IMAGGarment+: Efficient Attribute-Wise Diffusion for Garment Generation

Jian Yu[1*], **Fei Shen**[2*], **Cong Wang**[3], **Yanpeng Sun**[2], **Hao Tang**[4], **Qin Guo**[5], **Xiaoyu Du**[1†]

[1]Nanjing University of Science and Technology
[2]National University of Singapore
[3]Nanjing University
[4]The Hong Kong Polytechnic University
[5]The Hong Kong University of Science and Technology

Abstract

Diffusion models have advanced fine-grained garment generation, yet balancing controllability, efficiency, and texture fidelity remains challenging. Adapter-based methods often yield incoherent details, while full fine-tuning is computationally expensive and prone to overwriting pretrained priors. To address these limitations, we propose IMAGGarment+, an efficient diffusion framework for controllable and high-quality garment synthesis. It comprises two key modules designed for efficient and attribute-aware conditioning. First, we introduce an attribute-wise feature extractor (AFE) that disentangles key garment attributes, silhouette, logo, position, and color, into parallel latent streams. Each stream is optimized independently via LoRA, ensuring minimal parameter overhead while retaining expressive capacity. Second, we develop an attribute-adaptive attention (AA) module to inject attribute-specific cues into the generative process through a selective, layer-wise injection strategy. Specifically, silhouette and color features are injected into early decoder layers to guide structural and appearance formation, while logo features are propagated across all layers to ensure cross-scale consistency. Extensive experiments on fine-grained garment benchmarks demonstrate that IMAGGarment+ outperforms state-of-the-art baselines with less than 20% additional parameters, validating its effectiveness and efficiency.

Introduction

Generative artificial intelligence (Podell et al. 2023) has recently made substantial progress, driving significant advancements in various applications within the fashion industry. In particular, garment generation aims to rapidly synthesize high-fidelity apparel images based on multimodal design inputs (Tang, He, and Qin 2025; Tang et al. 2023, 2020, 2022), such as hand-drawn silhouettes, color schemes, graphic logos, and spatial attributes. This capability dramatically reduces the traditional time and labor costs associated with apparel design, enhancing user immersion and realism in virtual scenarios (*e.g.*, virtual dressing (Chen et al. 2024a; Shen et al. 2025a) and virtual try-on (Kim et al. 2024b; Baldrati et al. 2023)) through personalized garment visualization. Moreover, given the accelerating growth of online shopping, garment generation offers promising potential as an effective promotional tool, facilitating innovative marketing strategies for fashion brands and personalized e-commerce experiences.

Early garment generation methods (Zhang et al. 2023; Cao et al. 2024; Yan et al. 2025; Zhu et al. 2024) typically rely on a single or limited number of conditioning inputs. For example, DiffCloth (Zhang et al. 2023) aligns textual attribute phrases with garment regions and incorporates a semantic-binding attention mechanism to enable cross-modal generation. AnimeDiffusion (Cao et al. 2024) employs a two-stage hybrid training strategy to disentangle structure and color, significantly improving generalization in sketch-to-garment colorization. Similarly, ColorizeDiffusion (Yan et al. 2025) introduces a two-stage latent variable approach to mitigate distribution shifts, enhancing colorization quality and reducing artifacts. LogoSticker (Zhu et al. 2024) adopts a two-stage framework combining relational pretraining and identity disentanglement, enabling accurate and natural integration of logos into garments. Although these methods have progressed in controlling specific attributes, most support only a limited set of conditions and fall short of addressing the unified modeling of multiple visual attributes, such as silhouette, color, texture, and logo, required in real-world applications like virtual try-on and automated fashion design.

To address the increasing demand for fine-grained controllability in garment generation, recent research (Zhang, Rao, and Agrawala 2023; Li, Li, and Hoi 2023; Shen et al. 2025b) has leveraged advances in multi-attribute image generation. Existing approaches can be broadly categorized into two types: adapter-based methods and fine-tuning methods. For adapter-based methods, they inject efficient modules into frozen diffusion models for conditional control. For instance, ControlNet (Zhang, Rao, and Agrawala 2023), T2I-Adapter (Mou et al. 2024) and IP-Adapter (Ye et al. 2023) can be used to control different attributes individually. While a single adapter incurs minimal overhead, multi-attribute modeling typically requires training multiple adapters, substantially increasing overall computational cost. In contrast, fine-tuning methods retrain part or all of the model to enhance generation capabilities. BLIP-Diffusion(Li, Li, and Hoi 2023) improves visual fidelity by training on paired data, but at the expense of significant computational re-

[*]These authors contributed equally.
[†]Corresponding author.

sources and the risk of forgetting pretrained knowledge. IMAGGarment-1(Shen et al. 2025b) adopts a two-stage framework to model different attributes explicitly. Although it improves attribute-level controllability, it suffers from limited generalization and lacks efficient end-to-end optimization. Current methods lack a unified and flexible mechanism for disentangling and controlling multiple attributes, often leading to conflicts among attribute representations and degraded image fidelity and controllability. Besides, achieving a balance between efficiency and fine-grained controllability remains a core challenge that has yet to be fully addressed.

To address the aforementioned challenges, we introduce IMAGGarment+, an efficient, end-to-end framework tailored for garment synthesis under multiple attribute conditions. Our approach leverages pretrained, frozen diffusion models to exploit their intrinsic prior knowledge, enabling efficient and refined optimization. Specifically, we propose an attribute-wise feature extractor (AFE) module that simultaneously encodes multiple garment attributes (*e.g.*, silhouette, color, logo) in a fine-grained manner, while maintaining low parameter overhead through LoRA-based adaptation. Furthermore, we develop an attribute-adaptive attention (AA) module, which strategically injects attribute-specific representations into UNet layers. Guided by empirical analyses, silhouette and color semantics are selectively integrated into the most suitable layers, whereas logo features are consistently incorporated across all layers to ensure cross-scale consistency. By effectively mitigating representational conflicts inherent in multi-attribute conditioning, IMAGGarment+ achieves precise, controllable garment generation with minimal computational cost. To summarize, our main contributions can be outlined as follows:

- We propose IMAGGarment+, an efficient diffusion framework that achieves controllable garment generation through the integration of attribute-wise feature extraction and attribute-adaptive conditioning strategies.
- We design an efficient attribute-wise feature extractor (AFE), which simultaneously encodes garment attributes, including silhouette, logo, position, and color, in a parallel manner, optimized exclusively via LoRA to minimize parameter overhead.
- We present an attribute-adaptive attention (AA) module with a selective, layer-wise injection strategy that aligns attribute semantics with hierarchical decoder stages, facilitating precise structure-texture coordination while preserving efficiency.

Related Work

Controllable Diffusion-Based Generation

Diffusion models (Shen et al. 2024; Shen and Tang 2024; Shen et al. 2025a) have demonstrated impressive capabilities in controllable image generation across a wide range of domains, including portrait synthesis (Tao et al. 2025; Kim et al. 2024a), scene editing (Brooks, Holynski, and Efros 2023), and garment composition (Zhang et al. 2024). To enable attribute-level control, existing methods typically adopt one of two strategies: (i) designing condition-specific inputs such as layouts, semantic maps, instance images, and sketches, or (ii) modifying parts of the diffusion model (*e.g.*, decoder) to align generation with external constraints. For example, ARMANI (Zhang et al. 2022) employs garment sketches and partial clothing as visual priors to enhance controllability in fashion generation. AnyDoor (Chen et al. 2024b) fine-tunes the decoder of the denoising model to enable precise logo reconstruction. This limitation poses a challenge in real-world applications, where garment generation must jointly consider diverse and interdependent attributes such as silhouette, color, texture, and logo.

Efficient Conditioning and Parameterization

To reduce the computational overhead of adapting large diffusion models, recent works have explored parameter-efficient tuning strategies such as adapter modules, LoRA, and prompt tuning. Adapter-based approaches (Zhao et al. 2023; Zhang, Rao, and Agrawala 2023) insert small trainable modules into frozen backbones to enable flexible conditional injection without modifying pretrained weights. For instance, IP-Adapter (Ye et al. 2023) employs compact encoders and linear projections for visual conditioning, though it often struggles with fine-grained semantic alignment. LoRA (Hu et al. 2022) improves efficiency by approximating weight updates via low-rank decomposition, allowing only a small subset of parameters to be trained. DreamFit (Lin et al. 2025) applies LoRA within attention layers to support virtual try-on generation. Prompt tuning (Lester, Al-Rfou, and Constant 2021; Li and Liang 2021) further reduces training cost by guiding pretrained models through learnable continuous vectors. While effective, these methods typically assume uniform feature injection across network layers and overlook attribute-specific alignment. This limits their capacity to model semantic hierarchies or accommodate diverse, interdependent visual cues.

Methodology

Problem Definition. We formulate garment generation as a multi-conditional image synthesis task, aiming to produce photorealistic and structurally coherent apparel images under the guidance of multiple explicitly controllable visual attributes. The problem can be formally expressed as:

$$I_{gen} = G(S, M, H, L), \tag{1}$$

where I_{gen} denotes the generated garment image, and G represents the proposed generative model. The inputs include a structural silhouette S, a chromatic hint map H, a user-specific logo L, and its associated spatial mask M, which defines the intended logo placement region on the garment.

Architecture Overview. To ensure faithful attribute control and high-fidelity synthesis, our framework consists of two key components: the attribute-wise feature extractor (AFE) and the attribute-adaptive attention (AA) module, as shown in Figure 1. The AFE module disentangles multiple visual attributes and encodes them into parallel latent streams. These streams are adapted using LoRA to maintain fine-grained expressiveness with minimal parameter overhead. The resulting features are injected into the Denoising UNet through the AA module, which aligns attribute se-

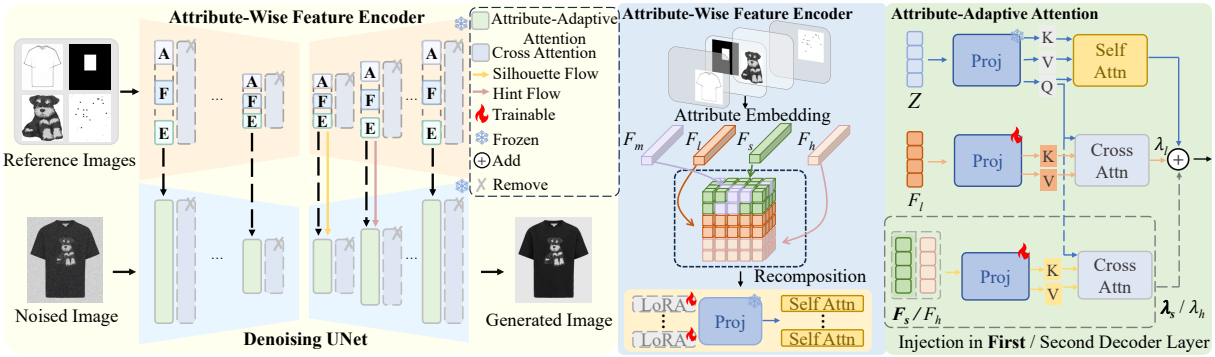


Figure 1: **Overview of our IMAGGarment+ framework.** We design an attribute-wise feature encoder leveraging LoRA layers. The features from multiple control conditions are simultaneously extracted by the attribute-wise feature encoder and seamlessly integrated into the denoising UNet via attribute-adaptive attention mechanisms. Silhouette flow and hint flow respectively represent the pathways through which silhouette and hint features are injected into the network.

mantics with decoder layers via a selective injection mechanism. Specifically, features corresponding to silhouette and color are injected into early decoder layers to guide structural layout and chromatic distribution, while logo features are propagated across all layers to preserve fine-grained consistency. This design forms the basis of our layer-wise injection strategy, which ensures effective cross-attribute coordination while preserving computational efficiency.

Attribute-Wise Feature Extractor

To ensure fine-grained control over multiple garment attributes, we design an attribute-wise feature extractor (AFE) that decouples conditional representations across attribute dimensions while maintaining parameter efficiency. Previous methods (Lin et al. 2025; Shen et al. 2025a) typically employ separate diffusion UNet encoders for each condition to preserve semantic specificity. However, such stacking strategies significantly increase model complexity and impose a heavy training burden. Observing that self-attention layers play a crucial role in enabling dynamic information interaction among features, we retain shared parameters for general layers and selectively adapt attention blocks using LoRA. This hybrid reuse-adapt strategy enables the model to extract expressive attribute-specific features without compromising generation fidelity.

To enhance spatial controllability, we first multiply the silhouette S with the spatial mask M to obtain a refined structural input. The processed silhouette, logo, and color hint are separately encoded into latent vectors via a VAE encoder. Then they are concatenated, and forwarded into the AFE module. During this process, the spatial mask feature denoted as F_m is integrated into the processed silhouette feature F_s to enable precise positional control. As shown in Figure 1, the AFE replaces the original single self-attention layer with multiple parallel attention branches, each dedicated to a distinct attribute stream. Formally, for each attribute $i \in \{l, s, h\}$ (logo, silhouette, and hint), we compute:

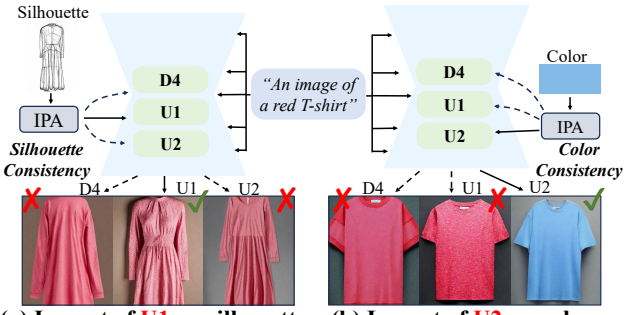
$$F_i^{new} = \text{Softmax}\left(\frac{\hat{Q}_i \hat{K}_i^\top}{\sqrt{d}}\right)\hat{V}_i\,, \tag{2}$$

where $\hat{Q}_i = F_i(\hat{W}_q + \Delta\hat{W}_q^i)$, $\hat{K}_i = F_i(\hat{W}_k + \Delta\hat{W}_k^i)$, and $\hat{V}_i = F_i(\hat{W}_v + \Delta\hat{W}_v^i)$. Here, F_i denotes the input feature for each attribute, and $\hat{W}_q$, $\hat{W}_k$, $\hat{W}_v$ are the frozen weights initialized from SDXL, while $\Delta\hat{W}_q^i$, $\Delta\hat{W}_k^i$, and $\Delta\hat{W}_v^i$ are LoRA-based residuals that provide attribute-specific adaptation. The resulting features F_i^{new} are concatenated across the batch dimension and passed into the downstream modules. To improve efficiency, we remove the redundant cross-attention layers from the original UNet and adopt a selective update strategy aligned with the layer-wise injection principle. Specifically, since silhouette and color features are only injected into the first two decoder layers, we omit their dedicated self-attention and LoRA layers beyond those layers. Consequently, the AFE introduces only 177.6M additional parameters, and training is restricted solely to these LoRA layers. This design reduces the number of trainable parameters by 93% compared to full fine-tuning while preserving strong attribute alignment and controllability.

Layer-Wise Injection Strategy

We observe that different conditional attributes may influence distinct aspects of the image generation process, and that injecting all attribute features uniformly across UNet layers can lead to suboptimal or conflicting effects. To verify this, we conduct a layer-sensitivity analysis using the IP-Adapter (Ye et al. 2023) framework, as shown in Figure 2. Specifically, we inject visual inputs (*e.g.*, a silhouette or a color hint) into one designated decoder layer, while applying a contradictory text prompt (*e.g.*, "a T-shirt") to the rest. This probing strategy reveals that silhouette cues most strongly affect generation when injected into the first decoder layer, where global structure is reconstructed, while color cues are most effective in the second decoder layer, related to color refinement. In contrast, logo features, which are detail-sensitive and spatially localized, require reinforcement across all layers to preserve fidelity and consistency.

Based on these findings, we formulate a layer-wise injection strategy that assigns each attribute to its most semantically compatible layer: silhouette features are injected into the first decoder layer to guide spatial layout, color features into the second layer for chromatic modulation, and logo



(a) Impact of U1 on silhouette (b) Impact of U2 on color

Figure 2: **Influence of UNet layers** on silhouette and color consistency. We observe that the first decoder layer governs silhouette, while the second modulates color.

features across all layers to ensure cross-scale detail preservation. This targeted injection not only enhances controllability but also reduces redundant computation, enabling efficient fine-tuning of attribute-specific LoRA modules. The layer-wise strategy serves as a foundational design prior for our attribute-adaptive attention, guiding where and how conditional features should be integrated during generation.

Attribute-Adaptive Attention

To fully leverage multi-attribute features as conditional guidance, we introduce an attribute-adaptive attention module that selectively injects attribute-specific representations into different layers of the denoising UNet. As illustrated in Figure 1, the features extracted by the AFE module are decomposed into individual attribute streams, each injected into the latent representation via dedicated cross-attention.

To align the spatial semantics of each attribute with the generative process, we employ a region-level alignment mechanism tailored to their visual characteristics. Specifically, logo features, which require high-detail preservation, are injected into all layers to ensure cross-scale consistency. In contrast, silhouette and color features, which primarily influence global layout and chromaticity, are selectively injected into the first and second decoder layers, respectively. This selective injection mitigates feature interference while enhancing attribute-specific guidance. Formally, the attention-augmented latent feature Z_{new} is computed as:

$$Z_{new} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V + \lambda_l \cdot \text{Softmax}\left(\frac{QK_l^\top}{\sqrt{d}}\right)V_l + I(x), \tag{3}$$

where $I(x)$ denotes the injection term corresponding to silhouette and color cues at decoder layer x, defined as:

$$I(x) = \begin{cases} \lambda_s \cdot \text{Softmax}\left(\frac{QK_s^\top}{\sqrt{d}}\right)V_s & , \quad \text{if } x = 1 \\ \lambda_h \cdot \text{Softmax}\left(\frac{QK_h^\top}{\sqrt{d}}\right)V_h & , \quad \text{if } x = 2 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{4}$$

Here, Z is the latent feature, and $Q = W_q Z$, $K = W_k Z$, and $V = W_v Z$ are projections using frozen weights from the pretrained SDXL backbone. For each attribute $i \in \{s, l, h\}$ (silhouette, logo, hint), we compute $K_i = W_k^i F_i$ and $V_i = W_v^i F_i$ via trainable linear projections applied to the corresponding attribute feature F_i. The hyperparameter λ_i modulates the injection strength of each attribute stream.

Method	CSS ↓	LLA ↑	FID ↓	LPIPS ↓
DiffCloth*	108.20	0.15	137.97	0.60
BLIP-Diffusion*	104.44	0.13	106.85	0.68
DreamFit*	81.99	0.33	72.25	0.42
AnyDoor*	68.24	0.65	40.90	0.17
IMAGGarment-1*	<u>57.29</u>	<u>0.77</u>	<u>16.51</u>	<u>0.12</u>
Ours	**38.30**	**0.90**	**16.16**	**0.11**

Table 1: **Quantitative results** compared with SOTA methods. IMAGGarment+ achieves the best performance, with the best results in **bold** and second-best underlined. * denotes methods re-implemented for fair comparison.

Training and Inference

During training, only the parameters of the LoRA layers and the linear projection layers within the attribute-adaptive attention module are updated. Let $\mathcal{C} = \{S, M, H, L\}$ denote the set of conditional inputs, including silhouette, mask, color hint, and logo. We adopt the mean squared error (MSE) loss, consistent with the original diffusion training objective (Rombach et al. 2022):

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,\mathcal{C},t}\left[\|\epsilon - \epsilon_G(z_t, \mathcal{C}, t)\|^2\right], \tag{5}$$

where ϵ denotes randomly sampled Gaussian noise, and ϵ_G refers to the noise predicted by the generative model G at timestep t during the denoising process.

During inference, we apply conditional classifier-free guidance (CFG) (Ho and Salimans 2022) to balance conditional and unconditional predictions:

$$\hat{\epsilon}(z_t, \mathcal{C}, t) = \omega \cdot \epsilon_G(z_t, \mathcal{C}, t) + (1 - \omega) \cdot \epsilon_G(z_t, t), \tag{6}$$

where ω controls the CFG strength, and z_t denotes the noisy latent features at timestep t.

Experiments

Implementation Details

Dataset. All experiments are conducted on the GarmentBench dataset (Shen et al. 2025b), which comprises 8,200 training samples annotated with sketch, color hint, logo, and spatial mask. A separate test set of 500 samples is used for evaluation. The dataset provides diverse attribute combinations and realistic garment appearances, enabling comprehensive evaluation of multi-condition controllability.

Metrics. We adopt four representative metrics to evaluate the generated results in terms of visual realism and conditional controllability. Color structure similarity (CSS) (Zeng et al. 2014) quantifies color fidelity by comparing spatial color patterns. Logo location accuracy (LLA) (Fujitake 2024) assesses the spatial controllability of logo placement. Fréchet inception distance (FID) (Heusel et al. 2017) and learned perceptual image patch similarity (LPIPS) (Zhang

(a) In-Domain Conditions (b) Out-of-Domain Combinations

Figure 3: **Qualitative comparisons** with SOTA methods on GarmentBench. IMAGGarment+ produces significantly more realistic garment images with faithful attribute control compared to existing approaches.

Method	Quality	Similarity	Controllability
AnyDoor	0.4	0.2	0.3
IMAGGarment-1	0.7	0.6	0.3
Ours	**1.9**	**2.2**	**2.4**

Table 2: **User study results.** Higher scores indicate better perceived realism and user preference across all metrics.

et al. 2018) are used to measure distributional similarity and perceptual quality, respectively.

Hyper-Parameters. IMAGGarment+ initialized with pre-trained weights from SDXL (Podell et al. 2023) to leverage prior visual knowledge. During training, we set the LoRA rank to 128 and optimize only the parameters of the LoRA of attribute-wise feature encoder and attribute-adaptive attention module. All models are trained on paired images resized to 640×512, using the AdamW optimizer (Loshchilov and Hutter 2019) with a fixed learning rate of 1×10^{-5}. Training is conducted for 130k steps with a batch size of 10 on 2 NVIDIA A800 GPUs. At inference time, we use the DDIM sampler (Song, Meng, and Ermon 2022) with 50 steps and set the classifier-free guidance (CFG) (Ho and Salimans 2022) scale $\omega = 2.5$. The attribute injection weights λ_s, λ_l, and λ_h are determined via cross-validation and fixed at 1.0 for all experiments. To enhance CFG robustness, we apply a 5% dropout rate to each conditioning input.

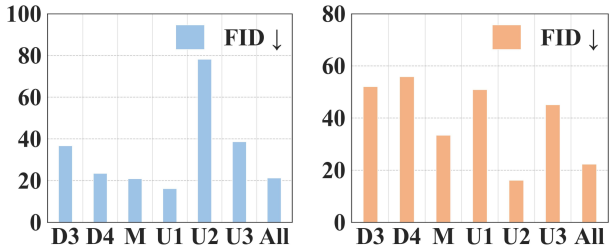
Main Comparisons

We conducted a comparative analysis of our proposed method against five relevant state-of-the-art approaches: DiffCloth (Zhang et al. 2023), BLIP-Diffusion (Li, Li, and Hoi 2023), DreamFit (Lin et al. 2025), AnyDoor (Chen et al. 2024b) and IMAGGarment-1 (Shen et al. 2025b). Since Dreamfit can only control single visual conditions, we em-

Figure 4: **Qualitative ablation results** of IMAGGarment+.

Method	CSS ↓	LLA ↑	FID ↓	LPIPS ↓
B0 (w/o AFE)	94.50	0.51	77.92	0.33
B1 (w/o AA)	54.34	0.77	25.35	0.20
B2 (w/o layer-wise injection)	61.22	0.89	21.30	0.12
Ours	**38.30**	**0.90**	**16.16**	**0.11**

Table 3: **Quantitative ablation results** of IMAGGarment+.



(a) Silhouette injection locations (b) Color injection locations

Figure 5: **Injection location analysis.** U1 and U2 achieve the best FID scores for silhouette and color, respectively.

ploy a concatenation approach to adapt its input for handling multiple conditions.

Quantitative Results. As shown in Table 1, we perform a comprehensive quantitative evaluation of our method against the baselines. Specifically, since they lack dedicated mechanisms for capturing diverse logo patterns, BLIP-Diffusion (Li, Li, and Hoi 2023) and AnyDoor (Chen et al. 2024b) struggle to preserve logo fidelity, resulting in degraded performance on CSS and FID. Compared to DreamFit (Lin et al. 2025), our method effectively disentangles and integrates features of different conditions through its attribute-wise feature extractor and attribute-adaptive attention mechanism. In conclusion, our IMAGGarment+ outperforms all compared methods across all evaluated metrics, demonstrating its robust capability in conditional control and faithful detail preservation.

Qualitative Results. Figure 3 illustrates the qualitative comparison between IMAGGarment+ and baseline methods. We observe that, despite producing visually commendable images, BLIP-Diffusion (Li, Li, and Hoi 2023) and Diff-Cloth (Zhang et al. 2023) demonstrate significant inconsistencies between generated attributes and the specified conditions. Moreover, due to intrinsic limitations in feature extraction granularity and interference among multiple conditioning factors, other methods often produce generated images containing substantial redundant artifacts and unwanted distortions. In contrast, our approach exhibits superior fidelity in logo reconstruction, accompanied by pronounced consistency in structural integrity and chromatic distribution.

User Study. We conducted a user study with 30 participants to evaluate garment generation from a perceptual perspective. For each sample, three results from different methods were shown side-by-side, and participants were asked to rank them with scores of 3 (best), 2 (fair), and 1 (worst) based on three criteria: quality, similarity, and controllability. Each participant evaluated 20 sets using a custom interface. As shown in Table 2, our method consistently receives the highest average ranking across all criteria, indicating superior visual realism and attribute alignment.

Ablation Study

Effectiveness of AFE. To assess the contribution of AFE, we replace it with the IP-Adapter (Ye et al. 2023) to integrate silhouette and color features. This modification constrains the model to coarse-grained feature representations. As illustrated in the second column of Figure 4, the generated image manifests attributes that are markedly incongruent with the original conditioning inputs. Furthermore, its inferior performance across all evaluation metrics in Table 3 also underscores the effectiveness of the AFE module in capturing fine-grained feature representations.

Effectiveness of AA. To demonstrate the effectiveness of the AA module, we substitute it with a feature fusion strategy that integrates latent and conditional features via element-wise addition. Visual results in Figure 4 clearly reveal that in the absence of this module, the model tends to produce disorganized and semantically inconsistent logo patterns with obvious decline in quality. As shown in the third column, the absence of the AA module impairs the model's ability to selectively attend to salient regions associated with distinct attributes and to allocate attention effectively, thereby exacerbating interference among conditional factors.

Injection Location Analysis. To assess the functional role of layer-wise injection, we perform an ablation study by removing it from the model. As shown in Figure 4, the lack of the layer-wise injection strategy causes the model to misclassify visually similar colors as target attributes and introduces inconsistencies in the delineation of clothing contours. Moreover, the reduced FID and CSS quantitative metrics reported in Table 3 indicate a decline in our model's ability

Figure 6: **Hyperparameter results** on the injection strength of logo λ_l, silhouette λ_s, and color hints λ_h.

to preserve structural integrity and color fidelity. To further investigate the influence of different layers, we systematically evaluate the effect of injecting attribute features into specific layers, including downsampling (D), middle (M), upsampling (U) layers and the entire UNet (All). Building on previous work (Voynov et al. 2023; Agarwal et al. 2025), we analyze six intermediate layers that are most likely to influence structural and color control. As shown in Figure 5 (a), the results indicate that injecting silhouette features into the first upsampling layer (U1) achieves the lowest FID score, markedly surpassing the performance of all other injection strategies. Furthermore, as presented in Figure 5 (b), injecting color features into the second upsampling layer (U2) delivers the most optimal performance, facilitating precise modulation of color distribution. This suggests that the structure-wise and color-wise compatibility of U1 and U2 helps mitigate the influence of adverse information, thereby facilitating the model's enhanced preservation of the coherence between garment contours and color distribution.

Hyperparameter Analysis. We investigate the impact of three critical hyperparameters defined in Equations 3 and 4: the injection strengths of the logo (λ_l), silhouette (λ_s), and color hint features (λ_h). As illustrated in Figure 6, the parameter λ_l regulates the completeness of the reconstructed logo. When λ_l is set too low, the output tends to produce cluttered or semantically incoherent patterns. We set $\lambda_l = 1.0$ to preserve fine-grained logo details. Similarly, parameters λ_s and λ_h govern the consistency of structural and color distributions, respectively. As their values increase, the stability of structural and chromatic consistency improves. We set both parameters to 1.0 to ensure optimal visual fidelity.

Comparison with Efficient Methods. Furthermore, we compare our method with three representative efficient adapter approaches: T2I-Adapter (Mou et al. 2024), ControlNet (Zhang, Rao, and Agrawala 2023) and IP-Adapter (Ye et al. 2023). We implement these methods on SDXL and adopt stacked configurations of their respective auxiliary modules to support multi-condition inputs. The quantitative results are presented in Table 4. Due to the absence

Method	CSS ↓	LLA ↑	FID ↓	LPIPS ↓	# Params (M) ↓	Inference Time (s) ↓
T2I-Adapter*	104.43	0.53	44.78	0.33	**316**	**3.86**
ControlNet*	103.48	0.27	55.20	0.52	5,004	10.32
IP-Adapter*	98.76	0.27	43.21	0.42	1,405	5.04
Ours	**38.30**	**0.90**	**16.16**	**0.11**	447	4.91

Table 4: **Quantitative comparison** with efficient methods. Our method achieves competitive performance.

Figure 7: **Qualitative comparison** with efficient methods. We achieve the most consistent results.

of specialized mechanisms for fine-grained feature extraction, these adapter-based methods struggle to preserve intricate details and maintain precise conditional control. This leads to inferior performance across all evaluation metrics. Additionally, compared to ControlNet (Zhang, Rao, and Agrawala 2023) and IP-Adapter (Ye et al. 2023), our model introduces a significantly reduced number of trainable parameters. Despite not having the fewest trainable parameters, our method possesses the second lowest count and achieves the best overall performance. This notable trade-off highlights the strong competitiveness and practical significance of our approach. Moreover, the visual results illustrated in Figure 7 reveal that these adapter-based methods predominantly focus on holistic feature representation, often compromising the retention of fine-grained details.

Conclusions

We present IMAGGarment+, an efficient and controllable diffusion framework tailored for garment generation under multi-attribute conditions. By introducing the attribute-wise feature extractor and attribute-adaptive attention modules, our method enables efficient and precise integration of diverse attributes into pretrained diffusion backbones with minimal parameter overhead. Through a selective layer-wise injection strategy, IMAGGarment+ effectively aligns conditional features with the structural roles of different decoder layers, resolving conflicts across heterogeneous inputs. Extensive experiments on garment benchmarks show that our framework achieves state-of-the-art controllability and visual fidelity, while remaining computationally efficient.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (62172226) and Jiangsu Provincial Science and Technology Major Project (No. BG2024042).

References

Agarwal, A.; Karanam, S.; Shukla, T.; and Srinivasan, B. V. 2025. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6053–6062. IEEE.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 23393–23402.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Cao, Y.; Meng, X.; Mok, P.; Lee, T.-Y.; Liu, X.; and Li, P. 2024. AnimeDiffusion: Anime diffusion colorization. *IEEE Transactions on Visualization and Computer Graphics*, 30(10): 6956–6969.

Chen, W.; Gu, T.; Xu, Y.; and Chen, A. 2024a. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6939–6948.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6593–6602.

Fujitake, M. 2024. Rl-logo: Deep reinforcement learning localization for logo recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2830–2834. IEEE.

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. *Advances in neural information processing systems*, 30.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kim, C.; Lee, J.; Joung, S.; Kim, B.; and Baek, Y.-M. 2024a. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024b. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8176–8185.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, D.; Li, J.; and Hoi, S. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36: 30146–30166.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lin, E.; Zhang, X.; Zhao, F.; Luo, Y.; Dong, X.; Zeng, L.; and Liang, X. 2025. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5218–5226.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025a. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Shen, F.; and Tang, J. 2024. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37: 6246–6266.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Shen, F.; Yu, J.; Wang, C.; Jiang, X.; Du, X.; and Tang, J. 2025b. Imaggarment-1: Fine-grained garment generation for controllable fashion design. *arXiv preprint arXiv:2504.13176*.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Tang, H.; He, S.; and Qin, J. 2025. Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning. In *IJCAI*, 6227–6235.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *ACM Multimedia*, 610–618.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition. In *ACM Multimedia*, 1719–1728.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognit.*, 130: 108792.

Tao, J.; Zhang, Y.; Wang, Q.; Cheng, Y.; Wang, H.; Bai, X.; Zhou, Z.; Li, R.; Wang, L.; Wang, C.; et al. 2025. InstantCharacter: Personalize Any Characters with a Scalable Diffusion Transformer Framework. *arXiv preprint arXiv:2504.12395*.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Yan, D.; Yuan, L.; Wu, E.; Nishioka, Y.; Fujishiro, I.; and Saito, S. 2025. ColorizeDiffusion: Improving Reference-Based Sketch Colorization with Latent Diffusion Model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5092–5102. IEEE.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zeng, K.; Wang, Z.; Zhang, A.; Wang, Z.; and Zhang, W. 2014. A color structural similarity index for image quality assessment. In *2014 IEEE International Conference on Image Processing (ICIP)*, 660–664. IEEE.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, S.; Chong, Z.; Zhang, X.; Li, H.; Cheng, Y.; Yan, Y.; and Liang, X. 2024. Garmentaligner: Text-to-garment generation via retrieval-augmented multi-level corrections. In *European Conference on Computer Vision*, 148–164. Springer.

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 *Proceedings of the 30th ACM International Conference on 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23154–23163.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 11127–11150.

Zhu, M.; Chen, X.; Wang, Z.; Zhao, H.; and Jia, J. 2024. Logosticker: Inserting logos into diffusion models for customized generation. In *European Conference on Computer Vision*, 363–378. Springer.