2024 年（第 17 届）
中 国 大 学 生 计 算 机 设 计 大 赛

人工智能实践赛作品报告

作品编号：

作品名称：基于生成式 AI 的个性化文创图像作品设计

填写日期：2025/3/30

填写说明:

1、 本文档适用于人工智能挑战赛预选赛；
2、 尽管预选赛仅完成部分工作，但是本文档需要针对决赛做出方案设计；
3、 正文、标题格式已经在本文中设定，请勿修改；标题#的快捷键为“Ctrl+#”，正文快捷键为“Ctrl + 0”；
4、 本文档应结构清晰，突出重点，适当配合图表，描述准确，不易冗长拖沓；
5、 提交文档时，以 PDF 格式提交；
6、 本文档内容是正式参赛内容的组成部分，务必真实填写。如不属实，将导致奖项等级降低甚至终止本作品参加比赛。

目 录

第1章 作品概述

1.1 主题创意来源与产生背景

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。

1.2 作品的用户群体

追求个性化体验的游客： 他们希望能够设计出独一无二的旅行纪念品。

文创产业从业者： 他们可以利用该工具提升产品设计的效率和创新性。

1.3 主要功能

本作品结合文字渲染框架，实现对图片上文字的删除、修改及添加，尤其专注于精准的中文文字生成。

1.4 应用价值

推动文旅融合： 丰富旅游产品供给，释放文化产业经济价值，助力文化旅游业转型升级。

促进文化传播： 赋能个体创造文化作品，让文物等文化元素以更生动有趣的方式传播。

满足用户需求： 解决了市场上个性化文创产品供给不足的问题，满足人们对美好生活的新期待。

第2章 问题分析

2.1 问题来源

当前市场上的文创产品大多采用预先设计制作的模式，难以满足游客日益增长的个性化需求。这种供需矛盾是本项目提出的主要问题来源。此外，现有的图像生成模型在文字控制方面存在明显的技术瓶颈，尤其是在处理中文时，容易出现字体扭曲、模糊和错误等问题，这限制了相关技术在中文文创领域的应用潜力。同时，对于旅行者等用户而言，缺乏便捷的、能够随时随地进行个性化文创设计和制作的工具也是一个亟待解决的问题。

2.2 现有解决方案

传统的图像编辑软件 (如 Adobe Photoshop, Illustrator)： 这些软件功能强大，可以实现对图片的文字进行修改和添加。但它们通常需要专业技能，且操作相对复杂，难以满足普通用户快速、便捷地进行个性化文创设计的需求。此外，这些软件在生成与背景自然融合的文字方面也存在一定的局限性。

在线设计平台 (如 Canva, 稿定设计)： 这些平台提供了丰富的模板和素材，用户可以进行简单的文字替换和排版。但其个性化定制程度相对较低，难以实现高度自由的创意表达。

已有的文字控制图像生成模型 (如 GlyphDraw, Textdiffuse, AnyText)： 这些模型在解决字体与背景融合方面取得了一定的进展，但正如前文所述，它们仍然难以完全避免文字生成中的错误，并且缺乏专门针对中文的优化。

当前模型通过集成大语言模型提升了文本生成的稳定性，然而，对文本生成位置的精细化控制以及基于图像内容的文本引导修改能力仍有待提升。

2.3 本作品要解决的痛点问题

无法满足用户日益增长的个性化文创产品需求： 现有方案在定制化程度和操作便捷性方面存在不足，难以让普通用户轻松设计出独特的文创产品。

现有文字控制图像生成模型在中文处理上的缺陷： 缺乏专门针对中文优化的高精度文字生成模型，导致在中文文创设计中应用受限。

缺乏便捷的移动端个性化设计工具： 用户在旅途中或其他不方便使用电脑的场景下，难以随时进行创意设计。

文字与图片背景融合度不高： 现有技术难以实现生成的文字与图片背景的自然、无缝融合，影响视觉效果。

2.4 解决问题的思路

【填写说明：作品的功能和性能需求；使用的数据集，包括数据格式，数据来源，数据获取方式，数据特点，数据规模等，并给出具体的数据样例。所提出的指标或要求必须在第5章得到印证】

2.4.1 作品的功能和性能需求

功能需求：

方式一：基于用户输入的文字

1. 文字输入: 用户可以在界面上提供的文本输入框中键入或粘贴想要使用的文字内容，例如一句诗词、一句格言、 一个祝福语、或者任何简短的文本。

2. 供参考的物品：用户可以在界面上看到所有进行加强训练的物品，模型在生成这些物品的准确率上有一定的保 证，同时也可以看到比较优美的一些例子，供用户参考。

3. 文字位置标注: 自由拖拽模式: 用户可以使用鼠标在预览区域内拖拽一个矩形框，以指定文字将要放置的位置和大致大 小。 预设位置选择: 系统可以提供一些常用的预设位置选项（例如：居中、左上角、右下角等），供用户快速 选择。

4. 图片生成: 用户点击“生成”按钮后，AI 产品将根据用户输入的文字、标注的位置、选择的文字样式和文创风格， 生成一张带有文字的创意图片。

5. 结果预览与调整: 生成的图片会显示在预览区域，用户可以查看效果。如果对结果不满意，可以返回修改文字内 容、位置、样式或风格，然后重新生成。

6. 保存与分享: 用户可以将生成的文创图片保存到本地设备，或者分享到社交媒体平台。

方式二：基于用户输入的成形图片

1. 图片上传: 用户可以通过文件上传功能上传一张已经存在的图片作为创作的基础。支持常见的图片格式，例如 JPG、PNG 等。

2. 供参考的物品：用户可以看到一些比较优美的例子，但仅供参考。

3. 文字位置标注: 与方式一类似，用户可以使用自由拖拽模式或预设位置选择在上传的图片上标注文字将要放置的位置和大 小。 用户可以添加多个文本框，标注多个文字的位置。

4. 文字内容输入与样式定制: 对于每个标注的位置，用户可以输入相应的文字内容。

5. 图片生成: 用户点击“生成”按钮后，AI 产品将在用户上传的图片上，根据标注的文字位置和样式，以及添加的其 他文创元素，生成最终的文创图片。

6. 结果预览与调整: 生成的图片会显示在预览区域，用户可以查看效果。用户可以调整文字的位置、样式，或者添 加/修改其他文创元素，然后重新生成。

7. 保存与分享: 用户可以将最终生成的文创图片保存到本地设备，或者分享到社交媒体平台。

总而言之，该 AI 产品旨在提供一个简单易用、功能丰富的界面，让用户能够通过输入文字或上传图片，并灵活地标 注文字的位置和定制样式，最终生成具有个性化创意的文创图片。（本项目保留的原本扩散模型具有的生成能力，可以不使用文字渲染生成，即退化成 sd1.5）

性能需求：

为确保卓越的性能表现，该模型应展现出较高的文本准确性；同时，也应该具备卓越的图像生成能力，能够根据明确的指令，细致地描绘出指定的各类

物品；此外，要求避免过拟合现象的发生，从而保证了模型在面对未知数据时依然能够保持优异的泛化性能。

2.4.2 数据集

本项目制作了两份数据集，第一份数据集用于微调 AnyText 模型，另一份数据集用于微调 stable diffusion v1-5。

数据集来源：

本项目利用爬虫技术从 Google、Bing、百度等搜索引擎抓取数据，自动化访问网站并提取网页内容（如文本、图片、链接等），筛选并保存有价值的信息。项目重点聚焦于文物图像数据的采集，旨在抓取并整合图文结合的信息，以构建高质量的数据集。

大规模多语言数据集 AnyWord-3M，于论文 AnyText 中提出。数据来源涵盖 Noah-Wukong、LAION-400M 及多个 OCR 任务数据集（如 ArT、COCO-Text、RCTW 等），覆盖街景、书籍封面、广告等多种文本场景。OCR 数据直接利用已有标注，其余图片经 PP-OCR 处理并由 BLIP-2 生成文本描述。经过严格筛选与后处理，最终获得 303 万余张图片，包含 900 万行文本及 2000 万余字符。

数据规模：

第一份基于 AnyWord-3M，按照水印、文字块、语言（保留大部分中文数据集，少部分英文数据集）等标准精筛后，保留约 40 万条数据。第二份数据集由爬虫采集约 1000 张与中华文化及文物相关的图片，并通过水印筛选和修整优化质量。随后，使用 wd14-convnextv2-v2 进行自动标注，并对标注结果进行了适当调整，以提升准确性和适用性。

数据样例：

第一份数据集（采用 json 存储图片信息）：

json 存储图片的信息方式如下：


```
{
    "img_name":"00b36ddd2bca24d0be115808e473fb5471bcc233.jpg",
      "annotations": [
        {
          "polygon": [
            [
              177,
              94
            ],
            [
              399,
              94
            ],
            [
              399,
              108
            ],
            [
              177,
              108
            ]
          ],
          "text": "struggleforabetterfuture。",
          "language": "Latin",
          "rec_score": 0.973587155342102,
          "valid": false
        },
        {
          "polygon": [
            [
              176,
              112
            ],
            [
```

```
            449,
            112
        ],
        [
            449,
            143
        ],
        [
            176,
            143
        ]
    ],
    "text": "将来的你一定会感谢",
    "language": "Chinese",
    "rec_score": 0.996906578540802,
    "valid": true
},
{
    "polygon": [
        [
            166,
            141
        ],
        [
            456,
            144
        ],
        [
            456,
            196
        ],
        [
            166,
            193
        ]
    ],
    "text": "现在奋斗的你",
    "language": "Chinese",
    "rec_score": 0.994633138179779,
    "valid": true
},
{
    "polygon": [
        [
            124,
            184
        ],
        [
            283,
            196
        ],
        [
            280,
            228
        ],
        [
            121,
            216
        ]
    ],
```

```
            "text": "Time",
            "language": "Latin",
            "rec_score": 0.9853572845458984,
            "valid": true
        }
    ],
    "caption": "a wall with a clock and a sign that says time is money",
    "wm_score": 0.0015544143971055746
}
```


对于图上的每一个文字块，标注着他们的坐标、文字信息、语言和置信度，此外还有图片的 caption 和水印的程度。

第二份数据集（采用 txt 存储图片的 caption）：

标注：Sun god bird gold ornament, yellow theme, abstract, solo

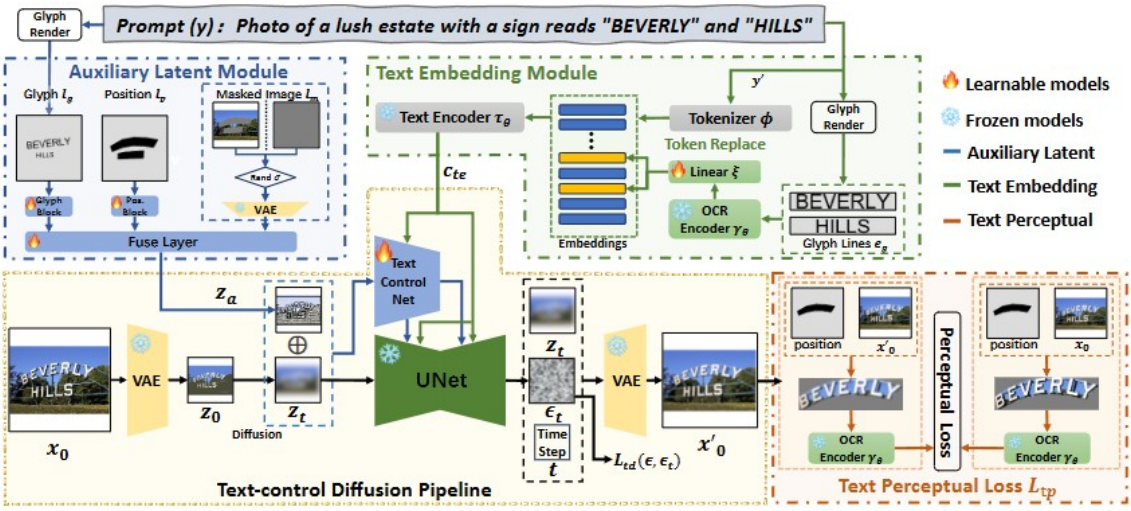
使用 wd14-convnextv2-v2 标注，部分标注有误，进行了略微地修改。

第3章 技术方案

【填写说明：从原理层面，详细介绍系统所采用的技术方案，先总体介绍，给出技术路线框架图，然后分模块详细介绍。着重介绍解决问题的思路，以及所涉及的模型、协议、算法等，以及可能的对算法的改进；原创工作详述，非原创工作简述，并尽可能标注引用文献】

3.1 技术框架

本项目使用的模型主要由三部分组成 **Text-control Diffusion Pipeline**、**Auxiliary Latent Module** 以及 **Text Embedding Module**。



3.1.1 Text-control Diffusion Pipeline

在这一部分，本项目通过变分自编码器（VAE）来生成潜在层特征z_0,潜在层的扩散算法逐步给z_0增加噪音并生成新的潜在层特征z_t，其中 t 代表时间步。辅助层特征z_α、文字嵌入层特征c_{te}和时间步被作为条件预测噪音ϵ_t,并将它加入到z_t。更详细地说，为了控制生成的文字，将z_α加入到z_t并将他输入到可训练的 TextControlNet 里（一个可训练的 UNet 编码层），这样就能使该部分生成z_α

3.1.2 Auxiliary Latent Module

由三个因素决定——glyphl_g、位置$llpp$和掩码后的图像$llmm$。glyph $llgg$使用 glyph render（使用 Arial Unicode）生成到相应的位置上，考虑到生成不规律的文本框有一定难度，所以该模块使用位置l_p,glyph render 文本框使用矩形，通过和$llgg$结合，该模块可以告知模型将文本生成到不规则的文本框上。此外该模块将掩码后的图像作为信息，告诉模型不要修改这些地方，并使用 VAE 下采

样。为了合并这些条件，该模块使用卷积层下采样 glyph l_g 和位置 l_p ，使他们跟 z_t 有相同的空间大小，最后使用卷积融合层来合并他们。

3.1.3 Text Embedding Module

文本编码器善于从描述中提取语义信息，但却会忽略需要渲染的文本的语义信息。此外，大多数预训练的文本编码器都是在基于拉丁字母的数据上训练的，因此无法很好地理解其他语言。在 AnyText 中，提出了一种新颖的方法来解决多语言文本生成的问题。具体而言，该模块将字形线条渲染为图像，编码字形信息，并用它们替换 token 的嵌入。然后，将替换后的嵌入作为 token 输入到基于 transformer 的文本编码器中，得到融合后的中间表示，这些表示随后通过交叉注意力机制映射到 UNet 的中间层。由于该模块的做法使用图像渲染文本，而不是仅依赖于特定语言的文本编码器，因此显著提升了多语言文本生成的效果。

3.2 技术重点（解决问题的思路）

3.2.1 数据集制作

本项目构建了两份数据集。第一份基于 AnyWord-3M，按照水印、文字块、语言（保留大部分中文数据集，少部分英文数据集）等标准精筛后，保留约 40 万条数据。第二份数据集由爬虫采集约 1000 张与中华文化及文物相关的图片，并通过水印筛选和修整优化质量。随后，使用 wd14-convnextv2-v2 进行自动标注，并对标注结果进行了适当调整，以提升准确性和适用性。

3.2.2 模型微调

针对中文场景，本项目将对 AnyText 模型先与 Realistic_Vision_V4.0（基于 sd1.5)进行权重的合并，在中文训练集上的专门训练，使模型更好地适应中文应用场景，确保其在实际应用中展现卓越的表现能力。此外为了更贴合主题，本项目在 Realistic_Vision_V4.0（基于 sd1.5)权重基础上使用 dreambooth 的方法进行微调，并与微调后的 AnyText 模型的权重进行整合。（第一步的 sd1.5

只是用于训练的，其权重并不会改变，而第二步才微调 sd1.5 然后进行权重的整合）第一步使用第一份数据集，第二部采取第二份数据集。（详情见 3.2.1)

3.2.3 部署与框架实现

本项目将使用 Gradio 搭建模型的交互式部署框架。结合任务需求，设计并开发直观易用的网页和交互界面，便于用户体验模型的功能和效果，同时支持进一步的迭代优化。

第4章 测试分析

【填写说明：通过测试与对比，论证系统的有效性，可包括验证数据的来源与规模、测试过程、分析与结论等等。各参赛队务必重视数据测试，所有对自己作品准确性、有效性、稳定性，甚至作品受欢迎的程度的宣称，都应该得到数据结果或对比实验的支持，否则评审人有理由怀疑其真实性】

首先是文字正确率的测试：

采用 OCR 识别生成图片中的文字，与标注的进行比对，完全正确才能被认为正确，此外还计算了编辑距离（均是越低越好）。

训练前：

文字到图生成的正确率：

文字修改正确率：

训练后：

文字到图生成的正确率：

```
/root/autodl-tmp/test1k.json loaded, imgs=1000
evaluate: 100%|                                        | 1000/1000 [09:57<00:00,  1.67it/s]
Done, lines=8368, sen_acc=0.6671, edit_dist=0.8298
```

文字修改的正确率：

```
/root/autodl-tmp/test1k.json loaded, imgs=1000
evaluate: 100%|                                        | 1000/1000 [02:55<00:00,  5.69it/s]
Done, lines=2092, sen_acc=0.6644, edit_dist=0.8282
```

发现效果并不好，由于该步训练需要大量的算力（至少 8 卡 V100），采取原先的模型。

接下来测试与训练好的扩散模型合并后的效果，采用 FID（越低越好）：

训练前：

```
100%|                                        | 799/801 [
100%|                                        | 800/801 [
100%|                                        | 801/801 [
02:44<00:00,  4.88it/s]
100%|                                        | 80/80 [00:10<00:00,  7.85it/s]
FID:  31.55861167111368
```

训练后：

```
(anytext) root@autodl-container-dfbc4bb0ff-09e94c68:~/autodl-tmp/AnyText# ./eval/eval_fid.sh
100%|                                        | 801/801 [03:09<00:00,  4.22it/s]
100%|                                        | 80/80 [00:09<00:00,  8.65it/s]
FID:  34.24209446278792
```

前后 FID 接近，证明模型并没有过拟合。

对比效果展示：

Prompt:太阳神鸟金饰上面写着“天路”

Prompt:太阳神鸟金饰上面写着“天路”

Prompt:卡通青铜树，上方写着“神树

Prompt:卡通青铜树，上方写着“神树

更多展示：

Prompt:长信宫灯，写着“平安”

Prompt：卡通青铜面具，头顶刻着“王”

第5章 作品总结

【填写说明：从创意、技术路线、工作量、数据和测试效果等方面对作品进行自我评价和总结，并对作品的进一步提升和应用拓展提出展望】

5.1 作品特色与创新点

5.1.1 作品特色

个性化定制： 用户可以根据自己的想法和创意修改图片上的文字，制作出独一无二的文创产品。

便捷性： 旨在为用户提供随时随地进行文字创作和设计的便利，尤其是在旅途中。

自然融合： AI 生成的文字能够自然地融入各种背景环境，效果优于简单的文字添加功能。

5.1.2 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。

5.2 应用推广

技术优化：提升模型性能，降低计算资源消耗，提高生成速度和文字准确性，增强复杂背景和艺术字体的表现力。

功能拓展：增加图案、纹理等设计元素的生成与编辑功能，优化界面，实现实时预览、智能提示，并提供个性化设计建议和模板推荐。

应用场景：深化旅游和文化领域合作，定制文创产品，拓展至教育行业，助力文化课程教学，并为企业品牌提供定制化设计服务。

市场与合作：通过线上线下推广扩大用户群，联手文创制造商、电商平台，实现设计到生产的无缝对接，并建立用户社区，促进创作生态发展。

5.3 作品展望

市场潜力巨大： 能够满足大量游客和文化爱好者的个性化定制需求。

技术优势明显： 专注于解决中文文字生成难题，具有差异化竞争优势。

应用场景广泛： 可应用于旅游纪念品、个性化礼品、文化宣传品等多种场景。

符合发展趋势： 契合文化和旅游融合发展的大方向。

参考文献

[1](2024-11-04 16:25)习近平总书记强调，把文化旅游业培育成为支柱产 业. https://www.sohu.com/a/823541100_234564

[2] Jingye Chen*13, Yupan Huang*23, Tengchao Lv3, Lei Cui3, Qifeng

Chen1, Furu Wei3. TextDiffuser: Diffusion Models as Text Painters. arXiv:2305.10855v5 [cs.CV] 30 Oct 2023

[3] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng*, Xuansong Xie. ANYTEXT: MULTILINGUAL VISUAL TEXT GENERATION AND EDITING. arXiv:2311.03054v5 [cs.CV] 21 Feb 2024

[4]Jonathan Ho, Ajay Jain, Pieter Abbeel. Denoising Diffusion Probabilistic Models. arXiv:2006.11239v2 [cs.LG] 16 Dec 2020

[5] 张 静，孙巧榆，刘珍兵. 基于深度学习的场景文本检测方法研究综述. 智 能计算机与应用.2024:2095-2163

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[7]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS) (pp. 5998-6008). Retrieved from https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf [8]Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character Region Awareness For Text Detection (CRAFT). [9]Baidu Inc.(2024). PP-OCR. https://github.com/PaddlePaddle/PaddleOCR

[9] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng*, Xuansong Xie. ANYTEXT: MULTILINGUAL VISUAL TEXT GENERATION AND EDITING. arXiv:2311.03054v5 [cs.CV] 21 Feb 2024

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint,abs/2301.12597, 2023.