OneIG-Bench: 全方位细致评估图像生成

Jingjing Chang[1,2] **Yixiao Fang**[2,†] **Peng Xing**[2] **Shuhan Wu**[2] **Wei Cheng**[2]
Rui Wang[2] **Xianfang Zeng**[2] **Gang Yu**[2,‡] **Hai-Bao Chen**[1,‡]

[1] SJTU [2] StepFun
[†] Project lead [‡] Corresponding author

Dataset **Code** **Project Page**

Abstract

文本到图像（T2I）模型因其能够生成与文本提示相匹配的高质量图像而受到广泛关注。然而，T2I 模型的快速进展暴露出早期基准测试的局限性，缺乏全面的评估，例如推理、文本渲染和风格的评估。值得注意的是，最新的最先进模型凭借其丰富的知识建模能力，在图像生成问题上展示了有希望的结果，这些问题需要强大的推理能力，但现有的评估系统尚未充分解决这一前沿问题。为了系统地解决这些差距，我们引入了 OneIG-Bench，这是一个精心设计的全面基准框架，用于对 T2I 模型进行多维度的细粒度评估，包括提示-图像对齐、文本渲染精度、推理生成内容、风格化和多样性。通过结构化评估，该基准能够深入分析模型性能，帮助研究人员和从业者确定图像生成全流程中的优势和瓶颈。具体来说，OneIG-Bench 通过允许用户专注于特定评估子集，实现灵活评估。用户可以仅针对与所选维度相关的提示生成图像，并相应地完成相应的评估，而无需为整个提示集生成图像。我们的代码库和数据集现已公开，以促进 T2I 研究社区中可重复的评估研究和跨模型比较。

近年来，文本到图像（T2I）模型在图像质量、语义对齐、文本渲染精度以及基于知识的图像生成推理方面取得了显著进展 [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]。然而，评估系统的发展显著滞后于模型的进步：现有的大多数基准测试仍局限于单一维度的评估，缺乏全面性。例如，T2ICompBench [**?**]、GenEval [**?**] 和 DSG-1k [**?**] 专注于短文本语义理解，而 DPG-Bench [**?**] 引入了密集提示评估，但在风格和文本等有限维度的覆盖上不足。尽管 WorldGenBench [**?**] 涉及世界知识和推理，但迫切需要一个更全面的多维度评估框架，以提供科学的模型评估并指导技术发展。

为了推动文本生成图像模型的发展，我们认为开发一个全面的基准框架，可以在多个相互关联的维度上评估模型，对于促进严谨而全面的评估至关重要。我们引入了图 1 所示的 OneIG-Bench，其中包含了超过 1000 个全方位的提示，主要来源于真实用户的输入，系统地设计以全面评估文本生成图像模型在各种生成能力上的表现。基于不同的生成主题，我们将评估数据集分类为六大核心评估类别：通用物体、肖像、动漫与风格化、文本渲染、知识与推理及多语言性。通过利用我们的框架进行生成主题的分类区分，我们可以更细致地评估模型的多维能力。因此，这种系统化的评估允许用户识别出具备针对性能力的模型，使其与特定应用需求精确对齐。

针对不同的评估维度，我们精心设计了定量指标，考虑到各种因素以确保评估的全面性和客观性。这些指标旨在精确衡量模型在与特定主题领域相关的不同方面的表现，从而实现更准确和深入的评估。具体而言，对于普通物体、肖像、动漫和风格化的评估，我们开发了一个评估系统，用于衡量模型对输入提示的响应能力。对于动漫和风格化，我们还将风格相似性纳入我们的评估系统，专门评估模型再现多样艺术风格的能力。在文本呈现评估部分，我们关注三个核心指标：生成文本与真实文本之间的编辑距离、一次视觉输出中的文本完成率，以及整体文本生成的准确性。关于知识和推理部分，我们的评估集中在模型是否具备必要的领域知识并能准确解读用户意图，从而生成语义连贯和逻辑一致的图像。对于多语

www.xueshuxiangzi.com

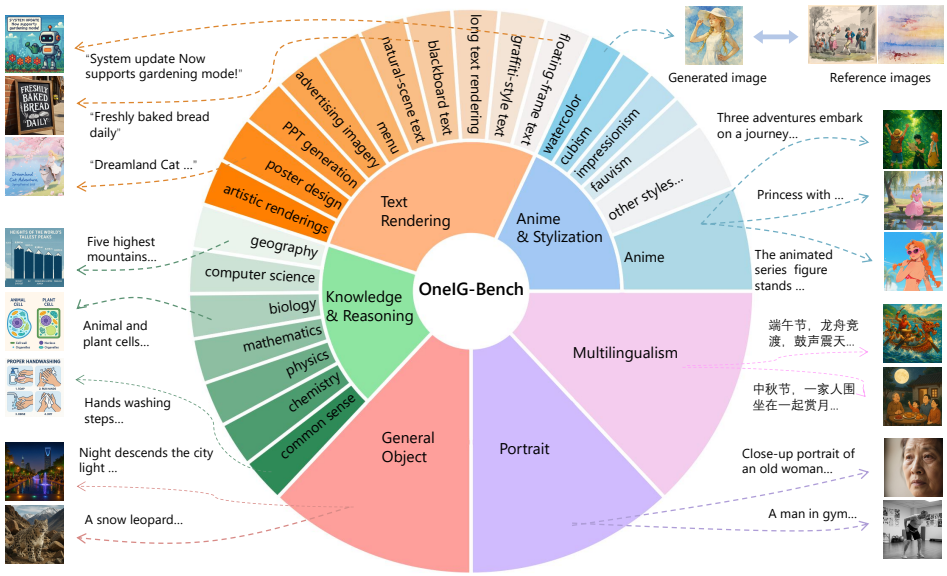

Figure 1: OneIG-Bench 概述。一 IG-Bench 包含六个核心类别，每个类别都旨在评估不同生成维度上针对性的能力，每个类别有大约 200 个精心策划的提示，以确保全面覆盖各种情境。

言部分，我们评估从文化元素提示生成的图像与相应文化元素之间的对齐程度。此外，我们还评估在所有六个维度上的多样性。

我们将我们的主要贡献总结为以下三点：

- 我们展示了 OneIG-Bench，它由六个提示集组成，其中前五个——245 个动漫和风格化，244 个肖像，206 个一般对象，200 个文本渲染，以及 225 个知识和推理提示——每个都提供英文和中文版本，还有 200 个多语言提示，旨在对当前的文本图像模型进行全面评估。
- 我们开发了一种系统的定量评估方法，通过标准化的指标促进能力的客观排序，从而实现不同模型之间的直接比较。具体来说，我们的评估框架允许 T2I 模型仅为与特定评估维度相关的提示生成图像，并相应地在该维度内评估性能。
- 我们基于提出的基准对最先进的开源方法以及专有模型进行评估，以促进文本到图像研究的发展。

1 相关工作

1.1 文本到图像模型

文本到图像（T2I）生成旨在开发生成与给定文本描述语义一致的图像模型。生成对抗网络（GAN） [?, ?] 时代的早期探索奠定了基础工作，但这些方法由于模式崩溃而面临显著局限，常常无法准确生成即使是简单的主体。在近年来，基于扩散模型范式的生成方法已成为主流趋势 [?, ?] 。值得注意的进展包括基于 Unet 的架构如 Stable Diffusion XL [?] ，基于扩散图像变换器（DiT）的模型 [?, ?, ?] ，双流 MMDiT 框架 [?, ?] ，以及结合双流和单流网络的混合设计 [?, ?, ?] 。这些模型在生成与文本语义高度一致的高质量图像方面取得了显著进展。随着技术进步，T2I 模型的评价框架必须从单一属性评价（例如颜色、形状）发展到多层次评价，涵盖语义对齐、风格一致性和文本渲染精确性。同时，基于自回归的模型 [?, ?, ?, ?, ?, ?, ?, ?] 在知识建模和复杂文本理解方面显示出独特优势，这需要在评估框架中纳入系统推理能力的评价。总之，T2I 生成的快速发展强调了建立全面而严格的评估体系的迫切需要，以准确衡量模型性能，识别优势和劣势，并促进该领域的可持续进步。

1.2 文本到图像评估

在文本生成图像的早期阶段，研究人员通常使用一些指标来评估图像生成的质量，例如 FID [?] （Fréchet Inception Distance）、SSIM [?] （结构相似性指数测量）、PSNR（峰值信噪比）等。然而，这些方法未能全面理解模型的能力，也未能捕捉模型理解高阶语义的能力。

www.xueshuxiangzi.com