Diffusion Models for Conditional and Personalized Generation

Miria Feng - miria00

1 Motivation and Project Description

The project proposes a personalized generative stylist application, which aims to allow users to upload 4 images of themselves, and visualize outfits on the individual. This allows users to virtually try on fashion pieces before committing to purchase, and reduces likelihood of shipping back for returns. Given 3–4 photos of themselves and a natural language text prompt describing a desired outfit (e.g., "My next Halloween costume for a Harry Potter themed party"), the model generates composite fashion images depicting the specific user wearing currently purchasable items from online shopping platforms. The user can browse the generated options, select their favorite, and click to purchase, after viewing how they themselves will look in the outfits.

This is a **personalization project, with a personalization video component**: beyond personalized images, we propose to also generates a 3-second video of the user doing a spin in their selected outfit, providing a more immersive sense of fit and appearance before purchase. By letting users see themselves in an outfit rather than relying on how it looks on the standard default model, we reduce the uncertainty that drives returns and create a more sustainable experience. Critically, the generated fashion items are immediately purchasable and not outdated inventory. This requirement motivates a custom, continuously updated dataset pipeline and a fine-tuned CLIP model.

2 Related Work

This project builds on generative modeling literature across three areas of diffusion models, personalization, and LoRA fine-tuning.

Diffusion Models. Diffusion models were introduced in Lecture 11. Podell et al. [2023] proposed Stable Diffusion XL (SDXL), which uses a larger UNet backbone and a two-stage refinement pipeline to improve high-resolution image generation. We had previously experimented with both Stable Diffusion v1.4 and SDXL, and note that the latter demonstrates greater expressiveness.

DreamBooth. Ruiz et al. [2023] proposed DreamBooth, a method for "personalizing" text-to-image diffusion models by fine-tuning them on a small set of subject images (3–5 photos) associated with a unique text input. The method optimizes a combined loss that preserves the model's prior knowledge while embedding the new subject into the output domain:

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

where the second term is a class-specific prior preservation loss. This enables each generated image to faithfully depict the specific user.

LoRA. Hu et al. [2022] introduced Low-Rank Adaptation (LoRA), which injects trainable low-rank decomposition matrices into existing layers of a pre-trained model. Rather than fine-tuning all parameters, LoRA dramatically reduces the number of trainable parameters for downstream tasks while maintaining competitive performance. We apply LoRA in conjunction with DreamBooth to efficiently fine-tune SDXL for user-specific generation.

Additional related work includes virtual try-on methods and pose-guided generation models. However, most existing virtual try-on approaches require garment-specific segmentation masks or paired training data, limiting their flexibility. Our approach instead leverages the semantic understanding of large-scale diffusion models to generate composite outfits from free-form text prompts, conditioned on real, purchasable items through a custom CLIP model.

3 Project Overview and Goals

The project consists of four main components:

1. Custom Dataset Pipeline. We build a continuously updated fashion dataset by scraping the e-commerce site Luisaviaroma, which features high-resolution product photos with daily updates. Using Selenium and BeautifulSoup, we extract images along with designer and description metadata to form image-text pairs. Multiprocessing accelerates the scraping process, and a SQLite database tracks previously scraped articles to prevent duplicates. Images are padded to 512×512 on neutral backgrounds, with each folder containing 3–5 sample images and the folder name serving as the text descriptor.

2. Personalized Generation. We fine-tune SDXL with DreamBooth and LoRA to generate images of the user wearing composite fashion items. The baseline uses Stable Diffusion v1.4 with DreamBooth alone. The full model uses SDXL with both DreamBooth and LoRA.

3. Conditional CLIP Model. To ensure generated outputs depict currently relevant and purchasable items, we fine-tune a CLIP model (built on the OpenCLIP foundation with additional layers) on our custom dataset. This model maximizes the cosine similarity between the user's text prompt and generated images, grounding the generation in real fashion items from specific brands. We combine the custom CLIP model with different baseline configurations for ablation studies.

4. Video Generation. The user selects their favorite generated image, and the system produces a 3-second video showing a spin/motion. We create short animations using OpenPose guidelines, label them as mini video-text pairs, and fine-tune a text-to-video model by updating projection matrices in attention blocks using standard diffusion training loss. All experiments run on an RTX4090, RTX5090, and A100s.

Final Goals: (1) a working end-to-end pipeline from user photos and text prompts to generated fashion images and short video, (2) a Flask-based web interface for real-time inference as a demo during poster presentation, (3) quantitative evaluation using FID, VRAM usage, and generation time metrics for the write-up.

4 Milestones and Timeline

- Build and validate the custom dataset scraping pipeline. Collect initial dataset of ∼5,000 image-text pairs from Luisaviaroma. Set up training infrastructure on RTX 4090. *(Feb 20 – Feb 23)*

- Setup the baseline model (SD v1.4 + DreamBooth). Establish baseline FID and qualitative results. *(Feb 24 – Feb 26).* Fine-tune SDXL with DreamBooth + LoRA. Examine sensitivity to hyperparameter tuning. Train the custom CLIP model on the scraped dataset. *(Feb 27 – Mar 1)*

- Ablation studies combining custom CLIP with different generative models. Implement the video generation pipeline using OpenPose and text-to-video fine-tuning. *(Mar 2 – Mar 4).* Build the Flask web interface for end-to-end real-time inference. Explore JAX/FLAX acceleration and benchmark speed vs. VRAM tradeoffs. *(Mar 5 – Mar 7).* Conduct real user studies with 10 participants. Compile final quantitative metrics (FID, generation time, VRAM). Prepare final report and poster. *(Mar 8 – Mar 10)*

References

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 1(2):3, 2022.

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.

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.