E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)03-0862-18

论文引用格式:Ma J N, Liu L, Fu X D, Liu L J and Peng W. 2026. Structured style enhancement learning for multimodal guided dress image generation. Journal of Image and Graphics, 31(3):0862-0879(马嘉妮, 刘骊, 付晓东, 刘利军, 彭玮. 2026. 多模态引导裙装图像生成的结构化风格增强学习. 中国图象图形学报, 31(3):0862-0879)[DOI:10.11834/jig.250338]

多模态引导裙装图像生成的结构化风格增强学习

马嘉妮[1],刘骊[1,2*],付晓东[1,2],刘利军[1,2],彭玮[1,2]

1. 昆明理工大学信息工程与自动化学院,昆明　650500; 2. 云南省计算机技术应用重点实验室,昆明　650500

摘　要: **目的**　针对多模态引导的裙装图像生成中存在的多角度文本注释信息冗余与冲突、跨区域风格传递能力有限以及语义与风格难以精细协同控制的问题,提出了一种结构化风格增强学习方法。**方法**　以文本描述作为输入,针对裙装特点设计动态属性模板生成策略,智能提取并重构7类关键裙装属性,构建消除冗余与冲突的结构化文本提示;建立文本反转语义融合机制,将裙装图像特征经文本反转生成伪词嵌入,与结构化提示融合,形成语义丰富的文本表示;构建跨域图像特征对齐模块,引入跳跃交叉注意力,实现草图结构与风格图像的选择性融合并实现跨区域风格关联;建立双重条件协同融合框架,将增强的文本表示与跨域风格表示分层注入潜在扩散模型,精细控制语义与风格以生成裙装图像。**结果**　实验在DressCode Multimodal数据集裙装子集上与目前较新的5种方法进行比较。结果表明,所提方法的弗雷歇起始距离(Fréchet inception distance,FID)和学习感知图像块相似度(learned perceptual image patch similarity,LPIPS)较对比方法提高2.131和0.193,对比语言图像预训练分数(contrastive language-image pre-training score,CLIPScore)和纹理分数(texture score,TS)分别提高17.57%和8.29%,说明本文方法具有更好的生成效果。**结论**　本文提出的多模态引导裙装图像生成的结构化风格增强学习方法,能有效聚焦语义内容与风格结构间的深层关联,在确保多模态一致性的同时,实现高质量的裙装图像生成。

关键词: 裙装图像生成;结构化文本提示;文本反转语义融合;跨域图像特征对齐;扩散模型

Structured style enhancement learning for multimodal guided dress image generation

Ma Jiani[1], Liu Li[1,2*], Fu Xiaodong[1,2], Liu Lijun[1,2], Peng Wei[1,2]

1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming　650500, China;

2. Computer Technology Application Key Laboratory of Yunnan Province, Kunming　650500, China

Abstract: **Objective** Multimodal image generation aims to produce high-quality images by jointly encoding and conditionally controlling diverse modalities such as text, sketches, and reference style images, while ensuring semantic consistency. With the digital transformation of the fashion industry, applications such as virtual try-on, digital garments, and online shopping are placing higher demands on the realism and controllability of generated images. Different clothing categories vary considerably in their structural characteristics and visual attributes. Among clothing categories, dresses, as a

收稿日期:2025-07-25;**修回日期**:2025-09-09;**预印本日期**:2025-09-16

*通信作者:刘骊　ieall@kust.edu.cn

基金项目:国家自然科学基金项目(62262036,62362043);兴滇英才支持计划项目(KKXY202203008);云南省科技计划项目(202503AA080013,202502AD080003)

Supported by: National Natural Science Foundation of China(62262036, 62362043); Xingdian Talent Support Project (KKXY202203008); Yunnan Provincial Science and Technology Support Project(202503AA080013,202502AD080003)

key category in the fashion industry, present greater challenges due to their coverage of both upper and lower body regions and their wide variety of styles, increasing complexity in structure modeling, texture synthesis, and cross-region style transition. Currently, existing approaches for dress image generation can be broadly classified into three categories: generative adversarial network (GAN) methods, autoregressive models, and diffusion methods. Early works predominantly rely on GANs, which leverage adversarial training between generators and discriminators to capture local textures and structural details. However, these models often suffer from training instability and mode collapse. Autoregressive models first discretize multimodal inputs into token sequences and progressively generate image regions within a latent token space. This approach enables effective fusion of multimodal conditions. However, their strictly sequential generation process makes it difficult to model long-range dependencies, leading to limited information flow and reduced scalability in complex image synthesis tasks. More recently, diffusion models have emerged as a new paradigm due to their stepwise denoising process, which enables stable training and stronger conditional control. These models iteratively reconstruct clean images from noisy representations, making them well suited for incorporating complex multimodal guidance. Despite recent progress in semantic consistency and style generation, dress image synthesis remains a challenging task due to its rich attribute variations and stylistic diversity. First, multi-angle textual annotations of dresses often contain redundancy and conflict and lack contextual coherence, making it difficult to provide deep semantic guidance for visual synthesis. Second, although sketches offer explicit structural cues, they fall short in conveying consistent cross-region styles, leading to local-global style mismatches. Moreover, effective dress generation requires fine-grained coordination between semantics and style, yet current guidance mechanisms lack a unified and efficient control strategy. In this paper, we propose a structured style enhancement learning method for dress image generation. **Method** First, a dynamic attribute template generation strategy is designed on the basis of the structural and stylistic characteristics of dresses. It enables the intelligent extraction and reconstruction of seven key dress attributes, thereby constructing structured textual prompts that mitigate redundancy and semantic conflicts in multiview descriptions. Second, a semantic inversion fusion mechanism is proposed, wherein visual features of dress images are projected into pseudo-token embeddings through a text-inversion process and subsequently fused with the structured prompts to form semantically enriched textual representations. Third, a cross-domain image feature alignment module is developed, in which a skip-cross attention mechanism is introduced to selectively integrate sketch structures and style references, facilitating consistent style transfer across spatial regions. Finally, a dual-branch conditional fusion framework is constructed, wherein the enhanced textual and style representations are hierarchically injected into a latent diffusion model. This feature enables fine-grained control over semantic content and visual style during the dress image generation process. **Result** Comparative experiments were conducted on the dress subset of the DressCode Multimodal dataset against five state-of-the-art approaches to evaluate the effectiveness of the proposed method. Quantitative results show that, compared with the second-best method, the proposed method achieves improvements of 2.131 in Fréchet inception distance and 0.193 in learned perceptual image patch similarity, indicating that the introduced structured style-enhanced learning mechanism provides clearer, richer, and more consistent multimodal guidance during training. The Contrastive Language-Image Pre-training score is improved by 17.57%, demonstrating that the unified semantic template focusing on attribute information effectively eliminates ambiguity and redundancy in natural language descriptions, enhances semantic coherence, and provides the model with clearer generation targets. The integration of visual features further enriches the textual semantics, resulting in better alignment between the generated dress images and their textual descriptions. An 8.29% increase in texture score indicates the model's strong ability to maintain style consistency and detailed fidelity even under complex sketch structures by effectively sharing and transferring style information across regions. The sketch distance score remains within a low and comparable range to existing methods, reflecting stable generation performance. Qualitative experimental results demonstrate that the proposed method is highly capable of recognizing and reconstructing key dress attributes, particularly color, material, and pattern. Furthermore, it achieves satisfactory performance in style transfer across regions, overall style consistency, image quality, and visual realism. These outcomes effectively fulfill the dual constraints of semantics and style required for high-quality dress image generation. The results of user studies further demonstrate that the proposed method achieved the highest preference rates across three evaluation dimensions: overall visual quality of the generated dress images, consistency between the images and the input text, and

cross-domain style consistency within the generated results. This finding indicates that the proposed approach can effectively enhance the perceptual quality of dress image generation. Moreover, compared with the relatively close performance reflected by automatic evaluation metrics, user preferences reveal more pronounced differences among competing methods in terms of generation quality and style fidelity. These findings highlight the necessity of subjective evaluation in perception-driven tasks such as dress image generation, as it provides valuable insights into the perceptual differences that may not be fully captured by objective metrics. **Conclusion** To address the challenges in multimodal-guided dress image generation, including redundant and conflicting multiview textual annotations, limited cross-region style transfer capability, and the difficulty of fine-grained coordination between semantics and style, this paper proposes a structured style enhancement learning method. The method effectively captures the deep correlations between semantic content and structural style, ensuring multimodal consistency while achieving high-quality dress image generation.

Key words: dress image generation; structured text prompt; text inversion semantic fusion; cross-domain image feature alignment; diffusion model

0 引　言

多模态引导图像生成(Hu等,2024)通过将文本、草图和风格图像等多种模态信息统一编码并条件控制,在保证语义一致性的同时生成高质量图像,广泛应用于内容创作、虚拟交互以及辅助设计等场景。现有主流方法可分为3类:1)基于生成对抗网络(generative adversarial network, GAN)的方法(Wang等,2024)通过构建条件生成器—判别器的对抗训练和多尺度特征融合捕捉局部细节,实现对多模态约束的响应;2)基于自回归模型的两阶段方法(Gafni等,2022)将多模态信息离散化为标记序列,并在标记空间逐步生成图像区域,灵活融合条件信息;3)基于扩散模型的去噪方法(Zhang等,2023a)以随机噪声为起点,借助多模态语义嵌入逐步逼近目标分布,生成语义一致且逼真的图像。

随着时尚产业的数字化转型,虚拟试衣、数字服装及在线购物等应用对生成图像的真实性和可控性提出更高要求。鉴于不同服装类别在结构特征和视觉属性上的差异,部分研究聚焦于特定类别,如Liu和Wang(2024)针对T恤图像的生成与交互式定制;Choi等人(2021)侧重对上装图像引入对齐感知的归一化,以提升高分辨率服装图像的生成质量。然而,作为时尚产业重要品类的裙装因同时覆盖上、下身且款式多样,在结构复杂度、纹理生成和跨区域风格过渡方面更具挑战(Shen等,2025)。现有大多数研究仍以文本为主导,如Shastri等人(2022)采用多模型并行提升生成图像的多样性和准确性;Lampe等人(2024)结合CLIP(contrastive language-image pre-training)文本嵌入和交叉注意力机制提高生成服装图像的质量。但单一文本引导易因多角度注释导致的冗余、冲突及语义歧义而难以提供精确控制。为此,部分研究尝试引入额外视觉信息:Zhang等人(2021)融合预测分割图整合服装形状,并利用空间风格损失细化纹理细节;Zhang等人(2022)通过MaskCLIP(mask guided contrastive language-image pre-training)建立服装部件级跨模态码本,提高生成结果与文本描述的一致性;Baldrati等人(2023)设计无分类器引导策略,以实现多条件可控生成。尽管上述方法在一定程度上增强图文条件之间的关联,但图像通常仅作为辅助模态,未能深度参与语义表达过程,难以捕捉裙装复杂的视觉特征,使得生成图像缺乏真实感。

扩散模型成为融合多模态的主流框架后,Jiang等人(2022)提出分层纹理感知码本和混合专家采样器,提升了生成图像的真实感;Zhang等人(2023b)构建多粒度语义单元,并引入语义绑定交叉注意力模块,以解决部件缺失和属性混淆问题;Zhang等人(2025)通过服装组件数量与语义信息构建检索先验,并结合多级校正策略实现细粒度结构对齐。然而,上述方法忽略了对服装风格控制,而裙装作为覆盖范围广泛、结构层次丰富的服装类别,无法确保裙装中关键的跨区域纹理一致性。

由于裙装图像生成既需利用文本蕴含的丰富属性信息,又应补充视觉形态和风格表达,因此,多模态引导裙装图像生成仍然具有以下难点:1)多角度文本注释冗余、冲突且缺乏上下文连贯性,难以提供深层次视觉形态理解;2)草图虽然能明确结构,但跨区域风格传递能力有限,导致局部与整体

风格不一致；3）裙装图像生成需语义与风格间实现精细协同，而现有引导机制缺乏统一、高效的控制策略。

因此，本文以文本、草图以及风格图像作为输入，提出多模态引导裙装图像生成的结构化风格增强学习方法，以提高裙装图像生成的质量。本文方法流程如图 1 所示，主要贡献如下：1）通过结构化文本增强模块，首次针对裙装提出动态属性模板生成策略，自适应重构 7 类关键属性，引入属性级冲突检测构建无冲突的结构化文本提示；同时将裙装图像视觉特征转化为伪词嵌入，以实现文本—视觉的深度协同增强。2）设计跳跃交叉注意力机制，构建跨区域风格学习模块，突破传统逐层传递的局限性，实现风格特征的选择性提取与远距离传递；并通过跳跃连接直接建立从胸部到裙摆不同区域间的风格关联，以实现裙装跨区域风格一致性的控制。3）提出分层协同的双重条件控制策略，分别在深层和浅层引入文本语义与跨域风格特征，并通过跨尺度条件一致性约束，确保生成的裙装图像语义与风格的精确统一。

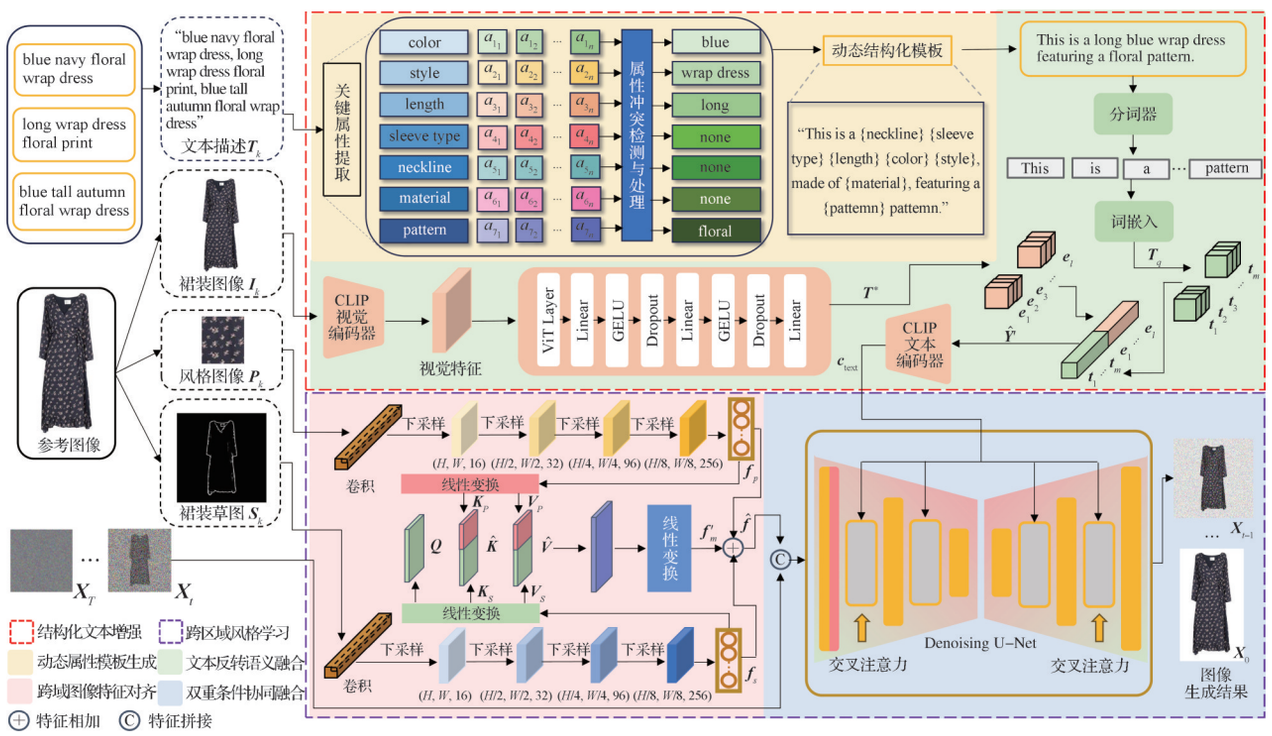

图 1　多模态引导裙装图像生成的结构化风格增强学习方法流程图

Fig. 1　The method framework of structured style enhancement learning for multimodal guided dress image generation

1　相关工作

在多模态引导图像生成的技术框架下，服装图像生成形成了多种研究路径。考虑到不同服装类别在结构特征和视觉属性上的差异，部分研究聚焦于单一服装类别以实现精准控制。Chen 等人（2020）提出基于边缘映射的解耦方法，通过自监督学习分离纹理与结构，实现上装领型和袖型的精准设计；Surya 等人（2020）通过循环架构和两阶段生成器，结合细粒度文本属性与颜色聚类，逐步生成视觉一致的裤装图像。不同服装类别在服装结构复杂度和视觉特征上存在显著差异，裙装尤为突出。Rico Gómez 等人（2024）通过局部熵量化分析发现，裙装款式多样、纹理复杂、褶皱丰富，其生成复杂度高于其他服装类别，且跨上下身区域的风格传递难度更大。

早期服装图像生成主要依赖 GAN 与文本条件。Jain 等人（2019）提出层次化文本处理机制，通过语义级联分步细化，实现服装特征精准映射；Ak 等人（2020）采用双向长短期记忆（bidirectional long short-term memory，BiLSTM）提取文本中单词级特征和句

子级特征，增强生成图像与文本描述的一致性；Moradian等人（2024）基于CLIP模型和小波损失，以解决因对抗训练不充分导致的细节模糊。然而，这些方法仅适用于相对简洁的文本描述，难以处理裙装复杂的多角度注释，如包含长度、领型、袖型和材质等多重属性的描述，导致细节与属性一致性不足。

为提升可控性，许多工作引入图像、姿态等辅助模态，崔怀磊等人（2022）设计融合空间与通道注意力的混合机制，强化对文本语义在图像空间的细粒度映射；Li等人（2020）引入文本—图像仿射组合模块，实现文本描述与图像区域精准对齐，同时保留原始视觉特征，有效降低语义错配风险；Pernuš等人（2025）利用姿态、语义及图像级别的多重约束进行反向优化，引导GAN捕捉更丰富的结构属性与时尚元素；徐正国等人（2025）将姿态引导与语义驱动的部件级增强相结合，通过对服装区域的独立增强，提升了生成图像真实感；Morelli等人（2023）则将图像作为引导信号，整合到全局文本语义表示中，以提升语义丰富性与表现力。尽管这些方法在多模态协同下提升了图像生成的结构完整性与语义一致性，但多数仍依赖于显式对齐，忽略了裙装多角度描述中所蕴含的丰富属性级语义，且缺乏从语义层面对文本表示进行结构化增强。现有研究中，属性信息多依赖于人工标注数据或规则模板提取，缺乏对多角度注释中隐含属性的智能理解。为此，本文设计裙装七维属性分解策略，从多角度描述中提取并整合裙装的7类关键属性，通过属性级冲突检测与语义重构机制，构建无冲突的结构化文本提示。受到Morelli等人（2023）的启发，但与之不同，本文侧重于裙装图像生成，引入裙装图像视觉特征，与属性级结构化文本提示融合，以进一步增强文本表示的语义丰富度和细节刻画能力。

扩散模型凭借其强大的生成能力与复杂条件的建模能力，已逐渐成为图像生成任务的主流框架。部分研究基于扩散模型，聚焦于风格表达与纹理控制能力上的提升。Sun等人（2023）通过文本编码器和图像编码器提取跨模态语义特征，结合交叉注意力机制和联合训练策略，提升了材质与风格的表达控制；Zhang等人（2024）采用两阶段扩散模型，先生成草图轮廓再优化纹理细节，实现精细服装设计；Yan等人（2023）设计多阶段特征融合机制，将风格与轮廓特征融合于潜在空间，增强输入条件的表达力；Baldrati等人（2024）通过多层交叉注意力捕获细节，实现文本与织物纹理的同步调节。尽管上述方法在多模态融合、纹理细节生成以及条件引导方面取得了一定进展，但对于裙装这类跨越上下身区域的服装类别的风格一致性控制方面仍存在局限。由于裙装从胸部的精致细节到腰部的收束设计，再到裙摆的飘逸形态，需要保证整体风格的协调统一，因此，与Sun等人（2023）方法不同，本文更侧重于草图结构与风格之间的协同，设计跨域图像特征对齐模块，引入跳跃交叉注意力机制实现风格特征的选择性提取与远距离传递，以提升生成裙装图像在风格传递上的区域一致性。

2 结构化文本增强

为解决多角度文本注释冗余冲突问题，本文首次对裙装提出结构化文本增强模块。首先，输入文本描述，利用动态属性模板生成策略提取裙装关键属性信息，通过属性级冲突检测机制构建结构化文本提示；然后，利用文本反转模块将输入的裙装图像特征转化为伪词嵌入并与文本提示词嵌入融合；最后，通过文本编码器得到增强后的文本条件表示。

2.1 动态属性模板生成

针对多角度文本注释导致的信息冗余、描述冲突问题，本文从属性级语义角度出发，考虑到裙装在款式、材质和图案等属性方面具有丰富的设计变化，设计裙装专门化的七维属性分解策略，基于大型语言模型（large language model，LLM）的语义理解与推理机制实现裙装关键属性提取与重构。输入多模态裙装数据集，$\boldsymbol{D}=\{\boldsymbol{D}_1,\cdots,\boldsymbol{D}_k,\cdots,\boldsymbol{D}_d\}$，其中，$d$表示样本个数；$\boldsymbol{T}_k\in\boldsymbol{D}_k$表示输入的多角度文本描述；$\boldsymbol{I}_k\in\boldsymbol{D}_k$表示输入的裙装图像；$\boldsymbol{P}_k\in\boldsymbol{D}_k$表示输入的风格图像；$\boldsymbol{S}_k\in\boldsymbol{D}_k$表示输入的裙装草图。

根据裙装专业设计知识和特点（Cheng等，2021），设计涵盖7个关键维度的属性提取策略。首先，为每一维度设计带有属性定义与典型表达示例的提示，引导LLM理解文本描述$\boldsymbol{T}_k$中各属性语义边界并提取候选值。如表1所述，每个属性有3～5个属性示例，涵盖该属性在裙装领域最常见的表达方式。对于未明确提及的关键属性，LLM综合多角度的文本描述进行推理填补。若推理结果无法有效确定该

属性的具体内容，则将该属性字段留空。第i个关键维度会得到候选属性集合$\boldsymbol{A}_i$，即

$$\boldsymbol{A}_i = LLM_{\text{extract}}(\boldsymbol{T}_k, \boldsymbol{p}_i, \boldsymbol{E}_i) = \{a_{i_1}, a_{i_2}, \cdots, a_{i_n}\} \quad (1)$$

式中，$\boldsymbol{p}_i$表示与第i个属性对应的提示词，$\boldsymbol{E}_i$表示与第i个属性对应的示例。

表 1　裙装关键属性示例

Table 1　Examples of key attributes for dress

关键属性	示例
Color	white、black、blue、red、pink
Style	wrap dress、tank dress、slip dress、shirt dress
Length	maxi、midi、mini
Sleeve type	sleeveless、short sleeved、long sleeved
Neckline	scoop neck、V-neck、high neck
Material	lace、knit、silk、mesh、leather
Pattern	printed、striped、embroidered

当候选属性集合$\boldsymbol{A}_i$包含多个不同的属性值，即$|\boldsymbol{A}_i| > 1$时，视为第i个关键维度存在冲突。为解决该问题，本文提出基于频率优先与语义推理的属性冲突处理机制，以提升属性提示的准确性。已有研究（Antol等，2015）通过频率作为最终预测的信号来源，表明多源描述中出现频率较高的词往往能反映更高的语义代表性与用户共识。因此，定义属性置信得分为候选属性在N段描述中出现的频率。具体为

$$Score_i(a) = \frac{count(a)}{N}, a \in \boldsymbol{A}_i \quad (2)$$

式中，$count(a)$表示a在N段描述中出现的次数。若存在唯一最大得分，则直接选择作为最终属性值。具体为

$$a^* = \arg\max_{a \in \boldsymbol{A}_i} Score_i(a) \quad (3)$$

若存在多个候选值具有相同最高频率，即$\exists a_1 \neq a_2, Score_i(a_1) = Score_i(a_2) = max$，则通过LLM进行二次推理，隐式计算候选属性对之间的语义一致性分数。若模型能够输出一致性较高的属性，则采用该值，否则，将该属性留空。最终得到关键属性集合$\boldsymbol{A} = \{\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_7\}$。

然后，为实现自适应结构化表示，设计动态结构化模板生成机制，根据提取到的关键属性自适应调整结构。基础模板P_{base}覆盖裙装的核心属性，包括领口、袖型、长度、颜色和款式，形式为

$$P_{\text{base}} = \text{“}This\ is\ a\text{”} + [Neckline] + [SleeveType] + [Length] + [Color] + [Style] \quad (4)$$

式中，[*Neckline*]、[*SleeveType*]、[*Length*]、[*Color*]以及[*Style*]表示占位符，对应提取到的关键属性。当检测到材质或图案属性时，模板会自动进行材质和图案扩展，形式为

$$\begin{cases} P_{\text{material}} = P_{\text{base}} + \text{“}made\ of + [Material]\text{”} \\ P_{\text{pattern}} = P_{\text{base}} + \text{“}featuring\ a\ [Pattern]\ pattern\text{”} \end{cases} \quad (5)$$

式中，[*Material*]和[*Pattern*]表示材质和图案的占位符。这种动态调整机制确保生成的文本提示既包含完整的属性信息，又避免了冗余描述。

最后，结合前述提取到的关键属性和所设计的结构化模板，生成最终的提示文本$\hat{P}$。将关键属性集合$\boldsymbol{A}$映射到动态结构化模板P，构造出最终的完整提示$\hat{P} = P(\boldsymbol{A})$。

2.2　文本反转语义融合

针对文本模态表达局限性，本文将文本反转技术扩展至裙装领域，设计跨模态语义对齐机制，实现视觉细节与文本语义的深度融合。基于第2.1节中得到的文本提示$\hat{P}$，将其输入到CLIP模型对应的分词器中进行词元化处理，得到离散词元序列$\boldsymbol{q}$，即

$$\boldsymbol{q} = \{w_1, w_2, \cdots, w_m\} \quad (6)$$

式中，w_i表示第i个词元，m表示词元的个数。然后，通过词嵌入查找将每个词元映射为其对应的嵌入向量，构建出文本提示$\hat{P}$的词嵌入序列$\boldsymbol{T}_q$，即

$$\boldsymbol{T}_q = \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m \in \mathbf{R}^{m \times d} \quad (7)$$

式中，$\boldsymbol{t}_i$表示第i个词元的词嵌入向量，$d = 768$，表示词嵌入的维度。

输入裙装图像$\boldsymbol{I}_k$，经标准化预处理后，输入CLIP视觉编码器中，提取其最后一层隐藏特征，得到一组具有丰富语义信息的图像嵌入向量，即

$$\boldsymbol{f}_{I_k} = VE(\boldsymbol{I}_k), \boldsymbol{f}_{I_k} \in \mathbf{R}^{d'} \quad (8)$$

式中，$VE(\cdot)$表示CLIP的视觉编码器，$\boldsymbol{f}_{I_k}$为裙装图像的视觉特征表示，d'为图像嵌入的维度。

由于CLIP的视觉与文本编码器存在模态间分布差异，本文设计面向裙装的文本反转模块$\mathcal{F}_\theta$，采用Transformer-MLP（multilayer perceptron）混合架构，在单次前向传播中，直接从图像特征生成伪词嵌入序列。首先，在特征增强阶段，输入维度为d'的视觉特征向量$\boldsymbol{f}_{I_k}$，先通过线性映射至Transformer输入维度，再送入多层Transformer编码器中。每一层Transformer计算输入序列中各元素间的依赖关系，捕捉全局上下文信息。通过多层堆叠的Transformer

编码器，视觉特征向量被逐层增强，逐步捕获更丰富的上下文语义和全局依赖，得到增强后的视觉语义特征$\boldsymbol{h}$。然后，在空间映射阶段，将增强后的视觉特征$\boldsymbol{h}$输入到设计的3层前馈神经网络中，其结构包含两层隐藏层，分别使用GELU（Gaussian error linear unit）激活与dropout正则化。具体为

$$\begin{cases}\boldsymbol{h}_1 = f_{\text{Dropout}}(f_{\text{GELU}}(\boldsymbol{W}_1\boldsymbol{h} + b_1)) \\ \boldsymbol{h}_2 = f_{\text{Dropout}}(f_{\text{GELU}}(\boldsymbol{W}_2\boldsymbol{h}_1 + b_2))\end{cases} \tag{9}$$

式中，$\boldsymbol{W}_i$、b_i为各层的权重和偏置。最后，经线性层投影至文本嵌入空间中，得到伪词嵌入序列$\boldsymbol{T}^*$。具体为

$$\boldsymbol{T}^* = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_l] = \boldsymbol{W}_3\boldsymbol{h}_2 + b_3 \in \mathbf{R}^{l \times d} \tag{10}$$

式中，$\boldsymbol{e}_i$表示伪词嵌入向量，$l = 16$表示所生成的伪词嵌入数量。

在伪词嵌入融合阶段，将得到的伪词嵌入序列$\boldsymbol{T}^*$和文本提示$\hat{\boldsymbol{P}}$的词嵌入序列沿序列维度进行拼接，得到融合的词嵌入表示$\hat{\boldsymbol{Y}}$。具体为

$$\hat{\boldsymbol{Y}} = f_{\text{Concat}}(\boldsymbol{T}_q, \boldsymbol{T}^*) \in \mathbf{R}^{(m+l) \times d} \tag{11}$$

由于本模块采用文本主导的跨模态融合架构，其中视觉信息作为文本语义的重要补充。虽然输入的视觉特征已映射变换至文本嵌入空间，实现了表示维度的统一，但融合的词嵌入表示$\hat{\boldsymbol{Y}}$仅完成了空间层面的特征组合，尚未实现语义层面的深度对齐。为了充分发挥跨模态信息的协同效应，融合后的词嵌入表示需要在CLIP文本编码器中进一步处理。通过多头自注意力机制建立视觉与文本信息间的复杂关联，实现跨模态特征的深度交互与语义对齐。

为了满足CLIP文本编码器对输入长度的要求，拼接后的嵌入$\hat{\boldsymbol{Y}}$会进行必要的截断或填充处理，以保证序列长度固定为$C = 77$。

$$\hat{\boldsymbol{Y}}' = \begin{cases}\hat{\boldsymbol{Y}}[1{:}C] & m + l \geqslant C \\ [\hat{\boldsymbol{Y}}; \boldsymbol{0}] & m + l < C\end{cases} \tag{12}$$

式中，$\boldsymbol{0}$表示填充向量。经过长度规范化的融合嵌入$\hat{\boldsymbol{Y}}'(Y' \in \mathbf{R}^{C \times d})$输入到CLIP文本编码器$TE$中，生成用于引导裙装图像生成的文本条件表示$\boldsymbol{c}_{\text{text}} = TE(\hat{\boldsymbol{Y}}') \in \mathbf{R}^{C \times d}$。

3 跨区域风格学习

基于2.2节得到的文本条件，结合输入的草图域和风格域图像，通过构建包含跨域图像特征对齐和双重条件协同融合的跨区域风格学习模块，实现结构、风格与语义的深度融合。其中，草图域图像提供显式结构控制，风格域图像提供区域纹理约束，文本条件提供全局语义信息。跨域图像特征对齐利用跳跃交叉注意力机制实现风格特征的选择性提取与跨区域传递，建立草图域与风格域之间的特征对齐关系；双重条件协同融合将跨域风格表示与增强的文本表示协同作为条件，提供精准的双重条件特征，实现语义与风格的精细协同。

3.1 跨域图像特征对齐

除文本提示外，草图作为显式结构控制，可为裙装图像生成提供轮廓约束。现有方法在处理草图引导的图像生成时往往独立处理不同草图区域，缺乏跨区域的信息交流，且风格特征常集中在某些显著区域，导致风格传递不均。因此，本文引入风格图像来对整个裙装草图区域进行调控。

采用层级式卷积特征提取模块，由多个3×3卷积层和SiLU激活函数组成，每层卷积操作为

$$\boldsymbol{X}^{l+1} = \sigma\left(f_{\text{Conv}}(\boldsymbol{X}^l) + b_l\right) \tag{13}$$

式中，$\boldsymbol{X}^l$表示第l层的输入特征图，$f_{\text{Conv}}(\cdot)$表示卷积操作，b_l表示该层的偏置，σ为SiLU（sigmoid linear unit）激活函数。在卷积过程中，通过步幅为2的卷积实现逐步下采样，提取不同尺度的特征表示。

$$\begin{aligned}C{:}\{&(H,W,16) \rightarrow (H/2,W/2,32) \rightarrow \\ &(H/4,W/4,96) \rightarrow (H/8,W/8,256)\}\end{aligned} \tag{14}$$

式中，H表示输入图像的高度，W表示输入图像的宽度。输入裙装草图$\boldsymbol{S}_k$和风格图像$\boldsymbol{P}_k$，提取对应的跨域特征表示为$\boldsymbol{f}_s = F(\boldsymbol{S}_k)$，$\boldsymbol{f}_p = F(\boldsymbol{P}_k)$。其中，$F(\cdot)$表示共享参数的特征提取网络，$\boldsymbol{f}_s$表示提取到的草图域特征，$\boldsymbol{f}_p$表示提取到的风格域特征。

为实现跨域特征的选择性提取与有效传递，本文设计跳跃交叉注意力机制。裙装草图特征作为查询向量，利用其结构划分信息指导风格特征的跨域匹配。将草图域特征$\boldsymbol{f}_s$线性投影映射为查询向量$\boldsymbol{Q}$和键值对$\boldsymbol{K}_S$，$\boldsymbol{V}_S$，该过程表示为

$$\boldsymbol{Q}, \boldsymbol{K}_S, \boldsymbol{V}_S = f_{\text{Linear}_s}(\boldsymbol{f}_s) \tag{15}$$

式中，$f_{\text{Linear}_s}(\cdot)$表示线性投影操作。

风格域特征$\boldsymbol{f}_p$映射为跨域键值对$\boldsymbol{K}_P$，$\boldsymbol{V}_P$，具体为

$$\boldsymbol{K}_P, \boldsymbol{V}_P = f_{\text{Linear}_p}(\boldsymbol{f}_p) \tag{16}$$

然后，将风格域和草图域的键值对进行拼接，实现跨域信息的直达传递，具体为

$$\hat{\boldsymbol{K}} = \boldsymbol{K}_P \Big\| \boldsymbol{K}_S, \quad \hat{\boldsymbol{V}} = \boldsymbol{V}_P \Big\| \boldsymbol{V}_S \tag{17}$$

式中，‖ 表示沿序列长度方向的拼接操作。交叉注意力通过融合不同域的键值对实现跨域特征关联。具体为

$$\boldsymbol{f}_m = f_{\text{CrossAttention}}(\boldsymbol{Q}, \hat{\boldsymbol{K}}, \hat{\boldsymbol{V}}) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\hat{\boldsymbol{K}}^{\mathrm{T}}}{\sqrt{d_k}}\right)\hat{\boldsymbol{V}} \tag{18}$$

为进一步建立从胸部到裙摆不同区域的直接跨域关联，采用跳跃连接进一步处理融合特征$\boldsymbol{f}_m$。在通道维度上，融合特征通过线性映射进行重新编码与信息流的调整，以优化通道间的语义分布，增强不同通道间的信息差异，从而提升融合特征的表征能力，得到优化后的融合特征$\boldsymbol{f}_m'$。具体为

$$\boldsymbol{f}_m' = f_{\text{Linear}}(\boldsymbol{f}_m) \tag{19}$$

最后，将优化后的融合特征、草图域特征和风格域特征进行特征整合，既保持各域特征的原有表达特性，又实现结构、风格与语义信息的多维度深度对齐，以增强跨域信息的互通性和互补性，生成综合保留裙装草图精确结构约束与风格图像丰富纹理信息的跨域风格条件表示$\hat{\boldsymbol{f}}$。具体为

$$\hat{\boldsymbol{f}} = \boldsymbol{f}_m' + \boldsymbol{f}_s + \boldsymbol{f}_p \tag{20}$$

3.2 双重条件协同融合

基于潜在空间扩散模型，构建双重条件协同融合机制，实现文本条件与风格条件的深度协同。首先，前向加噪过程向原始裙装图像的潜在表示$\boldsymbol{X}_0 \in \mathbf{R}^{H\times W\times C}$中逐步添加高斯噪声，其过程为

$$q(\boldsymbol{X}_t|\boldsymbol{X}_0) = \mathcal{N}\left(\boldsymbol{X}_t; \sqrt{\bar{\alpha}_t}\,\boldsymbol{X}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \tag{21}$$

式中，$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$为时间步$t$的累积噪声衰减系数，$\boldsymbol{X}_t$是每个时间步$t$生成的带噪样本。经过时间步$T$，得到标准高斯噪声$\boldsymbol{X}_T \sim \mathcal{N}(0, \boldsymbol{I})$。在反向去噪阶段，为实现语义与风格间的精细协同融合，引入双重条件协同机制。结合2.2节获得的文本条件表示$\boldsymbol{c}_{\text{text}}$与3.1节中跨域风格条件表示$\hat{\boldsymbol{f}}$共同作为协同条件输入。文本条件表示$\boldsymbol{c}_{\text{text}}$以交叉注意力形式逐层注入到U-Net的多个中间层，其中，文本条件表示$\boldsymbol{c}_{\text{text}}$作为查询向量，与U-Net多个中间层中的特征图进行交互，从而在深层语义空间中对全局结构进行引导，并为生成过程提供了一个语义框架，调控生成过程中的语义内容。跨域风格条件表示$\hat{\boldsymbol{f}}$与噪声潜在变量拼接后输入到U-Net编码器，在浅层网络协同引导局部纹理及色彩信息生成。

双重条件协同融合的噪声预测网络形式化为

$$\epsilon_\theta(\boldsymbol{X}_t, t, \boldsymbol{c}_{\text{text}}, \hat{\boldsymbol{f}})$$

模型训练目标最小化真实噪声$\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$与网络预测噪声之间的均方误差。其中，真实噪声是在扩散模型前向加噪过程中为构建带噪样本从标准高斯分布中独立采样的噪声变量。损失函数表示为

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{X}_0, \epsilon, t}\left\| \epsilon - \epsilon_\theta(\tilde{\boldsymbol{X}}_t, t, \boldsymbol{c}_{\text{text}}, \hat{\boldsymbol{f}}) \right\|_2^2 \tag{22}$$

式中，$\tilde{\boldsymbol{X}}_t$表示在训练过程中加入噪声的潜在表示。

通过上述双重条件协同融合训练，模型在潜在空间中恢复符合文本语义与跨域风格双重约束的裙装图像潜在表示。最终使用预训练解码器将该表示映射回图像空间，生成分辨率为512 × 512像素的裙装图像生成结果。

4 实验结果与分析

4.1 实验设置与评估指标

针对多模态图像生成任务，由于DressCode Multimodal数据集（Baldrati等，2023）能够提供高质量图文对，且具备同一服装样本的多角度文本注释，考虑到裙装在款式、风格以及结构上复杂多样（Shen等，2025），本文选取DressCode Multimodal的裙装子集作为研究对象。当前具备多角度文本注释的裙装图文数据仅在DressCode Multimodal数据集中公开可得，其他数据集在文本注释粒度或类别划分上不具备类似特性。从裙装类别下选取13 000个文本—图像对作为数据集。参考Sun等人（2023）的数据划分策略，随机选取其中11 700对样本用于训练，1 300对用于测试。由于DressCode Multimodal数据集原始设定面向虚拟试衣任务，提供的裙装草图存在变形且缺乏对应裙装风格图像，因此本文对草图与风格图像进行了重新提取，使用HED（holistically-nested edge detection）边缘检测算法（Xie和Tu，2015）从裙装图像中提取草图。参考Baldrati等人（2024）的服装风格图像提取策略，通过BASNet（boundary-aware salient object detection network）（Qin等，2019）模型对裙装图像进行前景分割，获得裙装掩码。结合裙装及其对应的裙装掩码，采用滑动窗口机制在裙装区域内截取大小为64 × 64像素的图像块，滑动步长为32像素，以保证每件裙装提取到有效的风格

图像。

实验基于PyTorch框架实现所提网络架构，在训练与测试阶段，实验平台搭载Intel Xeon Platinum 8352V 12核心CPU（2.10 GHz）和一个NVIDIA A100 GPU（32 GB显存）。采用两阶段训练策略以确保模型收敛稳定性：第1阶段对文本反转模块进行预训练，生成的伪词嵌入数量为16，使用AdamW优化器训练20 000步，优化器的超参数为$\beta_1 = 0.9$，$\beta_2 = 0.999$。初始学习率为0.000 01，其中前500步采用线性预热策略，批次大小为16，权重衰减为0.02。第2阶段是端到端联合训练，文本反转模块冻结，其他模块学习率为0.000 01，批次大小为16，共学习30个epoch。在扩散模型的反向过程中，时间步长设置为1 000。

本文采用5个评估指标：弗雷歇起始距离（Fréchet inception distance，FID）衡量生成图像与真实图像在特征分布上的差异，值越低表示图像质量越高。学习感知图像块相似度（learned perceptual image patch similarity，LPIPS）基于深度特征感知相似度度量，值越低表示视觉质量越好。对比语言图像预训练分数（contrastive language-image pre-training score，CLIPScore）用于衡量由CLIP模型评估的语义一致性，值越高表示图文匹配度越高。草图距离（sketch distance，SD）（Baldrati等，2023）衡量生成图像与输入草图之间的结构一致性，值越低表示结构遵循程度越高。纹理分数（texture score，TS）（Baldrati等，2024）评估生成图像与输入风格图像间的风格一致性，值越高越好。上述5个指标覆盖多模态引导裙装图像生成的图像质量、语义一致性、结构保真度和风格相似性4个维度，从模态一致性和感知质量两个层面评估本文方法性能。

4.2 实验结果与性能分析

4.2.1 定量对比分析

为全面评估本文方法的实现效果，选择了5类具有代表性的对比方法：纯文本引导的Stable Diffusion（Rombach等，2022）、结构控制的ControlNet（Zhang等，2023a）、服装专用的HieraFashDiff（Xie等，2025）、多模态融合的ControlNet（Zhang等，2023a）+ IP-Adapter（Ye等，2023）和多模态服装适配的IMAGGarment-1（Shen等，2025）。所有方法均在DressCode Multimodal数据集裙装子集上进行训练与评估，对比方法的结果均来源于开源代码的复现。

如表2所示。在图像整体质量上，与对比方法相比，本文取得了最优的FID值8.925，以及次优的LPIPS值0.205，这主要归因于本文引入的结构化风格增强学习引导机制，在训练过程中提供了更为清晰、丰富且一致的多模态信息，使得生成结果保留更多细节，增强了整体图像的质量和视觉上的真实感。在图文一致性方面，本文在CLIPScore指标上取得了32.874的最高得分，显著优于其他方法。通过对文本引入统一的语义模板且更侧重于属性信息，消除了自然语言描述中的歧义与冗余，提升了语义连贯性，为模型提供更明确的语义目标，同时引入视觉特征进一步丰富了文本语义，从而提升了生成裙装图像与文本描述之间的语义匹配。在风格呈现上，本文在TS上获得最优值0.692，表明跨区域风格学习模块有效捕捉并重构风格图像，在草图结构复杂的场景上，也能共享风格信息，保证良好的风格一致性与细节还原能力。在结构一致性度量指标SD上，虽然本文方法未达到最优值，但仍处于较低的范围内。相比于其他方法，这一差异主要源于本文方法在跨域图像特征对齐过程中对风格信息进行强化，一定程度上削弱了对草图结构的严格遵循，导致SD指标有所下降。综合来看，本文方法在大部分指标上均取得最优或次优表现，充分验证了本文方法在语义一致性、结构还原性及风格表现力上均较好地遵循了输入条件，能够有效地提升裙装图像生成的整体质量。

在评估裙装图像实现效果的基础上，本文进一步分析各方法的推理效率，结果如表3所示。与Stable Diffusion相比，ControlNet和HieraFashDiff由于引入结构信息，推理时间显著增加，分别达到8.22 s和6.96 s。ControlNet结合IP-Adapter虽然融合了风格图像，但由于其视觉引导模块较为轻量级，推理时间在2.75 s，表现出较高的运行效率。IMAGGarment-1在支持文本、草图与风格图像输入的情况下，推理时间为6.72 s，虽然不及ControlNet + IP-Adapter高效，但相较于ControlNet与HieraFashDiff仍具优势。本文方法在文本和风格方面引入了结构化文本增强和跨区域风格学习机制，虽然在一定程度上引入了计算开销，但仍将单幅图像的推理时间控制在7.77 s，与ControlNet、HieraFashDiff以及IMAGGarment-1复杂结构方法的推理时间相比，时间开销处于合理范围内。

表 2　本文方法与其他方法的定量对比

Table 2　Quantitative comparison between proposed method and other methods

模型	输入	FID ↓	LPIPS ↓	CLIPScore ↑	SD ↓	TS ↑
Stable Diffusion（Rombach 等，2022）	文本	13.095	0.737	<u>30.850</u>	0.810	0.674
ControlNet（Zhang 等，2023a）	文本 + 草图	11.937	0.397	27.902	0.194	0.656
HieraFashDiff（Xie 等，2025）	文本 + 草图	<u>11.056</u>	0.398	27.955	<u>0.173</u>	0.639
ControlNet（Zhang 等，2023a）+ IP-Adapter（Ye 等，2023）	文本 + 草图 + 风格图像	13.749	**0.197**	29.908	**0.164**	<u>0.676</u>
IMAGGarment-1（Shen 等，2025）	文本 + 草图 + 风格图像	22.577	0.342	27.893	0.544	0.604
本文	文本 + 草图 + 风格图像	**8.925**	<u>0.205</u>	**32.874**	0.176	**0.692**

注：加粗、下划线字体分别表示各列最优、次优结果。“↑”和“↓”分别表示值越高越好、值越低越好。

表 3　本文方法与其他方法的推理时间对比

Table 3　Inference time comparison between proposed method and other methods

模型	输入	推理时间/s
Stable Diffusion（Rombach 等，2022）	文本	**1.79**
ControlNet（Zhang 等，2023a）	文本 + 草图	8.22
HieraFashDiff（Xie 等，2025）	文本 + 草图	6.96
ControlNet（Zhang 等，2023a）+ IP-Adapter（Ye 等，2023）	文本 + 草图 + 风格图像	<u>2.75</u>
IMAGGarment-1（Shen 等，2025）	文本 + 草图 + 风格图像	6.72
本文	文本 + 草图 + 风格图像	7.77

注：加粗、下划线字体分别表示最优、次优结果。

4.2.2　定性对比分析

为了验证本文方法的有效性，在 DressCode Multimodal 数据集裙装子集上，以相同的输入与 Stable Diffusion、基线方法 HieraFashDiff 以及 ControlNet 和 IP-Adapter 结合方法进行定性对比，结果如图 2 所示。在颜色属性精确性上（第 1 行），文本描述明确要求“black”属性，对比结果显示 Stable Diffusion 生成了非预期的白色元素，反映出纯文本引导的局限性；HieraFashDiff 和 ControlNet + IP-Adapter 的生成结果色调偏向灰色，未能准确捕捉“black”的语义要求；本文方法准确还原了深黑色调，验证了动态属性模板生成的有效性。在材质纹理的表达能力上（第 2、3、4 行），对比方法生成的图像中均存在局部区域纹理缺失的问题，表现出对“lace”、“floral”以及“cady”等材质纹理的表达不足；HieraFashDiff 和 ControlNet + IP-Adapter 在整体结构还原上具备一定优势，但在材质纹理的细节还原上仍显粗糙；相比之下，本文方法能够准确还原材质纹理细节，表明引入文本反转语义融合与跨域图像特征对齐的有效性。在整体风格表现上（第 5 行），Stable Diffusion 生成结果整体结构简单且风格模糊，表明纯文本在特定服饰结构理解的局限性；HieraFashDiff 和 ControlNet + IP-Adapter 的生成结果尽管在局部风格上与参考图像存在一定相似性，但在风格统一性上仍表现不足。相比之下，本文方法生成的裙装图像在整体风格上与参考图像表现出更高的一致性与相似度，进一步验证了跨域图像特征对齐的有效性。上述实验结果表明，本文方法在裙装属性理解、区域风格一致性以及整体图像质量方面均展现出良好的有效性与表现力。

此外，为验证本文方法在其他类别上的适用性，图 3 展示了部分上装与下装的生成结果。实验结果表明，本文方法能够较好地满足不同类别的输入要求，实现相应的图像生成。但由于上装、下装与裙装在结构和细节表达上仍存在差异，部分细节表现仍有进一步提升的空间。

4.3　消融实验结果分析

4.3.1　动态属性模板生成的消融分析

为了验证本文方法各模块的有效性，在 DressCode Multimodal 裙装数据集上以 Xie 等人（2025）的方法为基础模型（简称为 Base），分别引入各模块进

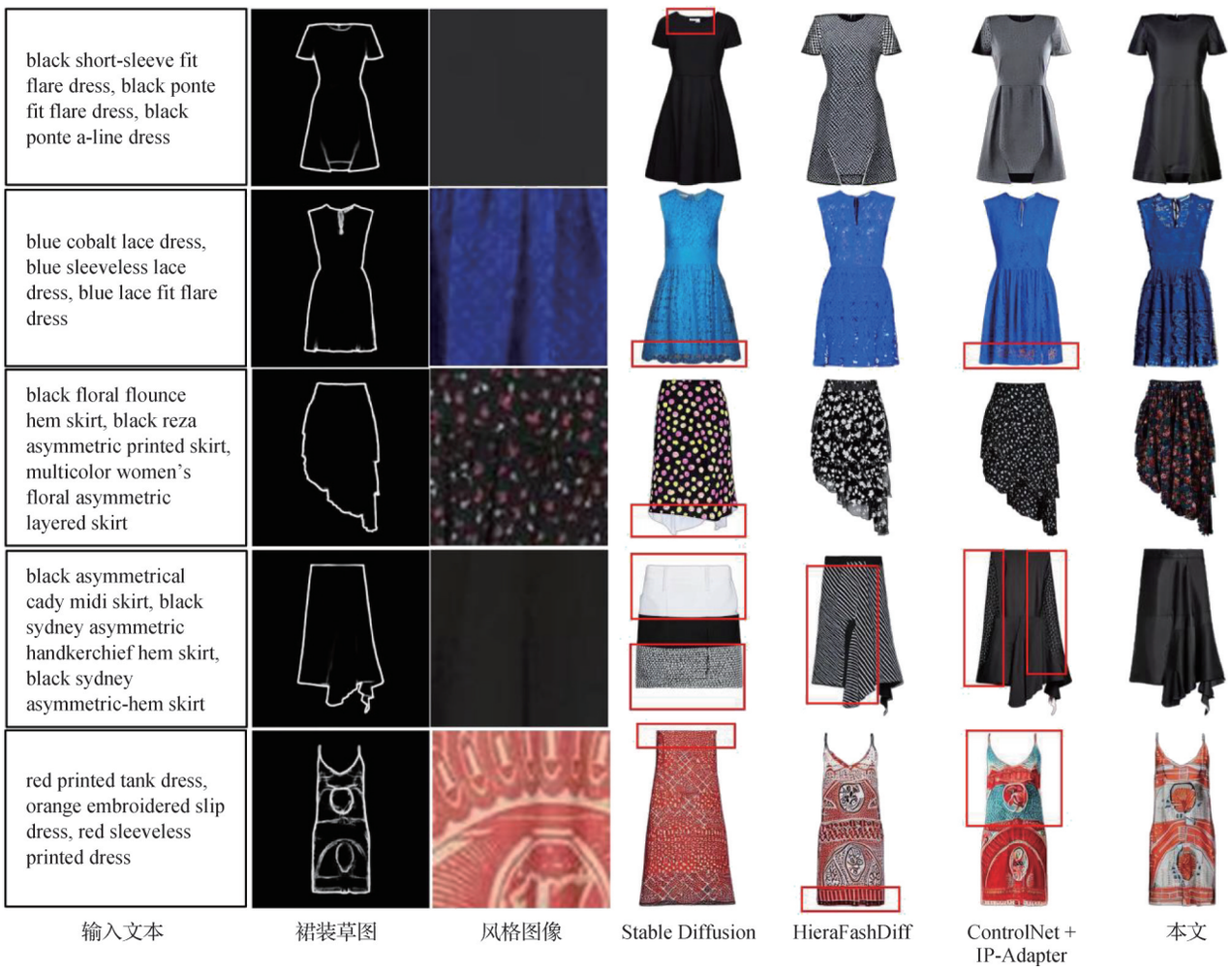

图2　本文方法与其他方法可视化结果对比

Fig. 2　The visual results of proposed method compared with other methods

图3　本文方法在其他类别上的生成结果

Fig. 3　The results of proposed method on other categories

行消融实验。为了分析本文方法中动态属性模板生成模块(简称DAT)的有效性,将该模块加入到Base中,结果如表4所示。引入DAT后,FID、CLIPScore和TS这3项指标均有提升,FID降低8.8%(从11.056降至10.083),表明生成图像的整体质量得到显著改善。这主要归因于DAT消除了原始文本中的歧义与冗余,为模型提供了更一致的引导信号。CLIPScore提升了7.6%(从27.955提升至30.072),表明统一的模板格式规范了输入文本,确保生成过程中语义连贯,显著提升了生成图像与文本的一致性。TS提升了2.4%(从0.639提升至0.654),DAT通过关键属性的提取有效强化了对文本中关键属性的理解,使得生成结果中对材质这一关键属性得到了更准确的表达。SD保持不变,表明DAT未影响整体结构表达的稳定性。LPIPS指标略有上升(从0.398升至0.406),这是因为DAT促使模型更关注关键属性的完整表达,在某些样例中生成更加丰富或细化的内容,从而与真实裙装图像之间的感知差

异上升。综合来看，动态属性模板生成模块能够提升生成裙装图像的准确性和质量。

为进一步展示动态属性模板生成模块的有效性，图 4 展示了动态属性模板生成模块的定性对比结果。由于 Base 方法已通过草图提供了款式、轮廓等结构性引导，本文重点选取颜色、材质与图案 3 个维度样例，验证 DAT 模块在处理文本特有属性方面的优势。在颜色属性冲突时（第 1 行），多角度文本注释存在“cream”和“white”以及“black”和“purple”等两个及以上颜色信息，Base 生成结果颜色表达模糊或偏离主色。DAT 基于 LLM 的语义理解与推理机制识别主导颜色信息，有效解决了颜色描述中的潜在冲突。在材质属性处理上（第 2 行），文本中涉及“lace”、“demin”等材质描述，由于 Base 缺乏对材质信息的强化，生成图像的材质粗糙或质感缺失。DAT 模块将材质视为关键信息并显式融入提示模板，显著增强了模型对细节的感知能力，实现更精确的材质表达。在裙装图案表现上（第 3 行），“floral”在多角度文本中存在高度概括性，Base 模型常出现图案与文本描述不符和变形等问题，引入 DAT 模块，通过结构化模板统一提示语义，强化图案作为视觉核心的信息表达，有效提升图案的规整性与语义一致性，实现复杂图案的还原。以上实验结果表明，动态属性模板生成模块显著提升了裙装图像生成中对文本属性的表达能力，特别在颜色冲突、材质准确表达以及复杂图案还原方面表现突出。

表 4　动态属性模板生成消融实验定量分析

Table 4　Quantitative results of the ablation study on dynamic attribute template generation

模型	FID ↓	LPIPS ↓	CLIPScore ↑	SD ↓	TS ↑
Base	11.056	**0.398**	27.955	0.173	0.639
Base + DAT	**10.083**	0.406	**30.072**	0.173	**0.654**

注：加粗字体表示各列最优结果。“↑”和“↓”分别表示值越高越好、值越低越好。

图 4　动态属性模板生成消融实验结果分析

Fig. 4　The visual results of the ablation study on dynamic attribute template generation

尽管动态属性模板生成模块针对裙装设计，图 5 展示了其在上装和下装的适用性。实验结果表明，对于上装（第 1、2 行），由于其在结构层次与属性描述方面（如袖型、领型等）存在一定共性，DAT 模块能够有效迁移并准确提取关键语义，实现高质量的图像生成；对于下装（第 3、4 行），在“beige”、“white”颜色属性存在冲突时，DAT 模块仍然能够有效解析颜色的主导信息，但由于下装在描述方式上更偏向于结构性剪裁与穿着形式的表达，像“pull-on”（无拉链穿脱）、“high-rise”（高腰）、“cropped trousers”（九分裤）以及“button”（系扣）等，与裙装以风格、材质和装饰性为主的表达方式存在差异。因此，尽管动态属

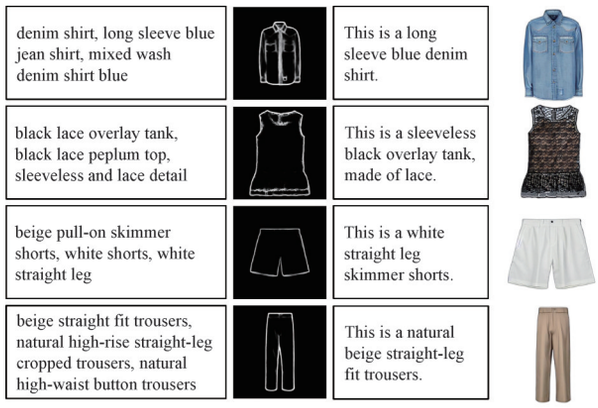

图5　动态属性模板生成在其他类别适用性分析

Fig. 5　The applicability of dynamic attribute template generation on other categories

性模板生成模块具备一定的迁移能力，但面对下装特定的一些属性解析时仍会出现属性缺失或者理解偏差的情况。

4.3.2　文本反转语义融合的消融分析

为验证所提出的文本反转语义融合策略（简称TSF）在裙装图像生成任务中的有效性，本文在Base基础上引入该模块，定量结果如表5所示。本文引入文本反转语义融合策略，能够提升裙装图像生成的整体视觉质量。在Base + TSF的情况下，各项评估指标均优于Base。其中，FID降低了约16.5%，LPIPS降低了28.6%，表明生成裙装图像在分布相似性与感知质量方面更接近真实图像；CLIPScore提升了约7.9%，说明将图像语义信息融入文本嵌入，使得生成裙装图像在语义层面与文本描述也更为一致；TS指标提高6.6%，说明视觉特征中包含的材质与纹理信息有效促进了裙装纹理的还原；SD指标虽略有下降但幅度较小，表明引入的视觉特征未对裙装图像的整体结构形态产生显著影响。实验结果表明，文本反转语义融合策略在保持生成裙装图像结构稳定性的同时，能够提升裙装图像的视觉保真度和语义一致性。

通过文本反转语义融合策略的定性消融实验，进一步展示该模块的有效性，生成裙装图像与Base对比如图6所示。可以看出，引入视觉特征后生成的裙装图像在裙装的面料、图案上更加丰富，且在语义上与输入文本描述更加契合，呈现出更高的视觉真实感与表达准确性。

为探究文本反转语义融合策略中伪词嵌入数量

表5　文本反转语义融合消融实验定量分析

Table 5　Quantitative results of the ablation study on text inversion semantic fusion

模型	FID ↓	LPIPS ↓	CLIPScore ↑	SD ↓	TS ↑
Base	11.056	0.398	27.955	0.173	0.639
Base + TSF	**9.231**	**0.284**	**30.174**	**0.172**	**0.681**

注：加粗字体表示各列最优结果。“↑”和“↓”分别表示值越高越好、值越低越好。

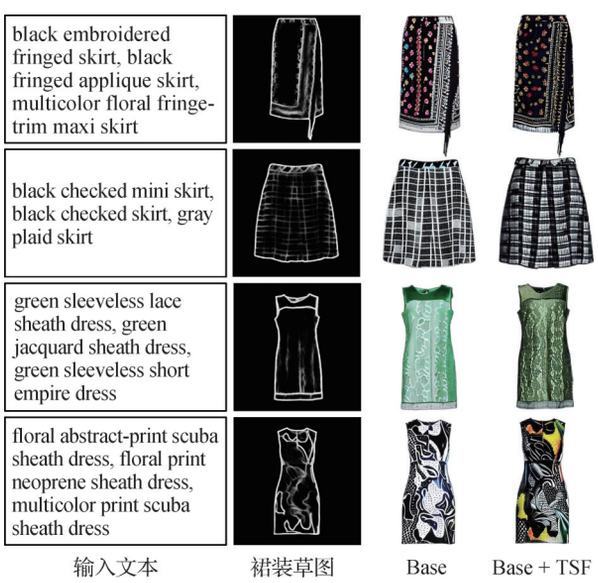

图6　文本反转语义融合消融实验可视化结果分析

Fig. 6　The visual results of the ablation study on text inversion semantic fusion

对生成效果的影响，本文设计了不同伪词嵌入个数的消融实验。在Base基础上，将原始的文本嵌入与融合视觉特征生成的伪词嵌入拼接经文本编码器编码后作为条件指导，Base框架和其余参数保持不变，结果如表6所示。其中，选择伪词嵌入个数为1，4，16，32进行对比，以分析不同程度的视觉信息对生成效果的影响。可以看出，当伪词嵌入个数较少（如1或4）时，视觉信息过少或不足，无法有效补充文本语义，限制了细节表达能力；当伪词嵌入个数为16时，FID、LPIPS、CLIPScore以及SD指标上均表现最优，TS指标相较于伪词嵌入个数少时也有明显提升，说明适量视觉信息引导能够有效增强文本信息的表达以及整体图像的质量；当伪词嵌入个数为32时，尽管在TS指标上略高于伪词嵌入为16时的效果，但其他指标有所下降，说明过多的视觉特征在一定程度上引入了冗余或噪声信息，干扰了文本语义的表达。

进一步，对不同个数伪词嵌入进行文本反转的模型参数量与单幅图像推理时间进行实验，结果如表7所示。随着伪词嵌入数量从1增加至32，模型参数量从56.4 M逐步上升至178.3 M，推理时间也由7.32 s增长至7.81 s，呈现出近似线性增长趋势，表明伪词嵌入数量直接会影响文本反转模型的复杂度与推理开销，虽然能提升模型表征能力，但同时带来了显著的计算资源消耗。综合考虑，本文实验设置伪词嵌入个数为16时可在实现质量与计算效率实现有效均衡。

4.3.3　跨域图像特征对齐的消融分析

将本文提出的基于跳跃交叉注意力机制的跨域图像特征对齐策略（Base + SCA）与Base、传统的特征相加（Base + add）以及普通的交叉注意力机制（Base + CA）进行消融实验，验证跨域图像特征对齐模块的有效性，定量对比结果如表8所示。可以看出，与简单的特征相加和传统交叉注意力机制相比，本文提出的基于跳跃交叉注意力机制的策略在FID、LPIPS、CLIPScore以及TS指标上均显著提升，说明该模块能够有效融合草图结构与风格图案信息，增强不同区域间的风格一致性表达，但SD指标有所下降，这是因为融合过程中牺牲掉了一部分对局部结构的强约束能力，使得生成裙装图像在结构上的还原精度有所下降，但该影响在视觉效果上并不明显。综合考虑各项指标的表现，说明跨域图像特征对齐模块能够提升整体风格的一致性，进而提升裙装图像的质量。

表6　不同伪词嵌入个数的结果对比

Table 6　The results of different number of pseudo-token embedding

伪词嵌入个数	FID ↓	LPIPS ↓	CLIP-Score ↑	SD ↓	TS ↑
l = 1	10.599	0.399	27.580	0.159	0.628
l = 4	11.118	0.417	27.660	0.165	0.632
l = 16	**9.934**	**0.342**	**28.072**	**0.150**	0.635
l = 32	10.029	0.350	27.661	0.157	**0.636**

注：加粗字体表示各列最优结果。“↑”和“↓”分别表示值越高越好、值越低越好。

进一步，对不同融合方式的定性结果进行实验，裙装图像生成结果如图7所示。可以看出，在Base + add条件下，尽管引入了一定的风格信息，但由于仅采用简单的特征相加方式，不同区域间仍存在一定的风格不一致情况。相比之下，Base + CA能够通过交叉注意力机制在一定程度上实现了结构与风格的初步融合，但部分区域仍存在风格信息传播不足，从而导致图像中整体风格图案分布不均。而本文提出的Base + SCA策略则能通过跳跃交叉注意力机制实现跨区域的信息流动与共享，显著地提升了风格一致性以及裙装图像整体的质量，整体效果都要优于Base + CA的策略。

表7　不同伪词嵌入个数的性能对比

Table 7　The performance of different number of pseudo-token embedding

伪词嵌入个数	参数量/M	推理时间/s
l = 1	56.4	7.32
l = 4	68.2	7.35
l = 16	115.4	7.51
l = 32	178.3	7.81

表8　跨域图像特征对齐消融实验定量分析

Table 8　Quantitative results of the ablation study on cross-domain image feature alignment

模型	FID ↓	LPIPS ↓	CLIP-Score ↑	SD ↓	TS ↑
Base	11.056	0.398	27.955	**0.173**	0.639
Base + add	9.511	0.421	30.568	0.180	0.667
Base + CA	9.113	0.391	30.668	0.193	0.670
Base + SCA	**9.068**	**0.346**	**30.794**	0.205	**0.685**

注：加粗字体表示各列最优结果。“↑”和“↓”分别表示值越高越好、值越低越好。

4.4　用户评估结果分析

为了进一步验证本文方法在裙装图像生成任务中的主观质量，参考Lampe等人（2024）的用户评估的方式，从DressCode Multimodal数据集裙装子集中选取了10组具有代表性的输入，5种方法基于此生成对应的裙装图像。每位参与者需从5幅候选裙装图像中选出其认为最优的一幅，评估参考3个维度：1）裙装图像的整体视觉质量；2）裙装图像对文本输入的一致性；3）裙装图像中跨域风格一致性。本次用户评估邀请80名参与者，大部分来源于高校学生群体，确保了评估结果的客观性与有效性。图8展示了不同方法在3项评估指标上的用户偏好统计结

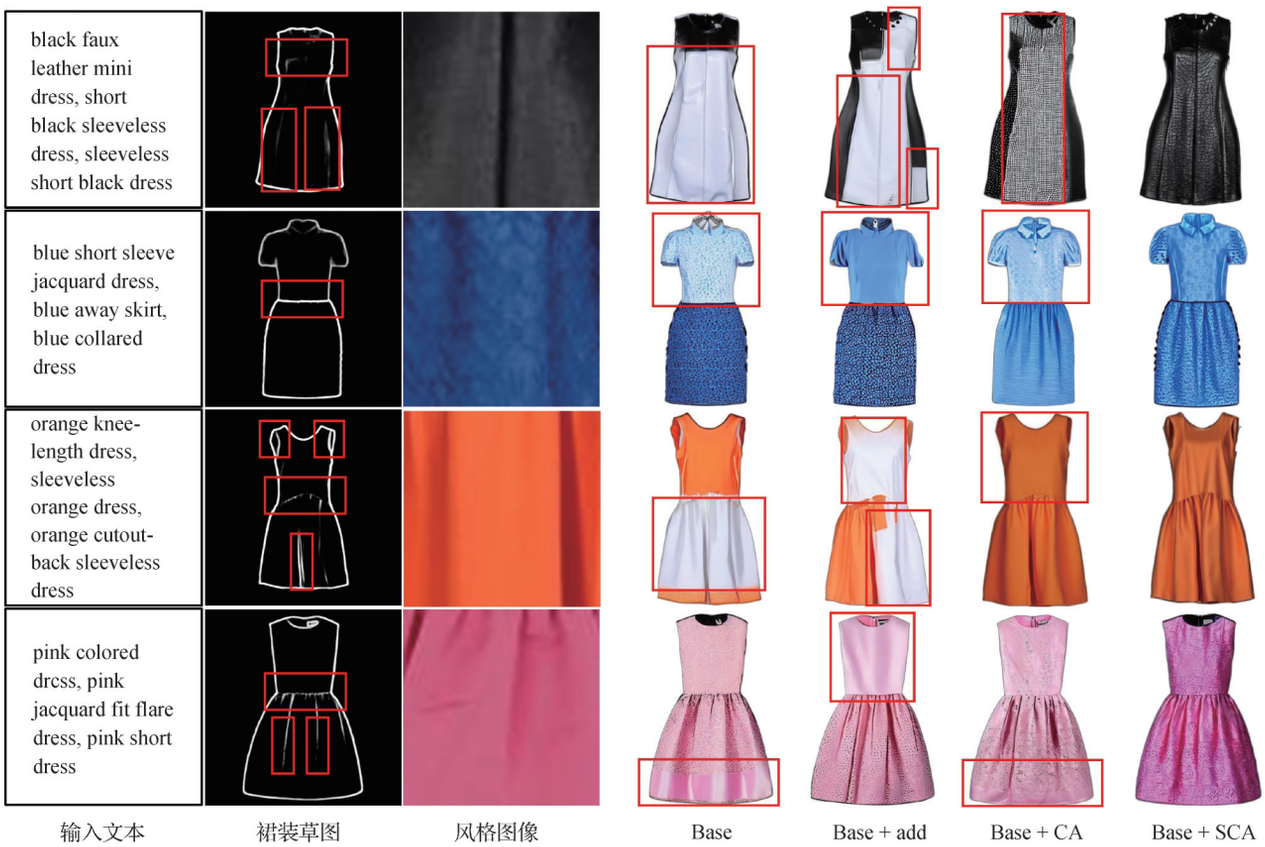

图7　跨域图像特征对齐消融实验的可视化结果分析

Fig. 7　The visual results of the ablation study on cross-domain image feature alignment

果。本文方法在3个维度上均获得最高的用户偏好比例,说明其能够有效提升裙装图像生成的主观感知质量。此外,从评估结果中还可看出,相较于自动评价指标所呈现的性能差异,用户评估更显著地区分了不同方法在裙装图像生成质量与风格表现方面的优劣,进一步说明在人类感知驱动的裙装图像生

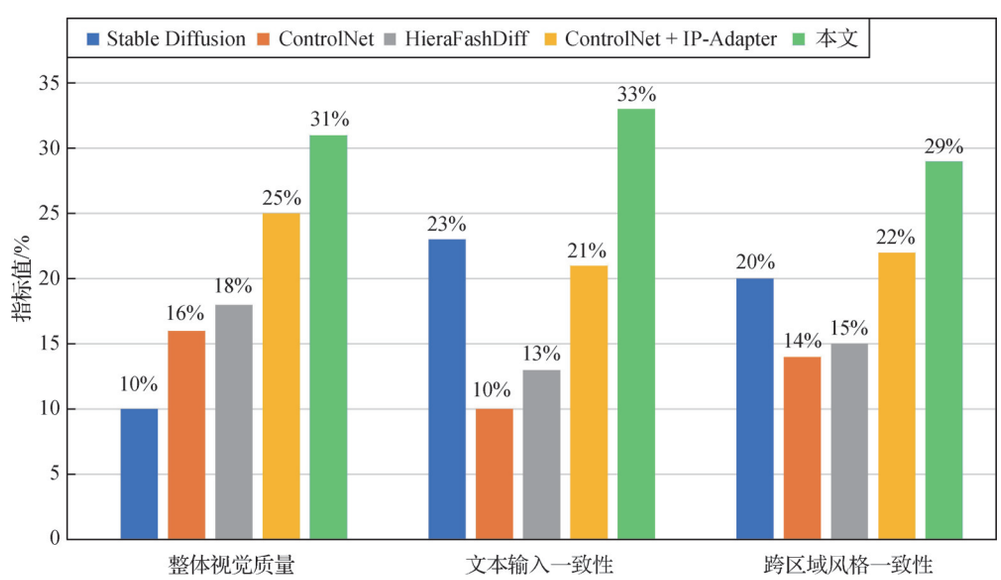

图8　用户评估结果分析

Fig. 8　The results of user study

成任务中，主观评价在揭示模型感知质量差异方面的必要性。

5 结 论

为进一步提升裙装图像生成的质量，本文通过构建结构化文本增强和跨区域风格学习模块，提出多模态引导裙装图像生成的结构化风格增强学习方法。其中，结构化文本增强模块通过动态属性模板生成和文本反转语义融合，提取与重构裙装七大关键属性，将裙装视觉特征与文本语义深度融合，实现文本语义的丰富表达；跨区域风格学习模块通过跨域图像特征对齐和双重条件协同融合，建立裙装不同区域结构间的风格关联，与文本条件分层协同，生成语义一致、风格协调的高质量裙装图像。实验结果显示，在 DressCode Multimodal 数据集裙装子集上，本文方法较其他方法在裙装图像整体质量和对输入约束的一致性方面具有明显优势。

本文方法虽然能够实现较为精细的语义对齐与跨域风格表达，但模块间复杂的信息交互增加了计算复杂度，限制其在移动端、在线服装设计等应用场景的实时性和部署效率。在草图结构较为复杂或边界密集的情况下，生成的裙装图像可能在边缘区域出现轻微的黑线伪影，且在处理不常见极端风格时，生成结果仍有待改善。此外，本文目前研究主要针对裙装类别展开，尚未对其他服装类别进行深入研究，未来将围绕上述问题开展研究，进一步优化模型并拓展提示模板与属性解析策略的通用性，以适应更多样化的服装类别和风格场景。

参考文献（References）

Ak K E, Lim J H, Tham J Y and Kassim A A. 2020. Semantically consistent text to fashion image synthesis with an enhanced attentional generative adversarial network. Pattern Recognition Letters, 135: 22-29 [DOI: 10.1016/j.patrec.2020.02.030]

Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick C L, et al. 2015. VQA: visual question answering//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/ICCV.2015.279]

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//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 23336-23345 [DOI: 10.1109/ICCV51070.2023.02138]

Baldrati A, Morelli D, Cornia M, Bertini M and Cucchiara R. 2024. Multimodal-conditioned latent diffusion models for fashion image editing [EB/OL]. [2025-07-08]. https://arxiv.org/pdf/2403.14828.pdf

Chen L L, Tian J, Li G, Wu C H, King E K, Chen K T, et al. 2020. TailorGAN: making user-defined fashion designs//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass, USA: IEEE: 3230-3239 [DOI: 10.1109/WACV45572.2020.9093416]

Cheng W H, Song S J, Chen C Y, Hidayati S C and Liu J Y. 2021. Fashion meets computer vision: a survey. ACM Computing Surveys (CSUR), 54(4): #72 [DOI: 10.1145/3447239]

Choi S, Park S, Lee M and Choo J. 2021. VITON-HD: high-resolution virtual try-on via misalignment-aware normalization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14126-14135 [DOI: 10.1109/CVPR46437.2021.01391]

Cui H L, Liu L, Zhang H X, Liu D M, Ma Y and Wang Z K. 2022. Detail preserving image generation method based on semantic consistency. Journal of Computer-Aided Design and Computer Graphics, 34(10): 1497-1505 (崔怀磊, 刘丽, 张化祥, 刘冬梅, 马跃, 王泽康. 2022. 基于语义一致性的细节保持图像生成方法. 计算机辅助设计域图形学报, 34(10): 1497-1505) [DOI: 10.3724/SP.J.1089.2022.19724]

Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D and Taigman Y. 2022. Make-A-Scene: scene-based text-to-image generation with human priors//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 89-106 [DOI: 10.1007/978-3-031-19784-0_6]

Hu H X, Chan K C K, Su Y C, Chen W H, Li Y D, Sohn K, et al. 2024. Instruct-imagen: image generation with multi-modal instruction//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4754-4763 [DOI: 10.1109/CVPR52733.2024.00455]

Jain A, Modi D, Jikadra R and Chachra S. 2019. Text to image generation of fashion clothing//Proceedings of the 6th International Conference on Computing for Sustainable Global Development (INDIACom). New Delhi, India: IEEE: 355-358

Jiang Y M, Yang S, Qiu H N, Wu W, Loy C C and Liu Z W. 2022. Text2Human: text-driven controllable human image generation. ACM Transactions on Graphics (TOG), 41(4): #162 [DOI: 10.1145/3528223.3530104]

Lampe A, Stopar J, Jain D K, Omachi S, Peer P and Štruc V. 2024. DiCTI: diffusion-based clothing designer via text-guided input//Proceedings of the 18th IEEE International Conference on Automatic Face and Gesture Recognition (FG). Istanbul, Turkiye: IEEE:

1-9［DOI：10.1109/fg59268.2024.10582026］

Li B W, Qi X J, Lukasiewicz T and Torr P H S. 2020. ManiGAN：text-guided image manipulation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA：IEEE：7877-7886［DOI：10.1109/CVPR42600.2020.00790］

Liu Y X and Wang L. 2024. MYCloth：towards intelligent and interactive online T-shirt customization based on user's preference//Proceedings of 2024 IEEE Conference on Artificial Intelligence (CAI). Singapore, Singapore：IEEE：955-962［DOI：10.1109/cai59869.2024.00175］

Moradian F, Azmi R, Darvishi H and Khademhossein M. 2024. Text to fashion image synthesis via CW-ControlGAN//Proceedings of the 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP). Babol, Iran：IEEE：1-7［DOI：10.1109/aisp61396.2024.10475289］

Morelli D, Baldrati A, Cartella G, Cornia M, Bertini M and Cucchiara R. 2023. LaDI-VTON：latent diffusion textual-inversion enhanced virtual try-on//Proceedings of the 31st ACM international Conference on Multimedia. Ottawa, Canada：ACM：8580-8589［DOI：10.1145/3581783.3612137］

Pernuš M, Fookes C, Štruc V and Dobrišek S. 2025. FICE：text-conditioned fashion-image editing with guided GAN inversion. Pattern Recognition, 158：#111022［DOI：10.1016/j. patcog. 2024.111022］

Qin X B, Zhang Z C, Huang C Y, Gao C, Dehghan M and Jagersand M. 2019. BASNet：boundary-aware salient object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA：IEEE：7471-7481［DOI：10.1109/CVPR.2019.00766］

Rico Gómez R, Lorentz J, Hartmann T, Goknil A, Pal Singh I, Halaç T G, et al. 2024. An AI pipeline for garment price projection using computer vision. Neural Computing and Applications, 36(25)：15631-15651［DOI：10.1007/s00521-024-09901-w］

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA：IEEE：10674-10685［DOI：10.1109/CVPR52688.2022.01042］

Shastri H, Lodhavia D, Purohit P, Kaoshik R and Batra N. 2022. VastrGAN：versatile apparel synthesised from text using a robust generative adversarial network//Proceedings of the 5th Joint International Conference on Data Science and Management of Data (9th ACM IKDD CODS and 27th COMAD). Bangalore, India：ACM：222-226［DOI：10.1145/3493700.3493721］

Shen F, Yu J, Wang C, Jiang X, Du X Y and Tang J H. 2025. Imaggarment-1：fine-grained garment generation for controllable fashion design［EB/OL］.［2025-07-08］. https://arxiv.org/pdf/2504.13176v1.pdf

Sun Z W T, Zhou Y H, He H H and Mok P Y. 2023. SGDiff：a style guided diffusion model for fashion synthesis//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada：ACM：8433-8442［DOI：10.1145/3581783.3613806］

Surya S, Setlur A, Biswas A and Negi S. 2020. ReStGAN：a step towards visually guided shopper experience via text-to-image synthesis//Proceedings of 2020 IEEE/CVF Winter Conference on Applications of Computer Vision. Snowmass, USA：IEEE：1189-1197［DOI：10.1109/WACV45572.2020.9093459］

Wang H Y, Wu P X, Rosa K D, Wang C and Shrivastava A. 2024. Multimodality-guided image style transfer using cross-modal GAN inversion//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA：IEEE：4964-4973［DOI：10.1109/WACV57701.2024.00490］

Xie S N and Tu Z W. 2015. Holistically-nested edge detection//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile：IEEE：1395-1403［DOI：10.1109/ICCV.2015.164］

Xie Z F, Li H, Ding H M, Li M T, Di X H and Cao Y. 2025. HieraFashDiff：hierarchical fashion design with multi-stage diffusion models//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA：AAAI：8762-8770［DOI：10.1609/aaai.v39i8.32947］

Xu Z G, Pu B C, Qin J M, Xiang Y P, Peng Z J, Song C F. 2025. Data benchmark and model framework for high-definition human image generation. Journal of Image and Graphics, 30(2)：375-390（徐正国，普碧才，秦建明，项炎平，彭振江，宋纯锋. 2025. 面向高清人体图像生成的数据基准与模型框架. 中国图象图形学报，30(2)：375-390）［DOI：10.11834/jig.240159］

Yan H, Zhang H J, Mu X Y, Fan J C and Zhang Z. 2023. FashionDiff：a controllable diffusion model using pairwise fashion elements for intelligent design//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada：ACM：1401-1411［DOI：10.1145/3581783.3612127］

Ye H, Zhang J, Liu S B, Han X and Yang W. 2023. IP-adapter：text compatible image prompt adapter for text-to-image diffusion models［EB/OL］.［2025-07-08］. https://arxiv.org/pdf/2308.06721.pdf

Zhang H X, Jiang S H and Fu Y. 2021. Stylized text-to-fashion image generation//Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). Jodhpur, India：IEEE：1-8［DOI：10.1109/fg52635.2021.9667042］

Zhang L M, Rao A Y and Agrawala M. 2023a. Adding conditional control to text-to-image diffusion models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France：IEEE：3813-3824［DOI：10.1109/ICCV51070.2023.00355］

Zhang S Y, Chong Z, Zhang X J, Li H H, Cheng Y H, Yan Y Q, et al. 2025. GarmentAligner：text-to-garment generation via retrieval-augmented multi-level corrections//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy：Springer：148-

164［DOI：10.1007/978-3-031-72698-9_9］

Zhang X J，Sha Y，Kampffmeyer M C，Xie Z Y，Jie Z Q，Huang C W，et al. 2022. ARMANI：part-level garment-text alignment for unified cross-modal fashion design//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon，Portugal：ACM：4525-4535［DOI：10.1145/3503161.3548230］

Zhang X J，Yang B B，Kampffmeyer M C，Zhang W Q，Zhang S Y，Lu G S，et al. 2023b. DiffCloth：diffusion based garment synthesis and manipulation via structural cross-modal semantic alignment//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris，France：IEEE：23097-23106［DOI：10.1109/ICCV51070.2023.02116］

Zhang Y M，Zhang T Y and Xie H R. 2024. TexControl：sketch-based two-stage fashion image generation using diffusion model//2024 Nicograph International（NicoInt）. Hachioji，Japan：IEEE：64-68［DOI：10.1109/nicoint62634.2024.00021］

作者简介

马嘉妮，女，硕士研究生，主要研究方向为图形图像处理。
E-mail：20232204356@stu.kust.edu.cn

刘骊，通信作者，女，教授，博士生导师，主要研究方向为计算机图形学与计算机视觉、图像处理。
E-mail：ieall@kust.edu.cn

付晓东，男，教授，博士生导师，主要研究方向为服务计算、决策理论与方法。E-mail：xdfu@kust.edu.cn

刘利军，男，副教授，主要研究方向为图像处理、云计算和信息检索。E-mail：cloneiq@kust.edu.cn

彭玮，女，教授，博士生导师，主要研究方向为机器学习和数据挖掘。E-mail：weipeng@kust.edu.cn