SVGDreamer：利用扩散模型生成文本引导的 SVG

摘要

最近，文本引导的可缩放矢量图形（SVG）在图像学和素描等领域展现了广阔的前景。然而，现有的文本到 SVG 生成方法在可编辑性方面存在不足，并且难以解决视觉质量和结果多样性等问题。为了解决这些局限性，我们提出了一种名为 SVGDreamer 的新型文本引导矢量图形合成方法。SVGDreamer 采用了一种语义驱动的图像矢量化（SIVE）过程，将合成任务分解为前景对象和背景，从而增强了生成图形的可编辑性。具体而言，SIVE 过程引入了基于注意力的基元控制和注意力掩码损失函数，能够有效地控制和操作单个元素。此外，本文提出了一种基于矢量粒子分数蒸馏（VPSD）的方法，通过将 SVG 视为由控制点和颜色分布组成的模型，解决了现有文本到 SVG 生成方法中的形状过度平滑、颜色饱和度过高、多样性有限以及收敛速度慢等问题。VPSD 方法还利用奖励模型对矢量粒子进行重新加权，从而提升了生成图形的审美吸引力并加速了收敛过程。本文进行了大量实验验证，结果表明，SVGDreamer 在可编辑性、视觉质量和多样性等方面均优于现有的基准方法。

关键词：SIVE；VPSD ； 文本引导

1 引言

1.1 选题背景

矢量图形（SVG）通过使用几何原语（如贝塞尔曲线、多边形、直线等）表示视觉内容，这种方式适合高质量的视觉设计，如标志、海报、插图等。与光栅图像（例如 JPEG、PNG 格式的图像）相比，矢量图像在存储和传输过程中具有显著的优势，主要表现为文件体积小，且由于其基于数学公式的特点，缩放时不会失真，这使得它在高分辨率设计、印刷以及网页设计中尤其重要。更为关键的是，矢量图像具有更高的编辑性，可以轻松地选择、修改和组合各个元素。这一特性为设计师提供了极大的灵活性，可以在设计过程中不断调整和优化细节，从而实现创意的自由发挥。然而，尽管 SVG 在视觉设计中的优势显而易见，当前的矢量图形生成方法依然面临不少挑战。尤其是近年来，文本到图像（Text-to-Image，T2I）生成技术 [17,24,27,29] 的飞速发展，推动了矢量图形生成领域的研究。例如，CLIPDraw [2]、VectorFusion [8] 等方法采用了 T2I 模型与扩散模型相结合的策略，通过光栅图像作为生成目标来优化矢量图形的参数。然而，尽管这些方法在生成质量上取得了一定的进展，生成的矢量图形仍然存在着缺乏编辑性、生成过程缺乏多样性和细节缺失等问题。因此，需要开发出新的方法来克服这些挑战，尤其是能够增强图形的编辑性和多样性，同时保持较高的视觉质量。

1.2 选题依据

基于文本的矢量图形生成（Text-to-SVG）技术，尤其是结合了 T2I 扩散模型的方案，近年来受到广泛关注。T2I 扩散模型，像 CLIPDraw 和 VectorFusion，通过结合 CLIP 模型 [21] 和 DiffVG 模型 [11]，使用图像生成的反馈来优化矢量图形的设计。然而，这些基于 T2I 的方法在生成矢量图形时存在一些无法忽视的缺点。T2I 方法生成的图像往往是一个整体，缺乏可编辑性，不像传统的矢量图形生成方法那样清晰地区分不同的元素。生成的图形中，前景和背景元素常常交织在一起，导致单个元素难以被独立修改或调整。管 T2I 模型能够生成高质量的光栅图像，但当这些图像转换为矢量图形时，通常会出现诸如过度平滑、颜色饱和过高、细节丢失等问题。这些问题导致生成的矢量图形在表现力和美观性上存在显著不足。有的优化方法，如 Vectorfusion 中基于得分蒸馏（SDS）的优化方式，往往通过集中优化控制点来产生符合文本描述的矢量图形。然而，这种方法通常导致生成的矢量图形缺乏多样性，生成的图形模式趋同，缺少细节与多样化的特征。为了解决这些问题，提出了一种新的模型——SVGDreamer，该模型通过创新的方法提高了矢量图形的质量、编辑性和多样性。具体来说，SVGDreamer 通过引入语义驱动的图像矢量化（SIVE）过程和基于粒子的矢量得分蒸馏（VPSD）方法，有效克服了现有方法中的种种不足。

1.3 选题意义

整体而言，论文的贡献主要有三点：1）引入了 SVGDreamer，这是一种用于文本到 SVG（Text-to-Vector，T2V）生成的新模型。这种新模型能够在保持可编辑性的同时生成高质量的矢量图形；2）提出了语义驱动的图像矢量化（Semantic-driven Image Vectorization，SIVE）方法，通过引入基于注意力的原语控制策略，使得生成的矢量图形中的前景和背景能够被有效区分。该过程不仅通过跨注意力图（cross-attention maps）来初始化控制点，还采用文本提示（text tokens）来精准地指导每个图形元素的生成。该方法确保了生成的矢量对象是独立的和灵活的编辑。3）此外，提出了基于粒子的矢量化分数蒸馏（Vectorized Particle-based Score Distillation，VPSD）损失，通过将矢量图形建模为控制点和颜色的分布，采用 LoRA 网络估计这些分布，精确对齐预训练的扩散模型。这一方法解决了现有得分蒸馏方法中存在的图形过度平滑、颜色饱和过高、缺乏细节等问题，以保证生成的矢量图形具有卓越的视觉质量和广泛的多样性；4）进行了全面的实验来评估提出的方法的有效性。结果表明，与基线方法相比，论文的方法具有优越性。此外，论文的模型在生成不同类型的矢量图形方面显示出强大的泛化能力。

2 相关工作

2.1 矢量图形生成

可扩展矢量图形（SVG）提供了一种声明式的格式，通过几何原语来表达视觉概念。生成 SVG 内容的一种方法是使用序列到序列（seq2seq）模型 [1, 3, 13, 20, 26, 39, 40]。然而，这些方法通常依赖于矢量形式的数据集，这限制了它们的泛化能力，并使得合成复杂矢量图形变得困难。与其直接学习一个专门用于 SVG 生成的网络，另一种替代方法是在评估阶段通过优化目标图像来进行图形合成。Li 等人 [11] 提出了一个可微分的光栅化器，成功地将矢量图形和光栅图像之间的差异进行了桥接。尽管传统的矢量图形生成方法通常需要基于矢量数据集，

但近期的研究表明，采用可微分渲染器的方法能够突破这一局限 [14,26,31]。此外，随着视觉文本嵌入对比语言-图像预训练模型（CLIP） [22] 的快速发展，许多基于文本引导的草图合成方法取得了显著进展，如 CLIPDraw [2]、CLIP-CLOP [16] 和 CLIPasso [37] 等。最近的研究工作 VectorFusion [8] 和 DiffSketcher [43] 结合了可微分渲染器与文本到图像扩散模型，用于矢量图形的生成，并在图标设计、像素艺术以及草图生成等领域取得了有前景的成果。

2.2 文本到图像扩散模型

去噪扩散概率模型（DDPMs） [5, 33–35]，尤其是那些以文本为条件的模型，在文本到图像生成领域表现出了很大的潜力。例如，无分类器引导（CFG） [6] 显著提升了视觉效果，并已广泛应用于多个大规模的文本条件扩散模型框架，如 GLIDE [17]、Stable Diffusion [28]、DALL · E 2 [25]、Imagen [30] 和 DeepFloyd IF [36] 等。文本到图像扩散模型的进步也促使了一系列新的文本引导任务的发展，比如文本到 3D 生成 [19]。在本研究中，我们使用 Stable Diffusion 模型作为监督信号，推动文本到 SVG 图形生成的研究。

2.3 分数蒸馏采样

自然图像建模领域的最新进展激发了研究人员的浓厚兴趣，尤其是在利用强大的 2D 预训练模型来恢复 3D 物体结构方面 [12,15,18,19,38,41],。如 DreamFusion [19]、Magic3D [12] 和 Score Jacobian Chaining [38] 等方法，探索了通过得分蒸馏采样（SDS）损失，借助 2D 文本到图像扩散模型 [28,30] 来实现文本到 3D 的生成，并取得了显著成果。这一进展也激发了文本到 SVG 生成的研究 [8,43]，但生成的矢量图形质量依然受限，并表现出与重建 3D 模型类似的过度平滑现象。为了应对这一问题，Wang 等人 [41] 提出了一种新方法，他们将 3D 模型视为随机变量而非 SDS 中固定的常量，并引入了变分得分蒸馏技术，从而有效解决了文本到 3D 生成中的过度平滑问题。

3 本文方法

3.1 本文方法概述

SVGDreamer 由两部分构成：语义驱动的图像矢量化（Semantic-driven Image Vectorization，SIVE）和基于矢量粒子的分数蒸馏（Vectorized Particle-based Score Distillation，VPSD）构成。其中 SIVE 根据文本提示矢量化图像，VPSD 则通过分数蒸馏从预训练的扩散模型中合成高质量、多样化并具有审美吸引力的矢量图。

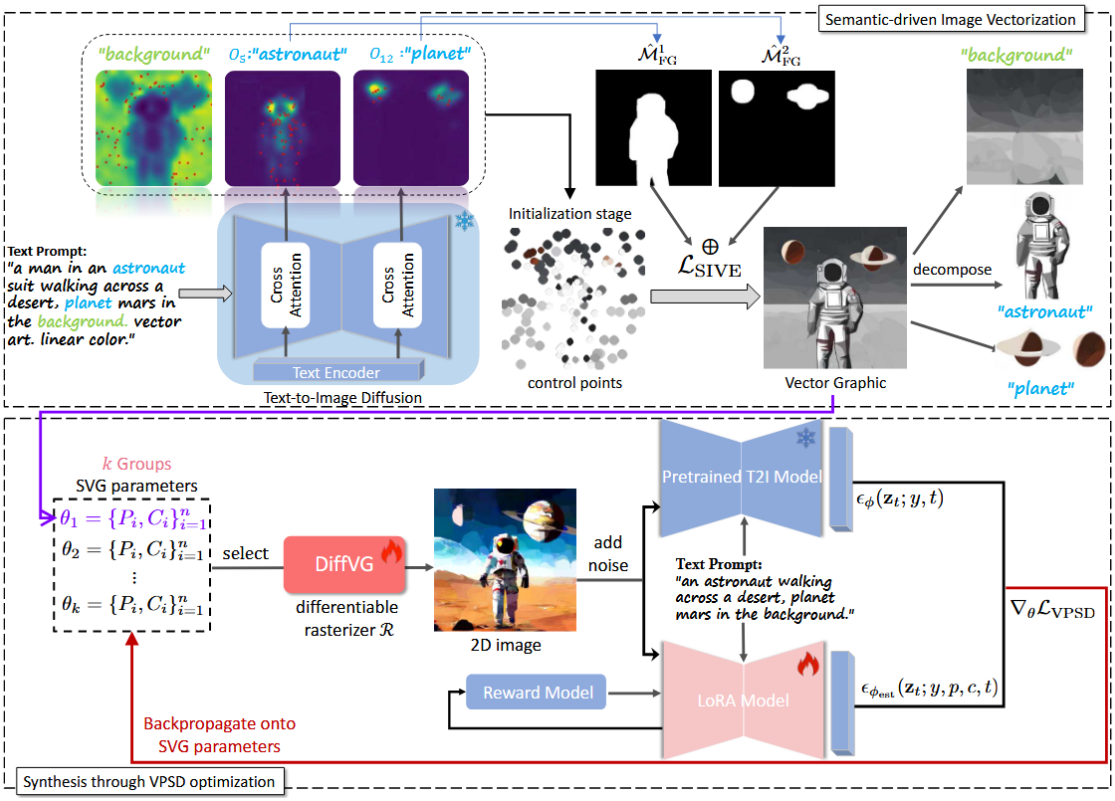

图 1. SVGDreamer 概述

3.2 语义驱动的图像矢量化 (SIVE)

SIVE 根据文本提示合成语义层次解耦的矢量图。它包括两个部分：(1) 矢量基元初始化 (Primitive Initialization)；(2) 基于语义级优化 (Semantic-aware Optimization)。如图 1 上半部分所示，文本提示中不同的词语对应不同的注意力图，这使得作者可以借助注意力图初始化矢量图控制点 (control points)。具体来说，作者对注意力图进行归一化，将它视为一个概率分布图，根据概率加权采样画布上的点作为贝塞尔曲线的控制点。然后，作者将初始化阶段获得的注意力图转换为可重复使用的掩码，大于等于阈值的部分设为 1，代表目标区域，小于阈值为 0。作者利用掩码定义 SIVE 损失函数从而精确地优化不同的对象。

$$\mathcal{L}_{\text{SIVE}} = \sum_{i}^{O} \left(\hat{\mathcal{M}}_i \odot I - \hat{\mathcal{M}}_i \odot \mathbf{x} \right)^2 \tag{1}$$

SIVE 确保了控制点保持在各自的语义对象区域中，从而实现不同对象的解构，最终结果如图 1 右上部分所示。

3.3 基于矢量粒子的分数蒸馏 (VPSD)

之前基于扩散模型的 SVG 生成工作，已经探索了使用分数蒸馏采样 (SDS) 优化 SVG 参数的方式，但这种优化方式往往会带来颜色过饱和、优化得到的 SVG 过于平滑的结果。受变分分数蒸馏采样的启发，作者提出了基于向量化粒子的分数蒸馏采样 (Vectorized Particle-based Score Distillation，VPSD) 损失来解决以上问题。相对于 SDS，这种采样方式将 SVG

建模为控制点和色彩的一个分布，VPSD 通过优化这个分布来实现对 SVG 参数的优化：

$$\nabla_\theta \mathcal{L}_{\text{VPSD}}\left(\phi, \phi_{\text{est}}, \mathbf{x}=\mathcal{R}(\theta)\right) \triangleq \mathbb{E}_{t,\epsilon,p,c}\left[w(t)\left(\epsilon_\phi\left(\mathbf{z}_t; y, t\right)-\epsilon_{\phi_{\text{est}}}\left(\mathbf{z}_t; y, p, c, t\right)\right)\frac{\partial \mathbf{z}}{\partial \theta}\right] \tag{2}$$

由于直接优化另一个模型 $\epsilon_{\phi_{est}}$ 的成本过大，所以引入 Lora 来减少被优化的参数量。

$$\mathcal{L}_{\text{lora}} = \mathbb{E}_{t,\epsilon,p,c}\left\|\epsilon_{\phi_{\text{est}}}\left(\mathbf{z}_t; y, p, c, t\right)-\epsilon\right\|_2^2 \tag{3}$$

最后，为了改善合成矢量图的美观评价，作者引入了一种奖励反馈学习方法（ReFL），将采样得到的样本输入到使用预训练的 Reward 模型中，共同进行对 LoRA 参数的优化：

$$\mathcal{L}_{\text{reward}} = \lambda \mathbb{E}_y\left[\psi\left(r(y, g_{\phi_{\text{est}}}(y))\right)\right] \tag{4}$$

最后完整的目标函数即为上述三个函数的加权组合

$$\min_\theta \nabla_\theta \mathcal{L}_{\text{VPSD}} + \mathcal{L}_{\text{lora}} + \lambda_r \mathcal{L}_{\text{reward}} \tag{5}$$

4 复现细节

4.1 创新点

SVGDreamer 是利用基于注意力的蒙版损失来分别优化前景和背景中的对象。这可确保控制点保持在各自的区域内，从而有助于对象分解。也就是说，层次结构仅存在于指定的对象中，不会与其他对象混淆。此策略为形成不同矢量图形的对象之间的排列和组合提供了动力，并增强了对象本身的可编辑性。

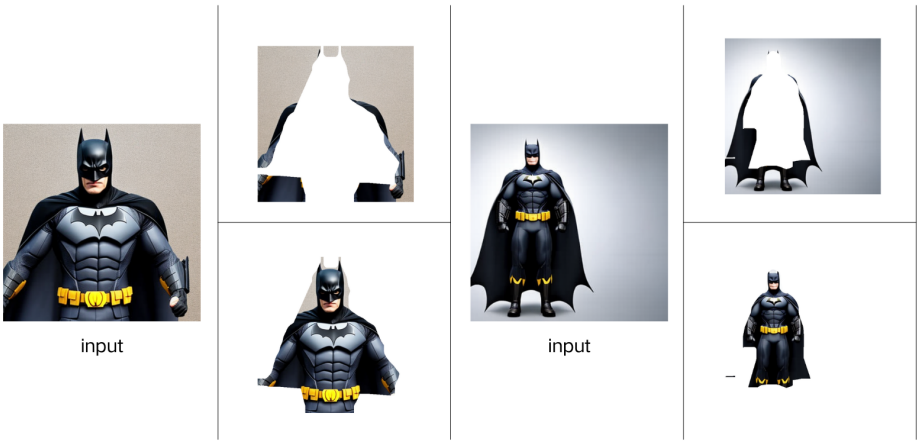

图 2. 基于注意力的前景对象与背景分割结果

但是实际，SVGDreamer 中基于注意力机制的分割效果并没有那么好，如图 2 所示，边缘区域的分割存在一定程度的不精确。例如蝙蝠侠的披风边缘位置，在某些视角下的前景背

Algorithm 1 Using Segment-Anything (SAM) to Generate Masks

Input: Model checkpoint path `checkpoint_path`, model type `model_type`, input image `image`
Output: Generated masks `masks` and visualization
Step 1: Initialize SAM model
Load `sam_model` from registry using checkpoint `checkpoint_path`
Move model to device: `sam_model.to(device)`
Step 2: Load input image
Load input image as a numpy array: `image`
Step 3: Create mask generator and generate masks
Initialize mask generator `mask_generator` with:
    `points_per_side = 45`, `pred_iou_thresh = 0.86`
    `stability_score_thresh = 0.92`, `crop_n_layers = 1`
    `crop_n_points_downscale_factor = 2`, `min_mask_region_area = 100`
Generate masks: `masks = mask_generator.generate(image)`
Step 4: Sort masks by area
if `len(masks) > 0` **then**
    Sort masks by area in descending order: `sorted_masks`
    Step 5: Visualize masks
    Create a transparent background image: `img`
    for all `mask` in `sorted_masks` **do**
        Extract segmentation: `segmentation = mask['segmentation']`
        Generate random color with transparency: `color_mask`
        Apply color to segmentation: `img[segmentation] = color_mask`
    end for
    Display `img` with visualization
end if

景区分不够清晰，可能出现小范围的背景残留或前景误分。前景分割虽然整体较为完整，但小细节的处理有一定丢失。例如蝙蝠侠披风的尖角、手部的精细轮廓等位置细节稍有损失。如果图像中前景存在遮挡（如蝙蝠侠被其他物体遮住），可能导致部分区域无法完整分割。

我的目标是希望生成的 SVG 具有可编辑性，每一条 SVG 的路径有意义一点，比如说一个带握把的 SVG 是由握把的路径和杯子的路径组成，而不是由一些没有意义的许多小路径组成的，因此改进思路是通过通过引入更精细的分割算法（如结合额外的边界检测模块）进一步提升边缘区域的分割精度。使用 Segment-Anything(SAM) [9] 模型来生成多个级别的分割掩码，而不再使用 SVGDreaner 的基于注意力机制的前景对象以及背景的分割，因此由 SAM 模型生成的 mask 不再是前景对象以及背景，而是由 SAM 模型生成的很多 mask，结果如下。

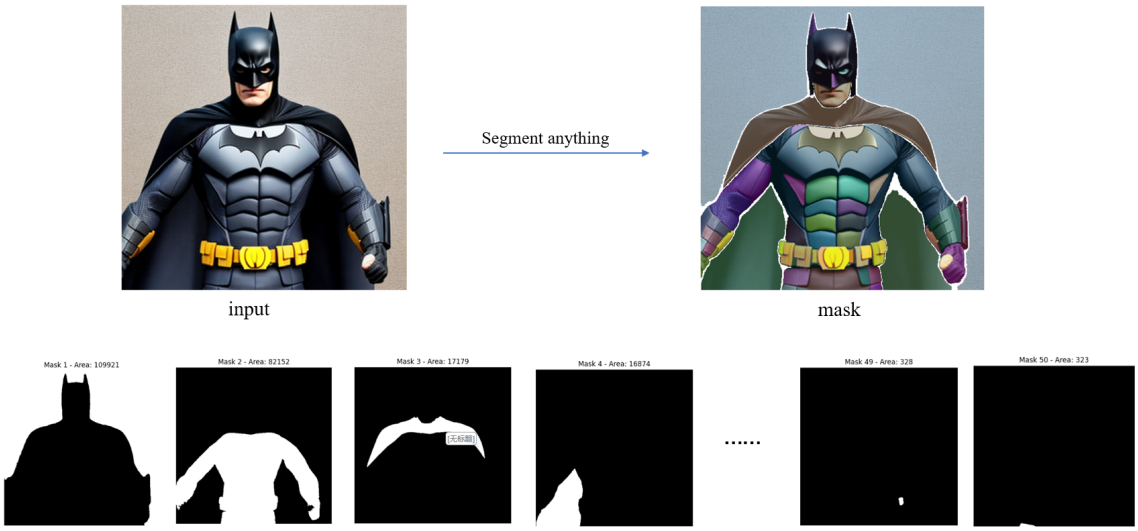

图 3. 基于 segment-anything 的分割结果（mask 中不同颜色代表不同的分割区域）

4.2 实验环境搭建

4.2.1 安装环境

您可以按照以下步骤快速启动并运行 **SVGDreamer**。这些步骤将允许您在本地运行快速推理。

在项目根目录运行 sh script/install.sh

或使用 Docker：docker run –name svgdreamer –gpus all -it –ipc=host ximingxing/svgrender:v1 /bin/bash

4.2.2 下载预训练的稳定扩散模型

通过首次运行时的设置下载预训练的 SD 模型：diffuser.download=True/conf/config.yaml diffuser.download=True

或者您仍然可以手动下载它：模型链接：https://huggingface.co/stabilityai/stable-diffusion-2-1-base

默认模型存储路径为：/home/user/.cache/huggingface/hub/models–stabilityai–stable-diffusion-2-1-base

5 实验结果分析

5.1 论文定性评估

图 4. 不同方法的定性比较

图 4 展示了 SVGDreamer 与现有文本到 SVG 方法之间的定性比较。与 CLIPDraw [2] 相比，SVGDreamer 生成的 SVG 在保真度和细节方面表现出更优的效果。此外，SVGDreamer 与基于 SDS 的现有方法进行了对比 [8, 43]，突出了在解决形状过度平滑和颜色过度饱和等问题上的优势。如第五列所示，SIVE 实现了语义上的解耦，但仍未能完全克服 SDS 方法固有的平滑问题。而在最后两列中，SVGDreamer 相比基于 SDS 的生成方法，展示了更为精细的细节，无论是从头优化模型，还是在优化过程中进行改进，这都显著提高了生成图形的美学评分。

表 1. Quantitative comparison of different methods.

Method / Metric	**FID↓**	**PSNR↑**	**CLIPScore ↑**	**BLIPScore ↑**	**Aesthetic↑**	**HPS↑**
CLIPDraw	160.64	8.35	0.2486	0.3933	3.9803	0.2347
VectorFusion (scratch)	119.55	6.33	0.2298	0.3803	4.5165	0.2334
VectorFusion	100.68	8.01	0.2720	0.4291	4.9845	0.2450
DiffSketcher (RGB)	118.70	6.75	0.2402	0.4185	4.1562	0.2423
SVGDreamer (from scratch)	84.04	10.48	0.2951	0.4311	5.1822	0.2484
+Reward Feedback	83.21	10.51	0.2988	0.4335	5.2825	0.2559
SVGDreamer	**59.13**	**14.54**	**0.3001**	**0.4623**	**5.5432**	**0.2685**

为了验证 SVGdreaner 提出方法的有效性，进行了实验，评估了模型在多个指标上的表现，包括 Fréchet Inception Distance（FID） [4]、峰值信噪比（PSNR） [7]、CLIPScore‘ [23]、BLIPScore [10]、美学评分（Aesthetic score） [32] 和人工评估得分（Human Performance Score, HPS） [42]。表 1 中展示了 SVGDreamer 与几种代表性的文本到 SVG 生成方法的对比，包括 CLIPDraw [2]、VectorFusion [8] 和 DiffSketcher [43]。在评估矢量图形的多样性和填充颜色饱和度时将 SD 采样结果作为基准（Ground Truth, GT），并分别计算了 FID 和 PSNR 指标。定量分析的前两列结果表明，SVGDreamer 在 FID 和 PSNR 得分上超越了其他方法，表明我们的生成方法在多样性方面优于基于 SDS 的合成方法 [8, 43]。为了评估生成的 SVG 与提供的文本提示之间的一致性，采用了 CLIPScore 和 BLIPScore 进行评测。为了衡量合成矢量图像的感知质量，使用 LAION 美学分类器 [32] 来进行美学评分。此外，还使用 HPS 从人类审美的角度对我们的方案进行了评估。

5.2 复现结果

我的目标是希望对 SVGDreamer 进行改进，使其生成的 SVG 每条路径更有意义一点。因此，我的工作是将第一个阶段语义驱动的矢量化（SIVE）中对图片前景对象与背景的分割进行修改，即将基于注意力机制的 mask 生成换成基于 segment-anything 的 mask 生成。但是经过实验后发现，结果并没有提升多少。这是因为第二阶段矢量粒子的分数蒸馏会对 mask 分割的区域去初始化一些 SVG 路径，然后再去优化这些路径，然后再继续初始化一些路径，再优化，不断重复以上过程，直至收敛。而这个初始化路径就是问题所在，为什么第二阶段称为矢量例子的分数蒸馏，因为初始化的路径是粒子形状的，相当于第二阶段初始化很多粒子形状的路径，如图 6 所示，然后再去做优化。

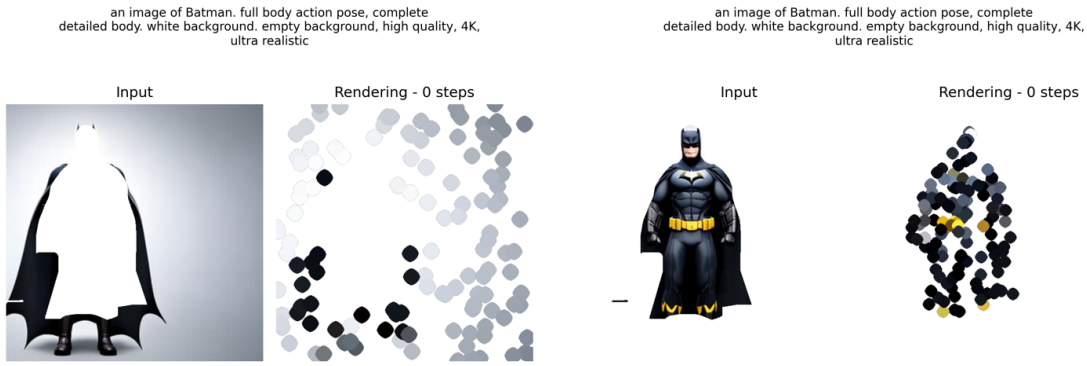

图 5. SVG 路径初始化

这样的初始化本生就产生了许多很小的路径，这些路径本身就没有实际意义，再去做优化，也没有意义，因为一开始初始化的路径就很小。由于时间太紧就还没对这一部分做改进。生成的 SVG 最终结果如图 7 所示。

图 6. “蝙蝠侠”的矢量图形生成结果

6 总结与展望

6.1 总结

在本文中，我复现了 SVGDreamer, SVGDreamer 由两部分构成, 第一部分是语义驱动的图像矢量化（SIVE）：基于注意力机制生成前景与背景的掩码（Mask），并通过控制点初始化分离不同的语义对象。第二部分是基于矢量粒子的分数蒸馏（VPSD）：将 SVG 建模为控制点和颜色的分布，采用 LoRA 网络估计这些分布。通过奖励模型引入审美优化机制，提升生成图形的视觉吸引力。但是经过实验发现，SVGDreamer 存在不足。SIVE 基于注意力机制的分割方法在边缘处理上存在精度不足，如前景与背景交界处模糊，导致掩码生成的不准确性。小细节处理能力有限，容易丢失精细的边缘信息（如蝙蝠侠披风尖角部分）。因此将注意力分割替换为 Segment-Anything 模型，但是发现尽管分割质量有所提高，但未能显著改善最终生成的 SVG 质量。经过实验还发现第二阶段的路径优化过程是导致 SVG 质量未能提升的问题所在。第二阶段的 VPSD 蒸馏过程中，路径初始化为粒子形状，导致生成许多小而无意义的路径。这些路径的存在浪费了优化资源，并降低了最终生成的 SVG 的实用性和审美性。不足之处在于没有对第二阶段做改进，导致生成的 svg 质量没有提高。

6.2 未来研究方向

对 SVGDreamer 中针对粒子形状路径初始化的问题，可以探索更加智能的路径生成方法，使每个路径具有明确的语义和设计意义。引入基于形状分析的路径合并和优化策略，减少冗余路径，提高 SVG 的整体简洁性和美观性。同时可以使用更先进的分割模型（如 Segment-Anything 的高级版本或多模态融合模型）以进一步优化前景与背景的分离，在分割阶段增加边缘检测模块，强化对复杂边界的处理能力。

参考文献

[1] Alexandre Carlier, Martin Danelljan, Alexandre Alahi, and Radu Timofte. Deepsvg: A hierarchical generative network for vector graphics animation. *Advances in Neural Information Processing Systems (NIPS)*, 33:16351–16361, 2020.

[2] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022.

[3] David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems (NIPS)*, 30, 2017.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 33, pages 6840–6851, 2020.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[7] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[8] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1911–1920, 2023.

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022.

[11] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

[12] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–309, June 2023.

[13] Raphael Gontijo Lopes, David Ha, Douglas Eck, and Jonathon Shlens. A learned representation for scalable vector graphics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Xu Ma, Yuqian Zhou, Xingqian Xu, Bin Sun, Valerii Filev, Nikita Orlov, Yun Fu, and Humphrey Shi. Towards layer-wise image vectorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16314–16323, 2022.

[15] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[16] Piotr Mirowski, Dylan Banarse, Mateusz Malinowski, Simon Osindero, and Chrisantha Fernando. Clip-clop: Clip-guided collage and photomontage. *arXiv preprint arXiv:2205.03146*, 2022.

[17] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[18] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d {gan}s know 3d shape? unsupervised 3d shape reconstruction from 2d image {gan}s. In *International Conference on Learning Representations (ICLR)*, 2021.

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[20] Qianru Qiu, Xueting Wang, and Mayu Otani. Multimodal color recommendation in vector graphic documents. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 4003–4011, 2023.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[26] Pradyumna Reddy, Michael Gharbi, Michal Lukac, and Niloy J Mitra. Im2vec: Synthesizing vector graphics without vector supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7342–7351, 2021.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NIPS)*, volume 35, pages 36479–36494, 2022.

[31] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing synthesis. *arXiv preprint arXiv:2111.03133*, 2022.

[32] Christoph Schuhmann. Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022.

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37, pages 2256–2265, 2015.

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NIPS)*, volume 32, 2019.

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[36] StabilityAI. If by deepfloyd lab at stabilityai. https://github.com/deep-floyd/IF, 2023.

[37] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[38] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12619–12629, June 2023.

[39] Yizhi Wang and Zhouhui Lian. Deepvecfont: Synthesizing high-quality vector fonts via dual-modality learning. *ACM Transactions on Graphics (TOG)*, 40(6), 2021.

[40] Yizhi Wang, Gu Pu, Wenhan Luo, Pengfei Wang, Yexin ans Xiong, Hongwen Kang, Zhonghao Wang, and Zhouhui Lian. Aesthetic text logo synthesis via content-aware layout inferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023.

[42] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2096–2105, October 2023.

[43] Ximing Xing, Chuang Wang, Haitao Zhou, Jing Zhang, Qian Yu, and Dong Xu. Diffsketcher: Text guided vector sketch synthesis through latent diffusion models. In *Advances in Neural Information Processing Systems (NIPS)*, 2023.