

GarmentAligner: Text-to-Garment Generation via Retrieval-augmented Multi-level Corrections

Shiyue Zhang¹, Zheng Chong^{1,3}, Xujie Zhang¹, Hanhui Li¹, Yuhao Cheng²,
Yiqiang Yan², and Xiaodan Liang^{1,3*}

¹ Shenzhen Campus of Sun Yat-sen University, Shenzhen, China

² Lenovo Research, Shenzhen, China

³ Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory, Shenzhen, China
xdliang328@gmail.com



Fig. 1: GarmentAligner is capable of producing high-quality garment images accurately depicting the quantity and spatial alignment of components specified in the provided captions.

Abstract. General text-to-image models bring revolutionary innovation to the fields of arts, design, and media. However, when applied to garment generation, even the state-of-the-art text-to-image models suffer from fine-grained semantic misalignment, particularly concerning the quantity, position, and interrelations of garment components. Addressing this, we propose *GarmentAligner*, a text-to-garment diffusion model trained with retrieval-augmented multi-level corrections. To achieve semantic alignment at the component level, we introduce an automatic component extraction pipeline to obtain spatial and quantitative information of garment components from corresponding images and captions. Subsequently, to exploit component relationships within the garment images, we construct retrieval subsets for each garment by retrieval augmentation based on component-level similarity ranking and conduct contrastive learning to enhance the model perception of components from positive and negative samples. To further enhance the alignment of components across semantic, spatial, and quantitative granularities, we propose the

* Corresponding author.

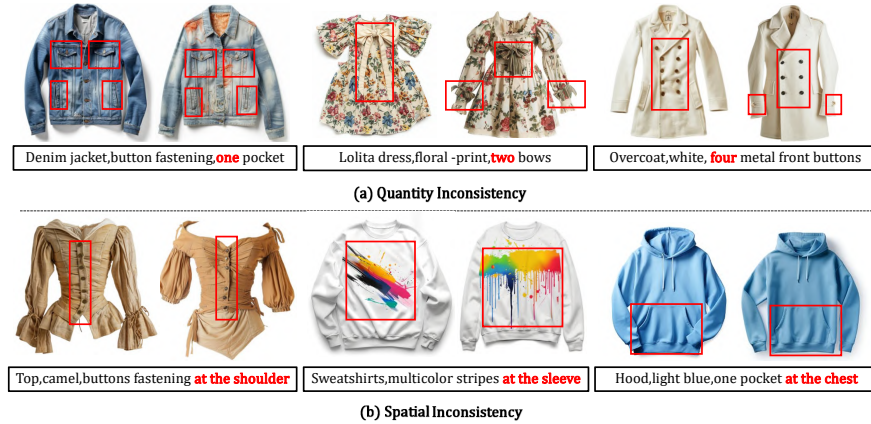


Fig. 2: The failures of state-of-the-art text-to-image model Midjourney in the text-to-garment task. The misalignment is primarily attributed to the quantities and spatial positions of components, thereby making it difficult to generate garments that meet the expected fine-grained details.

utilization of multi-level correction losses that leverage detailed component information. The experimental findings demonstrate that GarmentAligner achieves superior fidelity and fine-grained semantic alignment when compared to existing competitors.

Keywords: Garment generation · Diffusion model · Retrieval augmentation · Contrastive learning

1 Introduction

Text-to-garment generation, as one of the downstream tasks of text-to-image generation [24, 29, 31], has emerged as a promising technology in the field of computer vision, offering the potential to revolutionize the fashion industry by providing creative inspirations and accelerating the design process.

Despite the rapid developments in general text-to-image models [24, 29, 31], adapting these advanced models for garment generation tasks presents significant challenges. On the one hand, the visual semantics in the fashion domain significantly differ from those in general text-to-image generation. Garment captions possess specific textual structures and professional modifiers. Existing text-to-image models [24, 29, 31] struggle to achieve high consistency between textual descriptions and the generated garment images, as illustrated in Fig. 2. While some methods for garment generation [11, 33, 37, 43, 44] to capture the semantic structure of garment captions, thereby partially alleviating textual disparities, achieving fine-grained alignment still remains challenging. On the other hand, garment components, as crucial elements within garments, encompass distinct attributes and intricate interconnections, posing a challenge in the accurate generation of authentic garment images that satisfy the diverse requirements of

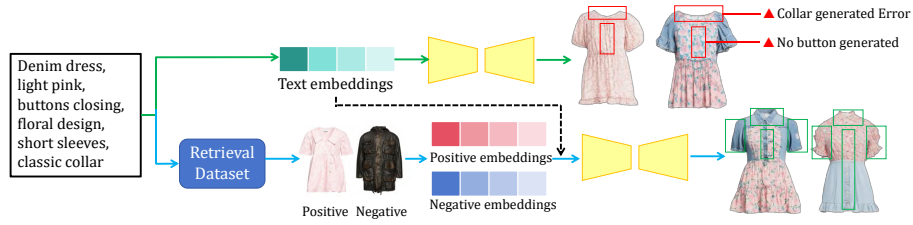


Fig. 3: The illustration of misalignment is addressed by retrieval-augmented contrastive learning. By assimilating insights from positive and negative samples retrieved via component-level similarity ranking, the model enhances its perception of component relationships.

these components. While several existing methods [37, 41, 43, 44] emphasize the importance of garment components, they primarily focus on the visual semantic aspect, neglecting the exploration of more extensive details such as component positioning and quantity.

Our key insight is to explore semantic information at various levels within garment images and their corresponding captions to enhance the overall quality and fine-grained details of generated garment images. To realize this objective, we initially devise an automated component extraction pipeline aimed at retrieving spatial and quantitative information of garment components from the associated images and captions. Nevertheless, optimizing the model directly using text-image pairs falls short of fully exploiting component-level information. This limitation confines the model to comprehend garment semantics solely at a holistic level.

To address this issue, we propose a training strategy integrating retrieval-augmented multi-level corrections to effectively adapt the general text-to-image diffusion model to the text-to-garment generation task. To enhance garment semantics and perception of component relationships, we implement a multi-level semantic similarity ranking approach for each sample. This approach utilizes extracted component-level information to construct its retrieval subset. Subsequently, pairs of positive and negative samples are extracted from the retrieval subset for contrastive learning, expanding the available training samples and learning the inherent relationship between components. The misalignment addressed by retrieval-augmented contrastive learning is shown in Fig. 3. Additionally, we propose corrections from three perspectives: visual perception, spatial alignment, and component quantity. Specifically, we utilize the CLIP Score [12] as the reward function for visual corrections, align spatial positioning using attention maps derived from text-conditioned cross-attention coupled with component segmentation, and ensure the generated components match the quantity specified in captions by conducting component detection during the training phase. The contributions of GarmentAligner can be summarized as follows:

- GarmentAligner introduces a retrieval-augmented contrastive learning strategy for the text-to-garment generation task, fully leveraging the rich informa-

tion extracted from garment components to enhance both the optimization of garment semantics and the perception of component relationships.

- GarmentAligner employs three component-level corrections, addressing visual, spatial, and quantitative aspects, thereby enhancing the fine-grained multi-level alignment of the text-to-garment model.
- Extensive comparative experiments conducted on the CM-Fashion dataset [43] with existing methods [5, 6, 8, 24, 28, 31, 34, 39, 43, 44] demonstrate the superiority of GarmentAligner in fine-grained multi-level semantic alignment for garment components, and detailed ablation studies validate the effectiveness of different parts of GarmentAligner.

2 Related Work

Text-guided image Generation. Many profound deep generative models [24, 28, 29, 31, 32] have emerged in the field of text-to-image generation, especially large-scale auto-regressive models and diffusion models. In the realm of autoregressive models, DALL-E [28] exhibited remarkable zero-shot capability, while Parti [42] illustrated the potential for scaling up autoregressive models. However, diffusion models, renowned for their capacity to progressively synthesize images from random noise, are predominantly favored due to their high diversity and sample quality. Notably, DALL-E 2 [27] operates within the image space of CLIP [25] to generate images, while Imagen [32] underscores the significance of leveraging pre-trained language models, employing a frozen T5 encoder [26]. The latent diffusion model [31] advocates encoding images with an autoencoder and subsequently utilizing diffusion models to generate continuous feature maps in latent space, thereby enhancing efficiency. Improving text-image consistency and enhancing image quality are fundamental objectives within the text-to-image generation domain. Numerous techniques have been explored for this purpose, including textual inversion [9], attention modulation [5], spatial control [1, 35], and human feedback [16, 36, 38] and so on, all yielding promising results. Despite these advancements, persistent fine-grained semantic imperfections remain in the generated outputs, particularly regarding the accurate positioning and quantity of depicted objects.

Garment Generation. As text-to-image generation emerges as one of the most impressive and promising applications in computer vision, downstream tasks in various scenarios are also rapidly evolving. Text-to-garment generation, as one of the downstream tasks, has garnered considerable attention due to its potential commercial value in the fashion design industry. Some text-to-garment methods [2, 18, 19, 22, 37, 41] focus on the generation and manipulation of garments on human bodies, while others like [11, 43, 44] directly convert given captions into in-shop garment images. Most of these methods [2, 11, 22, 33, 44] adopt pre-trained text-to-image latent diffusion model [31] as the backbone, adapting pre-trained models to the domain of text-to-garment through adjustments to captions or text embeddings or fine-tuning the backbone. Some efforts have focused on the significance of clothing components. Diffcloth [44] and ARMANI [43] utilize divided

captions and semantic segmentation of garment images to achieve component-level alignment. HieraFashDiff [37] partitions hierarchical design concepts to enable editing of components of garments worn by models. Fashion-Diffusion [41] introduces a dataset annotated at the component level. Despite numerous explorations in the domain of text-to-garment generation, results still exhibit granularity semantic flaws, with spatial positioning and quantity alignment being particularly severe.

Retrieval Augmentation. Retrieval augmentation is initially introduced to alleviate out-of-distribution performance degradation in language modeling and natural language processing. These methods [4, 10, 14, 17] typically employ a retriever to extract pertinent documents from an external database, subsequently employing a generator to reference these documents for prediction. By externalizing a model’s knowledge, retrieval augmentation offers advantages in accuracy, controllability, and scalability. Consequently, many methods have since incorporated this mechanism into computer vision tasks. For instance, RAFIC [20] utilizes retrieval augmentation to enhance few-shot image classification performance. RDMS [3] introduce the retrieval-augmented mechanism to diffusion models by training a conditional generative model through neighbor retrieval and flexible external conditioning inference. Re-Imagen [7] substitutes KNN neighbors for distance in latent space, allowing retrieved neighbors to consist of both images and text. RA-CM3 [40] proposes a unified approach to retrieve and generate any combination of text and images. Inspired by these methodologies, we implement retrieval augmentation at multiple levels on the text-to-garment dataset to generate samples for contrastive learning, thereby guiding holistic perception enhancement within the model.

3 Method

The primary objective of GarmentAligner is to enhance the alignment between input text prompts and generated garment images across multiple semantic levels in both holistic perception and fine-grained semantics. To achieve this, GarmentAligner integrates pre-trained latent diffusion models [31] as backbones to exploit their inherent knowledge and finetune the pre-trained text-to-image backbone with retrieval-augmented multi-level corrections to adapt it to the text-to-garment generation domain. Besides, an automatic component extraction pipeline (refer to Sec. 3.1) is adopted to attain in-depth component-level information from garment images with advanced open-domain detection and segmentation methods [15, 21, 30]. The overview of the proposed training strategy is depicted in Fig. 4.

The proposed GarmentAligner comprises two key parts:

- A retrieval-augmented contrastive learning approach (refer to Sec. 3.2) to utilize positive and negative samples retrieved from a subset constructed via semantic similarity ranking. This approach leverages an extensive sample pool to enhance the pre-trained model’s semantic perception capabilities.

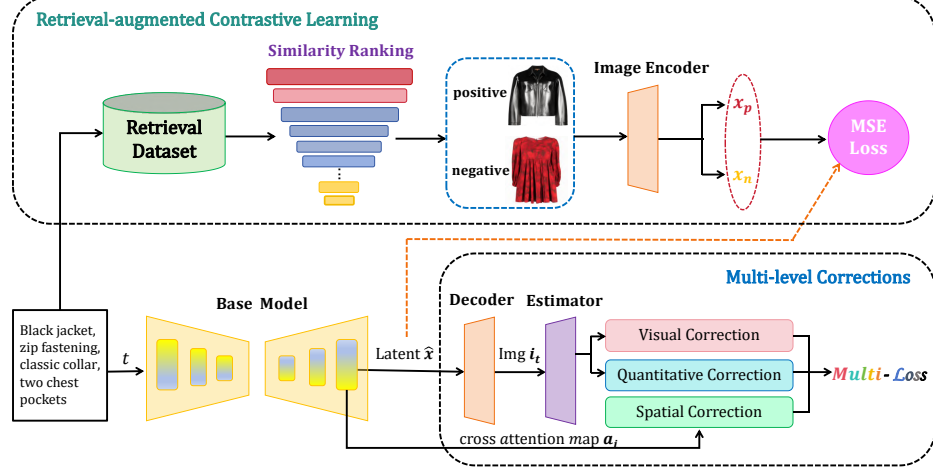


Fig. 4: The overview of the proposed GarmentAligner. During training, retrieval samples are systematically constructed, utilizing multi-level semantic similarity ranking for contrastive learning, with the objective of attaining global perceptual alignment. Simultaneously, multiple correction losses are employed to refine the visual semantics, spatial positions, and the quantity of garment components, thereby augmenting the granularity of details.

- Multi-level corrections (refer to Sec. 3.3) to address the deficiency in fine-grained semantic alignment of garment components across perceptual, spatial, and quantitative perspectives.

3.1 Automatic Component Extraction Pipeline

The CM-Fashion dataset [43] consists of images depicting in-shop garments accompanied by corresponding textual descriptions. However, these textual descriptions primarily emphasize overall characteristics, lacking detailed component-level quantity and spatial information. Consequently, models trained on this data often encounter misalignment issues when tasked with incorporating such specifics, thereby impeding the generation of high-quality garment images.

To attain comprehensive component-level insights, we utilize open-domain detection and segmentation models on garment images to meticulously identify and segment garment components. Specifically, we initially preprocess garment images with GroundingDINO [21] to obtain bounding boxes of the targeted components. Subsequently, we determine the quantity of each component by counting these boxes, and the position of each target component is discerned by computing the geometric center of the boxes. For the segmentation at the component level, we first deploy the garment parsing model [43] for primary segmentation results. We then apply SAM [15], in conjunction with the garment image and the acquired boxes, to enhance the segmentation of components not previously

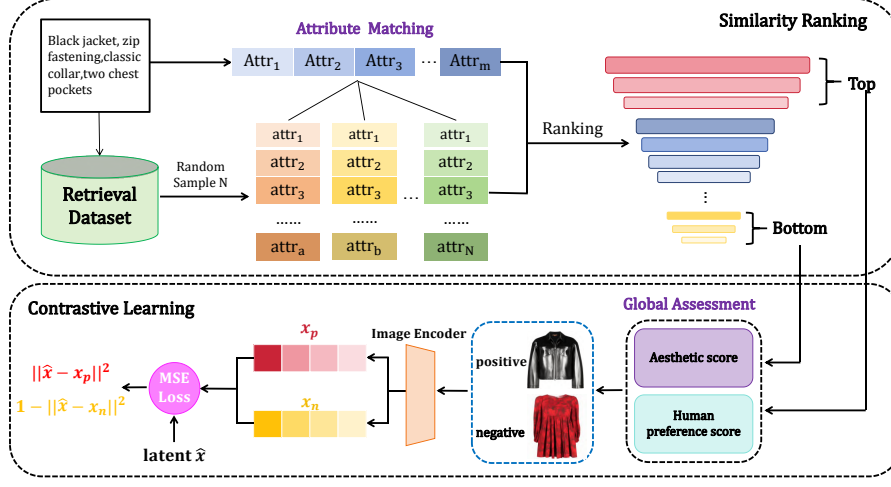


Fig. 5: The illustration of proposed Retrieval-augmented Contrastive Learning. Retrieval for each sample is performed within a randomly selected subset containing N samples, based on component-level semantic similarity ranking. Subsequently, the retrieval outcomes undergo global assessment filtering to acquire positive and negative samples for contrastive learning.

addressed by [43]. Finally, we align the information extracted from images with the textual descriptions to enrich the textual annotations with additional, precise component-level quantitative and spatial information.

Through this process, we establish a text-to-garment dataset based on the CM-Fashion dataset [43] comprising garment images, detailed captions, component segmentation, position, and quantity information.

3.2 Retrieval-augmented Contrastive Learning

While we have constructed a high-quality text-to-garment dataset, its scale falls significantly short in comparison to general text-to-image tasks. To address the issue of data scale, we propose a retrieval-augmented contrastive learning approach for fine-tuning the text-to-garment model with even limited available data. Fig. 5 illustrates the process and details of the proposed retrieval-augmented contrastive learning.

Given a garment sample, we conduct the retrieval among a random subset with N samples by semantic similarity ranking. The semantic similarity is computed utilizing extracted information concerning component-level quantity and semantics. Specifically, for a pair of samples (x, y) , comprising quantities (q_i^x, q_i^y) and text descriptions (t_i^x, t_i^y) for the i -th component, the similarity score for the i -th component between (x, y) can be expressed as:

$$S(x, y, i) = \frac{1}{|q_i^x - q_i^y| + Jaro(t_i^x, t_i^y)}, \quad (1)$$

where $Jaro(s_1, s_2)$ represents the Jaro distance between string s_1 and s_2 , and $|\cdot|$ denotes the absolute value. Consequently, the overall semantic similarity score of pair (x, y) across all k components can be expressed as:

$$S(x, y) = \sum_{i=1}^k S(x, y, i) - \alpha \cdot Jaro(t_x, t_y), \quad (2)$$

where t_x, t_y represent the garment captions of x and y , and α is the balancing weight.

The retrieval process generates a retrieval subset for each sample, enabling the expansion of a single garment sample to $N_p \times N_n$ sample pairs for contrastive learning, based on the similarity ranking of the retrieval subset and given positive and negative sampling numbers N_p and N_n . To enhance the efficacy of positive and negative sample comparisons, we incorporate aesthetic scores and human preference scores [36] for sample evaluation. This evaluation entails the selection of higher-scoring positive samples and lower-scoring negative samples, accentuating the disparity between them.

During the training phase, the optimization objective aims to minimize the Mean Squared Error (MSE) between the predicted latent result \hat{x} and positive samples x_p , while simultaneously maximizing the MSE between \hat{x} and negative sample x_n . Mathematically, this objective can be formulated as:

$$\mathcal{L}_{\text{RACL}} = \|\hat{x} - x_p\|^2 + 1 - \|\hat{x} - x_n\|^2, \quad (3)$$

where $\|a - b\|^2$ denotes the computation of the mean squared error between a and b .

3.3 Multi-level Corrections

To further enhance the fine-grained details at the component level, GarmentAligner involves three component-level corrections (see Fig. 6) into the training process of the diffusion model to provide feedback on the visual, spatial, and quantitative alignment for garment components.

Visual alignment. The CLIP Score [12] is employed as a reward function to align the text and image of the components. Specifically, upon obtaining a generated image \hat{X} , we identify the rough area of i -th component using the component mask m_i derived from the ground truth. Subsequently, we calculate the CLIP Score between the masked component area of \hat{X} and the corresponding text description t_i . The visual correction loss is computed as the following equation:

$$\mathcal{L}_{\text{visual}} = \sum_{i=1}^k \frac{1}{\text{CLIPScore}(m_i \odot \hat{X}, t_i)}, \quad (4)$$

where $\text{CLIPScore}(x, t)$ denotes calculating the CLIP Score [12] between image x and caption t .

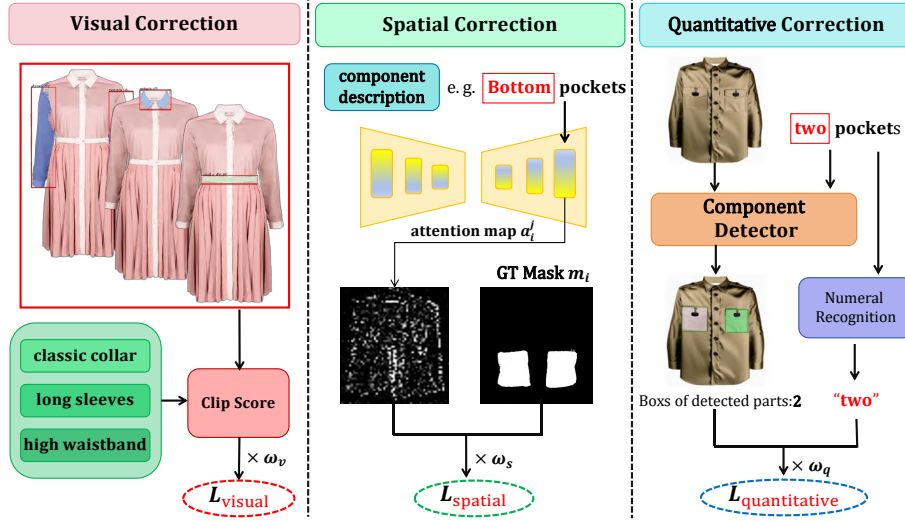


Fig. 6: The illustration of the proposed Multi-level Corrections. The generated garment images undergo decomposition at the component level, followed by correction procedures involving alignment to ensure text-image consistency, spatial cross-attention maps, and component quantities alignments.

Spatial alignment. We extract spatial attention maps $A_i = \{a_i^1, \dots, a_i^l\}$ corresponding to the i -th component description t_i within the diffusion model. Subsequently, we leverage the component mask m_i obtained from the ground truth to guide a_i^j by computing the MSE loss. This strategy facilitates the model in learning the correspondence from a component position caption to the respective spatial location on the generated garment. The spatial correction loss can be expressed as:

$$\mathcal{L}_{\text{spatial}} = \sum_{i=1}^k \sum_{j=1}^l \|a_i^j - I_j(m_i)\|^2, \quad (5)$$

where $I_j(\cdot)$ denotes the interpolation operation used to adjust the size of the component mask m_i to match that of a_i^j .

Quantitative alignment. To achieve quantitative alignment, we utilize a component detector [21] to determine recognition boxes for individual components, enabling the computation of component counts in the generated output. Subsequently, these counts $\hat{Q} = \{\hat{q}_1, \dots, \hat{q}_k\}$ are compared with the ground truth component quantities $Q = \{q_1, \dots, q_k\}$ to rectify quantitative alignment at the component level, as depicted by the following equation:

$$\mathcal{L}_{\text{quantitative}} = \sum_{i=1}^k |q_i - \hat{q}_i|. \quad (6)$$

The final loss function is determined by combining all correction losses and the retrieval-augmented contrasting loss:

$$\mathcal{L} = \omega_v \cdot \mathcal{L}_{\text{visual}} + \omega_s \cdot \mathcal{L}_{\text{spatial}} + \omega_q \cdot \mathcal{L}_{\text{quantitative}} + \omega_r \cdot \mathcal{L}_{\text{RACL}}, \quad (7)$$

where $\omega_v, \omega_s, \omega_q, \omega_r$ represent the weights used to balance all losses.

4 Experiments

4.1 Datasets

Our experiments are conducted on the CM-Fashion dataset [43], which consists of 500,000 garment images at a resolution of 512×512 , each accompanied by corresponding captions. We employ the proposed automatic component extraction pipeline outlined in Sec. 3.1 to extract component-level garment segmentation and count of components from the images. Subsequently, we enrich the captions with the extracted information. Consequently, we curate an augmented garment dataset derived from the CM-Fashion dataset [43], featuring optimized captions and component-level segmentation and quantity.

4.2 Implementation Details

GarmentAligner is implemented using PyTorch [23] and trained on 8 Tesla V100 GPUs with a batch size of 32. The mask prediction network is adapted from ARMANI [43], renowned for its high segmentation accuracy, making it suitable for garment datasets. During training, we introduced the prediction of a mixture of noise and images as the prediction type, replacing the previous method of solely predicting noise. This enhancement is anticipated to elevate the quality of the generated outputs. We utilized SD v2.1 [31] as the base pre-trained model and employed a learning rate of $1e - 6$ for 40 epochs, requiring approximately 70 hours to complete the training process.

4.3 Qualitative Results

We conduct a qualitative evaluation for GarmentAligner and state-of-the-art methods [6, 24, 39, 44]. The results depicted in Fig. 7 demonstrate GarmentAligner’s superiority in text-to-garment generation. While current state-of-the-art methods [6, 24, 39, 44] produce high-quality garment images aligned with the text description, they often overlook fine-grained information, particularly in quantity and position. SDXL [24] and Midjourney prioritize detail and richness excessively, sometimes exceeding semantic relevance. Conversely, Diffcloth [44], RAPHAEL [39], and PIXART- α [6] exhibit weaker semantic understanding of different garment components, hindering precise image generation. Comparatively, GarmentAligner excels in generating realistic fashion images aligned with textual descriptions and exhibits precise fine-grained alignment.



Fig. 7: Visual comparison with baselines [6, 24, 39, 44]. Red boxes indicate incorrect generated areas, green boxes denote correct ones, and yellow boxes signify absent ones. Our approach demonstrates exceptional performance in capturing the texture, positioning, and quantity of garment components, resulting in the generation of realistic fashion images with precise fine-grained alignment.

4.4 Quantitative Results

For quantitative comparisons, we employ several metrics: 1) FID [13] to assess the fidelity of the generated images, 2) Aesthetic Score to evaluate their aesthetic

Table 1: The quantitative comparison results with baselines [5, 8, 24, 28, 31, 34, 43, 44]. Our approach achieves the best performance in comparison with numerous methods, excelling in metrics for image quality and consistency between text and image.

Method	FID↓	CLIPScore↑	AestheticScore↑	HPSv2↑
DALL·E [28]	13.249	0.6423	4.8592	0.2137
ARMANI [43]	12.336	0.6988	5.3585	0.2237
SDv1.5 [31]	9.368	0.8911	5.2807	0.2419
SDv2.1 [31]	9.157	0.8818	5.3881	0.2426
ComposableDiffusion [34]	9.499	0.8306	5.0984	0.2346
StructureDiffusion [8]	9.238	0.8459	5.2148	0.2398
Attend-and-Excite [5]	9.351	0.8241	5.2452	0.2430
DiffCloth [44]	9.201	0.8974	5.3957	0.2440
SDXL [24]	9.091	0.8756	5.4299	0.2450
GarmentAligner(ours)	8.735	0.9245	5.8776	0.2648

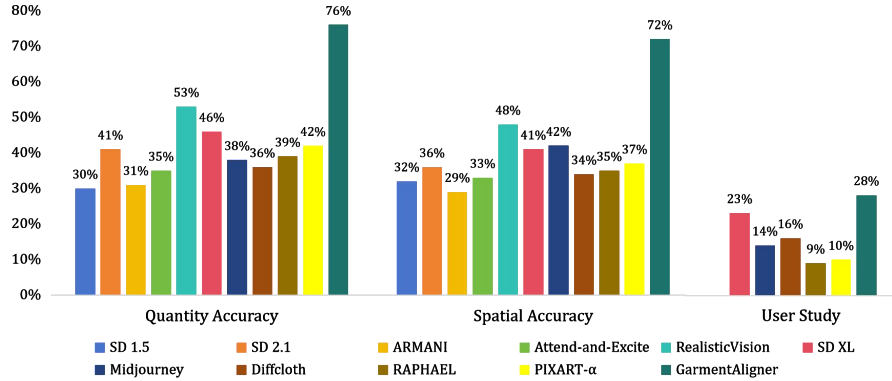


Fig. 8: The component-level quantity and spatial accuracy of our method and baselines [5, 6, 24, 31, 39, 43, 44] and user study result. Our approach demonstrates outstanding performance, significantly surpassing other methods in both quantity and spatial dimensions, and obtains user preferences.

quality, 3) CLIP Score [12] to measure the relevance between the given text and the generated images, and 4) HPS v2 [36] to gauge the degree of alignment with human preferences. The quantitative comparison results are presented in Tab. 1, juxtaposed with those of DALL·E [28], ARMANI [43], ComposableDiffusion [34], StructureDiffusion [8], Attend-and-Excite [5], DiffCloth [44], SD v1.5 [31], SD v2.1 [31], and SDXL [24]. Our proposed GarmentAligner achieves the lowest FID score and the highest scores in CLIPScore, Aesthetic score, and HPS v2 for garment synthesis, indicating superior performance across all metrics.

Additionally, we devised a component-level accuracy comparison to quantitatively assess the precision of component generation. Specifically, we employed our approach and baselines [5, 6, 24, 31, 39, 43, 44] to generate images based on 1000 captions, and then calculate the quantity and spatial accuracy. To ensure the

Table 2: The quantitative comparison of different parts of the model, where $[V]$ represents visual correction, $[S]$ represents spatial correction, $[C]$ represents quantitative correction, and $[R]$ represents retrieval-augmented contrastive learning.

Method	FID↓	CLIPScore↑	AestheticScore↑	HPSv2↑
GarmentAligner ^[V]	8.975	0.9136	5.4081	0.2459
GarmentAligner ^[S]	9.143	0.8976	5.4003	0.2447
GarmentAligner ^[C]	9.091	0.8840	5.3912	0.2433
GarmentAligner ^[V+S+C]	8.924	0.9183	5.4190	0.2462
GarmentAligner ^[R]	8.802	0.8984	5.7443	0.2639
GarmentAligner ^[V+S+C+R]	8.735	0.9245	5.8776	0.2648

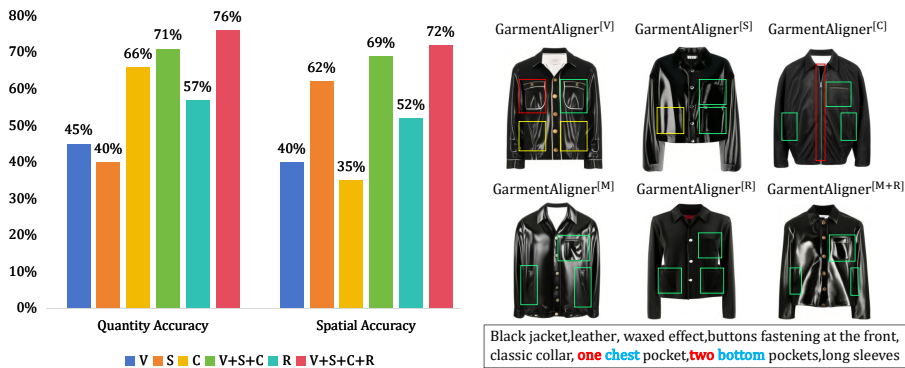


Fig. 9: The component-level quantity and spatial accuracy of different variants from the ablation study. The multi-level corrections exhibit improvements for target attributes, while the combination of different parts contributes to a more effective enhancement.

Fig. 10: The visual comparison of different variants from the ablation study. $[M]$ denotes the combination of $[V + S + C]$, while the rest share equivalent definitions as illustrated in Tab. 2. The boxes maintain identical semantics with Fig. 7.

validity of the experiment, we generate 100 images for each caption. As depicted in Fig. 8, our method outperforms others in accuracy by 20 ~ 45%, affirming the superiority of our approach in capturing fine-grained details of components.

4.5 Ablation Study

To validate the effectiveness of the proposed retrieval-augmented contrastive learning and the three corrections at the component level, we designed six variants of the proposed method and evaluated their performance based on the metric scores, accuracy and visual effects, as shown in Fig. 9, Fig. 10 and Tab. 2.

From Tab. 2, it is evident that retrieval-augmented contrastive learning makes the most significant contribution to the overall quality of generated garments, as evinced by its marked improvement across metrics such as FID [13], Aesthetic Score, and HPS v2 [36], which are indicative of image realism. On the other

hand, multi-level corrections prove notably effective in enhancing text-image consistency while concurrently improving performance on all metrics. The accuracy outcomes illustrated in Fig. 9 demonstrate the significant role of multi-level corrections in enhancing both the quantity and spatial accuracy at the component level, thus validating the effectiveness of our proposed method design. The combined application of these parts yields superior results, illustrating the complementary nature of the two pivotal components we propose. Fig. 10 illustrates the disparities among variant models in depicting detailed components of clothing, providing a more intuitive insight into the distinct roles played by different parts.

4.6 User Study

We executed a user study with 110 participants spanning a range of identities and age brackets to impartially evaluate our method against established baselines [6, 24, 39, 44], with a focus on fine-grained fidelity and overall image quality. We devise a questionnaire comprising 40 questions, each presenting shuffled outcomes generated from various methods associated with the same caption. Participants are instructed to select the image they perceive to be the most authentic and consistent with the provided textual description. The results, shown in Fig. 8, reveal that our method was the preferred choice for the majority of participants, garnering selections exceeding 28%.

5 Conclusion

In this study, we introduce GarmentAligner, a text-to-garment diffusion model aimed at rectifying fine-grained semantic misalignment issues inherent in garment generation. By combining retrieval-augmented contrastive learning with multi-level corrections, GarmentAligner effectively aligns the visual semantics, spatial positioning, and quantity of the generated garment components with provided captions. Furthermore, we devise an automatic component extraction pipeline to extract spatial and quantitative information about garment components from images, which can be applied to any garment dataset, thus facilitating the advancement of high-quality garment generation. Our experiments demonstrate that GarmentAligner produces superior garment images with improved semantic alignment compared to existing methods.

Social Impacts and Limitations. While producing high-quality garments, our method still encounters certain limitations. The training data utilized in our approach is generated through the proposed extraction pipeline, which heavily relies on the accuracy of semantic segmentation and detection models. Inevitably, errors may arise, particularly when dealing with large-scale datasets. Moreover, the biases inherited from pre-trained models potentially lead to outputs that lack robustness and user-friendliness. Future measures are necessary to address these ethical concerns through bias reduction and thorough scrutiny.

Acknowledgements

This work was supported in part by National Science and Technology Major Project (2020AAA0109704), National Science and Technology Ministry Youth Talent Funding No. 2022WRQB002, National Natural Science Foundation of China under Grant No. 62372482 and No. 61936002, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Mobility Grant Award under Grant No. M-0461, Nansha Key RD Program under Grant No.2022ZD014.

References

1. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18370–18380 (2023)
2. Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Multimodal garment designer: Human-centric latent diffusion models for fashion image editing (2023)
3. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Semi-parametric neural image synthesis (2022)
4. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J.W., Elsen, E., Sifre, L.: Improving language models by retrieving from trillions of tokens (2022)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* **42**(4), 1–10 (2023)
6. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023)
7. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator (2022)
8. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis (2023)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022)
10. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.W.: Realm: Retrieval-augmented language model pre-training (2020)
11. He, K., Yao, K., Zhang, Q., Yu, J., Liu, L., Xu, L.: Dresscode: Autoregressively sewing and generating garments from text guidance (2024)
12. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022)

13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
14. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M.: Generalization through memorization: Nearest neighbor language models (2020)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. *arXiv:2304.02643* (2023)
16. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192* (2023)
17. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021)
18. Li, N., Liu, Q., Singh, K.K., Wang, Y., Zhang, J., Plummer, B.A., Lin, Z.: Unihuman: A unified model for editing human images in the wild (2023)
19. Lin, A., Zhao, N., Ning, S., Qiu, Y., Wang, B., Han, X.: Fashiontex: Controllable virtual try-on with text and texture. In: *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings. SIGGRAPH '23, ACM* (Jul 2023). <https://doi.org/10.1145/3588432.3591568>, <http://dx.doi.org/10.1145/3588432.3591568>
20. Lin, H., Miao, L., Ziai, A.: Rafic: Retrieval-augmented few-shot image classification (2023)
21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023)
22. Ning, S., Wang, D., Qin, Y., Jin, Z., Wang, B., Han, X.: Picture: Photorealistic virtual try-on from unconstrained designs (2023)
23. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: *NIPS-W* (2017)
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
26. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
28. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation (2021)
29. Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D.: Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion (2023)
30. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
33. Seyfioğlu, M.S., Bouyarmane, K., Kumar, S., Tavanaei, A., Tutar, I.B.: Dream-paint: Few-shot inpainting of e-commerce items for virtual try-on without 3d modeling (2023)
34. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion (2023)
35. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
36. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Better aligning text-to-image models with human preference (2023)
37. Xie, Z., Li, H., Ding, H., Li, M., Cao, Y.: Hierarchical fashion design with multi-stage diffusion models (2024)
38. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* **36** (2024)
39. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems* **36** (2024)
40. Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., tau Yih, W.: Retrieval-augmented multimodal language modeling (2023)
41. Yu, J., Zhang, L., Chen, Z., Pan, F., Wen, M., Yan, Y., Weng, F., Zhang, S., Pan, L., Lan, Z.: Quality and quantity: Unveiling a million high-quality images for text-to-image synthesis in fashion design (2023)
42. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldrige, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022)
43. Zhang, X., Sha, Y., Kampffmeyer, M.C., Xie, Z., Jie, Z., Huang, C., Peng, J., Liang, X.: Armani: Part-level garment-text alignment for unified cross-modal fashion design. In: *Proceedings of the 30th ACM International Conference on Multimedia. MM '22, ACM (Oct 2022)*. <https://doi.org/10.1145/3503161.3548230>, <http://dx.doi.org/10.1145/3503161.3548230>
44. Zhang, X., Yang, B., Kampffmeyer, M.C., Zhang, W., Zhang, S., Lu, G., Lin, L., Xu, H., Liang, X.: Diffcloth: Diffusion based garment synthesis and manipulation via structural cross-modal semantic alignment (2023)