Fashion Chatroom: An Automated Pipeline for Fashion Dataset Construction

Yiwen Zhao, Huizhu Jia, Shanghang Zhang[*]

National Key Laboratory for Multimedia Information Processing,
School of Computer Science,
Peking University

Abstract

In recent times, Generative AI has made its mark in the fashion industry. Fashion model datasets play a crucial role in training generative models for the companies' design and sales processes. However, existing datasets encounter issues like inconsistent image quality and limited fashion attributes. Collecting these datasets can also be labor-intensive, and there are legal concerns surrounding their use. To address these challenges and alleviate the scarcity of high-quality fashion model data, we present an autonomous pipeline that combines the power of Large Language Models (LLMs), Visual Language Models (VLMs), and Diffusion Models for dataset construction, effectively reducing the need for extensive human labor. This approach facilitates better utilization of AI in the fashion domain.

1. Introduction

The fashion industry has witnessed a notable shift towards integrating generative AI within its design and sales processes. Traditionally, fashion businesses invested significant resources in hiring professional models and conducting elaborate photoshoots for promotional purposes. However, with the development of generative AI, they can now opt for more efficient approaches by showcasing clothing on mannequins or on the ground and employing reference-based generation to incorporate the missing human element. Nevertheless, the successful training of specialized models in this domain necessitates access to a large-scale dataset of human clothing data, which is currently lacking

There have been remarkable prior efforts dedicated to the construction of fashion datasets, primarily aimed at tasks such as virtual try-ons, human pose estimation, parsing, or segmentation. High-quality fashion images typically exhibit characteristics such as a clear foreground human, a clean background, full-body views, diverse poses, and various clothing items. One common approach for data acquisition involves scraping publicly available images from the internet; however, this method often results in an assortment of image quality. Additionally, ensuring the authenticity of the data presents a legal challenge. Alternatively, synthetic datasets have been used, offering cleaner backgrounds and full-body distributions, but existing datasets' limited pose variety and labor-intensive data generation process make them inconvenient and costly.

Motivated by the remarkable progress in generative models and the widespread utilization of Large Language Models (LLMs) and Visual Language Models (VLMs), we propose an innovative approach to dataset creation. Our method upholds image quality through a standardized generation process and avoids secondary modifications that may compromise quality, such as cropping human subjects from noisy backgrounds and pasting them onto a white setting. Importantly, our approach respects data privacy. Furthermore, it introduces an autonomous pipeline that minimizes human involvement, thus promoting efficiency.

Our proposed pipeline leverages the capabilities of LLMs, VLMs, and Diffusion models. Firstly, we implement community models that are fine-tuned to excel in human image generation, ensuring uniform quality across generated images through a predefined set of hyperparameters. Secondly, we set up a collaborative "chatroom", where large models and human users interact. Through prompt tuning and visual question-answering interactions, we extract attributes from reference images and incorporate them into our dataset, making it boast a rich diversity of textures, human poses, and relatively clean backgrounds. The "chat room" minimizes the need for human intervention in the pipeline, further enhancing its efficiency and effectiveness.

Our contributions can be summarized as follows:

- We introduce an automated pipeline that harnesses the prompt learning capabilities of LLMs and the image comprehension abilities of VLMs, effectively diminishing the human labor involved in the text-to-image generation process.
- This pipeline additionally facilitates the integration of valuable attributes from reference images into the generated outputs, thereby augmenting the volume of data within a user-defined domain while adhering to the expected distribution.
- Through the deployment of this automated pipeline, we have curated an expansive fashion dataset characterized by its diversity in fashion attributes and human poses. These attributes hold substantial promise for enhancing productivity and performance in downstream applications within the fashion domain.

[*] Corresponding author (shanghang@pku.edu.cn)

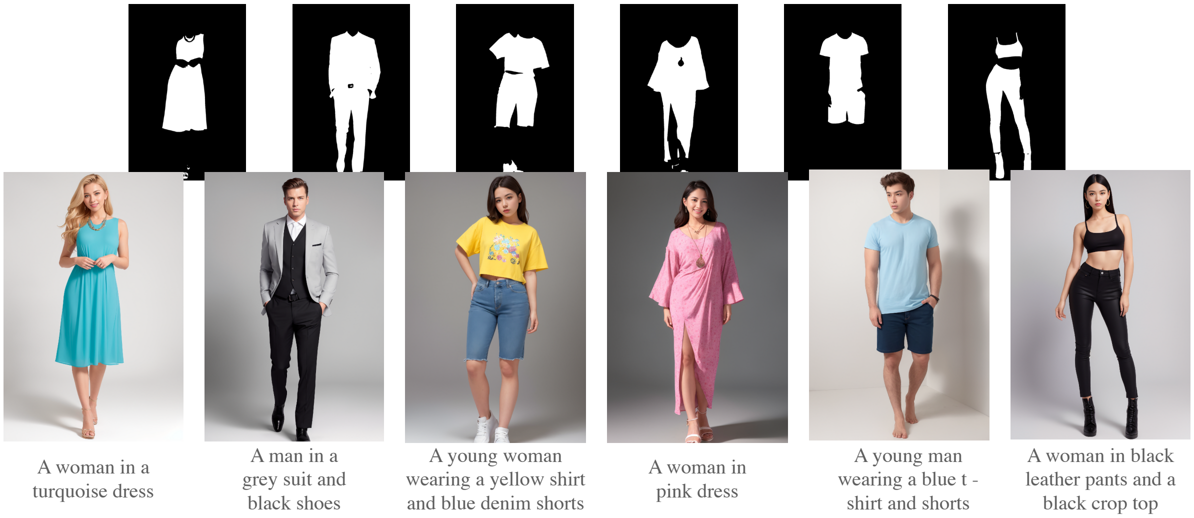
A woman in a turquoise dress
A man in a grey suit and black shoes
A young woman wearing a yellow shirt and blue denim shorts
A woman in pink dress
A young man wearing a blue t - shirt and shorts
A woman in black leather pants and a black crop top

Figure 1: An overview of the generated results in our dataset.

2. Related Works

2.1 Fashion Datasets

Existing fashion datasets can be categorized into two main groups: human-centric and clothing-centric. The SHHQ dataset (Fu et al. 2022) stands out as a realistic dataset with labeled fashion attributes and a hundred percent coverage of full-body images. However, its preprocessing involves substantial human labor. Certain operations, such as background removal, can negatively impact the quality of the human model. The Human-Art dataset (Ju et al. 2023) focuses on images within the art domain, contributing to bridging the gap between normal human settings and artistic human settings. Nevertheless, it falls short of achieving full realism and does not specifically cater to the fashion domain. Text2Human (Jiang et al. 2022) is a synthetic fashion dataset known for its high-quality textures. However, its human poses lack diversity due to its structure-dependent codebook-based generation pipeline.

Conversely, clothing-centric datasets (Surya et al. 2020; Sun et al. 2023) excel in offering a wide range of clothing types and detailed clothing attributes. However, these datasets primarily focus on clothing and do not fully account for the correlated distribution between humans and clothing. To tackle with this, we propose an autonomous pipeline for generating a human-centric dataset that includes diverse human poses, detailed clothing attributes, and clean backgrounds.

2.2 Text-To-Image Diffusion Models

Diffusion models are renowned for their ability to generate high-quality samples. Text-to-image diffusion models have seen significant advancements in recent years. Initially, classifier guidance diffusion was introduced by (Dhariwal and Nichol 2021). This approach involved training a classifier to predict class labels of noisy images, which can improve the performance of the original diffusion model. However, it relies on the gradient of the classifier for updates, which does not fully leverage the diffusion model's pre-trained capabilities. It may also struggle with complex text descriptions that are challenging to represent as a classification problem.

Subsequently, (Ho and Salimans 2022) proposed classifier-free guidance. They collectively conduct conditional training and unconditional training, pushing away the results of the text label and the null label, which leads to diverse results. As the generative AI community evolved, expert models in various domains emerged through fine-tuning the Stable-Diffusion Text-to-Image model. Nevertheless, the prompts for these models were entirely crafted by humans, whereas our pipeline incorporates language models(LMs) to reduce human intervention to generate a large-scale of images, thereby streamlining the process.

3. Automated Dataset Construction

In this section, we elucidate the various components of our pipeline, encompassing GPT (Brown et al. 2020) prompt tuning, attribute extraction, image generation, image selection, and the generation of masks and captions. The holistic pipeline is visually depicted in Figure 2.

3.1 GPT-3.5-turbo Prompt Tuning

In the realm of text-to-image human generating, several high-quality community models have emerged, fine-tuned from the stable diffusion model (Rombach et al. 2022). These models excel at generating intricate human appearance details, including features such as skin textures and hairstyles.

Traditionally, users manually design prompts for each image. However, as we aim to construct a dataset comprising hundreds of thousands of images, the reliance on manual prompt generation is not practical. Furthermore, community models are fine-tuned on different image-text data pairs, necessitating unique prompt formats for each. Obtaining high-quality prompts tailored to specific community

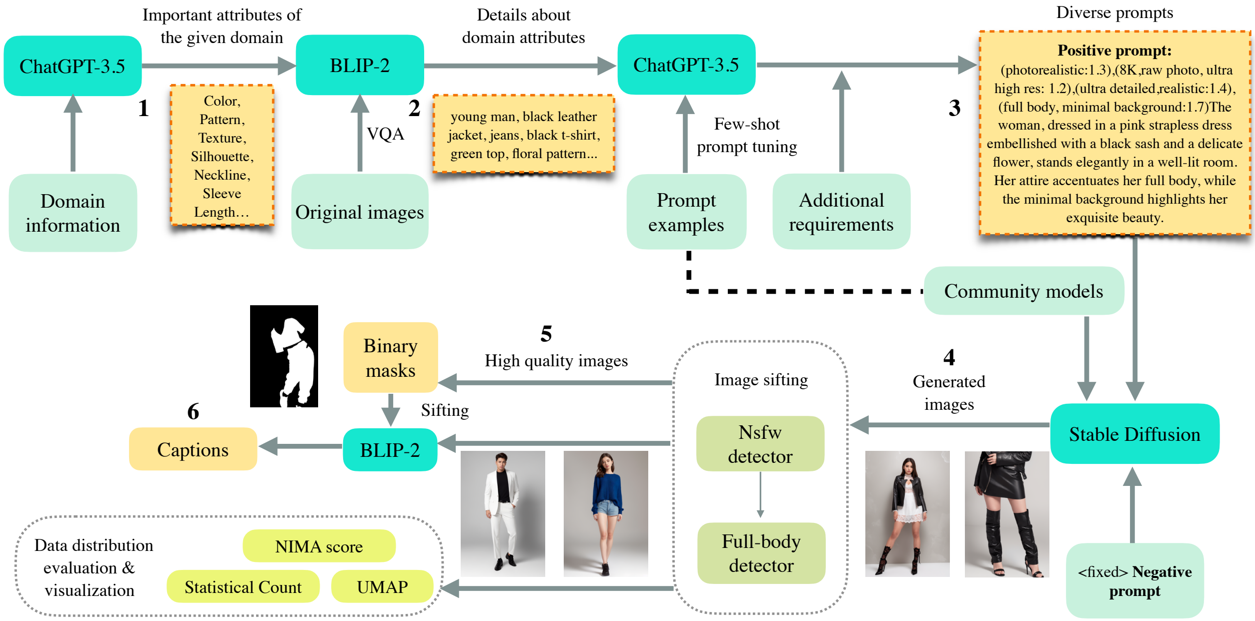
Important attributes of the given domain
Details about domain attributes
Diverse prompts
ChatGPT-3.5
BLIP-2
ChatGPT-3.5
1
Color, Pattern, Texture, Silhouette, Neckline, Sleeve Length…
2
VQA
young man, black leather jacket, jeans, black t-shirt, green top, floral pattern...
Few-shot prompt tuning
3
Positive prompt:
(photorealistic:1.3),(8K,raw photo, ultra high res: 1.2),(ultra detailed,realistic:1.4), (full body, minimal background:1.7)The woman, dressed in a pink strapless dress embellished with a black sash and a delicate flower, stands elegantly in a well-lit room. Her attire accentuates her full body, while the minimal background highlights her exquisite beauty.
Domain information
Original images
Prompt examples
Additional requirements
Community models
Binary masks
5
High quality images
Image sifting
4
Generated images
6
Sifting
Nsfw detector
Stable Diffusion
Captions
BLIP-2
Full-body detector
Data distribution evaluation & visualization
NIMA score
Statistical Count
UMAP
<fixed> Negative prompt

Figure 2: Automated pipeline for dataset construction.

models while preserving the diversity of fashion attributes becomes a challenge.

This is where prompt tuning comes into play. GPT-3.5-turbo and other large language models are renowned for their ability to be fine-tuned with just a few examples. Having been trained on vast datasets, they possess extensive knowledge and the capacity to learn new settings from minimal input. To facilitate this, we survey models from community websites and select a base model for our pipeline. We gather prompts and hyperparameters from high-quality model outputs, treating textual prompts as samples and utilizing widely adopted hyperparameters after conducting statistical analysis. We follow most community models and use Stable-Diffusion WebUI [1] API [2] for image generation.

3.2 Attributes Extraction

We randomly sample a subset from publicly accessible human fashion datasets as references and employ BLIP-2 (Li et al. 2023), a VLM, to extract attributes of the human subjects and their attire. As depicted in the upper half of Figure 2, we initiate this process by providing the VLM with a one-word description of the domain, using its capacity for summarizing essential attributes in the specified domain as references. Subsequently, we prompt the VLM to elucidate how the subjects and backgrounds in the reference images relate to these key attributes. This interactive exchange allows the VLM to furnish specific descriptions of the reference images within the user-defined domain. The results serve as the foundation for prompt generation.

3.3 Image Generation

We present textual prompts collected from the community to GPT-3.5-turbo, fine-tuning it to ensure its generated contents conform to the rules and preferences of the specific model. Notably, the negative prompts used in all instances remain consistent, with variations primarily confined to the positive prompts generated by GPT-3.5-turbo. This practice results in the stable quality of generated images characterized by a multitude of features.

3.4 Image Selection

Given our aim for the generated images to encompass diverse fashion attributes, a variety of human poses, and clean backgrounds that align with the requirements for human-clothing distribution learning, an image-sifting process becomes imperative. This process commences with the filtering of images containing violent or explicit content, which is accomplished through the use of the CLIP-based-NSFW-Detector [3]. Next, we employ Openpose (Cao et al. 2017; Simon et al. 2017) to detect human body joints, as well as to ascertain the number of humans in each image and whether all pre-defined body joints are visible. Based on the outcomes, we select and retain images featuring full-body human models. This step also provides the added benefit of concurrently saving the estimated human skeletons for potential downstream applications, such as serving as a reference for generating custom poses.

3.5 Masks and Captions

A key objective in our downstream tasks is inpainting different parts of the human model, which requires binary masks and image captions in training. While we experimented with

[1] https://github.com/AUTOMATIC1111/stable-diffusion-webui
[2] https://github.com/mix1009/sdwebuiapi
[3] https://github.com/LAION-AI/CLIP-based-NSFW-Detector

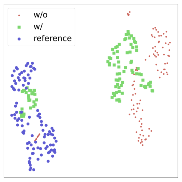

Figure 3: Data distribution by UMAP clustering using clip encoder (Radford et al. 2021) as the feature extractor.

Table 1: Attributes Statistics: the attributes' distribution in three distinct datasets: the original reference images, generated data without attribute extraction, and generated data with attribute extraction.

Data Source	Accessory	Pattern/ Texture	Minimal Background
reference	80%	62%	45%
w/o	43%	8%	51%
w/	68% ↑	24% ↑	97% ↑

various semantic segmentation methods, including Segment Anything (Kirillov et al. 2023), we ultimately opted for the human-parsing method (Gong et al. 2019) due to its relatively accurate delineation of clothing regions.

We utilize three template questions to query BLIP-2 for descriptions of the human models in the images, sequentially asking them in order. If a response is excessively brief or contains nonsensical information, we move on to the next question. Images that do not yield meaningful responses from any of the three questioning methods are simply discarded. Additionally, BLIP-2 evaluates whether the masks exhibit any unusual black holes, allowing us to filter out low-quality mask data. This process culminates in the acquisition of high-quality (image, mask, caption) tuples for our dataset.

4. Experiment

4.1 Inplementation Details

We selected the community model, majicmixRealistic [4], as the foundational fashion image generator. We employed the DPM++ 2M SDE Karras sampler, which was configured to operate at the original resolution of $w \times h = 512 \times 768$. Subsequently, we implemented the R-ESRGAN 4x+ upscaler and introduced a multiplier of 2 to enhance resolution. The classifier-free guidance scale is set to 4 for better diversity. We also followed community conventions by omitting the final block of cross-attention.

4.2 Quality Assessment

To gauge the contribution of our proposed pipeline to the generated images's adherence to the attribute distribution of reference images, we employed three distinct evaluation methods.

1) Statistical Count. We pick important fashion attributes and randomly sample the same amount of data from the original dataset and our generated dataset. We count the existence of each attribute manually and compare data generated with or without the attribute extraction step. As shown in Table 1, attribute extraction helps in preserving data diversity.

Table 2: Image quality evaluation: we use the non-reference evaluation score, NIMA to show the aesthetic quality.

Evaluation Metric	Deepfashion Multimodal	SHHQ	Ours
NIMA ↑	5.2618	4.9425	5.4244

2) UMAP (McInnes and Healy 2018). We utilize this clustering method to further investigate the distribution of the generated images. Image features were extracted using the CLIP-encoder (Radford et al. 2021). The outcomes, showcased in Figure 3, demonstrate the effectiveness of attribute extraction in approximating the distribution of the generated data to that of the original data.

3) NIMA (Esfandarani and Milanfar 2018) non-reference evaluation. We compared RGB images from our dataset with two other human-centric fashion datasets by quantifying their aesthetic quality. As shown in Table 2, our dataset exhibits the highest aesthetic performance.

5. Ethical Statement

It is important to protect data privacy and respect data authenticity. The human images in our dataset are generated using community models. Derived from a learned probability distribution, they do not infringe upon the legal rights of companies or individuals, showing better care in privacy protection compared to collecting public images online. Although bias in data still exists, it can be partly reduced by making the reference images culturally diverse and the attribute extraction step can inject more cultural elements into the generated data.

6. Conclusion

In conclusion, we have introduced an automated pipeline that seamlessly integrates Large Language Models (LLMs), Visual Language Models (VLMs), and Diffusion models to produce a comprehensive fashion dataset. Extracting references from sample images and infusing them into generating, the dataset is diverse in fashion attributes. The standardized generating pipeline and a collection of data-sifting methods ensure the quality of data and reduce ethical concerns. Although we target the fashion domain, this automated pipeline can generate images in other domains, e.g. furniture and architecture as well. The VLM can extract attributes from a small set of reference images and give them to LLM, then LLM combines the extracted attributes with its knowledge about the specific domain to generate novel text descriptions, which serve as prompts for the domain-specialized diffusion model to generate various images. It can be easily switched to the domain that cherishes novelty other than precision, thus better for art and design, to reduce human intervention in dataset construction.

[4] https://civitai.com/models/43331?modelVersionId=94640

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 8780–8794.

Esfandarani, H. T.; and Milanfar, P. 2018. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.*, 27(8): 3998–4011.

Fu, J.; Li, S.; Jiang, Y.; Lin, K.; Qian, C.; Loy, C. C.; Wu, W.; and Liu, Z. 2022. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. volume 13676, 1–19. Springer.

Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; and Lin, L. 2019. Graphonomy: Universal Human Parsing via Graph Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 7450–7459. Computer Vision Foundation / IEEE.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Ju, X.; Zeng, A.; Jianan, W.; Qiang, X.; and Lei, Z. 2023. Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. *CoRR*, abs/2304.02643.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. volume 202, 19730–19742. PMLR.

McInnes, L.; and Healy, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR*, abs/1802.03426.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10674–10685. IEEE.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

Sun, Z.; Zhou, Y.; He, H.; and Mok, P. Y. 2023. SGDiff: A Style Guided Diffusion Model for Fashion Synthesis. *CoRR*, abs/2308.07605.

Surya, S.; Setlur, A.; Biswas, A.; and Negi, S. 2020. ReStGAN: A step towards visually guided shopper experience via text-to-image synthesis. 1189–1197. IEEE.