

User Study 阅读文档

非常感谢您参与本次人脸生成、编辑主观评价测试，本次测试分为生成与编辑两部分，预计需要30分钟。请您在测试前仔细阅读本文档，以便于您了解本次测试的规则及评价标准。

生成部分测试

请打开Query_Generate.html作答

测试规则

请从第一行图片中选出“语义匹配度”最优的一个结果，从第二行图片中选出“图片真实度”最优的一个结果。单击图像一次即可选中图像，再次单击取消选中。当两行均选中图像时即翻页。选择时无需过多思考，跟从直觉即可。

指标介绍及选择原则

生成部分测试共包含两个指标的测试，下面将详细介绍。

语义匹配度

该指标衡量文本与图片之间的匹配程度。当图片表现出了文本中的每一个属性（性别、发色、表情等）描述并且与该文本没有语义冲突时即可认为该图片的语义匹配度高。

- 比如，文本中明确给出“He wears glass.”但是图片中的人脸没有眼镜，那么该图片语义匹配度低。
- 如果文本中没有明确说该人是否戴眼镜，那么图片中的人脸戴眼镜与否不影响该图片的语义匹配度。
- 您可以按您的理解选择性地忽略文本中某些与人脸不显著相关的描述（比如职业、教育经历等）。
- 您可以按您的理解代入想象或选择性忽略文本中某些与人脸不显著相关的描述（比如职业、国籍、教育经历等）。

注意：当您觉得超过一张图片的语义匹配度相似地高时，请您选择您看起来视觉体验最好（这张脸看起来很舒服，接近自然人脸）、最清晰（这张脸不模糊）的一张图片。如果您觉得视觉体验也难以分出高低，可以直接随机给出选择。

图片真实度

该指标衡量该图片反映一张真正人脸的质量高低，不考虑语义。请您直接选择“看起来最接近一张正常人脸”的图片。评价标准如下：

- 看起来确实像一个真人的照片。
- 这个人没有让我觉得不正常的属性，看起来很舒适。
- 图片质量高，细节不模糊。

注意：当您觉得有超过一张图片都有比较高的图片真实度时，请您选择“语义匹配度”最高的一项。如果您觉得“语义匹配度”也难以分出高低，可以直接随机给出选择。

编辑部分测试

请打开Query_Manipulate.html作答

测试规则

请从第一行图片中选出“语义匹配度”最优的**两个**结果，从第二行图片中选出“图片解耦度”最优的**两个**结果。单击图像一次即可选中图像，再次单击取消选中。当两行均选中两张图像时即翻页。选择时无需过多思考，跟从直觉即可。

指标介绍及选择原则

语义匹配度

该指标衡量编辑后的图像确实具有文本中描述的特征的程度。评价标准如下：

- 比如，给定一个没有刘海的人像，再给定一个文本“This is a human face with bangs.”，但是编辑后的图像中的人像没有刘海，那么该图像语义匹配度低。
- 如果两张编辑后的图像均有明显的刘海，那么无论两个图像中加上的刘海的多少，两图像的“语义匹配度”相当（也即只在乎是否成功编辑，不在乎编辑的强度）。但是注意前提是两者的刘海都必须“明显”，如果有一个编辑后的图像仅加上了若有若无、很不明显的刘海，则该图像“语义匹配度”不高。
- 如果编辑后的图像出现了显然不自然的、过度的特征，则认为“语义匹配度”也不高（例如文本是“A face with red lipsticks”，但是加上的红嘴唇过于鲜艳导致十分不自然）。

注意：当您觉得超过两张图片的语义匹配度相似地高时，请您选择您看起来视觉体验最好（这张脸看起来很舒服，接近自然人脸）、最清晰（这张脸不模糊）的两张图片。如果您觉得视觉体验也难以分出高低，可以直接随机给出选择。

图片解耦度

该指标衡量编辑后的图片对无关特征的保持程度以及编辑后的图像仍然接近一个真实自然人脸图像的程度。“高图片解耦度”应有如下的特征：

- 编辑后的图像中，文本没有描述的特征几乎没有变化。
- 编辑后的图像仍然是一个高质量的人脸图像（不考虑不合理的语义。比如为生理女性的人脸编辑加上络腮胡子后，尽管语义层面不符合现实，但只要人脸图像本身质量仍然很高即可）。
- 图片质量高，细节不模糊。

注意：当您觉得超过两张图片的图片解耦度相似地高时，请您选择您看起来语义匹配度最高的两张图片。如果您觉得语义匹配度也难以分出高低，可以直接随机给出选择。

测试结果反馈

完成测试后，请将自动下载的两份结果文档“Generate_result.csv”及“Manipulate_result.csv”发送给发布者。

再次感谢您的参与！