Supplementary Material of GarmentAligner

1 Introduction

In this supplementary material, we present additional implementation details, encompassing details about the component-level segmentation model, component-level detection, training details, and user study details, as delineated in Sec. 2. Furthermore, we furnish supplementary experimental findings, juxtaposing our proposed GarmentAligner with baseline methods and demonstrating more ablation studies of RACL, as expounded in Sec. 3.

2 Additional Implementation Details

Component-level Segmentation. The semantic segmentation network utilized for garment components is derived from ARMANI [7], while this network leverages PointRend [3] as its foundational segmentation architecture. Subsequently, it employs 100,000 image-mask pairs from the training dataset to refine the basic network for segmenting garment images into distinct components. The network demonstrates outstanding segmentation performance across diverse garment types, as evidenced by the segmentation results depicted in Fig. 1.

Component-level Detection. We utilize GroundingDino [4] for component detection and quantity estimation to facilitate quantity correction during training. However, GroundingDino is sensitive to the box threshold. As shown in Fig. 2, a low threshold may result in unrelated regions being misclassified as the target component, whereas a high threshold may lead to failure in detecting any area of interest. So we implement a box selection process based on the spatial information of detected boxes to ensure their alignment with the expected locations. Specifically, for components with relatively fixed positions, we integrate the centroids of the bounding boxes with the corresponding garment structure to refine the box selection process (e.g., the centroids of sleeves typically reside in the left center and right center, and the centroid of collar reside in the upper center). For finer components such as pockets, logos, and bows, which lack fixed positions, we employ a maximum area criterion for filtering.

Training Details. To further ensure training efficiency, we uniformly sample 100 timesteps from the interval of [0, 1000) for training purposes. And then We conduct multi-level corrections during these timesteps. Contrastly, for retrieval-augmented contrastive learning, we employ it at each sampling timestep due to its rapid execution.

User Study Details. Our user study is conducted in 4 different garment categories: tops, pants, skirts and overalls. And 10 garment captions are randomly selected for each category to generate garment images by our method and baselines. Then a questionnaire with 40 questions was distributed to 110 participants

(independent annotators), who were asked to choose the most photo-realistic item that matched the caption best for each question.

3 Additional Experimental Results

More Qualitative Results. In Fig. 3, We showcase more visual results of the text-to-garment task, comparing them with other baselines [2, 5, 6, 8], to demonstrate the superiority of our method. In Fig. 4, We present additional visual generation results by GarmentAligner across diverse garment categories.

More Quantitative Results. Tab. 1 provides additional experiment results on CC12M [1] dataset to further validate the generalizability of our approach.

Table 1: More quantitative results on CC12M [1] dataset.

Method	FID ↓	CLIPScore ↑	SSIM ↑	AestheticScore ↑	HPSv2 ↑
ARMANI	16.947	0.6431	0.4473	5.2846	0.2285
SDv1.5	12.999	0.8132	0.3047	4.9769	0.2511
SDv2.1	14.275	0.7994	0.3219	5.3505	0.2502
SDXL	14.471	0.8099	0.4325	5.3602	0.2518
Attend-and-Excite	12.948	0.7465	0.3130	5.1674	0.2483
DiffCloth	11.072	0.8359	0.5129	5.3183	0.2494
GarmentAligner(ours)	**9.392**	**0.8761**	**0.6227**	**5.4789**	**0.2595**

More Ablation Studies of RACL. The objective of contrastive learning is to specify crucial features for distinguishing between different categories of samples, thereby mitigating the effects of label noise and enhancing generalization capabilities. Using RACL, garment images with differences can be encoded more dispersedly in latent distribution, allowing the model to generate more diverse and fine-grained results. The retrieval dataset is constructed on the fine-tuning dataset based on the similarity of component information. The number of retrieved samples N for the RACL is a fixed value, which is determined through ablation studies as shown in Fig. 5 and Tab. 2. A larger N increases the probability of retrieving the most suitable samples, but it also raises computational costs. Additionally, for the weights of positive and negative samples α_1 and α_2, low values impact the training effect and high values could impact the texture and fabric of the garment. Besides using negative samples only may not learn correct information, while positive samples only risk adding unnecessary details.

Moreover, the motivation of the garment similarity score (EQ1, EQ2) is to evaluate garment similarity in 3 aspects: 1) component counts, 2) component semantic and 3) overall semantic. EQ1 calculates component-level similarity focusing on component descriptions and counts. EQ2 calculates the final score by summing component similarities and the similarity of the garment's overall description. As component or overall descriptions can be any text, Jaro distance is

adopted for its ability to measure two arbitrary strings. In Fig. 5, we conduct an ablation study for EQ1 and EQ2. Only EQ1 will degrade the garment color and texture, while only overall semantic in EQ2, accurate fine-grained information can not be correctly generated.

Table 2: More ablation studies of RACL.

Parameter	FID ↓	CLIPScore ↑	SSIM ↑	AestheticScore ↑	HPSv2 ↑
$\alpha_1 = 0.2, \alpha_2 = 0.4, N = 10$	12.906	0.8011	0.4133	4.9645	0.2487
$\alpha_1 = 0.2, \alpha_2 = 0.4, N = 5000$	10.048	0.8689	0.5183	5.2074	0.2548
$\alpha_1 = 0.2, \alpha_2 = 0, N = 5000$	10.203	0.8502	0.4585	5.1444	0.2534
$\alpha_1 = 0, \alpha_2 = 0.4, N = 5000$	9.407	0.8645	0.5070	5.2044	0.2561

References

1. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021)
2. 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)
3. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
4. 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)
5. 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)
6. 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)
7. 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
8. 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)

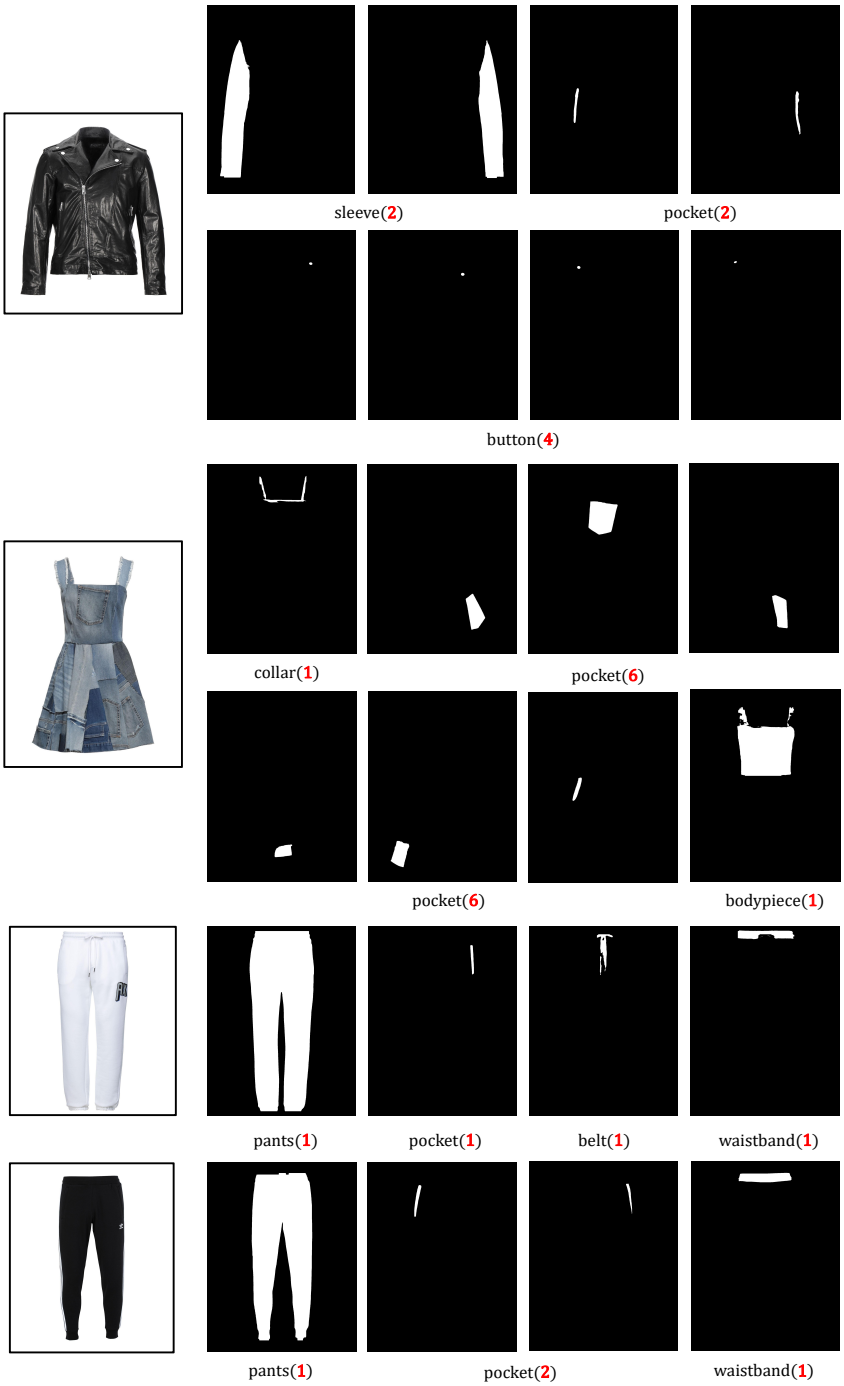

Fig. 1: Segmentation results across various types of garments. The red numbers indicate the quantities of each component.

Fig. 2: Detection results of components with different box thresholds of GroundingDINO [4]. When modified with the box threshold, the results significantly change, but incorporating box selection can effectively balance these variations.

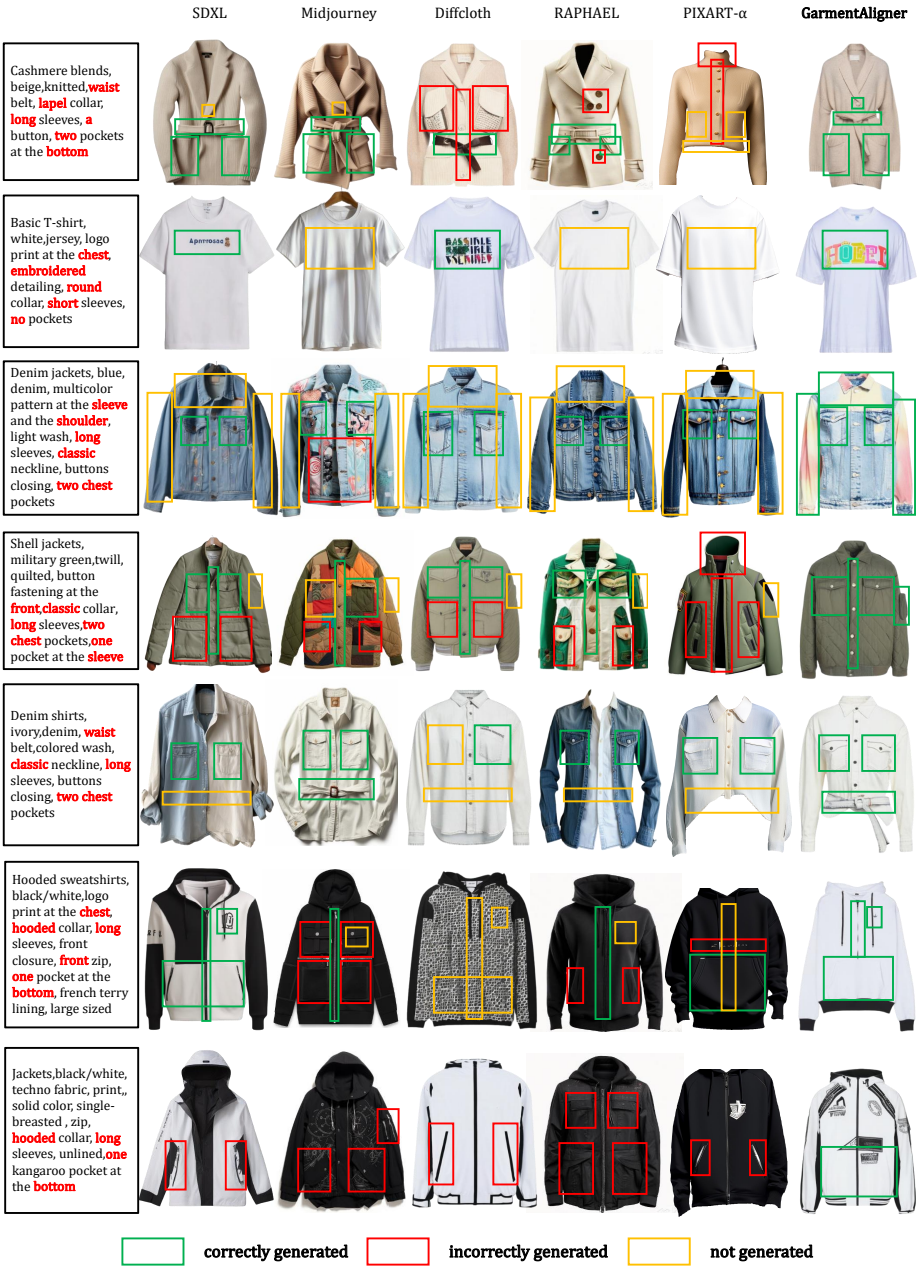

Fig. 3: More visual comparison with baselines [2, 5, 6, 8]. Red boxes indicate incorrect generated area, green boxes denote correct ones, and yellow boxes signify absent ones. Our approach demonstrates exceptional performance in capturing the texture, positioning, and quantity of garment components, resulting in the generation of realistic fashion images with precise fine-grained alignment.

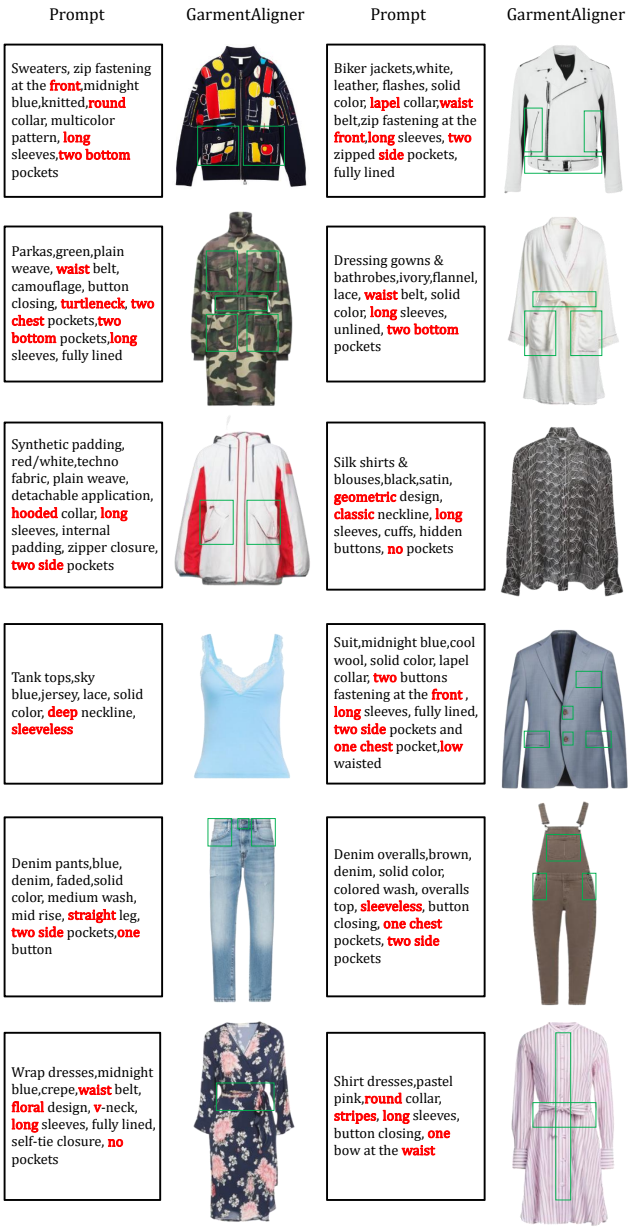

Fig. 4: Additional garment images generated by our method with given prompts. The red terms in texts emphasize semantic, positional, or numerical information, while their corresponding parts in images are highlighted by green boxes.

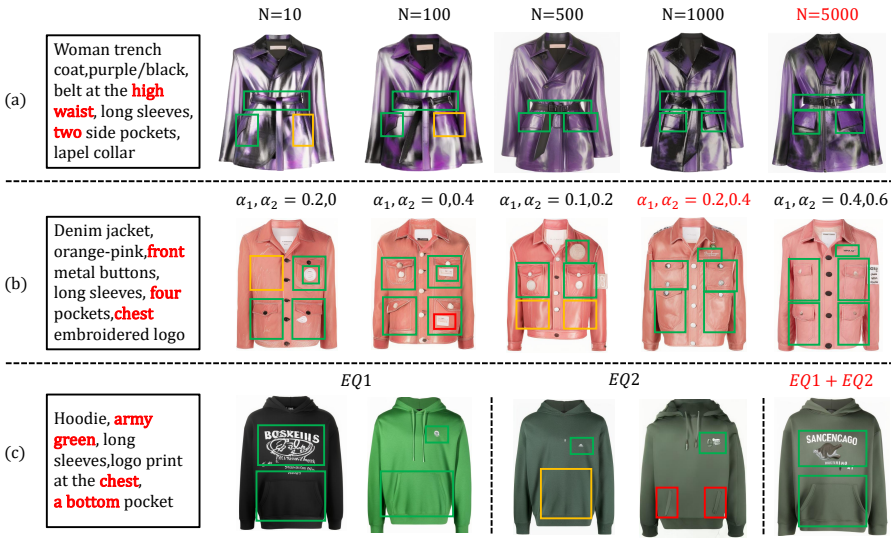

Fig. 5: Additional ablation studies of RACL. (a) N represents the number of retrieved data. (b) α_1 and α_2 represent the weights of negative and positive samples. (c) $EQ1$ and $EQ2$ respectively represent the component-level and overall similarity score respectively. The red numbers are the parameters we finally selected.