

HieraFashDiff: Hierarchical Fashion Design with Multi-stage Diffusion Models

Zhifeng Xie^{1,2}, Hao Li¹, Huiming Ding¹, Mengtian Li^{1,2}, Xinhan Di³, Ying Cao^{4*}

¹Department of Film and Television Engineering, Shanghai University

²Shanghai Engineering Research Center of Motion Picture Special Effects

³AI Lab, Giant Network

⁴School of Information Science and Technology, ShanghaiTech University

{zhifeng_xie, dinghuiming, mtl}@shu.edu.cn, hao.li.shu@outlook.com, dixinhan@ztgame.com, caoying59@gmail.com

Abstract

Fashion design is a challenging and complex process. Recent works on fashion generation and editing are all agnostic of the actual fashion design process, which limits their usage in practice. In this paper, we propose a novel hierarchical diffusion-based framework tailored for fashion design, coined as *HieraFashDiff*. Our model is designed to mimic the practical fashion design workflow, by unraveling the denoising process into two successive stages: 1) an ideation stage that generates design proposals given high-level concepts and 2) an iteration stage that continuously refines the proposals using low-level attributes. Our model supports fashion design generation and fine-grained local editing in a single framework. To train our model, we contribute a new dataset of full-body fashion images annotated with hierarchical text descriptions. Extensive evaluations show that, as compared to prior approaches, our method can generate fashion designs and edited results with higher fidelity and better prompt adherence, showing its promising potential to augment the practical fashion design workflow.

Introduction

Fashion design is an important and challenging activity, which requires navigating through a huge joint space of shape, color, material, pattern and layout in order to find solutions that satisfy aesthetic and functional requirements, and perhaps client-specific constraints. The typical workflow of fashion designers consists of two essential stages: *ideation* and *iteration*. In particular, fashion designers begin their creative process by coming up with some abstract design concepts in terms of theme, style and personality, and then translating these concepts into concrete design ideas (i.e., design drafts) through brainstorming or turning to reference examples for inspiration. Subsequently, they iterate on the initial idea by making small changes to the draft to finalize the design.

Recently, there is a growing interest in facilitating the fashion design process by building generative models that can synthesize realistic fashion images from user-specified hints and constraints (Chen et al. 2020; Dai et al. 2021; Cao et al. 2023b). However, these approaches is that they do not

consider (and thus fail to fit into) the practical fashion design pipeline, making them difficult to be directly adopted in practical scenarios.

Building models that can find wide adoption in the real fashion design process is non-trivial. First, a desirable model should learn the task of fashion design generation and editing jointly to separately tackle the ideation and iteration stages, so that it has capability to support the entire pipeline. This is in contrast to prior works that especially address either generation (Zhu et al. 2017; Jiang et al. 2022; Zhang et al. 2022; Sun et al. 2023) or editing (Ak et al. 2019; Kwon et al. 2022; Pernus et al. 2023; Baldrati et al. 2023; Wang and Ye 2024). Second, both generation and editing components should be conditioned on intuitive inputs at high and low levels, respectively, empowering users to express their design thoughts easily while facilitating efficient ideation and rapid iteration.

To address the aforementioned challenges, we propose *HieraFashDiff*, a novel framework specialized in facilitating fashion design, which instantiates a conditional diffusion model that learns to generate fashion images from given fashion-specific text prompts. Our key insight is that the typical fashion design workflow is reminiscent of the reverse process of diffusion models for image generation, which first denoises a pure noise into a coarse image (design draft) in the early stage, and then iteratively refine the coarse image to add more fine details to produce a realistic image (final design) over the remaining steps. Inspired by this, we propose to factor the reverse process of our diffusion model into two successive stages: an ideation stage spanning the earlier denoising steps and an iteration stage spanning the later denoising steps. Our model injects text descriptions at different levels into the two stages — the ideation stage is conditioned on high-level design concepts to produce noisy design drafts while the iteration stage is guided by progressively added low-level apparel attributes to refine the drafts towards complete designs. In this way, our framework naturally supports the concept-guided generation of fashion design proposals and the semantic editing of local fashion components in a unified framework to aid in the whole fashion design pipeline, as shown in Fig 1.

To train our model, we curate a new fashion dataset, coined as *HieraFashion*. Our dataset consists of more than 5k full-body apparel images, each of which is annotated with

*Corresponding author.



Figure 1: The proposed HieraFashDiff is capable of generating fashion design drafts from just abstract concepts (blue text), and allowing for local editing on the generated draft iteratively through a few apparel attribute descriptions (red text). Thus, our method can be used to facilitate typical fashion design workflow by enabling efficient ideation and rapid iteration.

a hierarchical text caption that is concise yet informative enough to capture the essence of fashion design. We evaluate our model on our newly collected dataset, demonstrating its state-of-the-art performance in terms of generation quality and prompt coherence, as compared to existing methods.

In summary, our contributions are as follows.

- We propose a fashion generation and editing framework that, for the first time, mimics the *whole* fashion design process explicitly, which can support efficient ideation and rapid iteration.
- We propose a novel hierarchical text-to-fashion diffusion model that decomposes the generation process into multiple stages conditioned on input prompts of different levels, which enables coarse-grained fashion draft generation and fine-grained modifications *jointly*.
- We curate a new dataset comprising full-body fashion images captioned with high-level design concepts and low-level local attributes.

Related Work

Text-Guided Fashion Image Generation. Text-to-image generation is a crucial and complex task that seeks to generate realistic images based on natural language descriptions. In the fashion domain, only a few works (Zhu et al. 2017; Zhang et al. 2022; Sun et al. 2023) attempt to generate fashion-related images (e.g. for apparel, accessories and fashion models) from textual description. Early approach (Zhu et al. 2017) to the text-guided fashions synthesis relied on Generative Adversarial Networks (GANs) presented a two-stage stylized image generation solution that generates realistic fashion images, conditioned on textual descriptions and semantic layouts. Zhang et al (Zhang et al. 2022) proposed the ARMANI framework for fashion synthesis focused on generating local details. Recent advances in diffusion models (Nichol et al. 2022; Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022) lead to more realistic generation. Sun et al (Sun et al. 2023) developed and applied the skip cross-attention module to integrate image and text modalities. Our method adopts a multi-stage framework to closely mimic the practical graphic design workflow, enabling it to support fashion generation and

editing simultaneously. Moreover, our method handles hierarchical text descriptions with explicit disentanglement of high-level concepts and low-level attributes, instead of captions that either encompass only low-level details or mix up high-level and low-level prompts together, to better facilitate ideation and iteration in fashion design.

Fashion Image Editing. Generative Adversarial Networks (GANs) have emerged as a cornerstone technology extensively applied in fashion-related image editing tasks (Ak et al. 2019; Kwon et al. 2022). FICE (Pernus et al. 2023) addressed text-conditional image editing with optimization-based GAN inversion guided by the CLIP model. Some recent efforts started to approach fashion image editing diffusion models. MGD (Baldrati et al. 2023) proposed a latent diffusion model to edit fashion images conditioned on multimodal inputs including text, pose and sketch. Zhang et al (Zhang et al. 2023) introduced a diffusion model incorporating structural semantic consensus guidance, utilizing a language structure parser to extract attribute words, thereby achieving fine-grained semantic alignment. TexFit (Wang and Ye 2024) predicted the editing area in an image based on the input text and used the predicted region to condition a diffusion model for local editing. Our method can perform local editing iteratively on full-body fashion images from apparel attribute descriptions, producing a sequence of high-quality, continuously evolving designs, which has not yet been demonstrated in the existing works. In addition, our method complements its editing functionality with a capability to generate full-body fashion designs from just high-level concepts, which is not available in the previous approaches.

Method

Hierarchical Multi-stage Diffusion Model

We build our model on the pre-trained Stable Diffusion (SD) (Rombach et al. 2022) and fine-tune the SD on a fashion dataset (Baldrati et al. 2023) and, therefore, our model operates in the latent space instead of pixel space. The key idea underlying our model is to abstract the common fashion design workflow as the denoising process of our model. As illustrated in Fig 2, the denoising process (i.e., the generation

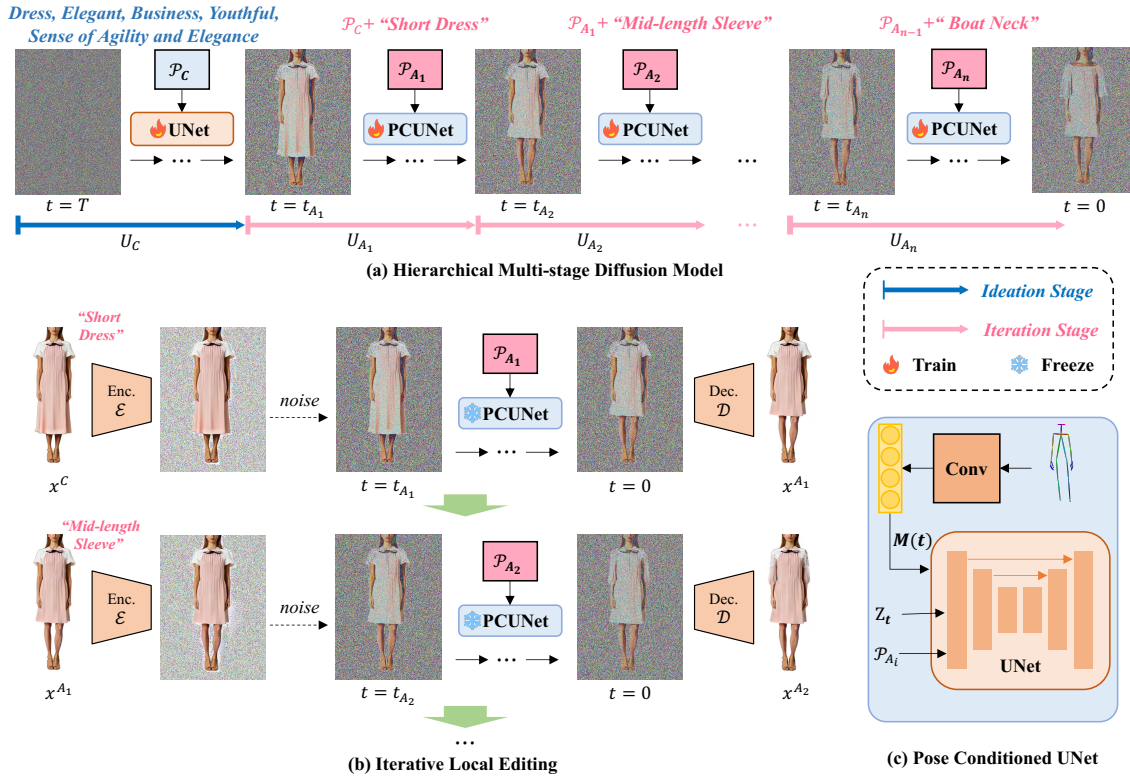


Figure 2: Overview of our method. (a) The denoising process of our model is decomposed into an ideation stage and an iteration stage, which are conditioned on high-level concepts and low-level attributes, respectively. (b) our editing method starts from the generated design draft x^C and produces a sequence of edited results $(x^{A_1}, x^{A_2}, \dots)$ given text prompts for different attributes (A_1, A_2, \dots) . (c) our UNet-based denoising network is conditioned on additional pose information.

process) is decoupled into two sequential, successive stages: an ideation stage and an iteration stage.

Denoising Process Decomposition. Let U denote the full timesteps of the denoising process. The ideation and iteration stages span the earlier steps U_C and the later steps U_A , respectively, so that $U = U_C \cup U_A$. The ideation stage is responsible for generating a design draft from high-level, vague concepts \mathcal{P}_C , and the iteration stage aims to iteratively refine the generated draft based on low-level attribute descriptions \mathcal{P}_A , by adding more and more fine-grained, local design components and details. Such decomposition formulation allows our model to naturally support automatic design generation and semantic design editing in an unified manner.

To enable more localized control in the editing scenario, we further decompose the iteration stage into a sequence of sub-stages, one for each apparel attribute — $U_A = U_{A_1} \cup U_{A_2} \cup \dots \cup U_{A_n}$, where U_{A_i} represents the time interval dedicated to i -th attribute and n is the number of attributes being considered. In this way, it is possible for users to iterate on the generated design by solely changing one attribute at a time. To arrange the apparel attributes sequentially through the denoising process, we need to determine their ordering. To this end, we leverage the inherent property of the diffusion model to maximize editability for each attribute. In particular, due to the noise variance schedule (Ho, Jain, and

Abbeel 2020), the amount of change to a generated image declines over denoising steps (Cao et al. 2023a; Yu et al. 2023). Therefore, we choose to order the attributes by the size of their affected areas in fashion images. For example, modifying dress length will cause a larger proportion of an images to be changed than modifying neckline type, and thus we rank dress length before neckline type. This will ensure that our model has sufficient ability to make desired edits to faithfully reflect different attributes. In our implementation, we consider 5 common apparel attributes, and order them as: “clothing length” > “sleeve length” > “sleeve type” > “collar type” > “hem type”.

Prompt Schedule. During the denoising process, the ideation stage is conditioned on the abstract design concepts \mathcal{P}_C , and the iteration stage are conditioned on a sequence of text prompts $\mathcal{P}_A = (\mathcal{P}_{A_1}, \dots, \mathcal{P}_{A_n})$, where each sub-stage i only generates local component given \mathcal{P}_{A_i} . We construct the iteration sub-stage prompts by starting with \mathcal{P}_C and adding one attribute description per sub-stage over time. Formally, the prompt for i -th iteration stage is defined as:

$$\mathcal{P}_{A_i} = \mathcal{P}_C + \sum_{k=1}^i A_k, \quad (1)$$

where the addition operator is overloaded to denote the concatenation of two prompt strings. This means that each iter-

ation sub-stage needs to consider not only its own attribute but also the given concepts and all the previous attributes, thereby helping better preserve what is already generated.

Conditioning on Pose Map. During the iteration stage, besides text prompts, we also input a 2D pose map, encoding 18 body joint positions, into our denoising network to further improve generation quality. The pose map is obtained by running a 2D pose detector (Cao et al. 2017) on the real image in training and the generated image by the ideation stage in testing. Providing joint position information to the model can help it better localize the regions to change for a given attribute. Furthermore, we find this additional input is beneficial to the preservation of the human pose during editing, as observed in (Baldrati et al. 2023).

Training. For training, we optimize the following objective:

$$\mathbb{E}_{z_0, t, \epsilon} \left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{C}(t), \mathcal{M}(t))\|_2^2 \right], \quad (2)$$

where $\mathcal{C}(t)$ is prompt schedule function:

$$\mathcal{C}(t) = \begin{cases} \mathcal{P}_C, & t \in U_C \\ \mathcal{P}_C + \sum_{k=1}^t A_k, & t \in U_{A_t} \end{cases}. \quad (3)$$

$\mathcal{M}(t)$ is a pose schedule function that returns the 2D pose map of the training image, if t falls within the iteration phase and empty if t is within the ideation phase.

Design Synthesis and Editing

Design Draft Generation. Given a high-level concept prompt \mathcal{P}_C , we aim to generate a design draft. To do this, we sample a base noise z_T^C and feed it into our denoising process to obtain z_0^C , which is then decoded by the SD decoder \mathcal{D} to generate a design draft $x^C = \mathcal{D}(z_0^C)$. Note that our sampling is run through all the timesteps, i.e, from $t = T$ to $t = 0$, instead of just through the ideation stage, in order to generate a *clear* design draft. This is possible because our denoising network sees the high-level concept description for all the timesteps during training according to our prompt schedule function in Eq. 3.

Iterative Local Editing. Given the generated draft x^C , we aim to edit it iteratively through a sequence of intuitive apparel attribute descriptions $(\mathcal{P}_{A_1}, \dots, \mathcal{P}_{A_n})$. As shown in Fig 2, to perform editing on the first attribute A_1 , we take the generated draft x^C as input and diffuse its latent obtained using the pre-trained image encoder to a noisy latent at the starting timestep, t_{A_1} , of the first attribute sub-stage using the forward process margin distribution $q(z_t|z_0)$. Then, we run the rest sampling steps to $t = 0$ from the noisy latent conditioned on \mathcal{P}_{A_1} to produce the edited result x^{A_1} . For each subsequent attribute A_k , we execute editing in a similar way except that the input is the edited result of the previous attribute and the sampling is run from the starting timestep of attribute A_k with \mathcal{P}_{A_k} as input prompt.

Typically, modifying an attribute, which only refers to a local *target* region in the fashion image, should ideally lead to localized changes. However, We find that the above method may cause undesirable non-local changes. To mitigate this issue, we observe that there exists a strong correlation between apparel attributes and human body parts

— e.g, changing sleeve type will primarily influence the region on and near the arm. This motivates us to leverage human body part masks associated with the corresponding attributes to enforce the preservation of non-target regions. To generate the body part masks, we apply the 2D pose detection method (Cao et al. 2017) to the input image to estimate joint locations. For a body part associated with attribute a , we construct a bounding box enclosing the relevant joint positions, and use the SAM (Kirillov et al. 2023) to segment the clothing region within the bounding box to form a binary mask, which labels the target and non-target regions with 1 and 0, respectively. Note that when increasing the length of an apparel component (e.g, from short dress to long dress), we directly use the entire region within the bounding box as the mask to cover the non-clothing region that needs to be modified (e.g, leg). Then, similar to the blended latent diffusion (Avrahami, Fried, and Lischinski 2023), we modify each denoising step within the period of attribute a using its mask \mathbf{m}_a (downsampled to the spatial resolution of the latent): $\tilde{\mathbf{z}}_{t-1} = \mathbf{m}_a \odot \mathbf{z}_{t-1} + (\mathbf{1} - \mathbf{m}_a) \odot \mathbf{z}_{t-1}^i$, where \mathbf{z}_{t-1} is sampled from the learned condition distribution $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$, \odot denotes element-wise multiplication and $\mathbf{1}$ is an all-ones image. \mathbf{z}_{t-1}^i is a noisy latent obtained by encoding the input image via \mathcal{E} into a latent and diffusing it to timestep $t - 1$. Intuitively, this keeps the non-target region unchanged, by replacing the values of \mathbf{z}_{t-1} in the non-target region with their counterparts from the corrupted input image encoding.

Hierarchical Fashion Dataset

To train our model, a fashion dataset captioned by both high-level concepts and low-level attributes is needed. Existing fashion datasets (Yang et al. 2020; Morelli et al. 2022; Jiang et al. 2022; Baldrati et al. 2023) is not sufficient to serve our purpose. Their texts captions either lack a complete capture of fashion design concepts or mix up the descriptions of different levels together without explicit separation. Therefore, we curate a dataset, *HieraFashion*, with 5200 full-body fashion images of high resolution (768×1024 pixels), spanning 8 common apparel categories including dress, coat, sweater, blouse, jumpsuit, pant, shirt and skirt.

The images in our dataset are meticulously selected from FACAD (Yang et al. 2020), DeepFashion-MM (Jiang et al. 2022) and Dress Code-MM (Baldrati et al. 2023) so that they have diverse design elements and rich variation in design pattern. To ensure the visual quality of the images and consistency across them, we cropped out the area above the eyes of the model in each image, and filter out the images with complex backgrounds, non-frontal body orientations, challenging poses, and half-body shots. One notable feature of our dataset is its hierarchical text descriptions — each image is annotated with a text description which comprises two parts: high-level design concepts and low-level apparel attributes. The textual annotations were performed by recruited *professional* fashion designers.

Overall, our dataset has several distinctive characteristics: 1) hierarchical text descriptions; 2) full-body fashion images with clear background; 3) high-resolution images. Please refer to the supplemental materials for more details.



Figure 3: Qualitative results of different methods for fashion draft generation from high-level design concepts.

Experiments

Datasets. We conduct experiments on the *HieraFashion* dataset. We divide the *HieraFashion* dataset into 4,000 training examples and 1,200 testing examples. Further, we use the Dress Code Multimodal dataset, with a total of 26, 400 image-text pairs to fine-tune the models.

Implementation Details. We fine-tune the stable diffusion model on the Dress Code Multimodal and *HieraFashion* datasets. We use 1000 timesteps for the reverse process, allocating the interval[1000, 900] to the ideation stage and setting the intervals for the 5 attributes in the iteration stage as [900, 800], [800, 710], [710, 630], [630, 560], [560, 0]. Note that we put most of the iteration sub-stages into the first half of the denoising process since the model has very restricted flexibility to change the generated image in the late denoising stage (Cao et al. 2023a; Yu et al. 2023). During training, we resize all the images to 512×704. We train our model for 117,299 steps on a single NVIDIA A6000 GPU on our *HieraFashion* dataset, employing a batch size of 4, a learning rate of 1e-5, and a linear warm-up for initial 500 iterations, with the AdamW (Loshchilov and Hutter 2019) optimizer. For image generation, we employ DDIM (Song, Meng, and Ermon 2021) with 100 steps as noise scheduler and use classifier-free guidance (Ho and Salimans 2022).

Compared methods. For design draft generation, we compare our method with state-of-the-art methods including Cogview (Ding et al. 2021), Stable Diffusion (Rombach et al. 2022), and Attend-and-Excite (Chefer et al. 2023). To ensure a fair comparison, all models are retrained using the Dress Code Multimodal (Baldrati et al. 2023) and *HieraFashion* datasets. For local editing, we consider general-

Method	FID ↓	Coverage ↑	CLIP-S ↑
Cogview	23.62	0.35	25.87
SD-finetune	18.45	0.48	27.92
Attend-and-Excite	15.33	0.63	28.46
Ours	10.27	0.76	30.61

Table 1: Quantitative evaluation of different methods for fashion draft generation. The best results are in bold.

purpose image inpainting methods, SD-Inpaint¹ and BrushNet (Ju et al. 2024), as baselines. We also compare with latest text-guided fashion image editing methods: FICE (Pernus et al. 2023), MGD (Baldrati et al. 2023), TexFit (Wang and Ye 2024). Both FICE and TexFit are retrained on the *HieraFashion* training set; MGD is not trained on our dataset since its training code is not accessible. For fair comparison, MGD, BrushNet and TexFit use the same body part masks as our method. We acknowledge that there are more generation and editing methods (Zhang et al. 2022; Li et al. 2022; Sun et al. 2023; Zhang et al. 2023) in the fashion domain. However, comparison with them is not feasible since their training code is not publicly available.

Evaluation metrics. To evaluate the realism and diversity of the generated images, we use the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Coverage (C) (Naeem et al. 2020) metrics. For both metrics, we employ the CLIP ViT-B/32 model as the feature extractor. Furthermore, to assess the coherence of the image to input prompts, we utilize the CLIP Score (CLIP-S) (Hessel et al. 2021). We fine-tune the CLIP ViT-B/32 model on the Dress Code Multimodal

¹<https://huggingface.co/runwayml/stable-diffusion-inpainting>

Method	A ₁		A ₂		A ₃		A ₄		A ₅	
	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑
FICE	22.72	27.94	23.15	28.07	23.69	28.13	24.28	28.36	25.53	28.47
SD-Inpaint	13.56	28.13	14.31	28.24	15.57	28.32	16.76	28.54	17.28	28.66
TexFit	13.98	28.01	14.82	28.17	15.73	28.15	16.50	28.42	17.11	28.64
TexFit-M	13.27	28.59	13.83	28.76	14.41	28.92	15.16	30.38	15.72	30.55
BrushNet	12.92	29.85	13.46	30.10	13.98	30.54	14.63	31.12	15.06	31.31
Ours	10.74	31.39	11.26	31.73	11.80	32.35	12.14	32.86	12.59	33.01

Table 2: Quantitative evaluation of different methods for iterative local editing on our *HieraFashion*. TexFit-M refers to TexFit with our body part masks rather than its predicted ones. The best results are in bold.

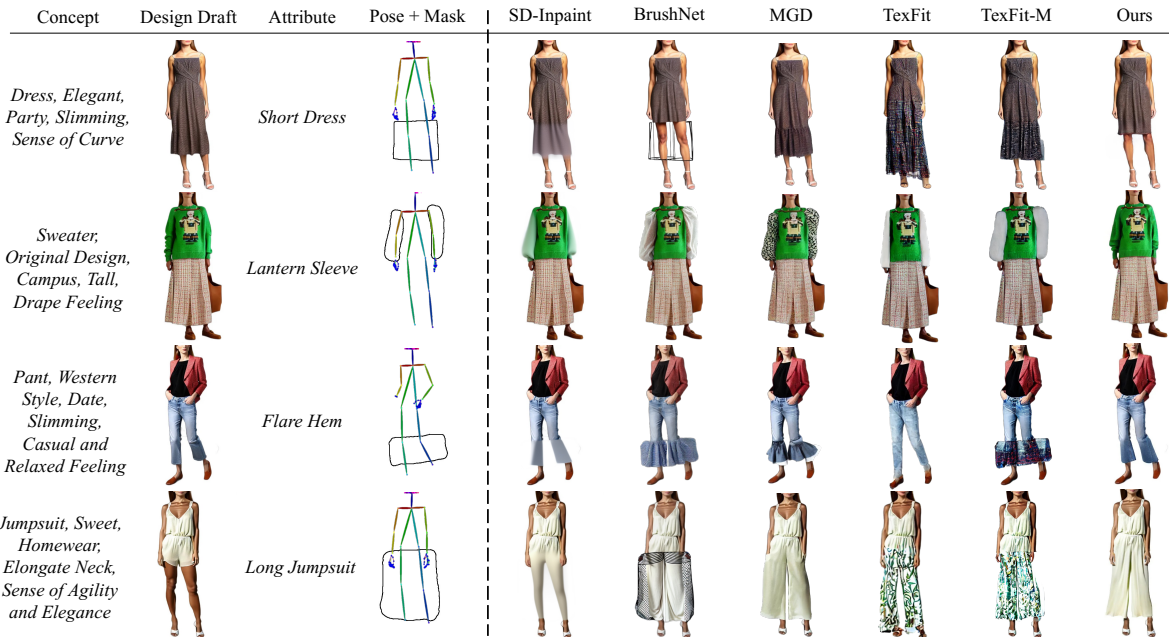


Figure 4: Qualitative comparison of iterative local editing. The latest editing methods often lack alignment with low-level attribute semantics, or cause undesirable global changes. Our method can precisely edit the corresponding regions according to the attribute descriptions while keeping the other regions unchanged, which is superior to other methods. TexFit-M refers to TexFit with our body part masks rather than its predicted ones.

dataset and then on *HieraFashion* dataset to adapt to images and text descriptions in the fashion domain. For fine-tuning the CLIP, we concatenate all the keywords for each image in our dataset into a single text description.

Comparison to Prior Methods

Quantitative Comparison. In Tab 1, we report the quantitative results of different methods on our *HieraFashion* dataset for design draft generation given high-level prompts. As can be seen, our method consistently outperforms the competitors, in terms of realism and diversity (i.e, FID and Coverage) and prompt coherency (i.e, CLIP-S).

Tab 2 shows the results of different methods for local editing. Our method outperforms the other methods across all the low-level apparel attributes. Moreover, as the number of editing iterations increases, the FID scores of all the methods gradually decline. This is because each editing iteration builds upon the previously generated image, resulting

in gradually degraded visual quality. Notably, our method is able to maintain consistently stronger performance through the iterations. In addition, TexFit-M using our body part masks is better than the original TexFit based on predicted editing regions, indicating that our approach of localizing target regions with body part masks is simple yet effective.

Qualitative comparison. In Fig 3, we provide a visual comparison of different methods in fashion draft generation. Our method can generate the designs that more faithfully convey the given concepts than the existing methods. Across different apparel categories, our method is able to synthesize designs with rich texture patterns and diverse apparel components. In contrast, CogView can synthesize simple apparels, but the generated results are blurry with many details lost. The results by the SD are sharp with sufficient textural details, but are not matching the given concepts properly. Attend-and-Exact improves upon CogView and the SD in terms of prompt alignment. However, there is still a notice-

	Draft Generation			A ₁		A ₂		A ₃		A ₄		A ₅	
	FID ↓	CLIP-S ↑	C ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑	FID ↓	CLIP-S ↑
Flat	14.29	28.17	0.43	16.79	29.21	18.13	30.09	19.39	30.36	20.48	30.55	21.34	30.83
RandAttrOrder	13.81	28.98	0.59	14.26	29.46	14.87	30.11	15.52	30.61	16.37	30.80	16.95	30.96
Ours	10.27	30.61	0.76	10.74	31.39	11.26	31.73	11.80	32.35	12.14	32.86	12.59	33.01

Table 3: Effect of hierarchical descriptions and attribute ordering. We compare with two variants of our method (Ours): one conditioning on flat descriptions that combine high-level concepts and low-level attributes (Flat), and one using random order of attributes (RandAttrOrder). The best results are in bold.

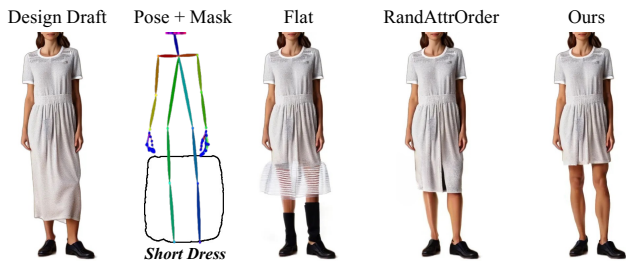


Figure 5: Comparison of our hierarchical model (Ours) against its flat (Flat) and random attribute ordering (RandAttrOrder) variants for local editing (long dress → short dress).

able gap between what its results convey and the given concepts. For example, the generated dress in the first row does not feel “elegant” and the result in the second row does not “elongate neck”.

Fig 4 compares the editing results of different methods. We observe two major shortcomings of the compared methods: first, they may struggle with producing the edits that precisely reflect the given attributes (e.g, the dress length is not reduced properly for all the other methods except BrushNet in the first row of Fig 4); second, they may generate the edited areas that are not harmonious with the rest part of the input design (e.g, the texture and color of the sleeve are changed in the second row of Fig 4). In contrast, our method does not suffer from these issues, producing harmonious results that can accurately reflect different input attributes.

Hierarchical Prompts and Attribute Ordering

Our model considers hierarchical text descriptions with a multi-stage framework. We evaluate our hierarchical model against a flat alternative that collapses high-level concepts and low-level attributes into a single description, and conditions on the same description throughout the denoising process (Flat). We also try using random ordering of attributes in the iteration stage (RandAttrOrder), rather than our proposed ordering based on the size of influenced areas.

The results on both generation and editing tasks are shown in Tab 3. Our full model is superior to the two alternatives across all the metrics for both tasks. The Flat performs the worst, giving very high FID scores and low CLIP-S scores, which confirms the importance of our hierarchical descriptions and multi-stage framework. Further, using random ordering of attributes also lead to inferior performance, sug-



Figure 6: Effect of pose conditioning and body part masks for local editing (short dress → long dress).

gesting the necessity of our proposed ordering strategy. Fig 5 shows a visual comparison of different methods in an editing scenario to quantitatively demonstrate the advantage of our full model over the two variants.

Ablation Study

We further ablate two design choices for our local editing: 1) conditioning on pose maps; 2) using body part masks. Fig 6 provides a visual comparison, which shows the importance of these two components to visual quality. When the pose information is not used, the model fails to make the desired edit. Without the body part mask, the edited result involves large global modifications to the input image, totally changing the original style. Removing one or both of the two components from our method lead to a degradation in both FID and CLIP-S across all the attributes (please see the supplemental for details).

Conclusion

In this paper, we propose a unified approach to facilitating fashion design based on a multi-stage diffusion model conditioned on hierarchical text descriptions. Our method naturally supports both design draft generation and iterative local editing, showing promising capability of fitting into and aiding in the typical fashion design workflow. We also contribute a fashion image dataset that comprise hierarchical text annotations, making it a valuable asset for training and evaluating various fashion design models. We hope that our idea of drawing analogy between the full fashion design pipeline and the denoising process of diffusion models, along with our dataset, can inspire and encourage future work in building practical systems for augmenting designers in the fashion domain and beyond.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62402306 and the Shanghai Natural Science Foundation of China under Grant No. 24ZR1422400.

References

- Ak, K. E.; Kassim, A. A.; Lim, J.; and Tham, J. Y. 2019. Attribute Manipulation Generative Adversarial Networks for Fashion Images. In *ICCV*, 10540–10549.
- Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended Latent Diffusion. *TOG*, 42(4): 1–11.
- Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *ICCV*, 23336–23345.
- Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023a. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *ICCV*, 22503–22513.
- Cao, S.; Chai, W.; Hao, S.; and Wang, G. 2023b. Image Reference-guided Fashion Design with Structure-aware Transfer by Diffusion Models. In *CVPR Workshops*, 3525–3529.
- Cao, Z.; Simon, T.; Wei, S.; and Sheikh, Y. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *CVPR*, 1302–1310.
- Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *TOG*, 42(4): 1–10.
- Chen, L.; Tian, J.; Li, G.; Wu, C.; King, E.; Chen, K.; Hsieh, S.; and Xu, C. 2020. TailorGAN: Making User-Defined Fashion Designs. In *WACV*, 3230–3239.
- Dai, Q.; Yang, S.; Wang, W.; Xiang, W.; and Liu, J. 2021. Edit Like A Designer: Modeling Design Workflows for Unaligned Fashion Editing. In *ACM Multimedia*, 3492–3500.
- Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; and Tang, J. 2021. CogView: Mastering Text-to-Image Generation via Transformers. In *NeurIPS*, 19822–19835.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 7514–7528.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 6626–6637.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.
- Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*.
- Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2Human: text-driven controllable human image generation. *TOG*, 41(4): 1–11.
- Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.
- Kwon, Y.; Petrangeli, S.; Kim, D.; Wang, H.; Swaminathan, V.; and Fuchs, H. 2022. Tailor Me: An Editing Network for Fashion Attribute Shape Manipulation. In *WACV*, 3142–3151.
- Li, Z.; Zhou, H.; Bai, S.; Li, P.; Zhou, C.; and Yang, H. 2022. M6-Fashion: High-Fidelity Multi-modal Image Generation and Editing. *arXiv preprint arXiv:2205.11705*.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.
- Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress Code: High-Resolution Multi-category Virtual Try-On. In *ECCV*.
- Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. In *ICML*, 7176–7185. PMLR.
- Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, 16784–16804. PMLR.
- Pernus, M.; Fookes, C.; Struc, V.; and Dobrisesk, S. 2023. FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion. *arXiv preprint arXiv:2301.02110*.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 10674–10685.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.
- Sun, Z.; Zhou, Y.; He, H.; and Mok, P. Y. 2023. SGDiff: A Style Guided Diffusion Model for Fashion Synthesis. In *ACM Multimedia*, 8433–8442.
- Wang, T.; and Ye, M. 2024. TextFit: Text-Driven Fashion Image Editing with Diffusion Models. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *AAAI*, 10198–10206.
- Yang, X.; Zhang, H.; Jin, D.; Liu, Y.; Wu, C.; Tan, J.; Xie, D.; Wang, J.; and Wang, X. 2020. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. In *ECCV*.
- Yu, J.; Wang, Y.; Zhao, C.; Ghanem, B.; and Zhang, J. 2023. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. In *ICCV*, 23117–23127.

Zhang, X.; Sha, Y.; Kampffmeyer, M. C.; Xie, Z.; Jie, Z.; Huang, C.; Peng, J.; and Liang, X. 2022. ARMANI: Part-level Garment-Text Alignment for Unified Cross-Modal Fashion Design. In *ACM Multimedia*, 4525–4535.

Zhang, X.; Yang, B.; Kampffmeyer, M. C.; Zhang, W.; Zhang, S.; Lu, G.; Lin, L.; Xu, H.; and Liang, X. 2023. DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cross-modal Semantic Alignment. In *ICCV*, 23097–23106.

Zhu, S.; Fidler, S.; Urtasun, R.; Lin, D.; and Loy, C. C. 2017. Be Your Own Prada: Fashion Synthesis with Structural Coherence. In *ICCV*, 1689–1697.