Fine-Tuning Stable Diffusion for Personalized Fashion Applications

Miria Feng - EE367, Stanford University

Abstract

Online shopping is a common denominator in the modern lives of many users, but most of us have purchased and returned items since reality did not meet our expectations. This project aims to reduce the tedious and wasteful cycle by assisting users in creatively and efficiently selecting current fashion items which suits their tastes! **Project summary and contributions:**

- User input: 3-4 photos of themselves plus a text prompt describing their desired outfit. Such as "My next Halloween costume for a Harry Potter themed party".
- Model output: 4 generated images of composite fashion items on the individual user. One 2-3 second video of the user doing a spin in their selected favorite outfit.
- The user will then be able to click and purchase any items directly from Luisaviaroma Online.
- Composite fashion items are generated conditioned on the user's input prompt, but are immediately relevant. This is critical since the creative area of fashion adapts at a rapid pace. What is relevant and available for purchase yesterday may not be available today.

Custom Dataset Pipline

Immediately relevant data: The custom dataset is made from e-commerce site Luisaviaroma[1], which features high resolution photos and quick daily updates of new items.

- The web scraping code continuously downloads images, thus populating a current and relevant fashion dataset.
- Selenium and BeautifulSoup libraries navigate the website and extract information about each item, including the designer and description, thus creating image-text pair descriptors.
- Multiprocessing speeds up the scraping process, and stores a record of all the articles that have been scraped in a SQLite database to prevent duplicates.
- Data is then padded to 512x512 on white or neutral backgrounds. Each folder now contains 3-5 sample images of new fashion items, with the folder name acting as its text descriptor.

Problem Setup and Related Work:

- Stable Diffusion: We experiment with 2 Stable Diffusion models - Stable Diffusion v1.4 and Stable Diffusion XL which has more parameters allowing more expressiveness. This component provides the generative semantic prior embeddings.
- Dreambooth [4]: enables "personalization" by implanting the individual user as a subject in the output domain via an unique identifier. Therefore each generated image or gif will be of the user themselves. The Dreambooth method works since its loss is:

$$\mathbb{E}x, c, \epsilon, t \left[|\epsilon - \epsilon\theta(z_t, t, c)|_2^2 + \lambda \, |\epsilon' - \epsilon_\theta(z't', t', cpr)|_2^2 \right] \quad (1)$$

- LoRA[2]: injects trainable rank decomposition matrices into layers of the architecture to reduce number of trainable parameters for downstream tasks.

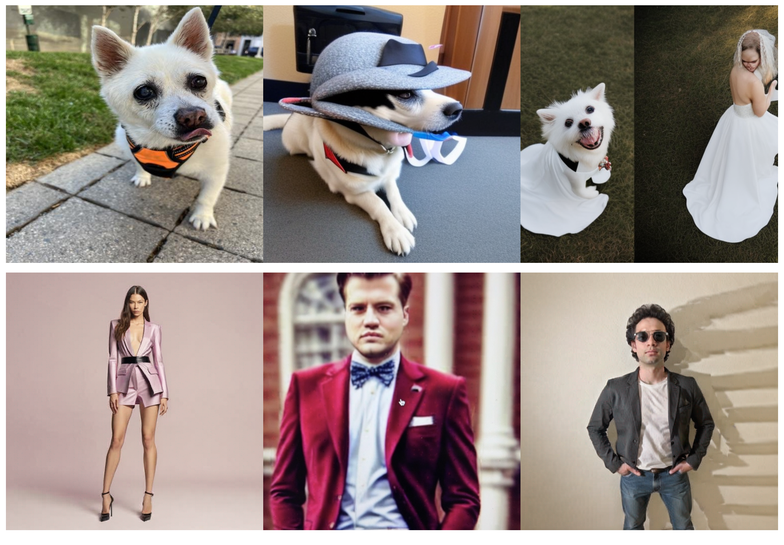
Methods and Experiments

The project is composed of two parts: the image generation then followed by the short video generation. Throughout all experiments we utilize the initial 4 photos of the same user to ensure fair comparisons. User1 is Lucy the dog, User2 is human.

1. **Image generation baseline:** Fine-tuned Stable Diffusion v1.4 with Dreambooth.
2. **Image generation model:** Fine-tuned Stable Diffusion XL[3] with Dreambooth and LoRA, followed by extensive hyperparameter optimization.
3. **Custom CLIP Model:** CLIP is a text and image encoder which aims to increase cosine similarity between text prompt and generated image. This is responsible for the conditional portion of the model. We utilize the Open-CLIP foundation and add additional layers to fine-tune CLIP on the custom dataset. This ensures the generated output is of currently relevant and purchasable fashion items even from user specific brands. We then combine the custom CLIP model with different combinations of baseline models to perform ablation studies.

We also conduct experiments in JAX and FLAX for acceleration, and note the speedup trade-offs with more VRAM consumption. All experiments ran on RTX4090.

Main Results and Examples of Output



- Row 1: the baseline User1 input image and initial baseline generated image. Third image row 1 shows the result of embedding 2 users with different tokens in Dreambooth on one single training run.
- Row 2: SDXL base fine-tuned generated images on User2, User3, and User4 with currently purchasable items. Notably, the Users remain facially recognizable but details (such as shadows and body width in User4's image) are distorted.

Evaluations

- We evaluate with 4 real life users providing prompts, since this is considered the current gold standard. Interface is deployed front-end with FLASK. Please see Ipad for video.
- Addition evaluation metrics include FID, VRAM usage, and generation times in paper.

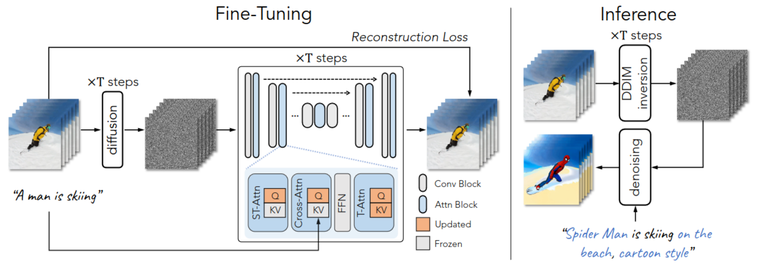
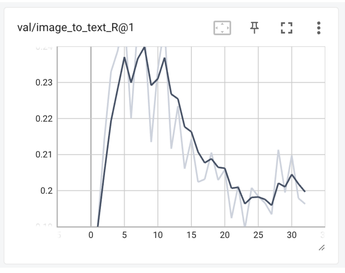


Figure 1. Low VRAM Video Genertaion



- Figure (1): shows the approach to training the video generation portion. This method works since the video motion for each user is the same (created in OpenPose)
- Left:
  Training the custom CLIP model on the custom scrapped dataset.

Conclusions and Next Steps

- Promising results of personalized fashion composite generations thus far. Images are faithful to the individual user.
- When models are fine-tuned on multiple users, there does exist some over-lap.
- Video generation results show promise, but require more fine-tuning to be faithful to the individual user. However this creates a feasible and cheap way to generate short videos without GPU VRAM shortage issues.
- Current front end developed with Flask on web server. Accepts user input, and performs inference in real time with the image generation and video generation models on RTX 4090 and 5090 backend.

References and Acknowledgements

Acknowledgements: Thanks to the EE367 teaching team!

[1] Dataset made from: https://www.luisaviaroma.com/.

[2] 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.

[3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[4] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.