



基于生成式 AI 的个性化文创图像作品设计系统 系统设计文档



任课教师 _____ 杨波

学 院 _____ 计算机学院

专 业 _____ 计算机科学与技术

组 别 _____ 第一组

组 长 _____ 郑仕博

成 员 _____ 陈奕嘉，苏泳豪

2025 年 4 月 24 日

目录

1	引言	2
1.1	目的	2
1.2	项目范围	2
1.3	文档概览	2
1.4	参考资料	2
1.5	术语与缩略语	2
2	系统概览	2
3	系统架构	3
3.1	架构设计	3
3.2	分解描述	3
3.3	设计原理	3
4	数据设计	4
4.1	数据说明	4
4.2	数据字典	4
5	组件设计	4
6	人机界面设计	5
6.1	界面概览	5
6.2	界面截图	5
6.3	界面控件与操作	5
7	需求矩阵	8
8	APPENDICES	8

1 引言

1.1 目的

本软件设计文档描述了“基于生成式 AI 的个性化文创图像作品设计系统”的架构与系统设计。面向开发、测试、维护本项目的工程人员及项目管理者，作为技术实现和系统集成的参考依据。

1.2 项目范围

该软件旨在利用生成式 AI 技术解决个性化文创产品供给不足的问题，核心功能包括：根据用户输入的文本和指定的位置生成创意图像，或编辑现有图像中的文本。重点目标是实现中文字符的高精度渲染，便于游客与文创从业者快速创作独特图像作品，助力文旅融合与传播。

1.3 文档概览

第 1 章介绍目的、范围、参考资料和术语；第 2 章提供系统概览；第 3 章详细阐述系统架构；第 4 章描述数据设计；第 5 章介绍各组件设计；第 6 章讲解人机界面设计；第 7 章为需求矩阵；第 8 章为附录。

1.4 参考资料

信息来源于网页https://www.sohu.com/a/823541100_234564。技术细节参考了 AnyText、TextDiffuser、DDPM 等文献。文档结构参考与https://github.com/SPM-PSP/SPM-PSP-Course-github/blob/main/SDD_Template.pdf。

1.5 术语与缩略语

AI（人工智能）、SDD（软件设计文档）、VAE（变分自编码器）、UNet（网络结构）、Stable Diffusion（SD，扩散模型）、AnyText（生成式模型）、Text-control Diffusion Pipeline、Auxiliary Latent Module、Text Embedding Module、Gradio（UI 库）、Prompt（文本提示）、OCR（光学字符识别）、FID（图像质量指标）、CFG-Scale（无分类引导因子）、 η （扩散采样参数）等术语在文中根据需要进一步解释。

2 系统概览

本系统是一个利用生成式 AI 的图像创作工具，支持文本生成图像和图像内文字编辑，专注于中文字符的精准渲染。系统基于 AnyText 并通过 Google 提出的 Dreambooth 方法微调 Stable Diffusion 模型，通过 Web 界面（Gradio）与用户交互，

后端使用 Python 与深度学习框架实现，支持 Docker 部署。系统的目的是解决文创产品同质化问题，赋能个体创作。

3 系统架构

3.1 架构设计

系统分为三层：用户界面层（Gradio 实现）、应用逻辑层。

用户界面层：负责输入（文本、图像、参数、坐标绘制）与结果展示。

应用逻辑层：解析输入、格式化参数、调用模型、处理输出与数据管理。以 AnyText 为核心，包括三大子模块：

1. 文本嵌入模块（Text Embedding Module）
2. 辅助潜变量模块（Auxiliary Latent Module）
3. 文本控制扩散管道（Text-control Diffusion Pipeline）

3.2 分解描述

Text Embedding Module: 接收用户提示词和需渲染文本，对需要生成的文字用占位符占位，生成对应字形图，用 OCR 提取特征后替换占位符，然后传入到 Clip 编码器。

Auxiliary Latent Module: 掩码、字形图，通过卷积处理生成与扩散模型匹配的空间向量。

Diffusion Pipeline: 以初始噪声为起点，联合文本嵌入与空间特征逐步去噪生成图像潜变量，最后 VAE 解码。

Gradio UI: 提供文本输入、图像上传、画布交互、参数调节、结果展示等供用户操作的界面展示等功能。

3.3 设计原理

采用 AnyText + Stable Diffusion v1.5 架构，针对中文文本渲染难题，结合字形信息与位置控制，以 Realistic_Vision_V4.0 作为底模保证图像质量和真实性，通过对 AnyText 框架进行中文语料微调提升中文文字的准确率，用 Dreambooth 对扩散模型进行微调，采取 Gradio 快速构建用户界面。

4 数据设计

4.1 数据说明

输入数据包括提示词（文本）、需渲染文本、位置坐标、参考图像（可选）、控制参数；训练后的权重以 ckpt 文件存储，约 5.73GB，训练数据包含两类：

1. AnyWord-3M 标注数据（JSON 格式），筛选后约 400k 张，用于文字渲染的训练；
2. 文创图像 + 文本描述（TXT 格式），约 1k 张，用于风格微调和物品的学习。

输出图像保存在服务器并将图像、debug 信息（可选）返回给用户。

4.2 数据字典

user_prompt: 字符串

text_to_render: 字符串列表

position_data: 坐标列表

edit_mask: 掩码图像/张量

reference_image: 上传图像

control_params: 参数字典，如 {'cfg_scale': 7.5}

generated_image: 最终生成图像

training_data_1: AnyWord-3M JSON 结构

training_data_2: TXT 列表与对应图像

model_weights: 模型权重文件

glyph_image、text_embedding、auxiliary_latent、image_latent: 中间张量

hehe98/wenchuang: 项目镜像，详情见 [dockerhub](#)

wenchuang.ckpt: 模型权重文件

strength: 文字渲染控制强度，可以为 0 即不使用文字渲染

CFG-Scale: 文字控制强度，低的话会导致生成图像与描述不符合，高的话图像会不自然

eta: 风格多样性，1 表示启用（更具变化），0 不启用（更保守）

5 组件设计

主要功能以组件化方式组织，核心函数如下：

generate_image: 解析提示词，生成字形图和文本嵌入，调用辅助模块生成空间信息，联合生成潜变量图像，再解码输出；

edit_image: 编码参考图像，加入掩码噪声生成初始状态，调用编辑流程生成新图像并融合原图。

每个过程中的局部变量包括潜变量，预测噪声，注意力图等张量，模块间依次传递处理。

6 人机界面设计

6.1 界面概览

提供 Web 端界面，两种主要操作模式：

1. 文到图像的生成：输入提示词，并将需要渲染的文本用“”标注，可以通过画布绘制文本位置、拖框选择文本位置或随机选择文本位置；
2. 图片文字编辑，手动掩盖需要修改区域，输入文本并进行编辑。

界面上有说明、参数设置、文本输入框、模式选择、文字位置标注、样例（点击即可）、运行按钮、图片结果展示和加强训练的物品，用户可调整 CFG-Scale、Steps 等参数，查看结果并保存。

6.2 界面截图

详情见图 1，图 2.

6.3 界面控件与操作

包括：

- 说明文本框
- 文本输入框（Prompt）
- 位置选择方式（单选按钮）
- 绘制画布（支持自由绘制、矩形、掩码）
- 参数调节控件（滑动条/输入框）
- “运行”按钮
- 图像展示区域
- 图片上传控件
- 示例加载按钮
- 参考生成物品展区
- 加强训练物品展区

操作：

用户可点击说明查看使用须知，点击参数调整控件调整参数，在文本输入框输入文字进行提示词输入，点击运行进行生成，点击样例进行生成，在模式选择框选择模式。

7 需求矩阵

详情见表 1。

需求	组件
文本输入（提示词）	文本输入框
图像上传	图片上传控件（或者绘制画布）
指定文字位置	绘制画布
参数调节	参数调节控件
结果预览	图像展示区域
保存分享	图像展示区域
Debug	图像展示区域和参数调节控件
模式选择	图片上传控件
示例与指导	说明文本框

表 1: 功能与需求表

8 APPENDICES

详见材料中的“项目注意事项”文档。