1e-4	5e-4
Avg loss	0.1596	0.1518	0.1959



- **Textual Inversion** → overall worse quality, many deformations, but good texture
- **DreamBooth** → well-defined body shape & facial features, sometimes anime style

Method	"<erica> on a floral-pattern towel"	"<erica> in a ramen-bowl"
Textual Inversion		
Dream-Booth		

References

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion.

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022