中国科学技术大学

PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation

姓　　名：	刘宇阳
学　　号：	SC24219012
学　　院：	微电子学院

期刊名称：	IEEE Transactions on Visualization and Computer Graphics
完成时间：	2024 年 10 月 10 日

解决的问题

文本到图像的生成框架已成为一种流行且有效的交互式范式，被广泛应用在学术界中。自然语言文本的无限空间允许艺术思想的自由表达，并显著降低了图像创作的门槛。随着自然语言处理(NLP)和计算机视觉(CV)技术的快速发展，最先进的生成模型如 Stable Diffusion 和 DALL·E 2 能够根据文本提示生成相关且高质量的图像，并在下游任务中展现出巨大的潜力，例如超现实视频生成和放射学图像合成。

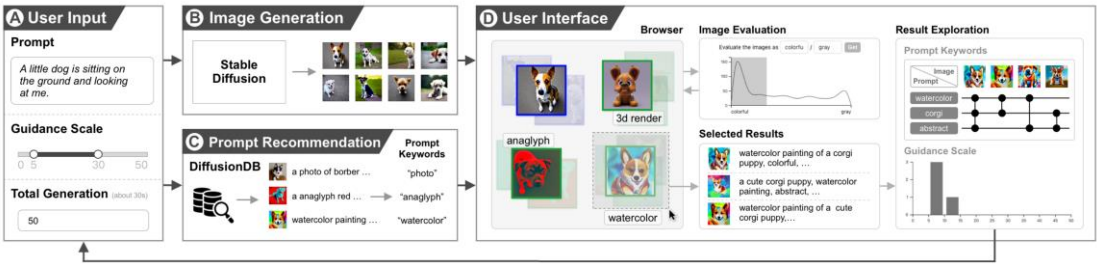
在以上模型成功的基础上，研究人员探索了一种被称为“提示”的人机交互技术。在创作过程中，用户通过自然语言提示来描述期望的图像特征（如主题和风格），调整模型的超参数（如引导比例）以获得所需的输出。然而，自然语言的复杂性和模糊性使得用户，尤其是初学者，难以开发出有效的提示，从而难以使模型生成所需的结果。此外，提示的不同超参数可能导致生成的图像存在显著差异。用户在试用有限的超参数值时，很难评估提示的质量。当生成的图像不符合预期时，用户可能会对如何调整提示或超参数感到困惑。虽然以往的研究提出了自动提示技术，但图像生成过程仍然高度依赖人类的主观判断，需要用户参与来优化生成过程。一些研究建议使用基于大型人工标注语料库的“魔法咒语”（如关键词）来制定提示，但这些指导原则可能过于通用，无法满足个性化图像创作的需求。

采用的方法

研究人员的工作针对文本到图像模型，结合数据库检索和即时生成，能识别有效提示关键词以进行个性化创作，并迭代优化提示。大量日常创建的图像为人工智能模型开发和内容检索等应用提供了丰富的信息，但总结和探索与图像相关的复杂语义是一个主要挑战，以往的研究集中于图像内容探索，提出了多种可视化技术，如节点链路图和增强散点图。而研究人员的工作则考虑了图像风格和模型超参数，利用预训练的视觉-语言模型 CLIP，帮助用户制定和优化提示，从而使用生成模型创建视觉吸引人的图像，并通过自然语言描述对图像进行评估和筛选。现代模型通常采用编码器-解码器架构，文本编码器为预训练的语言模型，而

图像解码器则使用生成对抗网络或扩散模型。尽管开源工具如 Stable Diffusion 和 DALL·E 为用户提供了创作便利，但生成质量依赖于用户的提示和主观判断，常常需要反复尝试不同措辞。为此，本文提出了一种交互式视觉分析系统，旨在通过总结和推荐关键词，帮助用户基于 DiffusionDB 优化其提示。

首先研究人员招募了 9 位参与者，通过不断的访谈与收集反馈，最终总结出系统的设计需求。设计的系统工作流程如下图所示。系统支持用户输入提示和模型超参数，包括引导比例范围和生成数量（针对不同的随机种子）。随后，系统利用这些提示和超参数生成图像集合（R1）。为了帮助用户改进提示，系统引入了一个提示推荐模型，从 DiffusionDB 中检索相似的创作结果（R1.2），并识别相关的提示关键词。生成的图像、检索到的图像和推荐的提示关键词根据语义共同嵌入到二维空间，并以多层次可视化呈现，以促进探索（R2）。基于此，系统允许用户指定美学评估标准（如美感），以高效评估和选择图像（R3）。用户可以选择感兴趣的图像子集，以进一步探索其提示关键词和引导比例，从而优化用户输入（R4）。



提示推荐模型从相似的图像创作中挖掘重要和相关的提示关键词。模型的流程如下图所示，包含五个步骤：

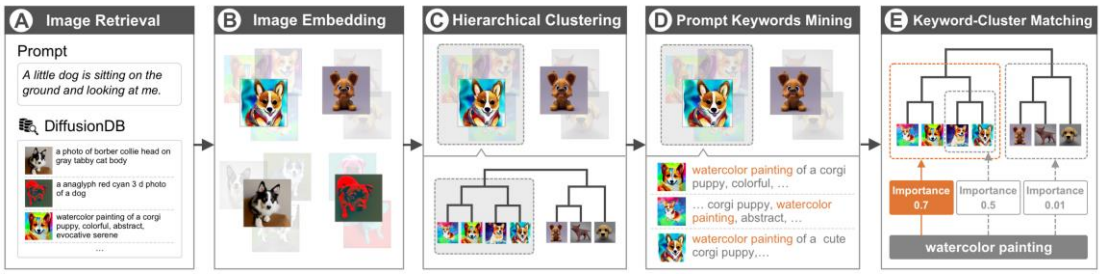
(A) 从 DiffusionDB 数据集中检索与用户输入提示相似的图像结果；

(B) 根据图像的语义特征对图像进行嵌入；

(C) 对图像进行层次聚类；

(D) 从图像聚类中识别重要和特殊的提示关键词；

(E) 将每个提示关键词匹配到其最相关的图像聚类。



为了检索与用户提示相似的图像，研究人员使用 DiffusionDB 中的图像及其原始提示作为搜索空间，并利用 CLIP 模型将图像和提示编码为 512 维向量，形成 1024 维的最终表示。为了减少视觉混乱并支持用户探索，研究人员采用 t-SNE 算法进行特征降维，并使用层次聚类将图像组织成树状结构。每个非叶子节点表示一个图像聚类，但只选择与提示关键词强关联的聚类，以确保有效挖掘关键词，并通过限制子节点数量和位置约束来避免冗余。再通过提示关键词挖掘与提示-聚类匹配。

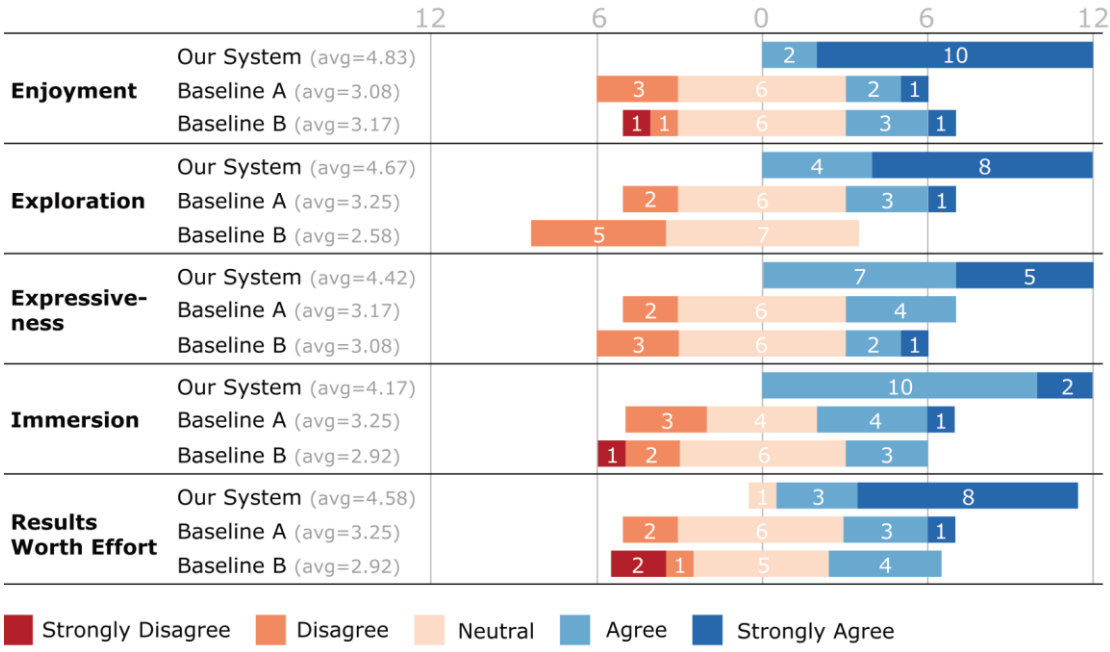
研究人员开发了一个视觉分析系统，利用 Stable Diffusion 模型和提示推荐模型，帮助用户进行交互式提示工程和文本到图像创作。用户界面包含四个视图：模型输入视图，图像浏览视图、图像评估视图、本地探索视图。

针对设计的系统，研究人员还进行了用户研究并和专家进行了访谈，以评估系统在促进交互式提示工程和图像创作方面的有效性和可用性。

得到的结论

图像评估视图能帮助大多数参与者评估和选择图像，本地探索视图帮助他们验证推荐的关键词，总之，所有参与者一致认为系统的工作流程直观，界面友好，易于学习和使用。

参与者还根据创造力支持指数对设计的系统和两个基准系统进行评分（不包含“协作”维度），如下图所示，总体而言，设计的系统在所有维度上都优于基准系统，这表明在促进文本到图像创作方面具有更高的有效性。



总的来说，本文介绍了 PromptMagician，一个用于文本到图像创作的交互式提示工程的视觉分析系统。该系统帮助用户生成和探索一系列图像结果，并迭代优化输入提示。研究人员设计了一个提示推荐模型，从 DiffusionDB 中以语义为基础进行检索和层次关键词提取，推荐与用户提示相关的重要提示关键词。提示关键词及其对应的图像在二维空间中共同嵌入，以便于交互式探索，并支持个性化评估。然后研究人员展示了系统的两个使用场景，并进行了用户研究和专家访谈，以验证其有效性和可用性。最终结果不仅表明 PromptMagician 推荐了有用的提示关键词并促进了交互探索，还为设计和改进文本到图像创作的提示方法及视觉系统提供了新的见解。