Personalized Image Generation and Inpainting
Team: Visionary Creators

Chi-yeh Chen chiyehc
Samuel Johnny sjohnny
Tyler Harp tharp
Oyeon Kwon oyeonk

1 Introduction and Motivation

The rise of personalized digital content has driven the demand for real-time image generation and customization tools. Current text-to-image generation technologies such as *Stable Diffusion*(Jiang et al., 2022a) and *DALL·E*(Ramesh et al., 2022) are powerful but often lack precision in composition and scene understanding. This project aims to build a real-time personalized fashion styling system that integrates image generation and inpainting to produce custom visuals based on user input.

One of the challenges we aim to tackle is the performance of the *AnyDoor*(Chen et al., 2024) model on fashion-specific datasets. For example, the *AnyDoor* model may not perform well on specific textures or small objects in fashion datasets. To address this, we propose targeted improvements such as data augmentation, segmentation modeling, and fine-tuning based on domain-specific assumptions.

2 Contributions

Our project makes the following key contributions:

- We propose a real-time image generation and inpainting system for personalized fashion styling using state-of-the-art models like *AnyDoor*.
- We benchmark the *AnyDoor* model on multiple fashion datasets and identify performance bottlenecks, such as handling specific textures and small objects. To address these, we propose enhancements such as data augmentation and segmentation modeling.
- We explore text-to-image generation by fine-tuning models like *DALL-E* on fashion-specific datasets, enabling users to input descriptive prompts and generate reference outfits.

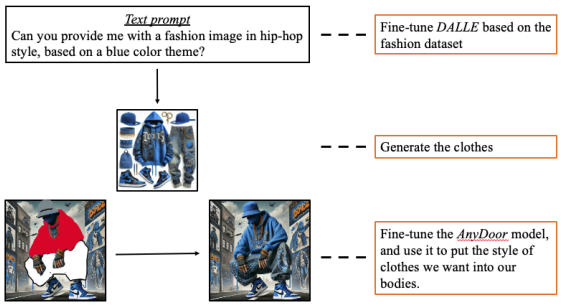

Figure 1: Workflow

3 Task Definition

3.1. **Data Modalities:** Inputs include textual descriptions and fashion-related images. Outputs will be AI-generated outfits or inpainted images based on fashion data.

3.2. **Tools, Libraries, Frameworks:** We plan to use state-of-the-art tools such as *AnyDoor*, along with custom modifications. The use of generative models like *DALL·E* for text-to-image and *AnyDoor* for real-time inpainting will serve as our core methodology.

3.3. **Datasets:** We will utilize several high-quality datasets focused on fashion products:

- iMaterialist Fashion Dataset: The iMaterialist Fashion dataset contains a large number of images and their fashion segmentations. This dataset contains images of people wearing a variety of clothing types in different poses.It contains 46 apparel objects and 92 related fine-grained attributes; the dataset contains around 50k clothing images(david shi, 2019).
- Fashion Product-Images: This dataset contains high-resolution images of different clothing used by humans; it contains their descriptions for each object in the

image and details the location it should be worn(Aggarwal, 2019).

- DeepFashion-MultiModal: is a large-scale, high-quality human dataset with rich multi-modal annotations. It has the following properties: It contains 44,096 high-resolution human images, including 12,701 full-body human images; each class label, keypoint, and cloth shape and texture were manually annotated, and there is a textual description for each image(Jiang et al., 2022b),(Liu et al., 2016).

3.4. **Evaluation Metrics:** Model performance will be evaluated using image generation quality metrics (e.g., FID, Inception Score), along with user satisfaction surveys and visual fidelity assessments. Benchmarking against the *AnyDoor* model will be critical.

4 Benchmarking

We plan to evaluate *AnyDoor*(Chen et al., 2024) on various fashion-specific datasets to identify potential areas of poor performance. Through this evaluation, we aim to find tackling points and suggest improvements. For example, the *AnyDoor* model may not perform well on specific textures or small objects in fashion datasets. To address this, we propose the following enhancements:

4.1. **Data Augmentation:** We will apply techniques like texture-based augmentation and jittering to improve model robustness to various fashion textures.

4.2. **Segmentation Modeling:** Focus on refining the model's ability to segment and differentiate between different fashion elements like fabrics, textures, and accessories.

4.3. **Custom Dataset Generation:** Using the existing public datasets, we will generate new training data to fine-tune the models specifically for fashion-based scenarios.

4.4. **Image Generation Improvements:** We will explore further improvements in text-to-image generation by fine-tuning model on fashion-specific data, enabling the generation of outfits from text prompts.

5 Project Steps and Timeline

Step 1: Benchmark *AnyDoor* performance using fashion datasets. We will access the performance of the dataset listed in Section 3.3 to assess where the model performs poorly.

Step 2: Identify performance bottlenecks such as handling textures, small objects, or intricate patterns.

Step 3: Implement improvements such as data augmentation, segmentation modeling, and custom dataset generation.

Step 4: Explore text-to-image generation and fine-tune it on fashion-specific datasets for improved results.

6 Optional: Text-to-Image Generation

Leveraging models like *DALL-E*, we will fine-tune them on fashion-specific datasets. This will allow users to input text descriptions and have the system generate a reference outfit, which can be refined by fitting it into the *AnyDoor* model for customization.

References

Param Aggarwal. 2019. Fashion product images dataset.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6602.

Menglin Jia Mihail Sirotenko Will Cukierski david shi, Maggie. 2019. imaterialist (fashion) 2019 at fgvc6.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022a. Text2human: Text-driven controllable human image generation. *Preprint*, arXiv:2205.15996.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022b. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.