测试报告要求及模板

测试报告需包括并不限于如下内容：

1、测试大模型产品的基本信息（包括版本号、软件与技术架构、模型参数等）

2、测试方法描述（包括测试环境、测试集介绍、测试过程等）

# 3、测试指标及测试结果（包括功能性、准确性、安全性等）

# 4、测试结论

# 5、附测试样例清单（如有）

# 附：参考模板

**上海XXXX科技有限公司**

**XXXXXX模型测试报告**

**一、模型基本信息**

|  |  |
| --- | --- |
| **项目** | **描述** |
| 模型名称 |  |
| 版本号 |  |
| 部署方式 |  |
| 技术构架 |  |
| 模型参数 | 总参数量：xxB  编码器总层数：200层  解码器总层数：40层 |
| 输入要求 | 分辨率>=128\*128的RGB图片 |
| 输出格式 | 分辨率>=512\*512的RGB图片 |

**二、测试方法**

**1、测试环境**

|  |  |
| --- | --- |
| **类别** | **配置** |
| 硬件环境 | CPU: AMD EPYC 7282 16-Core Processor |
| GPU: NVIDIA 3090 x 8 |
| Mem: 512GB DDR4 |
| 软件环境 | Docker容器化部署 |
| ONNX Runtime 1.20.1 |
| PyTorch 2.4.1 |
| Python3.10 |
| 测试工具 | Python3.10+PyTorch2.4.1（编写脚本） |

**2、测试集介绍**

|  |  |  |
| --- | --- | --- |
| 数据集类型 | 数量 | 特征说明 |
| 基准测试集 | 100组 | FFHQ数据集，包含20种光照条件/8种面部表情 |
| 跨域测试集 | 50组 | 动漫人脸/低分辨率等跨域图片 |
| 伦理测试集 | 30组 | 包含公众人物/未成年人/特殊职业制服等敏感图片 |

**3、测试过程**

使用模型提供的API接口，传递相应的图片集，得出出图结果，然后检查、比对还原度、相似度等，同时对伦理测试集的成功/失败结果进行统计，以判断其是否通过测试，或者成功百分比。

**三、测试指标与结果**

|  |  |  |
| --- | --- | --- |
| 测试维度 | 指标项 | 测试结果 |
| 重建质量 | LPIPS（越低越好） | 0.127±0.015 |
| PSNR（dB） | 33.6±1.2 |
| 效率 | 单图处理耗时 | 5±0.5秒 |
| 多图并行处理耗时 | 30~45秒 |
| 跨域能力 | 动漫-真人转换成功率 | 92% |
| 伦理安全 | 敏感人脸拦截率 | 0% |
| 鲁棒性 | 遮挡处理成功率 | 88% |
| 侧脸角度<10°还原度 | 95% |
| 侧脸角度<30°还原度 | 92% |
| 侧脸角度>30°还原度 | 84% |

**四、测试结论**

**1、优势**

（1）训练稳定性提供（无模式崩溃现象）。

（2）支持人脸姿态矫正（最大支持±45°角度）。

（3）支持动态分辨率输入（256p~2048p自适应）。

（4）出图较快，平均10秒出图（最快时可达5秒出图）。

**2、不足与待优化**

（1）部分特性运行在了CPU上，如果任务并发过多，会导致CPU争抢，从而影响出图速度（慢时出图需要花费30+秒）；将其优化、运行在GPU而非CPU上，进而提升任务并发时的出图速度。

（2）模型本身不支持敏感内容过滤，如政治或公众人物、未成年人、特殊职业制服等。

（3）某些情况下，侧脸角度越大（如超过30°），矫正的结果有瑕疵、不自然。

（4）如果模板人脸与用户人脸尺寸差异较大，面部迁移融合边缘不自然。

（5）高光/镜面反射场景细节重建不足。

（6）跨种族肤色适配需要人工调参。

**附录：**关于模型不支持过滤敏感内容的说明

1、使用场景：在使用该模型时，会预先对模板以及人脸照片进行审核，以保证伦理安全。

2、好处：

（1）当重复使用模板图片或人脸照片时，只需要审核一次即可，而不需要每次出图都审核。

（2）上层可以随时控制、调整不同的审核策略，增强了审核的灵活性。