

基于生成式AI的个性化 文创图像作品设计



项目创意来源



本作品的核心创意来源于当前市场上文创产品同质化严重，难以满足游客日益增长的个性化需求的痛点。习近平总书记关于推动文化和旅游融合发展，将文化旅游业培育成为支柱产业的指示，以及《如果国宝会说话》等成功案例，激发了通过创新方式“激活”文化遗产，赋能个体创造独特文创作品的想法。



青铜立人像



青铜神鸟鼠标垫

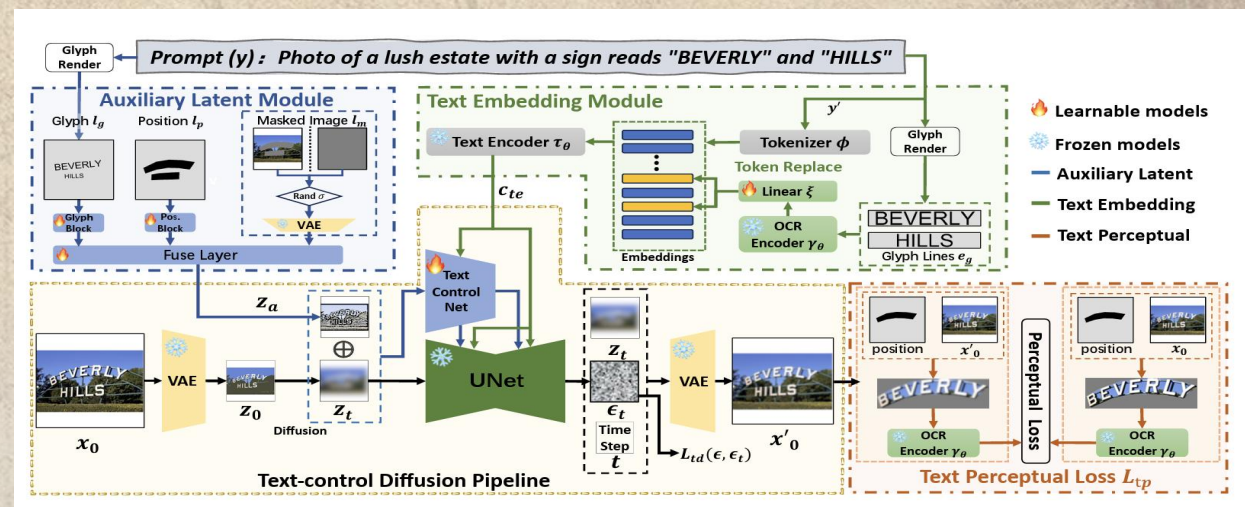
系统架构与模型设计



架构

使用阿里云的AnyText模型，主要由三部分组成——Auxiliary Latent Module, Text Embedding Module和 Text-control Diffusion Pipeline。

本项目将其分为文字控制框架和扩散模型两部分，并分别进行调整。



截自论文 "ANYTEXT: MULTILINGUAL VISUAL TEXT GENERATION AND EDITING"

功能

文字生图：通过文字说明进行图片生成，可以指定文字生成的位置。

文字编辑功能：对图片进行文字添加、修改和去除。

特点：能够图文融合。



图文融合示例

训练流程与数据处理



文字控制框架训练

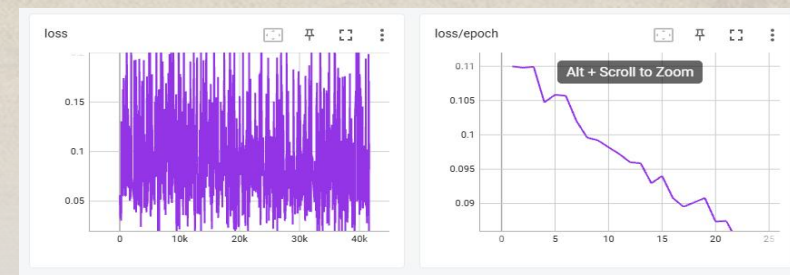
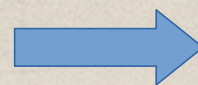
对开源数据集AnyWord-3M进行筛选（根据水印，有效性，同时保证中文训练数据集远大于英文数据集），保留了约400k的数据量，并用8卡V100进行训练。



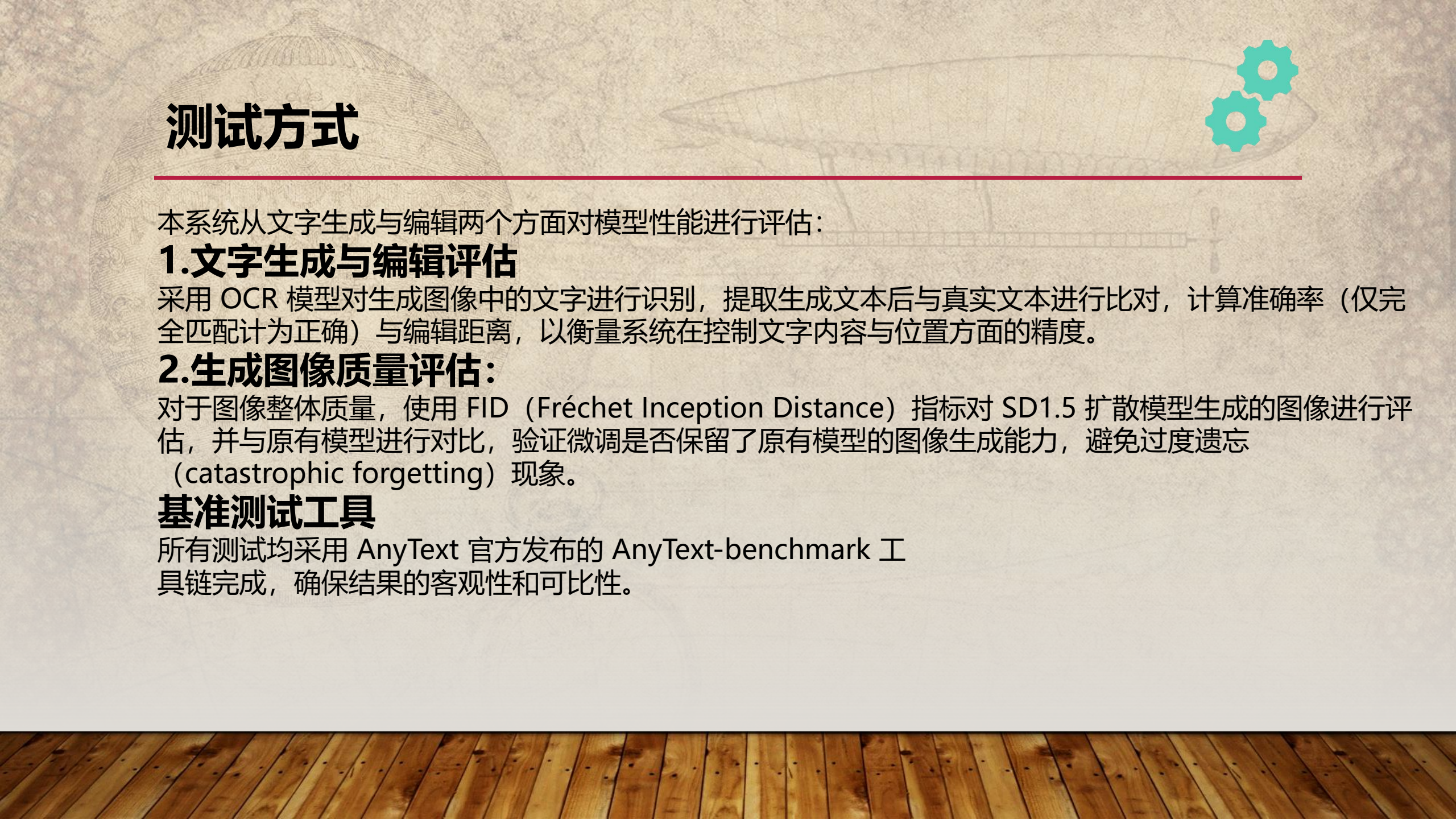
```
{  "data_root": "/root/autodl-tmp/AnyText/data/data/imgs",  "data_list": [    {      "img_name": "043517ab1c1560c0178a34e9108c33f467868286.jpg",      "annotations": [        {          "polygon": [            [              72,              99            ],            [              430,              97            ],            [              431,              210            ],            [              73,              212            ]          ],          "text": "KDG",          "language": "Latin",          "rec_score": 0.9999164041519165,          "valid": true        },        {          "polygon": [            [              198,              261            ],            [              322,              261            ],            [              322,              278            ],            [              198,              278            ]          ],          "text": "KEDAGROUP",          "language": "Latin",          "rec_score": 0.9964058995246887,          "valid": true        }      ]    }  ]}
```

扩散模型训练

在google、edge、百度等搜索引擎上用python爬取图片，经过筛选后，对保留的约1千张的图片进行去水印，调整分辨率为512*512，用wd14-convnextv2-v2进行标注，进行简要修改后，从hugging-face上下载Realistic_Vision_V4.0作为起始权重，采取DreamBooth的方式进行权重的微调。



测试方式



本系统从文字生成与编辑两个方面对模型性能进行评估：

1.文字生成与编辑评估

采用 OCR 模型对生成图像中的文字进行识别，提取生成文本后与真实文本进行比对，计算准确率（仅完全匹配计为正确）与编辑距离，以衡量系统在控制文字内容与位置方面的精度。

2.生成图像质量评估：

对于图像整体质量，使用 FID (Fréchet Inception Distance) 指标对 SD1.5 扩散模型生成的图像进行评估，并与原有模型进行对比，验证微调是否保留了原有模型的图像生成能力，避免过度遗忘 (catastrophic forgetting) 现象。

基准测试工具

所有测试均采用 AnyText 官方发布的 AnyText-benchmark 工具链完成，确保结果的客观性和可比性。

接口需求



应用程序接口 (API) 需求:

并不调用其它应用程序接口。

数据接口需求

:Gradio 本身主要负责用户界面的展示和与后端 Python 代码的交互，它不直接处理与外部数据源的连接。

硬件接口需求:

本项目不涉及硬件接口。



安全性与可维护性

安全性:

- 数据隐私保护: 确保对话发生在本地, 不收集用户数据。
- 防止恶意攻击: 强化防御体系, 防止恶意攻击。
- 访问控制: 对于进行过恶意访问的 IP 进行拉黑, 减少恶意访问量。

可维护性:

- 代码的可读性与可扩展性较好, 详细注释了大部分功能。
- 代码易于部署, 容易维护。



运行环境要求

1. 硬件要求:

GPU 4060, 8GB 显存, 32GB 内存。

2. 环境要求:

详见 https://github.com/SPM-PSP/PSP25-1/blob/main/codes/front_back_ends/environment.yaml

3. 网络要求:

需要连接互联网以下载模型和数据集

技术难点



多语言文字嵌入与位置对齐困难

文字生成与编辑一致性不足

训练资源消耗大，优化困难

评估方法易受 OCR 误差干扰

图像扩散模型易遗忘原始生成能力

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。**

组件设计及功能



1.Text-control Diffusion Pipeline

使 TextControlNet 控制文字的生成并且保证在模型没有文字生成需求时正常地生成图片

2.Auxiliary Latent Module

可告知模型生成文本到不规则的文本框中

3. Text Embedding Module

用于解决多语言文本生成的问题

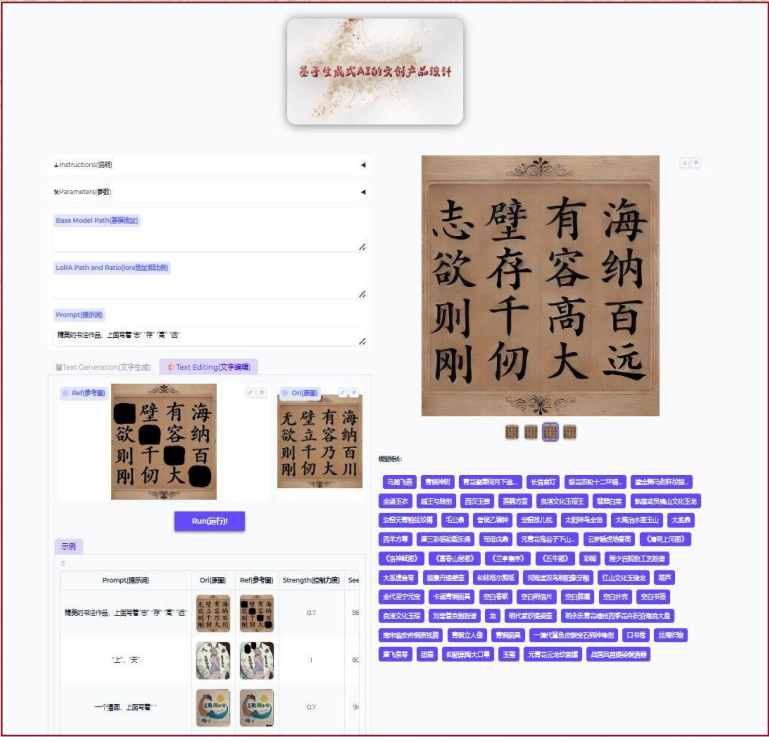
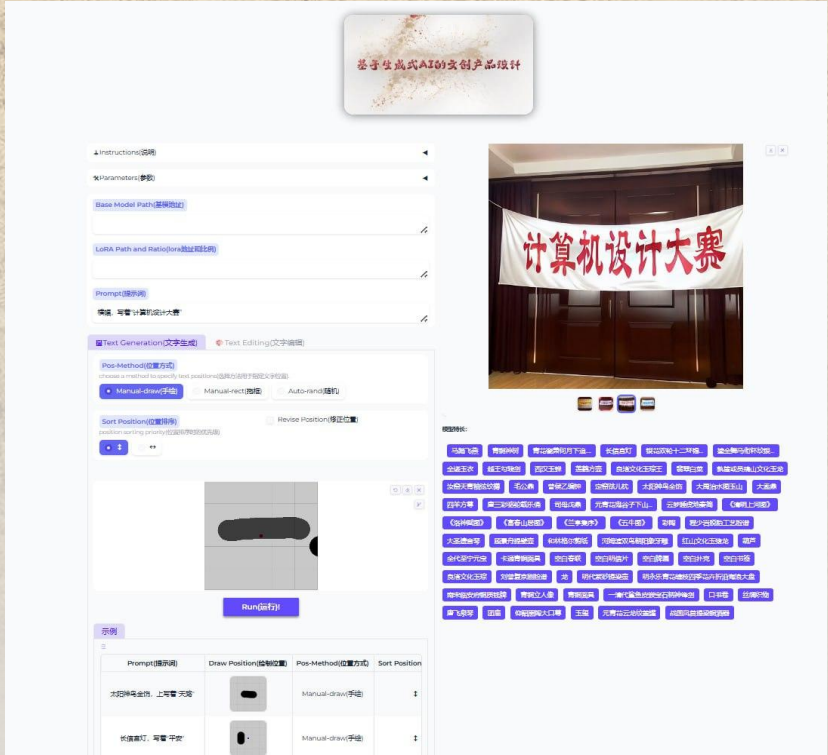
4 Gradio UI

提供了一个简单的界面，可以轻松地与模型进行交互。用户可以通过 Gradio UI 上传图像、输入文本、调整参数等操作，并实时查看生成的结果

人机界面实例截图



文字到图片到生成



图片文字编辑

需求矩阵



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

生成对比



Ours

提示词：太阳神鸟金
饰上写着“天路”



Ours

提示词：卡通青铜树，
上方写着“神树”



文
字
生
成
功
能

文
字
编
辑
功
能



提示词：“天”，“上”



更
多
例
子



项目测试



	状态	正确率 ↑	编辑距离 ↑
文字生成	训练前	0.6957	0.8402
	训练后	0.6644	0.8282
文字修改	训练前	0.6671	0.8298
	训练后	0.6644	0.8282

文字控制框架文字生成和文字编辑评估正确率通过使用 ocr 对文字进行识别，与正确的进行比对,计算正确率（完全正确才行）和编辑距离。

发现训练效果并不理想，与算力、参数以及训练批次有关（经费问题）。

	FID ↓
训练前	31.558
训练后	34.242

sd1.5 扩散模型评估采用 FID (评估时与框架合并了)。证明了扩散模型经过训练，并没有遗忘掉太多先前学到的知识。

均采用 AnyText-benchmark!

因此目前项目使用官方 AnyText 的权重并与训练好的扩散模型进行合并，得到最终的模型权重

创新点



本项目致力于通过创新的文字渲染模型，革新图片修改的应用场景，为用户提供丰富多样的选择，轻松设计独具个性的文创产品。此外，项目更支持便捷的文字编辑功能，实现图片中文字的修改与创意贴图，赋能用户更自由的创作表达。

本项目将模型划分为文字渲染框架和扩散模型框架，对扩散模型进行了更改并且微调，相比于如今许多的文生图大模型，该模型算力要求相对较低，容易部署，具备精确指定文字位置的功能和文字编辑的功能，同时该模型在文字生成的正确率上有一定的保证。



演示视频



首先通过gradio搭建并进入演示网站



**感谢老师倾听
恳请批评斧正！**

补充

FID 的核心思想是利用预训练的 Inception 网络来提取真实图像和生成图像的特征，然后比较这两组特征在特征空间中的统计距离。

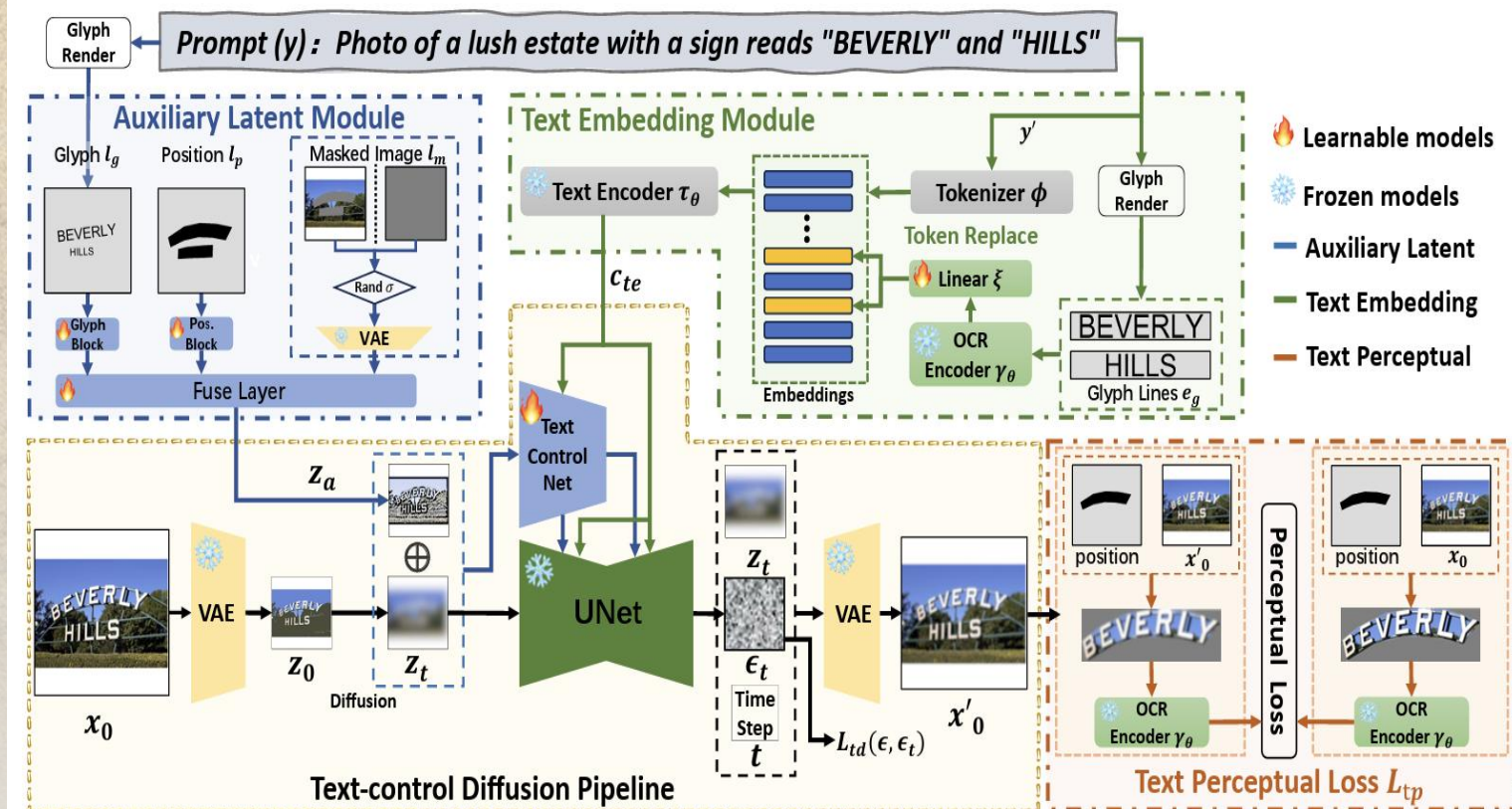
公式为：
$$FID = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2 * \text{sqrtm}(\Sigma_r @ \Sigma_g))$$

编辑距离指一个字符串修改成另一个字符串所需要的最少操作，操作包括删除、添加和修改一个字符串。

DreamBooth 是由 Google Research 和波士顿大学于 2022 年提出的一种个性化微调技术，旨在增强文本到图像生成模型（如 Stable Diffusion）的能力，使其能够生成包含特定主体的图像。通过对预训练模型进行微调，DreamBooth 能够在保持模型原有能力的同时，学习并再现特定主体的视觉特征。

补充

wd14-convnextv2-v2是一种基于 ConvNeXtV2 的预训练模型，通常是用于计算机视觉任务中的深度学习模型。ConvNeXt 是一个基于卷积神经网络 (CNN) 的模型，专为图像分类和其他计算机视觉任务（如物体检测、分割等）设计。ConvNeXtV2 是 ConvNeXt 的改进版本，它采用了现代卷积网络架构中的一些新技术，例如与 Vision Transformer (ViT) 相结合的理念，使得其在性能上有所提升。



截自论文 "ANYTEXT: MULTILINGUAL VISUAL TEXT GENERATION AND EDITING"