FINE-GRAINED CUSTOMIZED FASHION DESIGN WITH IMAGE-INTO-PROMPT BENCHMARK AND DATASET FROM LMM

*Hui Li[1], Yi You[1], Qiqi Chen[1], Bingfeng Zhang[2], George Q. Huang[1]**

[1]The Hong Kong Polytechnic University, China
[2]China University of Petroleum (East China), China

ABSTRACT

Generative AI evolves the execution of complex workflows in industry, where the large multimodal model empowers fashion design in the garment industry. Current generation AI models magically transform brainstorming into fancy designs easily, but the fine-grained customization still suffers from text uncertainty without professional background knowledge from end-users. Thus, we propose the Better Understanding Generation (BUG) workflow with LMM to automatically create and fine-grain customize the cloth designs from chat with image-into-prompt. Our framework unleashes users' creative potential beyond words and also lowers the barriers of clothing design/editing without further human involvement. To prove the effectiveness of our model, we propose a new FashionEdit dataset that simulates the real-world clothing design workflow, evaluated from generation similarity, user satisfaction, and quality. The code and dataset: https://github.com/detectiveli/FashionEdit.

***Index Terms*—** Image Editing, LMM, Fashion

1. INTRODUCTION

Generative AI (GenAI) aims to execute complex workflows for humans. As one of the key components in GenAI, the development of the Large Multimodal Models (LMMs) enable new capabilities of AI agents in industry workflows, such as financial analysis [3], industrial solutions [4], specialized assistants [5], and fashion design [6], which are attributed to their rich understanding and execution ability.

In the garment industry, an order begins with customer needs, then goes through the designer, pattern maker [7], tailor [8, 9], and finally ends with delivery [10]. Current LMMs mainly focus on analyzing customers' needs to recommend items as their preferences [11]. However, with the growing demand for personalized clothing, the customer is also willing to be their own designer, who creates and adjusts the design until satisfaction.

AI-generated fashion design focuses on customization based on natural language description (*e.g.*, Stable Diffusion 3 [1], DALL-E 2 [12], which easily transform the sparklings into visual demonstrations [13]. However, pure text description struggles with ambiguity, as shown in Fig.1, the description "white blazer" omits a detailed collar style that normally comes from a designer's professional skill. This raises the challenge: *How to instruct the fashion generation to understand customer desires beyond simple description?*

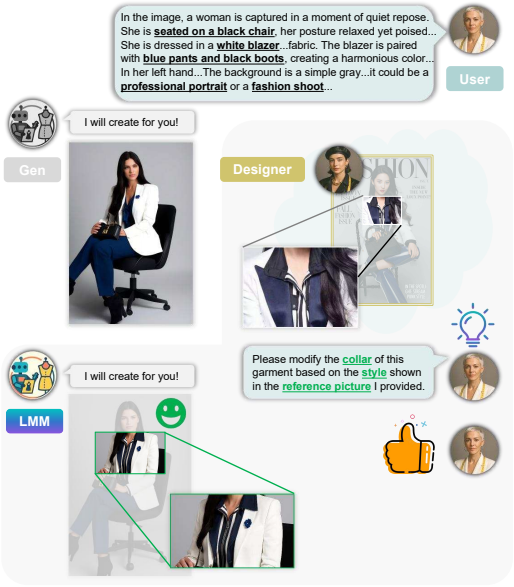

Fig. 1. Example of the real-world fine-grained customization fashion design.

In real-world scenarios, the customer shows a sample image as reference (*e.g.*, an existing design from a fashion magazine) where the designer translates the principle into precise fashion elements. This inspires us to propose a new benchmark: *Better Understanding Generation (BUG)* by showing AI the Image-into-Prompt, to meet the request of fine-grained customization in fashion design. BUG initializes a draft design image first, then continuously modifies the image not only following the user's text-prompts but also referring to image-prompts. Different from previous LMM image editing, our approach is more challenging than editing from real image [14, 15], on object level modification [16, 17] or need-

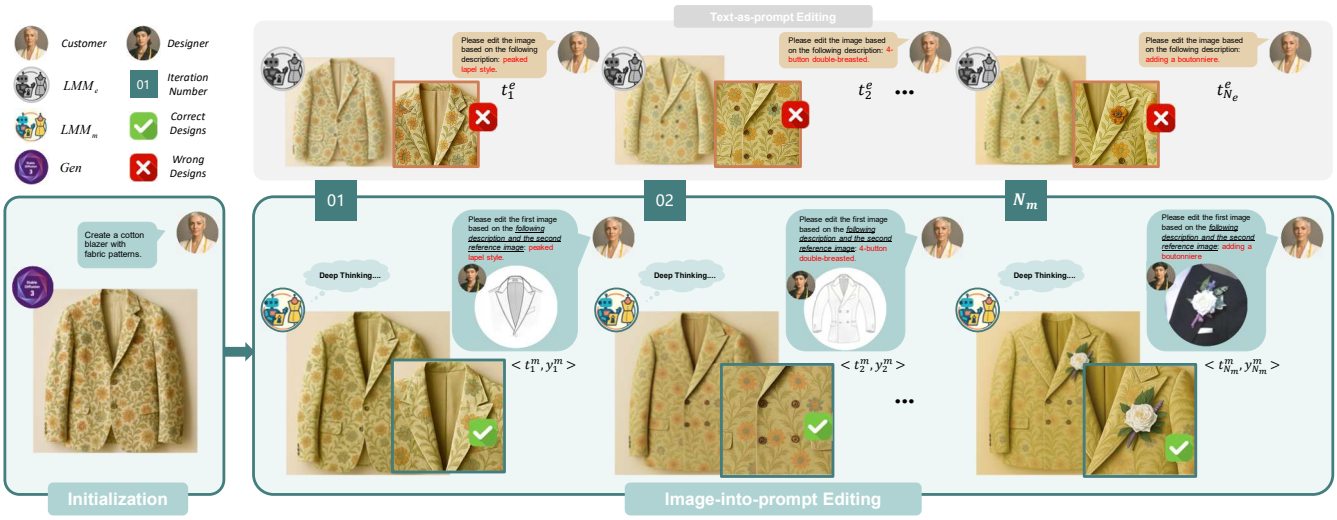

Fig. 2. The workflow example of fine-grained customized fashion design. (1) **Initialization** generates a draft design from the user request based on SD3 [1]. (2) **Image Editing** refine the design iteratively via text-as-prompt/image-into-prompt editing, with GPT4.1-mini [2]. The three iteration results are modified referring to the sketch, the details, and the example image.

ing further human involvement [18].

To evaluate the performance of this task, we propose a new FashionEdit dataset modified from DeepFashion-MultiModal [19]. FashionEdit uses LMM to analyze two fashion fine-grained components, including the generated images and the differences between the generated and ground-truth images. The differences comprise the descriptions and cropped regions from the original images, corresponding to the user's desires and the referring images. We evaluate the performance of models using BUG on this dataset from content similarity (CLIP [20]), user satisfaction (our CLIP*), and quality (PSNR). The CLIP* score increased 20.3% between pure text and our BUG after three modifications, proving the effectiveness of our method.

2. METHODOLOGY

2.1. Initialization

The initialization of our benchmark uses a standard text-based image generation model (*e.g.* SD3 [1]), which inputs a fuzzy text generation-prompt t^g and generates a design image y^g defined as: $y^g = Gen(t^g)$.

2.2. Image Editing

2.2.1. Text-as-prompt Editing

The vanilla text-based image editing takes the initial design from Sec.2.1, following an N_e-rounds of text editing-prompt T^e=$\{t_i^e\}_1^{N_e}$ from user to change the design. Correspondingly, the edited images are defined as Y^e=$\{y_i^e\}_1^{N_e}$ and each y_i^e is generated by an editing LMM_e:

$$y_i^e = LMM_e(y_{i-1}^e, t_i^e), \tag{1}$$

where $y_0^e = y^g$ and $y_{N_e}^e$ is the ready-made design.

2.2.2. Image-into-prompt Editing

Different from the vanilla method, the input of N_m-rounds customized editing is a combination of text editing-prompt T^m=$\{t_i^m\}_1^{N_m}$ and image editing-prompt Y^m=$\{y_i^m\}_1^{N_m}$, defined as $\{< t_i^m, y_i^m >\}_1^{N_m}$. Correspondingly, the edited images are defined as $\hat{Y}^e$=$\{\hat{y}_i^e\}_1^{N_m}$ and each $\hat{y}_i^e$ is generated by an customized editing LMM_m:

$$\hat{y}_i^e = LMM_m(\hat{y}_{i-1}^e, < t_i^m, y_i^m >). \tag{2}$$

Each t_i^m is modified from the original text-prompt t_i^e in Sec.2.2.1, where the referring prompt changes to "Please edit the first image based on the following description and the second reference image" plus the description and referring image, such as the example in Fig.2. It is worth noticing that N_m and N_e can be the same or different.

2.2.3. Applications

We analyze three applications in Fig.2:

Sketch Image: Sketch images typically derive from hand drawing. Such images contain the core concepts of fashion design but appear relatively rough (*e.g.*, line drawing). As shown in the first iteration, language often struggles to convey the professional design's core concepts, where image references tend to be more precious.

Detailed Image: Detail images are typically needed from fine-grained modification requests, which provide precision

Fig. 3. Example of our FashionEdit dataset from original DeepFashion-MultiModal. **(a)** shows the manual text annotations of the original dataset; **(b)** presents the original image; **(c)** is the image generated by SD3 [1] based on (a); **(d)** contains the image patches, descriptions and location derived from the differences between (b) and (c).

changes (*e.g.*, number, color, or layout). As shown in the second iteration, foundation LMMs struggle to capture these details, even is provided text descriptions (*e.g.*, 4 buttons arranged in a specific location).

Example Image: Sample images typically derive from designs by professional designers. Such designs generally include complex details that are difficult to describe from normal people. As shown in the last iteration, the user desires a specific item (*e.g.*, boutonniere) from a real image.

2.3. FashionEdit Dataset

2.3.1. Dataset Generation

DeepFashion-MultiModal [19] is a large-scale, high-quality fashion-oriented dataset containing rich multi-modal annotations. It provides human-annotated descriptions with fine-grained labels on two dimensions: clothes colors and clothes fabrics. One example is shown in Fig.3(a) and (b).

To precisely validate the fine-grained customized control fashion design task, we create a subset from DeepFashion-MultiModal called FashionEdit. Two more components are further developed: generated images and the differences between generated and original images, as shown in Fig.3(c) and (d). The creation process is as follows:

Method	Prompt	CLIP (↑)	CLIP* (↑)	PSNR (↑)
SD2$_{(train+val)}$	text	69.25	0.00	6.93
SD3$_{(train+val)}$	text	82.85	1.75	9.94
Vanilla$_{(1)}$	text	85.77	13.6	9.45
Vanilla$_{(2)}$	text	85.78	14.9	9.29
Vanilla$_{(3)}$	text	85.60	15.1	9.30
BUG$_{(1)}$	text+image	87.27	26.4	9.75
BUG$_{(2)}$	text+image	87.77	30.9	9.76
BUG$_{(3)}$	text+image	**87.91**	**35.4**	**9.96**

Table 1. Experiences of initialization methods, vanilla methods (different steps), and image-into-prompt methods (different steps) according to the CLIP, CLIP*, and PSNR scores on FashionEdit datasets. ↑: Higher is better.

(1) Images Generation: To simulate the current real-world clothing design processes, we employ the latest SD3 [1] to obtain AI-generated design images from pure descriptions. After the initial generation, we further filtered the top 11,546 images based on CLIP similarity to minimize the noise of the generation process (*e.g.* multiple humans). The train/val separate proportion is 10546/1000 in experience.

(2) Differences Analysis: To simulate human clothing modification (instruction + image input), we need the descriptions of the change and image parts between the generated and original images. Thus, we implement GPT4.1-mini [2] to analyze the Top-3 differences, and output the description with coordinates from the original images. The prompt is structured as follows:

```
# 1. Task Definition
    Detect the three detailed differences in the clothes
    between the two images and return a JSON style.
# 2. Problem Definition
    For each result, the description should only contain
    the difference of the first image, and give the
    bounding box of the box_2d should be [ymin, xmin,
    ymax, xmax], normalized to 0-1000.
# 3. Example of output
    The output format is limited to:"[{'description': '
    Wearing ...', 'box_2d': [0, 250, 600, 750]}"
```

3. EXPERIMENTS

3.1. Experimental Settings

To evaluate the generated image, we use the CLIP [20] similarity scores for generation similarity, the CLIP* for user satisfaction, and PSNR for quality. (1) **CLIP** is computed via cosine distance, which measures the similarity between high-dimensional image embeddings via the CLIP encoder; (2) **CLIP*** is calculated by counting the number of generated images with a high CLIP score (>90%), deriving the number of validation images; (3) **PSNR** is computed by the logarithmic ratio of peak reference signal power to reconstruction error power.

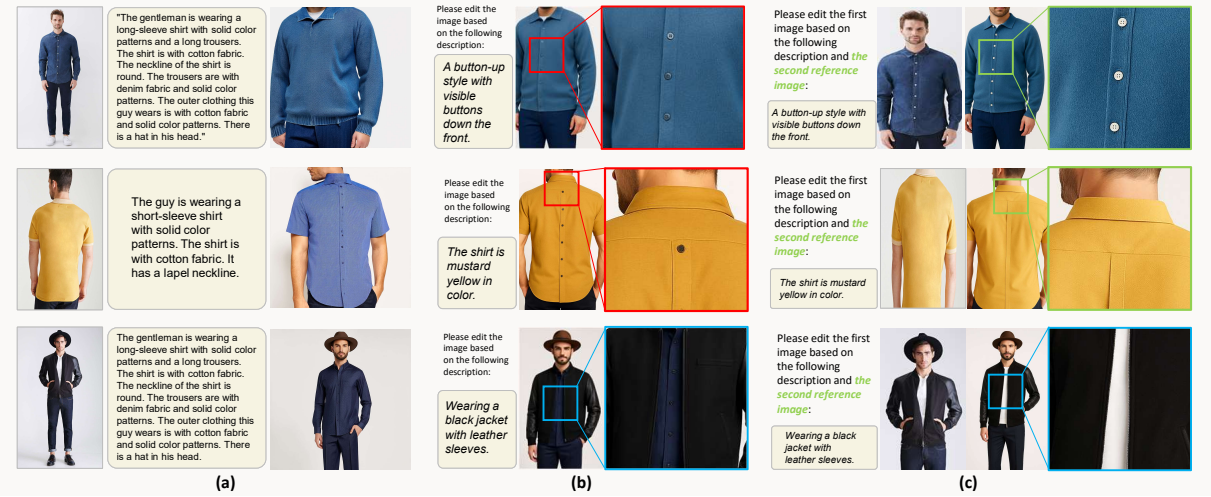

Fig. 4. Visualization of the results based on the validation set of FashionEdit. For each example from left to right, **(a)** combines the ground truth image, ground truth description, and initial generated image; **(b)** is the result of text-as-prompt (Sec.2.2.1) combines the text-prompt, modified image with detail; **(c)** is the result of image-into-prompt (Sec.2.2.2) combines the text-prompt, image-prompt, and modified image with detail.

3.2. Experimental Results

3.2.1. Comparisons with different methods

Through comparisons of different image generation models (upper Tab.1), the latest SD3 [1] outperforms SD2 [21] in the entire "train+val" sets of FashionEdit, where the most significant improvement is on the CLIP score from 69.25% to 82.85%. It is worth noticing that the CLIP* score is low for both SD2 and SD3, which indicates the dissatisfaction of the first generation, proving the necessity for image editing in the fashion design task.

Analyzing the different settings for vanilla methods (central Tab.1) with "(1)" to "(3)" modification steps in the "val" set of FashionEdit, CLIP* score continually increases from 13.6% to 15.1% with a 1.5% improvement, which demonstrates the effectiveness of multiple editing. Similar results are also reported for BUG settings with 9.0% improvement. It is worth noticing that the CLIP* score between "Vanilla(3)" and our "BUG(3)" after three modifications increased 20.3%, proving the potential of BUG under multiple modifications.

The best result comes from our BUG method with three modifications (lower Tab.1), where the CLIP score reaches 87.91%, CLIP* score reaches 35.4%, and PSNR score reaches 9.96%, respectively. This demonstrates the effectiveness of our image-to-prompt benchmark for a better understanding of generation. (GPT4.1-mini [2])

3.2.2. Visualization of Different Methods

In Fig.4, we provide three sets of visualization results to compare the text-as-prompt and image-into-prompt methods (one-reference setting), based on the validation set of our FashionEdit.

Our method provides a more detailed image generation (Case 1). For example, though both methods add visible buttons on shirts following the request, the image-into-prompt result matches the referring cropped image on button color in "white". This demonstrates the necessity of referring image, where the text description misses details.

Besides, our method handles the physical conflict to generate a more realistic image (Case 2). For example, the draft design is weird for creating a front-side shirt with a back-side human position. The text-as-prompt result changes the color of the shirt following the instruction, but ignores the conflict, while the BUG result flips the shirt into the right position. This demonstrates the importance of physics laws in the referring images.

Our method prioritizes the referring image over the text description (Case 3). For example, the image-into-prompt result corrects the underlayer of "cotton white shirt", while the text editing only adds the missing "jacket" based on the original wrong T-shirt. Our method enhances the design workflow where the user can choose from multiple outputs.

4. CONCLUSION

Our work analyzes the core challenge in fashion design with GenAI, and proposes a BUG benchmark that adopts both text-prompts and image-prompts under iterative image editing. Experience proves that our benchmark dramatically improves user satisfaction under the following instructions. The new FashionEdit dataset, which simulates the real-world clothing design workflow, provides a new possibility to further employ AI in the arts industry.

5. REFERENCES

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *ICML*, 2024.

[2] OpenAI, "Introducing gpt-4.1 in the api," 2025.

[3] Adil Nygaard, Ashish Upadhyay, Lauren Hinkle, Xenia Skotti, Joe Halliwell, Ian Brown, and Glen Noronha, "News risk alerting system (nras): A data-driven llm approach to proactive credit risk monitoring," in *EMNLP (Industry Track)*, 2024.

[4] Zhi-Qi Cheng, Yifei Dong, Aike Shi, Wei Liu, Yuzhi Hu, Jason O'Connor, Alexander G Hauptmann, and Kate Whitefoot, "Shield: Llm-driven schema induction for predictive analytics in ev battery supply chain disruptions," in *EMNLP (Industry Track)*, 2024.

[5] Mirae Kim, Kyubum Hwang, Hayoung Oh, Min Kim, Chaerim Park, Yehwi Park, and Chungyeon Lee, "Mild bot: Multidisciplinary childhood cancer survivor question-answering bot," in *EMNLP (Industry Track)*, 2024.

[6] Xujie Zhang, Binbin Yang, Michael C Kampffmeyer, Wenqing Zhang, Shiyue Zhang, Guansong Lu, Liang Lin, Hang Xu, and Xiaodan Liang, "Diffcloth: Diffusion based garment synthesis and manipulation via structural cross-modal semantic alignment," in *ICCV*, 2023.

[7] Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung, "Garmentcodedata: A dataset of 3d made-to-measure garments with sewing patterns," in *ECCV*, 2024.

[8] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han, "Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images," in *ECCV*, 2020.

[9] Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, et al., "Clothesnet: An information-rich 3d garment model repository with simulated clothes environment," in *ICCV*, 2023.

[10] Wenda Shi, Waikeung Wong, and Xingxing Zou, "Generative ai in fashion: Overview," *ACM Transactions on Intelligent Systems and Technology*, 2025.

[11] Han Liu, Xianfeng Tang, Tianlang Chen, Jiapeng Liu, Indu Indu, Henry Zou, Peng Dai, Roberto Galan, Michael Porter, Dongmei Jia, et al., "Sequential llm framework for fashion recommendation," in *EMNLP (Industry Track)*, 2024.

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

[13] Cheng Zhang, Yuanhao Wang, Francisco Vicente, Chenglei Wu, Jinlong Yang, Thabo Beeler, and Fernando De la Torre, "Fabricdiffusion: High-fidelity texture transfer for 3d garments generation from in-the-wild images," in *SIGGRAPH Asia*, 2024.

[14] Tim Brooks, Aleksander Holynski, and Alexei A Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[15] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al., "Hive: Harnessing human feedback for instructional visual editing," in *CVPR*, 2024.

[16] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," *NIPS*, 2023.

[17] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al., "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting," in *CVPR*, 2023.

[18] Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar, "Pixeltone: A multimodal interface for image editing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[19] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, 2022.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.