中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-14

论文引用格式： Bai Yufan, Qian Wenhua, Yang Rongyi. Portrait generation of Dongba paintings via integrating anatomical reward and style constraint[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260168. (白宇帆, 钱文华, 杨荣懿. 融合躯体奖励与风格约束的东巴画人像生成[J/OL]. 中国图象图形学报, XXXX:1-14. DOI: 10.11834/jig.260168.)[DOI:10.11834/jig.260168]

融合躯体奖励与风格约束的东巴画人像生成

白宇帆，钱文华*，杨荣懿

云南大学信息学院，昆明 650504

摘 要：目的 东巴画作为纳西族传统视觉文化的重要载体，具有古朴写意的笔触与鲜明的色彩特征。利用图像生成技术辅助东巴画创作，对其数字传播、风格理解与活态传承具有重要意义。然而，通用扩散模型仅以噪声预测损失为训练目标，缺乏针对艺术化人体结构的显式语义约束，直接应用于东巴画人像生成任务时，生成的结果存在人体结构伪影与风格退化现象，严重影响视觉质量。为此，本文提出一种融合躯体奖励与风格约束的东巴画人像生成方法。**方法** 针对扩散模型生成东巴画人像的结构伪影问题，设计东巴画躯体伪影奖励函数，将人体结构伪影显式建模为奖励信号，引导生成模型向结构合理分布收敛，减少人体结构伪影；针对抑制伪影过程中的风格退化问题，设计东巴画风格约束奖励函数，将风格一致性转化为奖励约束，引导生成模型保持东巴画风格分布，缓解风格退化。在此基础上，提出自适应权重调度策略，动态平衡躯体奖励损失、风格约束损失与噪声预测损失，实现生成过程中结构合理性与风格一致性的协同优化。**结果** 在自建东巴画人像图文数据集上的实验表明，本文方法的风格一致性指标FID(Fréchet inception distance，数值越低表示风格分布匹配度越高)降至0.1164，躯体合理性指标MAF(mean artifact frequency，数值越低表示人体结构伪影越少)降至0.6767，相对基线模型Stable Diffusion v1.5分别降低1.27%和2.39%。客观指标与主观视觉质量评估结果表明，本文方法能有效抑制人体结构伪影并保持东巴画风格一致性。**结论** 本文首次针对东巴画人像生成任务开展系统性研究，有效解决了伪影抑制与风格退化难以兼顾的矛盾，实现了结构合理性与风格一致性的协同优化，为东巴画人像的高保真数字化生成提供了新的技术路径，对东巴画的数字化保护与活态传承具有重要意义。

关键词： 东巴画；扩散模型；图像生成；人体伪影；风格约束；偏好对齐

Portrait generation of Dongba paintings via integrating anatomical reward and style constraint

Bai Yufan, Qian Wenhua*, Yang Rongyi

School of Information Science and Engineering, Yunnan University, Kunming 650504

Abstract: Objective Dongba painting, a representative visual form of the Naxi ethnic culture in southwestern China, features expressive brushwork, simplified symbolic human forms, and distinctive color compositions. With the rapid advance-

收稿日期：2026-04-01；**修回日期：**2026-07-29

* **通信作者：**钱文华 whqian@ynu.edu.cn

基金项目：国家自然科学基金项目(62162065)；云南大学“双一流”建设联合专项(202401BF070001-023)；云南省“视觉与文化计算创新团队”(202505AS350009)；云南省科技厅应用基础研究计划项目(202201AT070167)

Supported by: the National Natural Science Foundation of China under Grant(62162065); Joint Special Project Research Foundation of Yunnan Province(202401BF070001-023); Yunnan Province Visual and Cultural Innovation Team(202505AS350009); Yunnan Fundamental Research Projects(202201AT070167)

ment of generative models, text-to-image synthesis offers new avenues for the digital preservation, dissemination, and reinterpretation of such traditional art forms. However, when diffusion-based models are directly applied to Dongba painting portrait generation, they exhibit limited capability in modeling high-level semantic constraints, particularly human anatomical plausibility and artistic style consistency. This limitation results in two major challenges: (1) human structural artifacts such as distorted limbs and anatomical inconsistencies, and (2) style degradation, where generated images deviate from the characteristic Dongba visual distribution during artifact suppression. These issues significantly compromise both visual quality and cultural fidelity. To address these challenges, this study proposes a unified Dongba painting portrait generation framework that jointly optimizes anatomical plausibility and stylistic consistency through anatomical rewards and style consistency learning. **Method** A reward-driven diffusion optimization framework is developed, with three core components: an anatomical artifact reward, a style consistency reward, and an adaptive weight scheduling strategy. First, a Dongba anatomical artifact reward (DAAR) function is designed. It is built on a gated artifact segmentation network for structural anomaly detection. This network identifies distorted limbs, duplicated body parts, and implausible anatomical configurations, generates spatial confidence maps, and converts them into reward signals to guide the model toward structurally plausible human representations. Second, to mitigate style degradation during artifact suppression, a Dongba painting style consistency reward (DSCR) is introduced. A style prototype library is constructed from representative Dongba painting samples to capture canonical stylistic distributions. Based on this library, a dual-branch style evaluation mechanism is designed. The global branch extracts semantic-level features to align overall color schemes and compositional patterns, while the local branch focuses on fine-grained texture representations, including brushstroke characteristics and line stylization. By jointly optimizing these two levels of style representation, the proposed method ensures that generated images remain consistent with authentic Dongba painting aesthetics throughout the generation process. Third, to accommodate the varying optimization priorities of structural correction and style preservation across diffusion stages, an adaptive weight scheduling strategy is proposed. This mechanism dynamically adjusts the relative contributions of the anatomical artifact reward, style consistency reward, and diffusion denoising loss during training according to validation performance. Compared with static weighting schemes, this strategy enables more stable multi-objective optimization and avoids overfitting to either structural correctness or stylistic fidelity, thereby achieving balanced and robust generation performance. **Result** Extensive experiments are conducted on a self-constructed Dongba painting portrait image-text dataset comprising high-quality annotated samples. The proposed method is built upon Stable Diffusion v1.5 and compared with baseline and existing optimization strategies. Both objective and subjective evaluations are performed. For objective metrics, Fréchet Inception Distance (FID, lower values indicate better style consistency) is adopted to measure style consistency, while mean artifact frequency (MAF, lower values indicate fewer structural artifacts) is used to quantify structural artifacts. Experimental results demonstrate that the proposed method significantly improves both aspects. Specifically, the FID score is reduced to 0.1164, indicating closer alignment with the real Dongba painting distribution, while the MAF score decreases to 0.6767, reflecting effective suppression of human anatomical structural artifacts. These correspond to relative improvements of 1.27% and 2.39%, respectively, over the Stable Diffusion v1.5 baseline model. In addition, qualitative comparisons and user studies further validate the effectiveness of the proposed approach. Visual results show that the method successfully eliminates common artifacts such as limb distortion and duplication while preserving essential stylistic elements, including symbolic abstraction, line expressiveness, and color contrast. User preference evaluations indicate that the generated images achieve higher perceptual quality, stronger cultural authenticity, and improved aesthetic appeal compared with competing methods. **Conclusion** This study addresses the challenges of human anatomical artifacts and undesired stylistic degradation, achieving a synergistic improvement in artifact suppression and style preservation in Dongba painting portrait generation. By integrating an anatomical artifact reward, dual-branch style consistency learning, and adaptive weight scheduling into a unified framework, the method produces images characterized by sound anatomical plausibility and highly consistent artistic expression. The proposed approach provides a new and reliable technical pathway for the high-fidelity digital generation of Dongba painting portraits. This advancement overcomes key bottlenecks in image generation and is of significant practical value for the digital preservation, cultural propagation, and living transmission of the Dongba painting heritage.

Key words: Dongba painting; diffusion model; image generation; human anatomical artifact; style consistency; preference alignment

论文引用格式:【DOI:10. 11834/jig. 260168】

0 引 言

图像生成旨在以文本描述、语义标签或图像草图为输入,生成与输入信息匹配的视觉内容。其核心在于对高维视觉数据分布进行建模,并合成新的视觉样本。目前,该技术已广泛应用于艺术创作、电子商务、教育与医疗等领域(Zhang 等,2024)。

东巴画是纳西族文化的核心视觉载体,反映纳西族的古代社会生活,具有独特的笔触纹理与色彩搭配,人物画像栩栩如生(杨杰宏 等,2015)。如图1所示,东巴画以线描为造型基础,色彩艳丽而多变,人像造型写意凝练、神态生动传神,具有鲜明的文化内涵和艺术特征。目前,针对东巴画的研究主要集中于数字展示(杨志琴 等,2024)和图像理解(罗霜等,2025)领域,缺乏对东巴画人像进行图像生成的工作。同时,东巴画现存传世作品数量有限,可用于模型训练的高质量人像样本稀缺,难以支撑大规模生成模型的全量训练,属于典型的低资源生成场景。因此,开展低资源条件下的东巴画人像生成技术研究,不仅能辅助匠人创作,批量生成符合东巴画风格的作品,还能为东巴画活态传承提供技术支撑,具有重要的文化价值与应用意义。

图1 东巴画实例

Fig. 1 Example of Dongba painting

艺术风格生成是图像生成的重要研究方向,也是非遗数字化的核心技术支撑,现有研究主要形成两条成熟技术路线,已在多个经典画派的数字化模拟中得到验证。第一条路线是非真实感渲染(non-photorealistic rendering,NPR)技术,核心是对输入图像进行风格化转换。早期基于笔触渲染、图像类比、图像滤波的图像建模方法,已实现水墨画(毛国红,2006)、线描图(Spicker 等,2015)、油画(Semmo 等,2016)、云南蜡染(张可师,2017)等特定风格的数字化模拟;随着深度学习的发展,神经风格迁移(Gatys 等,2015)、生成对抗网络(Goodfellow 等,2014)等数据驱动方法进一步拓展了艺术风格生成范围,可直接生成莫奈、梵高、浮世绘等多种经典画派风格的图像(Zhu 等,2017)。但传统NPR技术以图像到图像的风格转换为核心路径,不具备基于文本从零生成图像的能力,且多数方法仅能实现颜色与局部纹理的迁移,难以对复杂人体结构与语义布局进行建模;在东巴画样本稀缺的场景下,其泛化能力和适配性受到限制。

第二条路线是文本驱动的图像生成(text-to-image generation,T2I)技术,基于文本描述从零合成完整图像,为艺术人像生成提供了新的技术路径。其中,以稳定扩散(Stable Diffusion,SD)为代表的潜在扩散模型(Rombach 等,2022),利用数十亿图文对构建潜在空间,通过迭代去噪实现高质量图像生成,成为当前图像生成领域的主流方法。然而,将此类通用模型直接应用于特定艺术风格人像生成时,生成结果难以体现目标风格特征,存在明显的领域偏移现象(Kouw 等,2018)。针对这一问题,Hu 等人(2022)提出低秩适配(low-rank adaptation,LoRA)技术,通过在注意力层中注入低秩矩阵实现高效微调,仅需少量目标风格数据即可实现目标风格生成,有效缓解领域偏移问题,为低资源场景的风格适配提供了可行方案。但现有文本驱动的艺术风格生成研究主要面向油画、水彩、动漫等主流艺术风格(Hertz 等,2024),针对东巴画人像的研究仍较少,未能解决风格保真与人体结构合理性的协同优化问题。

随着研究的深入,相关学者发现,传统扩散模型训练方式仅依赖扩散噪声预测损失,难以满足生成质量提升与人体结构伪影抑制的实际需求(Xu 等,2023)。为使生成结果更好地契合人类偏好,生成模型对齐(model alignment)成为研究热点(Ouyang 等,2022; Lamba 等,2025)。以 DDPO(denoising diffusion policy optimization)(Black 等,2023)和 Diffusion-DPO(diffusion direct preference optimization)(Wal-

lace 等,2024)为代表的相关方法,分别从奖励建模、策略优化、直接偏好优化与推理阶段噪声优化等角度入手,针对性实现模型生成质量的综合提升。此外,针对人像结构伪影问题,Wang 等人(2024)提出人体伪影检测模型(human artifact detection models, HADM),并基于大规模人体伪影数据集(human artifact dataset, HAD)实现多生成域的伪影识别;Wang 等人(2025)提出的 DiffDoctor(DiffDoctor: diagnosing image diffusion models before treating)则通过对伪影缺陷进行诊断定位并据此引导去噪过程,有效降低生成结果中的伪影频率。

然而,将上述方法应用于东巴画人像生成任务时存在以下问题:(1)人体结构伪影。东巴画以线描为造型基础,人像造型写意传神,具有鲜明的民族艺术特征。而通用扩散模型的学习先验主要源于真实自然图像,对此类艺术化人体结构的建模能力有限。同时,扩散模型在生成过程中缺乏显式的人体结构约束,难以保证肢体结构的合理性(Wang 等,2024)。在训练数据噪声与跨域风格差异的共同影响下,模型在生成东巴画人像时容易产生肢体畸形、错肢、多肢等人体结构伪影。(2)风格退化。现有伪影抑制方法仅对伪影进行惩罚,缺乏针对东巴画风格特征的显式约束,导致模型在抑制人体结构伪影的过程中,忽略了东巴画独特的手绘笔触与厚重色彩,逐渐偏离东巴画的风格分布,产生色彩暗沉和纹理不一致的风格退化现象。

为此,本文提出融合躯体奖励与风格约束的东巴画人像生成方法,主要贡献在于:(1)设计东巴画躯体伪影奖励函数。构建东巴画生成人像伪影分割数据集,设计东巴画门控伪影分割网络,将人体结构伪影转化为奖励信号,降低人体结构伪影的发生比例。(2)设计东巴画风格约束奖励函数。基于东巴画风格参考数据集构建风格原型库,将东巴画风格的全局语义与局部纹理特征量化为奖励信号,保持抑制伪影过程中东巴画风格的一致性。(3)提出自适应权重调度策略。动态调节风格约束损失与噪声预测损失在优化过程中的作用强度,在保持风格一致性的同时提升人体结构生成的稳定性。

1 本文方法

本文提出一种融合躯体奖励与风格约束的东巴画人像生成方法,减少扩散模型生成东巴画人像的人体结构伪影的同时,保持生成结果与东巴画风格的一致性,方法总体结构如图2所示。

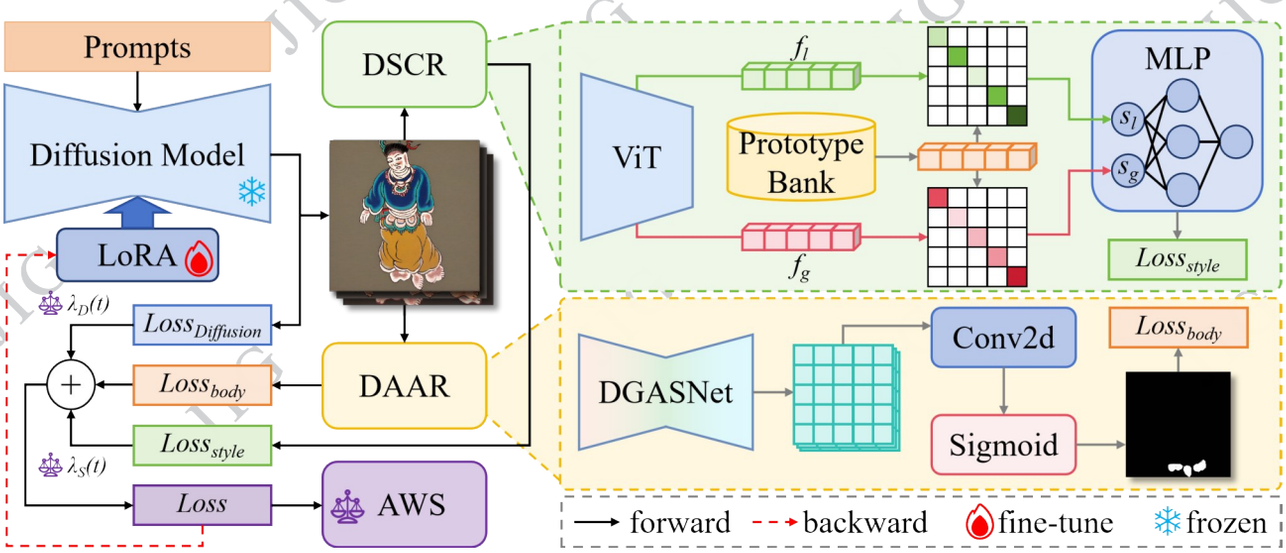


图2 方法总体结构图

Fig. 2 Overall structure diagram of the method

首先,将文本提示(prompts)输入预训练的SD1.5模型,并在其注意力层中引入LoRA模块,实现高效少样本风格适配,兼顾训练效率与模型稳定性。随后,为弥补扩散模型难以显式建模高层语义约束的不足,构建双奖励约束函数,从躯体结构合理性与风格分布一致性两个方面引导生成过程:其中

东巴画躯体伪影奖励函数(Dongba painting anatomical artifact reward，DAAR)通过构建东巴画门控伪影分割网络(Dongba painting gated artifact segmentation network，DGASNet)处理人体结构伪影，通过特征分割与置信度计算构建惩罚信号，显式修正结构伪影；而东巴画风格约束奖励函数(Dongba painting style consistency reward，DSCR)基于CLIP(contrastive language-image pre-training)(Radford等，2021)中的ViT(vision transformer)(Dosovitskiy等，2021)模型提取图像特征，并与预构建的风格原型库(prototype bank)进行对齐，解决抑制伪影引发的风格退化问题，保持东巴画风格一致性。最后，为协调结构约束与风格约束之间的潜在冲突，提出自适应权重调度策略(adaptive weight scheduling，AWS)，根据训练阶段动态调整风格约束损失与扩散预测损失的权重，联合优化总损失函数，实现躯体结构合理性与风格分布一致性的协同提升。

1.1 东巴画躯体伪影奖励函数

针对扩散模型生成东巴画人像时存在的人体结构伪影问题，本文设计东巴画躯体伪影奖励函数，通过构建东巴画门控伪影分割网络(DGASNet)，实现对东巴画人体结构伪影的像素级定位与量化评估。

1.1.1 东巴画门控伪影分割网络(DGASNet)

通用人体结构伪影分割模型难以区分东巴画中合理的艺术表达与真实的结构伪影(Wang等，2025)。为此，本文首先构建东巴画生成人像伪影分割数据集，对多余肢体、肢体畸形、面部模糊等人体结构伪影进行标注。在此基础上，以U-Net(U-net：convolutional networks for biomedical image segmentation)(Ronneberger等，2015)为伪影分割骨干网络，设计DGASNet，如图3所示。

分层编码器(encoder)：用于同时表征笔触纹理等浅层细节信息与肢体结构等深层语义信息。人体结构伪影通常表现为局部边缘断裂、异常纹理及不合理的几何连接，仅依赖单尺度特征易使模型混淆东巴画的艺术化表达与真实结构伪影。因此，本文设计分层编码器，从浅层细节与深层结构两个层面联合建模伪影特征。

在浅层特征提取阶段，构建标准注意力编码块(standard attention encoding block，SAEB)，模块结构如图4所示。双卷积逐步提取局部纹理与边缘信息，提升浅层细节表征的鲁棒性；CBAM(convolutional block attention module)(Woo等，2018)通过通道注意力与空间注意力对特征进行重加权，突出异常纹理和边缘断裂区域的响应，提升浅层伪影特征的辨识能力。

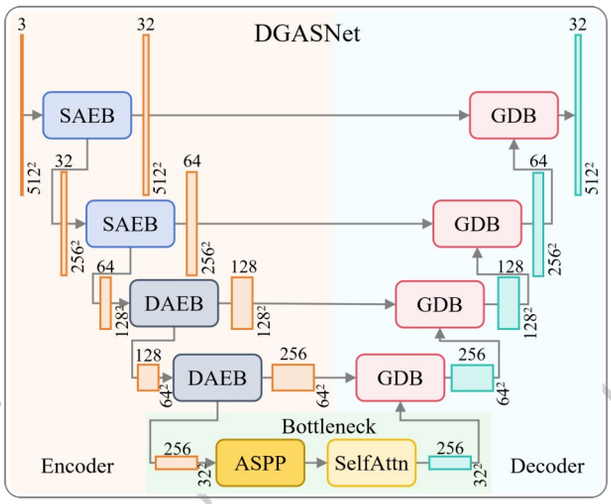


图3 DGASNet结构图

Fig. 3 Structure diagram of DGASNet

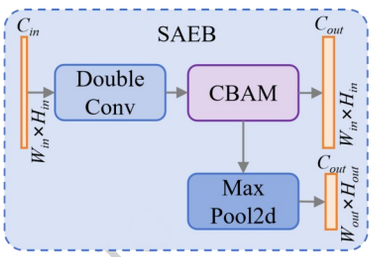


图4 SAEB结构图

Fig. 4 Structure diagram of SAEB

在深层语义建模阶段，设计可变形注意力编码块(deformable attention encoding block，DAEB)，模块结构如图5所示。东巴画人物常呈现艺术形变与线条夸张的特征，固定采样网格难以准确建模复杂的肢体结构。可变形卷积(Dai等，2017)通过学习采样偏移，感受野能够自适应匹配不规则结构，增强对异常结构的感知能力；结合CBAM进一步强化结构伪影区域的响应，提升深层语义特征的判别能力。

多尺度上下文瓶颈层(bottleneck)：东巴画人像的伪影区域常伴随局部结构异常与全局语义不一致，单一尺度卷积无法兼顾局部细节与全局上下文，导致伪影表征不足。为此，本文设计多尺度上下文瓶颈层，在统一特征空间协同建模多尺度信息与全局依赖。首先通过改进ASPP(atrous spatial pyramid pooling)模块(Chen等，2017)，以不同膨胀率的空洞

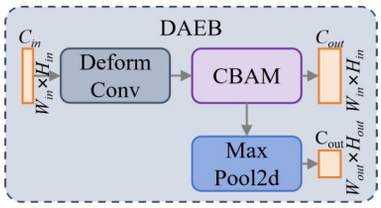


图5　DAEB结构图

Fig. 5　Structure diagram of DAEB

卷积并行提取多尺度上下文特征，提升多尺度伪影感知能力；再引入多头自注意力机制（Vaswani 等，2017）完成全局关系建模，弥补卷积在长距离依赖建模上的不足。该瓶颈层可强化全局语义一致性，最终提升网络对东巴画人像中复杂伪影的表征能力。

门控解码器（decoder）：为实现伪影区域的像素级分割，本文设计由四个门控解码块（gated decoding block，GDB）（如图6所示）组成的门控解码器：

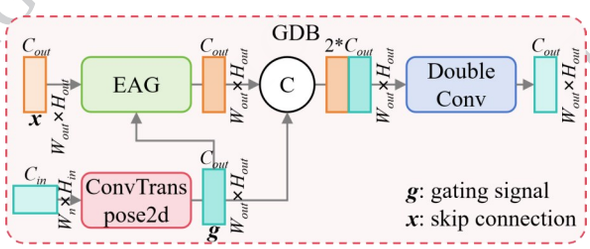


图6　GDB结构图

Fig. 6　Structure diagram of GDB

首先，对来自上一解码块或瓶颈层的特征进行核大小为2、步长为2的转置卷积上采样，使输出特征图尺寸扩大为输入的2倍，得到上采样特征。随后，采用增强注意力门（enhanced attention gate，EAG）机制对跳跃连接特征进行门控筛选，其结构如图7所示。

具体地，通过1×1卷积将门控信号（即上采样特征）和编码器跳跃连接特征映射到中间特征空间，生成空间注意力权重，计算为

$$\alpha_S = \sigma(\psi(\mathrm{ReLU}(\boldsymbol{W}_g * \boldsymbol{g} + \boldsymbol{W}_x * \boldsymbol{x}))) \tag{1}$$

式中，$\boldsymbol{g}$为门控信号，$\boldsymbol{x}$为跳跃连接特征，$\boldsymbol{W}_g$和$\boldsymbol{W}_x$为1×1卷积权重，σ为Sigmoid函数，α_S表示空间位置的重要性。注意力系数生成函数ψ的计算为

$$\psi(z) = \mathrm{BN}(\mathrm{Conv}_{1\times1}(z)) \tag{2}$$

式中，z为输入特征，BN为批归一化层。

之后，对跳跃连接特征施加CBAM以获得增强特征，计算为

$$\boldsymbol{x}_C = \mathrm{CBAM}(\boldsymbol{x}) \tag{3}$$

在EAG完成门控筛选后，将门控注意力与CBAM增强特征进行加权融合，计算为

$$\boldsymbol{x}_F = \boldsymbol{x}_C \odot \alpha_S \tag{4}$$

最后，将筛选后的特征$\boldsymbol{x}_F$与上采样后的解码器特征$\boldsymbol{g}$进行拼接融合，计算为

$$\boldsymbol{F}_D = \mathrm{DoubleConv}(\mathrm{Concat}[\boldsymbol{x}_F, \boldsymbol{g}]) \tag{5}$$

通过上述DGASNet的设计，可实现对东巴画人像人体结构伪影区域的像素级分割。

1.1.2　躯体奖励损失

为将DGASNet的像素级伪影分割能力转化为可微奖励信号，引导生成模型减少人体结构伪影，DAAR利用DGASNet解码器输出的多通道特征图，经由1×1卷积映射为单通道对数几率（logits），计算为

$$logits = \mathrm{Conv}_{1\times1}(\boldsymbol{F}_F) \tag{6}$$

式中，1×1卷积将C维特征压缩为1维，$\boldsymbol{F}_F$为多通道特征图。随后，经Sigmoid激活函数得到像素级伪影分割概率图，计算为

$$\boldsymbol{M} = \sigma(logits) \tag{7}$$

式中，σ为Sigmoid函数，$\boldsymbol{M} \in [0,1]^{H\times W}$。

本文将躯体奖励损失定义为整幅伪影概率图的平均值，计算为

$$L_A = \frac{1}{H \times W}\sum_{i,j}\boldsymbol{M}(i,j) \tag{8}$$

式中，H和W分别表示伪影概率图的高度与宽度，$\boldsymbol{M}(i,j)$表示位置(i,j)属于伪影区域的置信度。

DAAR通过对伪影区域进行像素级惩罚，引导生成模型在潜空间中远离易导致人体结构伪影的生成路径，从而减少人体结构伪影的发生。

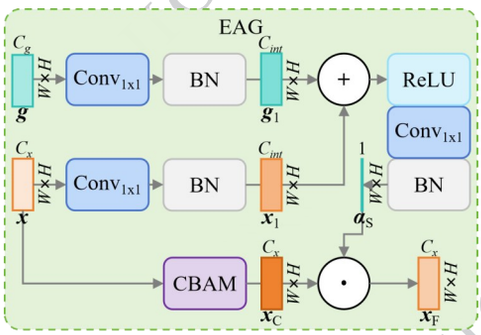


图7　EAG机制结构图

Fig. 7　Structure diagram of the EAG mechanism

1.2 东巴画风格约束奖励函数

针对扩散模型缺乏东巴画风格约束所引发的风格退化问题，本文提出东巴画风格约束奖励函数，确保生成图像在全局语义与局部纹理层面均能保持东巴画的艺术风格，方法框架如图8所示。该方法通过构建基于风格原型库的评估网络，结合CLIP模型的语义表征能力，从全局与局部两个层面对生成图像的东巴画风格进行量化评估。

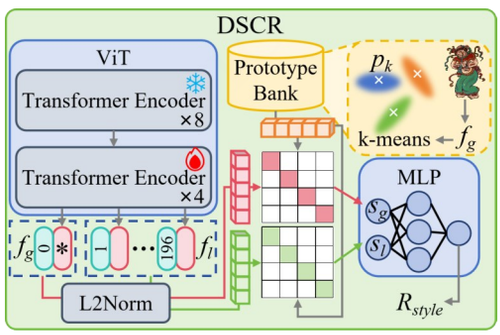


图8 DSCR方法框架图

Fig. 8 Framework diagram of the DSCR method

1.2.1 风格原型网络架构

为有效量化生成图像与东巴画风格的契合程度，同时兼顾全局语义与局部纹理的一致性，本文提出基于原型匹配的风格度量网络，主要由CLIP特征编码器、风格原型匹配机制和风格判别器三部分组成。

CLIP特征编码器：采用预训练CLIP中的ViT模型提取图像的语义视觉特征。首先对生成图像进行预处理，并输入ViT模型，提取模型最后一层特征序列并对其进行欧几里得范数归一化；随后将特征分解为分类token对应的全局风格特征与各patch的局部纹理特征，分别用于全局风格对齐与局部纹理一致性建模。

风格原型匹配机制：为精准捕捉东巴画独特的艺术风格，本文基于东巴画人像风格参考数据集构建可学习的风格原型库

$$\boldsymbol{P}=\{\boldsymbol{p}_k\}_{k=1}^{K} \tag{9}$$

式中，K为原型数量，$\boldsymbol{p}_k$为第k个风格原型中心。网络通过双分支路径计算生成图像与原型库的相似度。其中，全局风格对齐分数采用最大余弦相似度衡量图像整体与东巴画风格的契合程度，计算为

$$s_g=\max_{k=1}^{K}(\boldsymbol{f}_g\cdot\boldsymbol{p}_k) \tag{10}$$

式中，$\boldsymbol{f}_g$为ViT提取的全局特征，符号$(\cdot)$表示向量点积。最大池化操作使模型只需匹配原型库中最相近的一种风格原型，在保证风格一致性的同时保留生成结果的多样性。局部纹理一致性分数则通过计算各patch与原型库之间的平均最大匹配度，对图像的笔触和纹理特征评估，计算为

$$s_l=\frac{1}{N}\sum_{i=1}^{N}\max_{k=1}^{K}(\boldsymbol{f}_l^i\cdot\boldsymbol{p}_k) \tag{11}$$

式中，$\boldsymbol{f}_l$为ViT提取的局部特征。该分数能够约束图像各局部区域的纹理特征与原型库中的东巴画风格特征保持一致。

风格判别器：为融合全局风格对齐信息与局部纹理一致性信息，实现对东巴画风格的综合评估，本文将全局分数与局部分数拼接为特征向量$\boldsymbol{v}=[s_g,s_l]$，并输入MLP(multilayer perceptron)进行加权融合和奖励计算，计算为

$$R_s=\sigma(\boldsymbol{W}_4\cdot\phi(\boldsymbol{W}_3\cdot\phi(\boldsymbol{W}_2\cdot\phi(\boldsymbol{W}_1\boldsymbol{v}+b_1)+b_2)+b_3)+b_4) \tag{12}$$

式中，$\boldsymbol{W}_1\in\mathbf{R}^{128\times2}$，$\boldsymbol{W}_2\in\mathbf{R}^{64\times128}$，$\boldsymbol{W}_3\in\mathbf{R}^{32\times64}$，$\boldsymbol{W}_4\in\mathbf{R}^{1\times32}$为各层权重矩阵，$\sigma$为Sigmoid激活函数，$b_1,b_2,b_3,b_4$为对应偏置项，$\varphi$为ReLU非线性激活函数。最终输出$R_S\in[0,1]$为东巴画风格约束奖励，其值越接近1，表示生成图像的风格越符合东巴画特征。

1.2.2 风格约束损失

为最大化生成图像的风格契合度，本文将风格约束损失定义为奖励分数的负对数形式，以此引导模型向高分区域优化，计算为

$$L_S=-\log(R_S(\boldsymbol{I}_C)) \tag{13}$$

式中，R_S为东巴画风格约束奖励，$\boldsymbol{I}_C$为生成图像。该损失函数通过梯度回传推动生成模型的潜空间分布向东巴画风格分布靠拢，促使模型在抑制伪影过程中保持东巴画的风格一致性。

1.3 自适应权重调度策略

在训练过程中，躯体奖励损失、风格约束损失与噪声预测损失在不同阶段具有不同的优化需求。若采用静态权重，各分量的重要性在整个优化过程中固定不变，易导致阶段性优化失衡。为此，本文提出自适应权重调度策略：每隔固定训练步数T_V在验证集上进行一次模型评估，并依据性能变化趋势动态更新各分量权重；在相邻两次验证之间，权重保持不变，作为公式(17)中$\lambda_S(t)$和$\lambda_D(t)$的取值。

首先，在每个验证步t($t=T_V,2T_V,3T_V,\cdots$)计算近

期性能变化趋势，计算为

$$\begin{cases}\Delta_S(t)=R_S(t)-R_S(t-T_V)\\ \Delta_B(t)=R_B(t)-R_B(t-T_V)\end{cases}\tag{14}$$

式中，$R_S(t)$和$R_B(t)$分别为验证步t时DSCR与DAAR输出的平均风格奖励和躯体奖励，$\Delta_S(t)$和$\Delta_B(t)$分别为风格和躯体奖励的变化趋势。

当$\Delta_B(t)<-\delta$(δ表示恶化阈值，且$\delta>0$)时，说明当前人体结构伪影有加剧趋势，此时通过乘性衰减同时降低风格约束损失权重与噪声预测损失权重，使优化过程优先关注人体结构合理性，计算为

$$\begin{cases}\lambda_S(t)=\lambda_S(t-T_V)\cdot(1-\alpha)\\ \lambda_D(t)=\lambda_D(t-T_V)\cdot(1-\alpha)\end{cases}\tag{15}$$

式中，α为调整步长，$\lambda_S(t)$和$\lambda_D(t)$分别为随验证步t动态更新的风格约束损失权重和噪声预测损失权重。若伪影未加剧，但风格奖励出现退化或停滞，即$\Delta_S(t)\leqslant 0$，则适当提升风格约束权重，增强模型对东巴画风格特征的保持能力，计算为

$$\lambda_S(t)=\lambda_S(t-T_V)\cdot(1+\alpha)\tag{16}$$

通过上述策略，模型能够依据训练状态自适应调节风格约束损失与噪声预测损失的作用强度，在保持东巴画风格一致性的同时提升人体结构生成的稳定性。

1.4 总损失函数

为在东巴画人像生成中兼顾伪影抑制、风格保持与生成稳定性，本文将各项损失整合为端到端的优化目标，并引入AWS动态平衡风格约束损失与噪声预测损失的分量权重，总损失计算为

$$L_T=L_A+\lambda_S(t)L_S+\lambda_D(t)L_D\tag{17}$$

式中，L_T为总损失，L_A为躯体奖励损失，L_S为风格约束损失，L_D为扩散模型的噪声预测损失，$\lambda_S(t)$和$\lambda_D(t)$分别为训练过程中自适应更新的风格约束损失权重和噪声预测损失权重。

2 实验结果与分析

2.1 实验数据

为实现东巴画风格人物图像的高质量生成，本文构建东巴画人像图文数据集。从丽江市博物馆馆藏实物及相关文献采集高清扫描图像。针对人物与环境交叠的问题，本文采用人工标注分割提取人物区域，共获得93幅高质量、无遮挡的东巴画人像样本，如图9所示。由熟悉纳西东巴文化的专家为每幅图像撰写提示词，主要描述动作、服饰及色彩等关键信息，最终构建可用于生成模型训练与评估的东巴画人像图文数据集，示例如图10所示。

(a) 原始扫描件 (b) 透明背景人像样本

((a) original scan copy; (b) transparent-background portrait)

图9 东巴画人像数据集提取示例

Fig. 9 Example of Dongba portrait painting dataset extracting

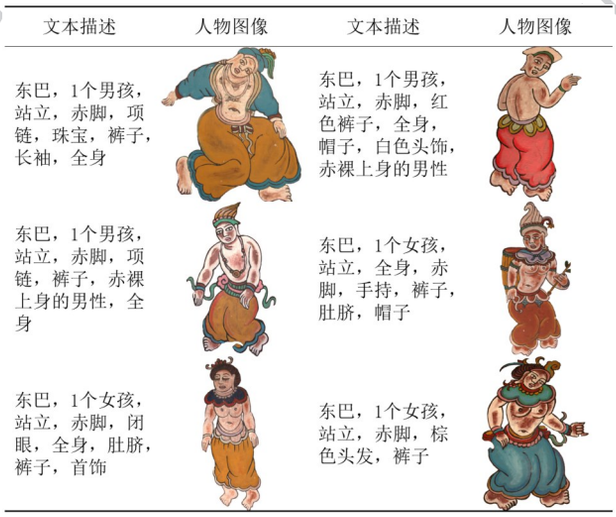


图10 东巴画人像图文数据集示例

Fig. 10 Example of Dongba painting portrait image-text dataset

针对人体结构伪影抑制与风格约束两类目标，本文构建功能化数据集：(一)东巴画生成人像伪影分割数据集。为训练DGASNet，借助SAM(segment anything model)(Kirillov 等，2023)对人体伪影区域进行像素级标注，构建包含1000幅东巴画生成人像样本的数据集，覆盖正常结构与多类结构伪影，如图11所示。(二)风格参考数据集。为构建DSCR风格原型库，选取62幅代表性东巴画人像作为风格参考样本，覆盖典型东巴画色彩特征与线条笔触，示例样本如图12所示。

2.2 实验设置

2.2.1 详细设置

实验基于PyTorch深度学习框架，在单张24GB NVIDIA GeForce RTX 3090显卡上实现，以SD1.5为

(a) 生成图像 (b)伪影区域标注图像
((a) generated image;(b) image with artifact area annotation)
图11　东巴画生成人像伪影分割数据集示例
Fig. 11　Example of Dongba painting generated portrait artifact segmentation dataset

图12　风格参考数据集示例
Fig. 12　Example of the style reference dataset

基础模型，引入LoRA进行参数微调。训练与推理分辨率均设为512×512，优化器采用AdamW，初始学习率设为1e-7。DAAR中U-Net的通道数设为[32,64,128,256]。DSCR采用CLIP ViT-B/16编码器，特征维度D为512，风格原型数量K为16。AWS中，权重调整步长α设为0.1。数据集按照8:1:1划分为训练集、验证集和测试集。

2.2.2　低资源场景优化策略

东巴画公开可获取的人像样本数量极其有限，若直接对预训练扩散模型进行全量微调，在小样本条件下极易出现过拟合和风格崩溃问题。针对这一低资源场景，本文采用以下策略提升模型训练稳定性与数据利用效率：(1)采用基于大规模图文数据预训练的Stable Diffusion v1.5作为生成基础模型，充分继承通用视觉语义知识；(2)采用LoRA参数高效微调策略，仅优化注意力层中的低秩参数矩阵，在保留预训练知识的同时降低模型参数更新规模，减少过拟合风险；(3)利用DAAR与DSCR构建显式高层语义约束，通过奖励信号弥补低资源条件下监督信息不足的问题，引导模型向结构合理且风格一致的方向优化；(4)采用风格原型库对有限东巴画样本进行统计建模，以原型匹配方式增强模型对东巴画风格分布的刻画能力，提高小样本条件下的风格保持能力。

2.3　评价指标

2.3.1　风格分布一致性指标

弗雷歇初始距离(Fréchet inception distance, FID)(Heusel等,2017)：在风格特征空间中计算，衡量生成分布与真实分布差异，计算为

$$\mathrm{FID}=\left\|\mu_r-\mu_g\right\|^2+\mathrm{Tr}\left(\sigma_r+\sigma_g-2\sqrt{\sigma_r\sigma_g}\right) \tag{18}$$

式中，μ_r、σ_r和μ_g、σ_g分别为参考集与生成集在风格特征空间中的均值与协方差。

核初始距离(kernel inception distance, KID)(Bińkowski等,2018)：基于最大均值差异(maximum mean discrepancy, MMD)，在小样本场景下更稳定，衡量生成分布与真实分布差异，计算为

$$\mathrm{KID}=\mathrm{MMD}_u^2\left(\mathrm{F}_\theta(P_\mathrm{R}),\mathrm{F}_\theta(P_\mathrm{G})\right) \tag{19}$$

式中，F_θ为Inception网络(Szegedy等,2015)的特征提取函数，θ为其网络参数，P_R和P_G分别表示真实分布与生成分布。

能量距离(energy distance, E-Dist)：在风格特征空间中计算生成分布与参考分布之间的能量距离(Székely等,2013)，评估两者的分布接近程度。

2.3.2　躯体结构指标

平均伪影频率(mean artifact frequency, MAF)：采用DiffDoctor(Wang等,2025)中使用的评价指标，根据图像最大伪影置信度阈值(0.5)判断图像是否包含伪影，并据此统计测试集中含有明显人体结构伪影的图像比例，计算为

$$\mathrm{MAF}=\frac{\sum_\mathrm{A}}{\sum_\mathrm{T}} \tag{20}$$

式中，Σ_A表示超过阈值的图像数量，Σ_T表示测试集图像总数。

2.4　客观质量评价

东巴画人像与自然人像在躯体比例与形态上存在差异，若仅对比躯体结构指标，则可能忽略风格一致性的核心约束。因此，本文以测试集中的文本提示词为输入，先从风格一致性方面选取基础扩散模型(SD1.5, SDXL)及通用领域的偏好对齐方法(DDPO, DiffusionDPO, DiffDoctor等)进行比较，再从躯体合理性方面选取伪影修复方法DiffDoctor和原始模型SD1.5进行对比分析。

表1结果表明，所提方法在FID、KID与E-Dist三项指标上均取得最优表现，相较于基线模型SD1.5，三项指标分别降低1.27%、0.11%和1.21%，整体优

表1　本文方法与其他方法的风格一致性指标对比

Table 1　Comparison of metrics of style consistency between the method in this paper and other methods

Method	FID ↓	KID ↓	E-Dist ↓
SD1.5(Rombach 等,2022)	<u>0.1179</u>	<u>0.0870</u>	<u>0.3299</u>
SDXL(Podell 等,2024)	0.1539	0.2806	0.4197
DDPO(Black 等,2023)	1.6025	0.2350	1.8157
DiffusionDPO(Wallace 等,2024)	2.3160	0.3424	2.6414
ReNO(Eyring 等,2024)	2.3743	0.3327	2.6853
DiffDoctor(Wang 等,2025)	0.4768	0.2118	0.7544
Ours	**0.1164**	**0.0869**	**0.3259**

注:加粗字体为每列最优值,下划线字体为次优值。表中各方法所用扩散模型均经过东巴画人像图文数据集微调。

于SD1.5及多种偏好对齐方法。尽管DDPO、DiffusionDPO、ReNO等方法在通用领域图像生成中可通过偏好对齐提升质量,但在东巴画风格生成任务中仍存在较大风格分布偏差,指标表现甚至劣于基线模型。而所提方法通过针对性的东巴画风格约束,能在优化中更稳定地保持风格分布、缓解风格退化,使生成结果在特征空间更贴近真实东巴画分布,验证了其在东巴画风格保持上的有效性。

表2　本文方法与其他方法的躯体合理性指标对比

Table 2　Comparison of metrics of anatomical plausibility between the method in this paper and other methods

Method	MAF ↓
SD1.5(Rombach 等,2022)	0.6933
DiffDoctor(Wang 等,2025)	**0.2100**
Ours	<u>0.6767</u>

注:加粗字体为每列最优值,下划线字体为次优值。表中各方法所用扩散模型均经过东巴画人像图文数据集微调。

表1与表2结果表明,所提方法在保持风格一致性指标最优的同时,将MAF由0.6933降至0.6767,相对降低2.39%,可在不破坏东巴画风格分布的情况下,有效降低生成人像的人体结构伪影率,提升结构合理性。虽然DiffDoctor在MAF上数值更低,但其风格一致性指标劣于所提方法,表明其在抑制结构伪影时难以维持稳定的东巴画风格分布。

2.5　主观质量评价

本文从人类视觉感知层面对各方法的生成效果开展主观视觉对比实验,结果如图13所示。

对比结果表明,SD1.5和SDXL虽能较好地保留东巴画的整体视觉风格,但易产生明显的人体结构伪影。例如,如图13第二行所示,SD1.5生成结果中出现多余手臂、脚趾的伪影,SDXL则出现多余腿部结构。相比之下,DiffDoctor、DiffusionDPO及ReNO等方法在一定程度上降低了人体结构伪影率,但同时伴随不同程度的风格退化。具体来看,DiffusionDPO的生成结果明显偏向写实摄影风格,与东巴画的艺术表现形式存在较大差异;DiffDoctor在结构形态上更接近东巴画人像,却呈现色彩变暗与纹理不一致的现象,并且部分样本通过仅生成人物上半身来降低伪影率,影响人像的完整性。相较而言,所提方法生成的人物肢体结构更加完整,未出现明显结构伪影,同时能够较好地保持东巴画的人物造型特征与艺术风格。

为从统计层面验证上述观察结果,本文邀请20名志愿者开展人类偏好实验。志愿者在观察不同方法生成的图像后,分别从风格偏好、结构偏好和综合偏好方面选择较优的多张图像,结果如图14所示。实验表明,所提方法在风格偏好与综合偏好上排名第一,在结构偏好上位列第二,说明其在保持东巴画风格一致性的同时兼顾了人体结构合理性,整体生成效果更符合人类视觉偏好。

2.6　消融实验分析

为验证DAAR、DSCR与AWS对东巴画人像生成任务的有效性,本文构建了不同模块组合的消融模型。

表3中的消融实验结果揭示了各模块在模型优化过程中的具体作用机制。当去除DSCR时,风格相关指标整体出现退化,说明该模块在约束生成结果的东巴画风格分布方面发挥了关键作用;当去除DAAR时,MAF指标明显上升,表明人体结构伪影抑制能力显著下降,说明该模块在减少人体结构伪影方面具有重要作用;在同时保留DSCR与DAAR但去除AWS时,模型在风格指标上相较完整模型有所退化,说明AWS通过动态权重调节能够增强多目标优化过程中的协调性,从而提升整体优化效果。综合来看,完整模型在FID和MAF指标上最优,其余指标也保持较强竞争力。与不包含任何模块的基线模型相比,完整模型在KID、FID、E-Dist及MAF指标上均取得不同程度的提升,验证了各模块设计的有

效性。

图13　各方法视觉质量比较

Fig. 13　Comparison of visual quality of various methods

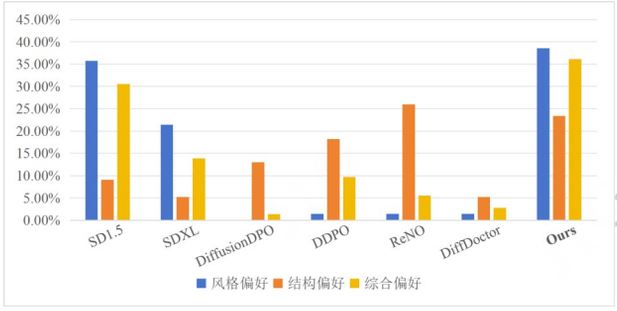


图14　各方法人类偏好比较

Fig. 14　Comparison of human preferences among various methods

表3　消融实验指标对比

Table 3　Comparison of metrics in ablation experiments

A	B	C	KID ↓	FID ↓	E-Dist ↓	MAF ↓
×	×	×	0.0870	0.1179	0.3299	0.6933
√	√	×	0.0893	0.1221	0.3356	0.6915
√	×	√	**0.0868**	<u>0.1165</u>	**0.3215**	0.7000
×	√	√	0.0885	0.1194	0.3324	<u>0.6800</u>
√	√	√	<u>0.0869</u>	**0.1164**	<u>0.3259</u>	**0.6767**

注：加粗字体为每列最优值，下划线字体为次优值。A表示DSCR，B表示DAAR，C表示AWS。

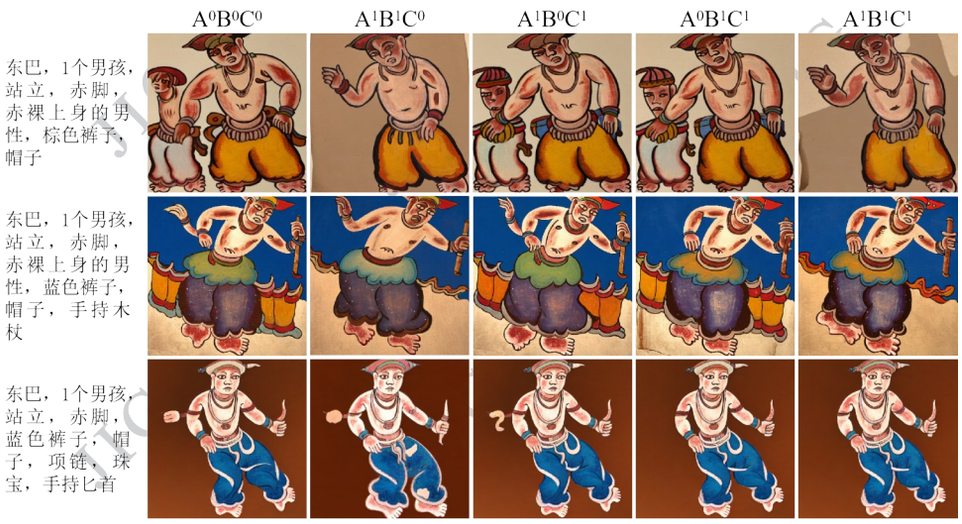
图15的视觉对比结果与表3的客观指标分析结论一致，基线模型（$A^0B^0C^0$）存在明显的人体结构伪影；消融DSCR（$A^0B^1C^1$）时，整体风格无明显退化，但结构优化效果弱于完整模型；消融DAAR（$A^1B^0C^1$）时，风格保持良好但未能消除结构伪影；消融AWS（$A^1B^1C^0$）时，出现明显的风格退化，且仍存在部分结构伪影；仅完整模型在有效减少人体结构伪影的同时保持东巴画风格一致性，在两大核心目标间实现最优均衡。

2.7　可视化分析

为进一步验证所提方法的有效性，本文对优化后的模型进行可视化分析。

图16展示了不同方法在相同提示词条件下生成结果的伪影热力图，其中热力图区域表示各方法输出的伪影置信度。可以观察到，DiffDoctor与所提方法均能在一定程度上抑制人体结构伪影，但DiffDoctor的生成结果出现了明显的风格退化现象，所提方法在抑制伪影的同时仍能保持东巴画风格的一致性。

图17展示了DSCR模块提取的风格特征在二维空间中的分布情况。通过UMAP（uniform manifold approximation and projection）（McInnes 等，2018）对



A表示DSCR，B表示DAAR，C表示AWS。上标⁰表示消融对应模块，上标¹表示保留对应模块。

图15　消融实验视觉质量比较

Fig. 15　Comparison of visual quality in ablation experiments

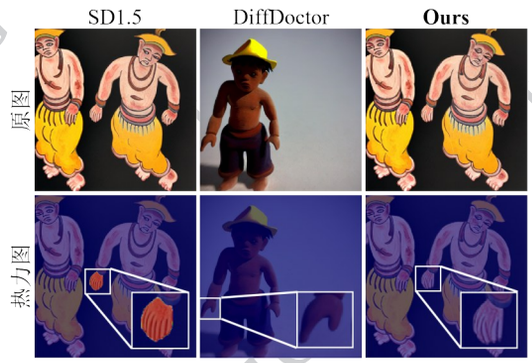


图16　伪影热力图

Fig. 16　Artifact heatmap

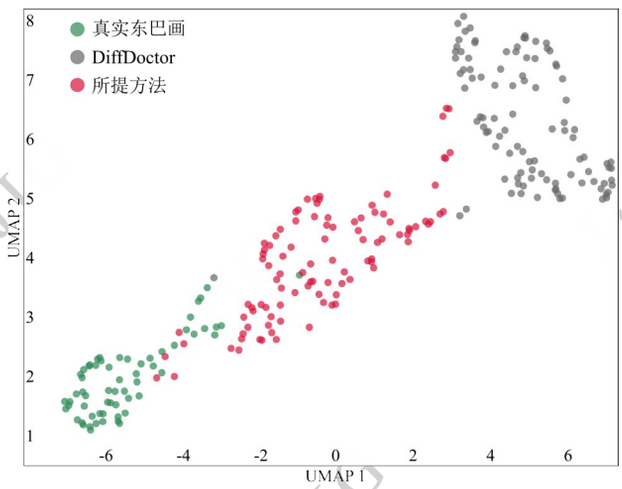


图17　风格特征分布UMAP降维可视化

Fig. 17　UMAP dimensionality reduction visualization of style feature distribution

风格特征进行降维可视化，可见所提方法生成图像的特征点更紧密地聚集在真实东巴画特征簇附近；相比之下，DiffDoctor生成结果与真实东巴画特征簇存在明显偏移，表现出一定程度的风格退化现象。该结果进一步验证了DSCR在约束生成结果保持东巴画艺术风格一致性方面的有效性。

3　结　论

本文提出一种融合躯体奖励与风格约束的东巴画人像生成方法，通过设计东巴画躯体伪影奖励函数，有效降低了生成东巴画人像的平均伪影率；设计东巴画风格约束奖励函数，有效缓解了抑制伪影过程中的风格退化问题，保持了东巴画的风格一致性；提出自适应权重调度策略，实现躯体合理性与风格一致性的协同优化。在本文构建的东巴画人像图文数据集上的实验结果表明，相较于基线方法，所提方法在风格一致性与躯体合理性方面均取得更优表现，并在主观视觉质量评估与用户偏好测试中获得更高认可度，验证了所提方法在东巴画人像生成任务中的有效性。

然而，本文方法仍存在一定局限。首先，东巴画人像中手部与服饰纹样存在密集交叠的视觉特征，模型对该区域的结构和纹理特征区分能力不足，易出现手部轮廓与装饰纹理边界模糊的问题。其次，受限于当前伪影标注数据的规模与精度，模型对细

粒度结构约束的学习仍有提升空间。未来工作将引入人体结构先验与跨模态语义引导，结合自动化标注技术扩展训练数据集，进一步提升复杂场景下的生成质量，并探索方法在更多民族艺术生成任务中的泛化能力。

参考文献（References）

Bińkowski M, Sutherland D J, Arbel M and Gretton A. 2018. Demystifying mmd gans// International Conference on Learning Representations (ICLR). Vancouver, Canada. [DOI: 10.48550/arXiv. 1801.01401]

Black K, Janner M, Du Y, Kostrikov I and Levine S. 2023. Training diffusion models with reinforcement learning[EB/OL]. [2026-2-22]. https://arxiv.org/pdf/2305.13301

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 40(4): 834-848. [DOI:10.1109/TPAMI.2017.2699184]

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H, et al. 2017. Deformable convolutional networks// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE Xplore: 764-773. [DOI:10.1109/ICCV.2017.89]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale// International Conference on Learning Representations (ICLR). Virtual-only (Global). [DOI: 10.48550/arXiv.2010.11929]

Eyring L, Karthik S, Roth K, Dosovitskiy A and Akata Z. 2024. Reno: Enhancing one-step text-to-image models through reward-based noise optimization// Advances in Neural Information Processing Systems (NIPS). Vancouver, Canada: NIPS′24: 125487-125519. [DOI:10.48550/arXiv.2406.04312]

Gatys L A, Ecker A S and Bethge M. 2015. A neural algorithm of artistic style. Journal of Vision. [DOI:10.1167/16.12.326]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets// Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: NIPS′14: Vol. 2: 2672 - 2680. [DOI:10.48550/arXiv.1406.2661]

Hertz A, Voynov A, Fruchter S and Cohen-Or D. 2024. Style aligned image generation via shared attention// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Xplore: 4775-4785. [DOI:10.1109/CVPR52733.2024.00457]

Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium// Advances in Neural Information Processing Systems (NIPS). Long Beach, USA: NIPS′17. [DOI: 10.48550/arXiv.1706.08500]

Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. 2022. Lora: Low-rank adaptation of large language models// International Conference on Learning Representations (ICLR). Virtual-only (Global): iclr: 1(2): 3. [DOI:10.48550/arXiv.2106.09685]

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. 2023. Segment anything// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Xplore: 4015-4026. [DOI: 10.1109/ICCV51070.2023.00371]

Kouw W M and Loog M. 2018. An introduction to domain adaptation and transfer learning[EB/OL]. [2026-2-22]. https://arxiv.org/pdf/1812.11806

Lamba P, Ravish K, Kushwaha A and Kumar P. 2025. Alignment and safety of diffusion models via reinforcement learning and reward modeling: A survey[EB/OL]. [2026-2-22]. https://arxiv.org/pdf/2505.17352

Luo S, Qian W H and Liu P. 2025. Fusion of prompt learning and visual semantic generation for image captioning of dongba paintings. Journal of Image and Graphics, 30(10): 3361-3376 (罗霜, 钱文华, 刘朋. 2025. 结合提示学习和视觉语义生成融合的东巴画图像描述. 中国图象图形学报, 30(10): 3361-3376) [DOI:10.11834/jig.240601]

Mao G. 2006. Research on an image-based ink wash painting rendering method. Wuhan: China University of Geosciences (毛国红. 2006. 一种基于图像的水墨风格绘制方法研究. 武汉: 中国地质大学)

McInnes L, Healy J, Saul N and Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29): 861. [DOI:10.21105/joss.00861]

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. 2022. Training language models to follow instructions with human feedback// Advances in Neural Information Processing Systems (NIPS). New Orleans, USA: NIPS′22: 27730-27744. [DOI: 10.48550/arXiv.2203.02155]

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis// International Conference on Learning Representations (ICLR). Vienna, Austria. [DOI: 10.48550/arXiv.2307.01952]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision// International Conference on Machine Learning (ICML). Vienna, Austria: PmLR 139: 8748-8763.

Rombach R, Blattmann A, Lorenz D, Esser P, and Ommer B. 2022. High-resolution image synthesis with latent diffusion models// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Xplore:

10684-10695.［DOI:10.1109/CVPR52688.2022.01042］

Ronneberger O, Fischer P and Brox T. 2015. U-net: convolutional networks for biomedical image segmentation// International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich, Germany: Springer: 234-241.［DOI: 10.1007/978-3-319-24574-4_28］

Semmo A, Limberger D, Kyprianidis J E and Dollner J. 2016. Image stylization by interactive oil paint filtering. Computers and Graphics, 55:157-171［DOI:10.1016/j.cag.2015.12.001］

Spicker M, Kratt J, Arellano D and Deussen O. 2015. Depth-aware coherent line drawings// Proceedings of SIGGRAPH Asia 2015 Technical Briefs. New York, USA: ACM: 1-5［DOI: 10.1145/2820903.2820909］

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. 2015. Going deeper with convolutions// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE Xplore: 1-9.［DOI: 10.1109/CVPR. 2015. 7298594］

Székely G J and Rizzo M L. 2013. Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference, 143(8): 1249-1272.［DOI:10.1016/j.jspi.2013.03.018］

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A D, et al. 2017. Attention is all you need// Advances in Neural Information Processing Systems (NIPS). Long Beach, USA: NIPS'17.［DOI:10.48550/arXiv.1706.03762］

Wallace B, Dang M, Rafailov R, Zhou L, Lou A, Purushwalkam S, et al. 2024. Diffusion model alignment using direct preference optimization// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Xplore: 8228-8238.［DOI:10.1109/CVPR52733.2024.00786］

Wang K, Zhang L and Zhang J. 2024. Detecting Human Artifacts from Text-to-Image Models［EB/OL］.［2026-2-22］. https://arxiv.org/pdf/2411.13842

Wang Y, Chen X, Xu X, Ji S, Liu Y, Shen Y, et al. 2025. DiffDoctor: Diagnosing Image Diffusion Models Before Treating// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, United States: IEEE Xplore: 18917-18926.［DOI:10.48550/arXiv.2501.12382］

Woo S, Park J, Lee J Y and Kweon I S. 2018. Cbam: Convolutional block attention module// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19.［DOI:10.1007/978-3-030-01234-2_1］

Xu J, Liu X, Wu Y, Tong Y, Li Q, Ding M, et al. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation// Advances in Neural Information Processing Systems (NIPS). New Orleans, USA: NIPS'23: 15903-15935.［DOI: 10.48550/arXiv.2304.0597］

Yang J H. 2015. A study on the stylized characteristics of dongba paintings. Journal of Art College of Inner Mongolia University, 12(2):8 (杨杰宏. 2015. 东巴画的程式化特征研究. 内蒙古大学艺术学院学报, 12(2):8［DOI:10.3969/j.issn.1672-9838.2015.02.017］

Yang Z Q. 2024. Design and implementation of Dongba painting digital exhibition platform based on Django + Ajax. Modern Information Technology, 8(19):109-114 (杨志琴. 基于Django+Ajax的东巴画数字展示平台设计与实现. 现代信息科技, 8(19):109-114)［DOI:10.19850/j.cnki.2096-4706.2024.19.021］

Zhang K S. 2017. Research on Interactive Digital Synthesis Technology of Yunnan Batik Painting. Kunming: Yunnan University (张可师. 2017. 交互式云南蜡染画数字合成技术研究. 昆明: 云南大学)

Zhang N and Tang H. 2024. Text-to-image synthesis: a decade survey［EB/OL］.［2026-2-22］. https://arxiv.org/pdf/2411.16164

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE Xplore: 2223-2232.［DOI:10.1007/978-3-030-58545-7_46］

作者简介

白宇帆，男，硕士研究生，主要研究方向为计算机视觉和文化计算。E-mail:yufanbai@stu.ynu.edu.cn

钱文华，通讯作者，男，教授，博士生导师，CCF高级会员(38886D)，主要研究方向为计算机视觉、文化计算等。E-mail:whqian@ynu.edu.cn

杨荣懿，女，硕士研究生，主要研究方向为深度学习和计算机视觉。E-mail:12024115003@stu.ynu.edu.cn