arXiv:2304.09337v1 [cs.HC] 18 Apr 2023

Promptify：借助大语言模型，通过交互式提示词探索实现文本到图像生成

Stephen Brade 多伦多大学 加拿大多伦多
stephen.brade@mail.utoronto.ca

Bryan Wang 多伦多大学 加拿大多伦多
bryanw@dgp.toronto.edu

Mauricio Sousa 多伦多大学 加拿大多伦多
mauricio@dgp.toronto.edu

Sageev Oore 达尔豪斯大学 加拿大哈利法克斯
sageev@dal.ca

Tovi Grossman 多伦多大学 加拿大多伦多
tovi@dgp.toronto.edu

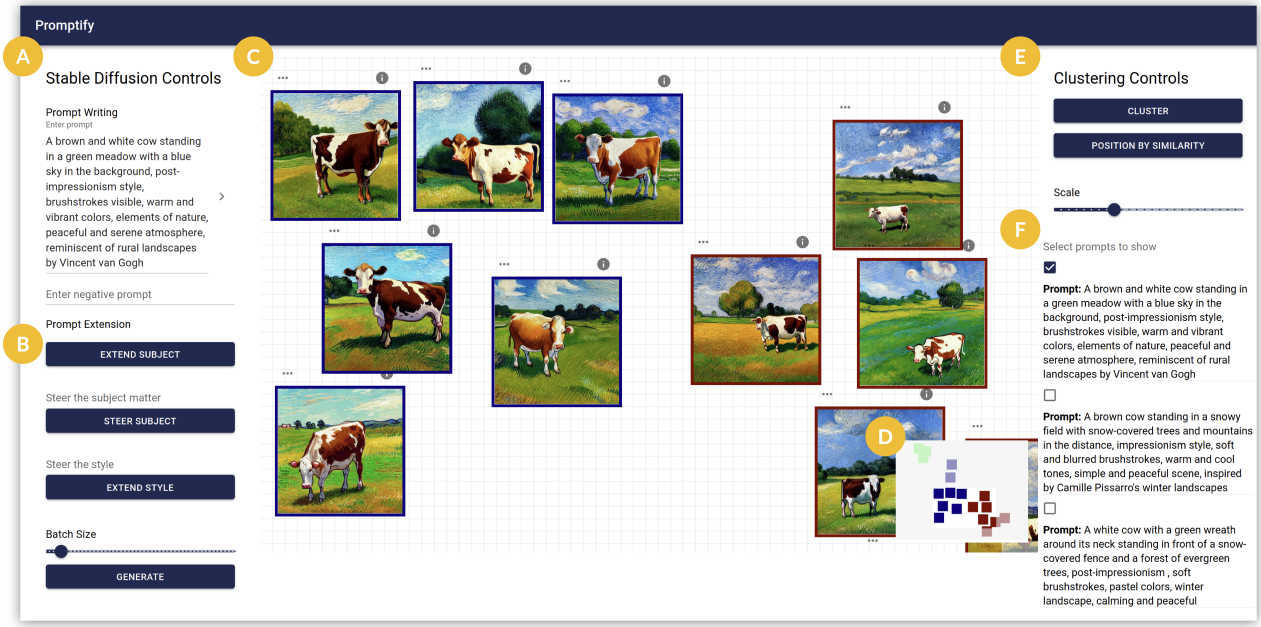

图 1：Promptify 系统界面。A）用于编写提示词的 Stable Diffusion 控件。B）自动提示词建议，帮助用户构思主题内容，并通过自然语言以可控的方式获取风格关键词建议。C）图像布局与聚类功能允许用户查看、整理和聚类生成图像。用户可以通过拖动、放置和缩放来操作图像与聚类。他们还可以根据单张图像或聚类获取提示词优化建议。D）小地图为聚类提供视觉提示，并帮助用户浏览可视化内容。E）聚类控件用于按相似性定位和聚类图像，并通过“Scale”滑块调整图像间距。F）提示词历史记录允许用户通过提示词历史记录菜单切换显示或隐藏所有先前的提示词，以查看或移除图像。

摘要

文本到图像生成模型已展现出根据文本提示词生成高质量图像的卓越能力。然而，编写能够准确体现用户创作意图的提示词仍然很有挑战性。为确保模型对提示词的理解符合用户意图，往往需要经历费时的反复试错过程。为应对这些挑战，我们推出了 Promptify，这是一个支持文本到图像生成模型进行提示词探索与优化的交互式系统。Promptify 利用一个由.....驱动的建议引擎借助大语言模型，帮助用户快速探索并构思多样化的提示词。我们的界面允许用户灵活组织生成的图像，并根据用户偏好，Promptify 会建议对原始提示词进行潜在修改。这一反馈循环使用户能够迭代优化提示词，强化所需特征，同时避免不需要的特征。我们的用户研究表明，Promptify 能够有效促进文本生成图像工作流，使用户首次尝试便能创建出具有视觉吸引力的图像，同时比广泛使用的基线工具显著降低认知负荷。

关键词
文本生成图像、提示工程、大语言模型

1 引言

Stable Diffusion (SD) [43]和 DALL-E-2 [39] 等文本到图像生成模型能够根据自然语言描述生成高质量图像。这些模型可以支持多个领域的创作过程，例如创作新闻插图 [26] 以及为工业设计构思创意 [22]。然而，近期研究发现，撰写提示词——这是引导图像生成的主要方式——可能是一项颇具挑战的任务 [10, 22, 26]。提示词编写之所以困难，是因为用户需要学习如何撰写提示词，使模型能够按照符合其预期创作成果的方式进行解读。这通常涉及反复试错：用户撰写提示词、观察生成输出，并在必要时完善提示词。当前模型缺乏内置支持，无法帮助用户发现有助于提升输出质量和探索其他可能性的实用关键词，这进一步增加了这一过程的难度。

以往的提示工程研究通过提供有效的提示词编写策略来应对这些挑战[24, 25, 36, 44, 56, 54]。[25]

例如，Liu 和 Chilton 建议优先识别主体和风格的关键词，并发现对相同关键词采用不同措辞并不会带来显著差异。此外，r/StableDiffusion 子版块[1] 等在线社区也已形成，供用户分享提示词示例和撰写策略。

尽管这些研究提供了宝贵见解，但它们只提出了较为宏观的建议，无法为追求特定美学效果的用户提供个性化推荐。因此，缺乏提示词编写经验且不熟悉相关关键词的新手用户，仍可能难以实现预期效果。为更好地了解提示文本到图像模型的实践与挑战，我们对在线社区的六位活跃成员进行了形成性用户访谈。参与者指出了若干共性，包括社区资源对学习如何编写提示的重要性、提示词优化的迭代性质，以及生成大型图像集合以获取灵感和进行探索的必要性。

本文介绍了 Promptify，这是一个交互式系统，旨在促进针对 Stable Diffusion（SD） [43], 这一热门开源文本到图像扩散模型的迭代式提示词探索与优化。Promptify 通过支持多个步骤来促进迭代工作流程，包括构思主题内容、编写风格修饰词、生成和浏览图像，以及优化原始提示词（图 1）。

使用我们的系统时，用户首先输入一个主题，Promptify 会建议不同的构想来拓展该主题。用户还可以进一步调整主题建议，以探索其他创意。接下来，用户输入简短的风格描述；Promptify 随后会扩展相关修饰词 [2]，以更清晰地表达风格。提示词建议引擎能够根据用户简短的输入生成复杂的提示词，从而生成视觉效果出色的图像。Promptify 会在图像布局与聚类界面中展示生成图像，让用户灵活地整理和浏览生成的图像集。最后，Promptify 会根据生成图像提供关键词建议，帮助用户完善原始提示词，强化期望的特征并避免不需要的特征。

我们使用针对聊天优化的大语言模型（LLM）GPT-3 开发了 Promptify 的提示词建议引擎。5 [7], 该引擎允许用户通过自然语言指令反复改进主体内容建议。对于风格描述，我们采用少样本提示技术，引导大语言模型生成模仿社区用语的关键词建议。

为评估 Promptify 的有效性，我们开展了一项有 14 名参与者参加的用户研究，将 Promptify 与 Automatic 进行比较；Automatic 是社区广泛使用的一款 Stable Diffusion 工具。参与者认为，与基线工具相比，Promptify 在辅助文本到图像生成工作流程方面更有用，并且首次尝试时就能创作出更具视觉吸引力的图像。Promptify 还帮助用户管理和比较大量图像，以识别并强化所需特征，同时在后续迭代中忽略不需要的特征。此外，用户报告称，使用 Promptify 时心理需求和挫折感显著降低。综上所述，我们的工作作出了以下贡献：

- Promptify 的设计与实现：这是一个全栈系统，支持交互式提示词探索与优化，旨在帮助新手用户进行文本到图像生成。
- 一种基于大语言模型的提示词建议引擎，可协助用户构思主题内容，并根据社区数据为文本提示词推荐风格描述。
- 一个用于按相似性对生成图像进行布局与聚类的界面，使用户能够发现主题趋势并优化提示词，从而改进后续生成效果。
- 一项涉及 14 名参与者的用户研究结果，展示了 Promptify 系统的有效性及其相较于常用基线工具的优势。

2 相关工作

2.1 文本生成图像生成模型

文本到图像生成模型是一类机器学习模型，它们根据输入的语言描述生成与之匹配的图像 [28, 32, 39, 40, 42, 43, 58]。Align-Draw [28] 是最早的一批模型之一，它学习估计文本与生成画布之间的对齐关系。Reed 等人使用卷积生成对抗网络（GANs）根据文本描述为鸟类和花卉生成逼真的 64×64 [42] 128x128 [41] 图像。StackGAN [58]通过堆叠两个 GAN 进一步提高了生成分辨率至 256×256：一个用于生成初始形状和基本颜色，另一个用于超分辨率。除基于 GAN 的模型外，OpenAI 还开发了 DALL-E [40],，这是一种基于 Transformer 的模型，也是最早引起广泛公众关注的文本到图像模型之一。其后继模型 DALL-E-2 [39]使用 CLIP（对比语言-图像预训练）嵌入 [38] 和扩散模型 [46]，生成更精细、分辨率高出 4 倍的图像。SD [43] 使用潜在扩散模型生成 512x512 图像，并向公众开放源代码。Promptify 以 SD 为基础，因为它广受欢迎且易于获取。

[1] r/StableDiffusion：https://www.reddit.com/r/StableDiffusion/ [2] 本文中，提示词的关键词与修饰词可互换使用。

2.2 提示工程

近年来，使用文本作为输入并生成输出的生成模型迅速发展，其中包括 GPT 系列等自回归大语言模型 [13, 34], ，以及 SD 等文本到图像模型 [43]。输入文本称为提示，它引导模型生成特定的输出。SD 等模型还使用负面提示，引导模型避开不理想的属性。提示工程 [24, 25, 36, 44, 50, 53, 54] 指构造提示以有效地与模型沟通，并引导生成结果达到预期目标的方法。这些方法因模型模态不同而有所差异，例如文本和图像。比如，少样本提示 [13, 24, 60] 是大语言模型中一种重要的提示技术。这种技术会向模型提供某项任务的一小组示例数据，例如翻译，以引导模型学习如何执行该任务（上下文学习）。对于文本到图像模型，相关研究主要关注识别关键词或关键短语 [36] ，并测试不同措辞如何产生不同的输出 [25]。3DALL-E [27] 使用 GPT-3 为生成三维设计提供关键词建议。Wang 等人 [50] 提出了一种自动方法，用于优化文本提示，以便让生成图像呈现准确的情感表达。Prompt-to-Prompt [21] 通过编辑初始提示来支持图像编辑。CLIPInterrogator [37] 是一款开源工具，它尝试通过生成可能与给定图像相匹配的文本提示词，对图像进行逆向分析。Promptify 在既有研究的基础上提出了一种新颖的提示方法，将大语言模型和文本到图像模型的提示工程技术联系起来。我们采用少样本提示，引导大语言模型协助 SD 模型的提示工程。Promptify 还集成了 CLIPInterrogator [37] ，帮助用户识别可能提升生成效果的关键词。

2.3 设计探索与语义编辑

计算机图形学和人机交互领域的大量研究都致力于帮助生成、管理和探索设计变体 [14, 29, 30, 45, 48, 55–57]。Design Galleries [29]利用可视化帮助用户整理不同图像渲染参数所生成的结果。Dream Lens [30]提供了一种可视化方式，用户可以叠加生成的设计，并使用凿子工具移除不需要的元素。GEM-NI [57] 是一款基于图的界面生成式设计工具，可让用户同时探索各种创意备选方案。其他研究则专注于为视觉内容生成创建语义控制 [14, 56] ，或利用降维技术直观地控制高维设计空间 [45, 55]。Promptify通过提供支持用户探索文本生成图像领域中各种变化的工具，为这一领域作出了贡献。

2.4 图像生成的用户界面

随着生成模型能够生成高质量图像，人们越来越关注如何设计界面，将这些模型的强大功能带给用户 [17, 18, 26, 27]。例如，GANzilla [17] 利用散布/收集技术，让用户能够反复探索不同方向并实现编辑目标。在作者后续的研究中，他们介绍了 GANravel [18] 它提出了一种由用户驱动的方向解缠工具，允许用户反复改进编辑方向。Opal [26]是一种交互式系统，旨在帮助用户利用文本到图像生成技术创作新闻插图。3DALL-E [27] 将设计师的意图转化为多模态（文本和图像）提示词，从而生成 3D 设计灵感。学术研究之外的贡献包括开源工具，例如 GANInterface [33], 、Easy Diffusion [15], 和 AUTOMATIC1111 的 Stable Diffusion WebUI [11], ，在本文后文中我们将其称为 Automatic1111。Promptify 通过提供一种新颖的界面，为这一研究领域增添了新的贡献：该界面根据 CLIP 嵌入的相似性对图像进行定位和聚类，提供比传统文件夹视图更高效的浏览体验。

3 形成性访谈

为了了解 Stable Diffusion 用户生成理想图像时的现行做法及其局限，我们采访了六位曾多次在 r/StableDiffusion 发帖的 Reddit 用户。我们尤其关注他们刚开始使用 Stable Diffusion 时的经历，因为用户通常会在这一阶段遇到最大的挑战。访谈通过视频会议进行，每次约半小时，参与者获得了 20 加元的报酬。参与者的经验水平各不相同。其中两位自述为数字艺术专业创作者（P1、P4），两位是业余爱好者（P2、P3），另两位则是新手（P5、P6）。所有参与者都表示，他们在个人图形处理器上使用 Automatic1111。参与者解释说，Automatic1111 之所以受欢迎，是因为它易于安装，并支持多种开源工具，这些工具可用于改进和处理生成图像。下文将讨论参与者面临的常见挑战，以及他们为克服这些挑战所采用的策略。

3.1 对初学者而言，提示词编写十分困难

不同经验水平的参与者都表示，作为初学者，要写出有效的提示词非常困难。为角色扮演战役创作视觉内容的 P2 提到，他们花了三个月才做出令自己满意的作品。此外，转入新的战役可能需要花上好几个晚上反复尝试提示词，才能取得满意的结果。P4 也谈到了为新的主题和风格生成图像时遇到的类似情况，并表示“对于每一种类型的输出，我的流程都会有所不同。”P2 和 P5 还指出，了解 Stable Diffusion 所用的训练数据很有帮助，因为这能让人对哪些关键词有效形成直觉。然而，这也为缺乏技术背景的用户设置了障碍。

由于难以发现有效的提示关键词，用户不得不向在线社区寻求指导。P6 指出“自 *SD* 发布以来，我一直在 *Reddit* 上关注它，并尝试模仿大家在那里所做的事情”。P5 提到，Reddit 上的用户“花了很多努力记录哪些方法有效、哪些无效”，并分享提示词和其他超参数。P4 提到，他们通过 YouTube 视频和 Discord 论坛学习，并从 Midjourney 转向 Stable Diffusion [6] “因为 *Stable Diffusion* 是开源的，而且大家会分享自己的提示词”；而 P2 在对提示不满意时“*[会]* 在 *Discord* 上求助”。这表明，参与者通过查看和模仿他人发布的

提示词，或在自己的提示词未能取得满意结果时搜索相似的提示词，提升了自身能力。一些工具允许用户检索提示词，例如*lexica.art* [4]，，或生成提示词，例如*MagicPrompt* [5]。然而，P3 评论说，这些模型往往会生成具有特定风格的提示词，因此他们不得不使用正则表达式，筛除包含与目标风格相关关键词的提示词，这十分耗时。

3.2 迭代实验是必要的

参与者（P1、P2、P6）表示，要获得令人满意的结果，往往需要反复撰写并完善提示词，尤其是在他们希望偏离社区的风格趋势时。例如，P2会维护一个收录有效关键词的存储库。为了扩充这个集合，P2会生成两组图像：一组使用包含新关键词的提示词，另一组则不包含该关键词，以确定它对生成结果的影响。其他参与者则描述了他们反复完善提示词的过程，以强化偏好的结果并抑制不希望出现的结果。例如，P1提到，当他们希望主体有一头粉色头发，但Stable Diffusion生成的主体却是绿发时，他们可能会在负面提示中添加 “绿色” 或 “绿头发” 。其他人指出，Stable Diffusion 可能会产生超出预期提示词的意外惊喜。P2谈到，在创建随机的非玩家角色时，他们乐于迎接偶然发现。他们利用Stable Diffusion生成大量对模糊概念的不同演绎，并从偏好的诠释中汲取灵感来完善提示词。不过，他们也指出，要完善提示词，使生成结果与先前生成内容的风格一致，并非易事。

3.3 管理模型随机性是一项挑战

参与者提到，管理 SD 根据单个提示生成图像时的随机性是一项困难。P3 表示，这种随机性可能“让我抓狂，因为我可能会失去我一直在寻找的本质。”P6 指出，“当提示过于简单时，*[*生成图像*]* 往往会在不同风格之间跳转。”而 P5 认为，“随机性是图像质量的重要因素。”这凸显了意外发现在文本到图像生成中的重要性。为克服这些挑战，所有参与者都提到会生成大批量图像（30-50 张），以找到具有所需主题内容和风格、同时避免异常的图像。在进行迭代实验时，使用性能较弱 GPU 的用户会采用较小的批次（2-10 张）来完善提示词，只有在取得满意结果后才生成大批量图像。参与者克服模型随机性的策略印证了 Liu 等人 [26] 的发现：生成更多图像能够为用户带来更好的结果，并帮助他们理解人工智能能力。

4 设计目标

在形成性访谈的指导下，我们确立了三个目标，以指导 Promptify 的设计与开发，从而促进使用文本到图像模型创建图像的工作流程。我们的系统旨在服务于新手用户，因为研究发现，在使用这些模型时，该用户群体面临着最为显著的挑战。

D1. 协助用户进行提示工程。有效的提示词编写或提示工程已在我们的形成性研究和先前研究中被确定为一项重大难题 [25, 44]。Promptify 旨在通过协助用户生成潜在的主题内容，并建议有助于实现其创意目标的有效关键词，来缓解这一挑战。

D2. 支持文本到图像生成的迭代实验。Promptify 应支持迭代工作流程，包括构思主题内容和风格、构建能够体现用户创作意图的有效提示词，以及根据用户需求，基于先前生成的结果优化初始提示词。

D3. 促进对生成图像的检查与整理。与以往研究一致， [26]，我们形成性研究的参与者强调了意外发现的重要性，以及生成大量图像以获得最佳结果的必要性。因此，Promptify 旨在帮助用户比较、对照和管理大量图像。

5 PROMPTIFY 系统

Promptify 通过提供三项主要功能改进文本到图像生成工作流程，包括 *1*）自动扩展提示词并提供建议；*2*）根据相似性进行图像布局和聚类；*3*）自动提供提示词优化建议。通过整合这些功能，Promptify 建立了一个反馈循环，使用户能够根据初始提示词生成高质量图像，检查模型如何解读其提示词，并获得编辑初始提示词以改进输出的建议。图 2 展示了使用 Promptify 的用户工作流程。接下来，我们将介绍各项功能，并将其与第 4 节所述的设计目标联系起来，以阐明其设计依据。

5.1 自动提示词建议

使用文本到图像模型的一项关键挑战在于构建提示，使模型能够正确理解用户的创作意图。为支持提示词编写（**D1**），我们开发了一个提示词建议引擎，如图 3 所示。系统中的提示词建议过程分为主题扩展 和风格扩展，分别基于一个原子输入进行扩展。在此语境中，原子输入指对主题或风格的基本描述。请注意，用户可以手动编辑系统提供的任何建议，也可以根据需要灵活选择不进行其中任一种扩展。

5.1.1 主题构思。用户首先在提示词编写字段中输入一个原子主题，例如 “狮子”。随后，他们可以点击“扩展主题”按钮来扩展主题，Promptify 会为提示添加更多细节。系统会显示一个弹出菜单，其中包含三个建议的扩展选项。例如，用户可能会收到这样的建议：“一头狮子站在岩石露头上，俯瞰广阔的沙漠景观，背景中有一轮落日。”如图 1-B 所示，用户还可以在“引导主题内容”文本字段中提供更多指令，进一步完善提示。例如，他们可以请求生成冬季主题提示词，

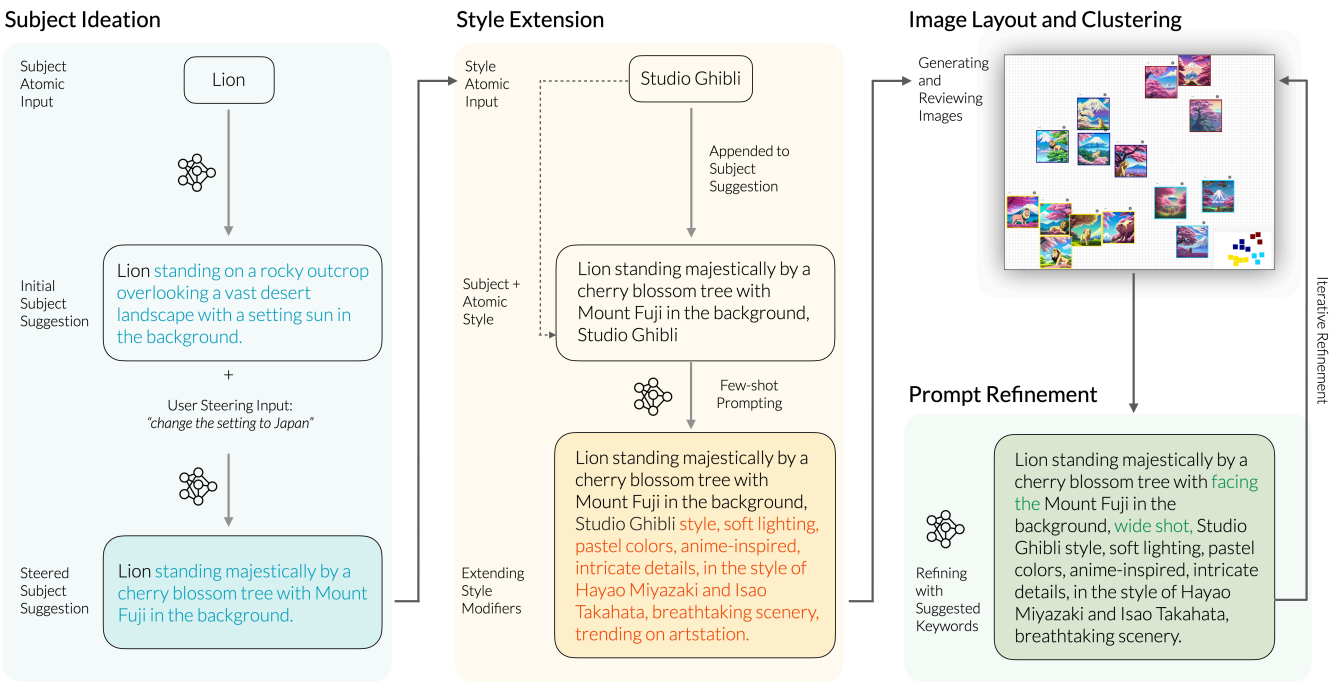

图 2：用户使用 Promptify 系统的工作流程。用户首先输入一个原子主题（例如 狮子），通过可引导的主题构思功能获取各种主题创意。接下来，用户输入一个原子风格，例如 “吉卜力工作室”，并使用风格扩展功能，按照所需的主题内容扩展提示中的风格部分。最后，用户生成图像，按相似性排列并聚类，从而轻松发现图像之间的主题。为了进一步完善作品，用户可以通过 CLIP Interrogator 搜索针对每张图像和每个聚类建议的修饰词，以强化所需特征，并弱化不想要的特征，为下一轮迭代做准备。用户也可以随时跳过 Promptify 的建议功能，自行编写提示的部分内容。神经网络图标表示该组件由大语言模型辅助。

输入 “将场景改为日本。”。随后，系统会生成一个融合用户指令的新提示词，例如“狮子威风凛凛地站在樱花树旁，背景是富士山。”

5.1.2 风格扩展。 用户选定所需的主题后，可以在“风格引导”字段中输入简短的风格描述，用于概括他们希望创作的高层次风格。随后，Promptify 会将该风格应用于每个推荐主题，并为提示补充更多相关细节。例如，如果用户输入“吉卜力工作室，”Promptify 会将提示扩展为“吉卜力工作室风格、柔和光线、粉彩色、动漫风格、精致细节、宫崎骏和高畑勋的风格、壮丽景色、*ArtStation* 热门”。我们采用 {主题，风格修饰词 1，风格修饰词 2.....} 的提示格式，以主题关键词开头，后接多个风格修饰词 [36, 51]。风格扩展功能旨在生成类似于 Stable Diffusion 在线社区所使用语言的提示建议，从而提高生成高质量图像的可能性（**D1**）。

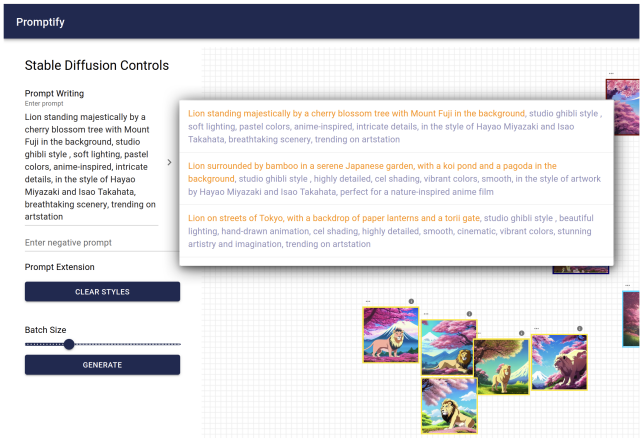

图 3：Promptify 的提示词建议菜单。建议的主题短语以橙色显示，而风格修饰词以紫色突出显示。第一条建议会自动填入提示词输入框。用户还可以进一步编辑建议的提示词。

5.2 图像布局与聚类

完成提示词编写后，用户可以使用 Promptify 批量生成图像。我们的系统通过将生成图像可视化呈现在二维交互式可缩放画布上，支持用户对其进行查看和管理（**D3**）。如图4所示，Promptify 允许用户重新排列生成图像，并根据相似性对其进行聚类。这些图像簇会在小地图上以不同颜色加以区分，为用户提供视觉提示。此外，用户还可以使用缩放滑块调整图像之间的距离，以避免产生干扰性的重叠。通过拖动和重新排列图像，用户可以比较不同图像的各个方面。该界面还支持跨多轮迭代进行查看（**D2**）。如果用户喜欢某个图像簇，可以将其记住，以便进行下一轮生成，并关注位于该图像簇附近的新图像。图像布局与聚类功能帮助用户找到所需图像，忽略不需要的图像，并发现生成图像中的主题元素。该功能有助于进一步编辑和完善提示词。

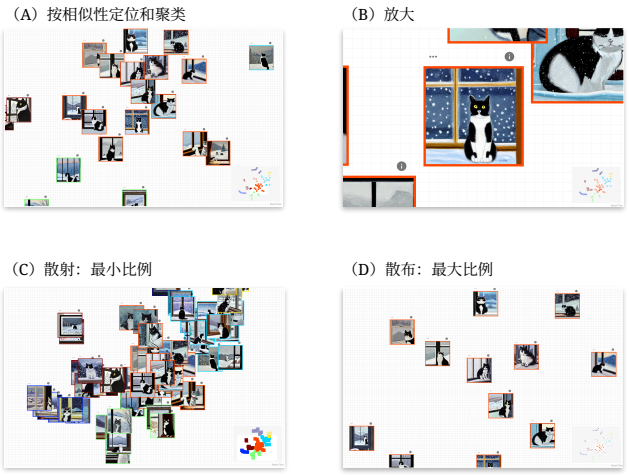

图4：Promp-tify 界面支持的图像浏览技术。（A）按图像相似度定位并聚类；图像的边框颜色表示其所属聚类，并显示在小地图上。（B）用户可以放大查看完整尺寸（512x512）的生成图像。（C）图像布局的最小比例。图像聚集在一起，彼此重叠，留白极少。（D）图像布局的最小比例。图像分布稀疏，彼此不重叠。

5.3 根据先前生成结果推荐修饰词

生成图像后，用户可以通过编辑提示词，在后续生成中增强或避免某些特征。Promptify 通过提供细化修饰词来支持这一做法；这些修饰词是与生成图像及其对应聚类相关的关键词或关键短语（**D2**）。如图5所示，用户可以点击图像上的信息按钮来查看修饰词，随后会显示一个包含三个可折叠子菜单的弹出菜单：一个列出与该图像相关的修饰词，一个列出与图像所属聚类相关的修饰词，另一个列出该聚类独有的修饰词。用户可以点击“添加”按钮来添加其中任意提示词，该操作会自动将修饰词纳入提示词中。推荐的修饰词旨在为用户提供与特定图像或聚类主题相关的描述性词语和短语，帮助用户调整输入提示词，以强化偏好的特征或弱化不希望出现的特征。

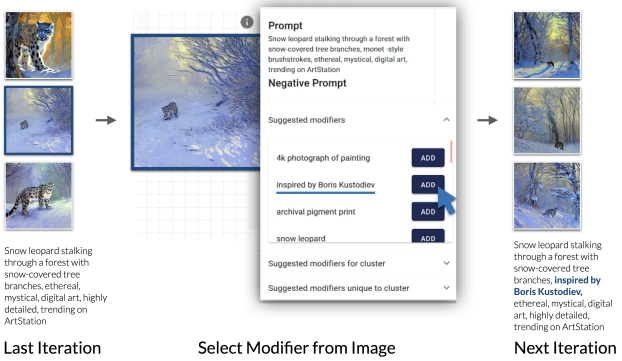

图5：我们的研究参与者使用 CLIP Interrogator 的关键词建议来优化提示词，使其聚焦于特定风格。用户根据前一次生成结果中间的图像添加了建议的修饰词，从而生成更多具有该风格的图像。

6 后端实现

6.1 自动提示词建议

提示词建议引擎通过向由 OpenAI 开发的名为 GPT-3.5（ChatGPT）的大语言模型发送提示来实现 [

7]。主题建议与风格建议的实现方式有所不同：前者通过零样本指令完成，后者则借助社区数据进行少样本提示。我们的大语言模型方案具有灵活性；未来若有适用于主题/风格扩展的零样本或少样本提示词证明有效，便可立即集成到系统中。每项大语言模型功能所用的提示词均可在补充文件中找到。

6.1.1 主题构思。 我们采用零样本提示，引导大语言模型生成用于扩展原子主题的建议。零样本是指除指令外，不向模型提供具体示例。我们发现，只需给出提示，大语言模型就能很好地构思出各种主题建议

仅提供指令。提供额外示例可能会引入偏见并缩小可能建议的范围，这不利于我们所需的探索性构思。为了让大语言模型了解文本到图像生成，我们在提示中提及 DALL-E [40], , 这是一个知名的文本生成图像模型，其出现时间早于 GPT-3.5 于 2021年9月 的知识截止时间。我们指示模型针对给定的原子主题（例如，“一头奶牛”）生成三个详细的提示建议。为了让用户能够通过语言引导

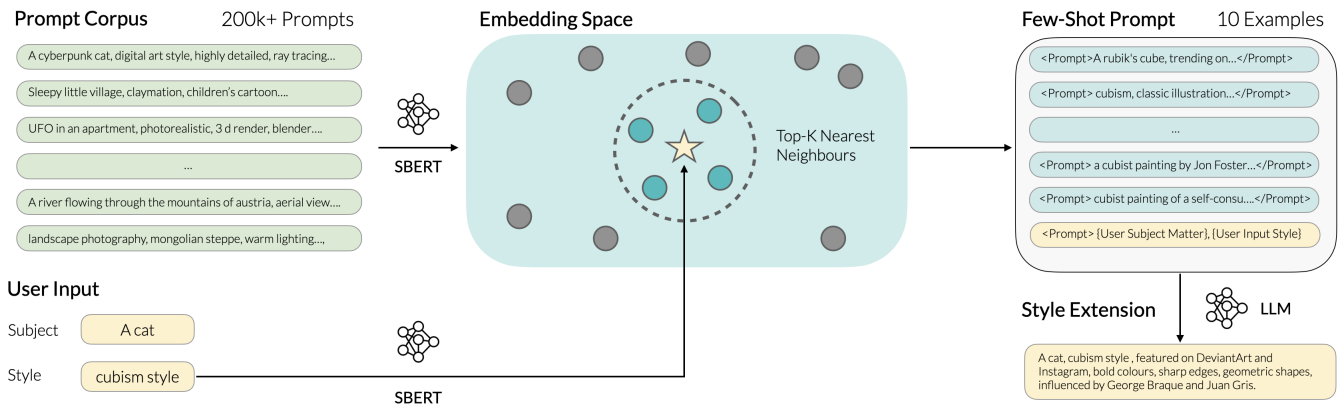

图 6：我们用于风格扩展的少样本提示流程示意图。我们使用 Sentence-BERT 嵌入预先对提示词语料库进行嵌入。当用户输入一种原子风格以请求扩展时，Promptify 会对输入进行嵌入，并将其作为查询执行 K近邻搜索，从语料库中检索语义相似的示例。随后，我们将检索到的示例和用户输入用于构建少样本提示词，供大语言模型执行风格扩展。

建议，我们会记录生成初始建议集的查询和回复。当用户提供要求修改的额外指令时，我们会将这些指令与初始指令一并发送给大语言模型，以生成新的建议。

6.1.2 风格扩展。我们使用少样本提示来教大语言模型如何根据用户输入的原子风格扩展提示的风格。与零样本学习相比，少样本提示会使用一小组特定任务示例，引导大语言模型生成遵循相似模式的输出。这种方法有助于我们设计风格扩展功能，推荐社区常用语言中的有效风格修饰词 [31] （**D1**）。

用户输入原子风格并请求扩展后，我们的系统会构建一个少样本提示词，其中包含从社区共享的 Stable Diffusion 提示词语料库中采样的 10 个与输入风格相关的提示词示例 [51]。这些示例通过基于句子相似性的 K近邻算法进行采样，具体如下。我们首先从 DiffusionDB [51] 数据集中随机采样 500k 条提示词，构建提示词语料库。每条提示词数据都带有一个不宜在工作场所查看（NSFW）评分，我们过滤掉评分大于 0.1 的数据。为了引导模型按照 主题内容、风格修饰词 *1*、风格修饰词 *2*.....格式生成提示词建议，我们仅选择至少包含 6 个修饰词（以逗号分隔的短语）的提示词。经过筛选，最终得到包含 238584 条提示词的语料库。随后，我们使用 Sentence-BERT [9] 将所有提示词嵌入为句子嵌入。当用户输入诸如 印象主义这样的原子风格时，Promptify 会使用相同的嵌入模型，将用户输入的短语（即查询）进行嵌入。然后，它使用 K近邻算法（k=10）检索语料库中与查询具有最高余弦相似度的 10 个示例。Liu 等人 [24], 的先前研究证明了 K近邻方法的有效性：在嵌入空间中选择更接近查询的示例，能够持续提升少样本示例选择的性能。如图 6 所示，少样本提示词由 10 个查询示例组成。在此基础上，用户当前的主题内容和原子风格将按照指定格式与这些内容拼接，并追加到少样本示例的末尾。随后，将最终提示词输入大语言模型，由其在少样本示例的引导下自动补全其他风格修饰词。

6.2 图像布局与聚类

图像布局与聚类功能基于 CLIP 嵌入实现。CLIP [3] 是一种神经网络，使用各种（图像、文本）配对数据进行训练，旨在获得语言与图像的联合多模态嵌入空间。我们选择 CLIP 嵌入，是因为它也用于 Stable Diffusion 和 DALL-E 2 等扩散模型中的去噪过程条件控制。为了实现图像布局与聚类，我们首先计算画布上所有生成图像的 CLIP 图像嵌入。接下来，我们使用 TSNE [49]将嵌入降至二维，以便轻松地可视化图像。降维后的二维嵌入将作为绘制图像的坐标，为用户直观呈现其图像集合。为防止图像重叠，我们对这些坐标进行缩放，使图像之间的最小距离为 128 像素。我们使用亲和传播 [19], 根据图像在屏幕上的位置对其进行聚类；这种方法无需用户预先确定聚类数量。缩放滑块允许用户在画布上拉开图像间距以避免重叠，其实现方式是将图像坐标乘以滑块值；滑块默认值为 1，范围为 0.5 到 3。

6.3 从生成图像中提出修饰词建议

6.3.1 CLIP Interrogator 提示词修饰语建议。 我们使用 CLIP Interrogator 为生成图像添加一系列提示词修饰语，供用户选择，以完善上一轮迭代中的提示。CLIP I nterrogator 是一种开源工具，它利用 BLIP [23] 为图像生成图像说明，随后识别出与图像相关的一系列高分关键词。这些关键词选自一个包含

约 100,000 个单词和短语以及 5,000+ 个艺术家姓名 [53]的语料库。我们选择 CLIP Interrogator，是因为它在用于 Stable Diffusion[53]时，能够有效生成可复现图像主题内容和风格的提示词。CLIP Interrogator 的一个局限是，它依赖于范围有限的艺术家和关键词，可能无法涵盖所有可能的艺术家。

6.3.2 整合选定的修饰词。当用户选择 CLIP Interrogator 推荐的提示词修饰语时，Promptify 会将其整合到提示中。此整合由大语言模型完成，方法是使用包含八个经人工整理的关键词整合示例的少样本提示词。我们编写这些示例，指导大语言模型在修改提示时避免冗余。例如，如果源提示提到“一头棕色的奶牛，”而 CLIP Interrogator 提供的修饰词建议为“一头奶牛，背景中有一座红色谷仓，”那么这个少样本提示词会将两者整合为“一头棕色的奶牛，背景中有一座红色谷仓。”每个少样本示例都由三个部分组成：从扩散数据库获取的源提示、从 CLIP Interrogator 获取的修饰词，以及两者的整合结果。当用户选择一个修饰词时，Promptify 会按照相同格式，将用户当前的提示和该修饰词附加到少样本示例中，让大语言模型自动补全整合结果。八个少样本示例中，有一半涉及与主题内容相关的 CLIP Interrogator 提示词修饰语，另一半则与风格相关。我们加入这项自动整合功能，以减轻用户的认知负担；否则，用户需要自行辨别冗余内容并编辑提示词。

6.4 系统架构

我们使用 Python Flask 和 React Flow [8]实现了 Promptify 全栈系统。该系统运行在一台配备 Nvidia Titan RTX GPU 的 Linux 机器上，使用 SD 生成一张 512x512 图像需要 4 秒。我们通过 OpenAI API[7] 使用 GPT-3.5-Turbo 模型提供提示建议和编辑功能。我们使用 Sentence- Transformers pythonlibrary [9] 中的 all-MiniLM-L6-v2 模型进行句子嵌入。我们使用 GitHub 上各自开源的代码，获取用于图像布局与聚类的 CLIP 嵌入以及 CLIP Interrogator。[3, 37]我们使用 scikit-learn 进行亲和传播。

7 用户评估

我们开展了用户评估，以评估 Promptfiy 在实现促进文本生成图像工作流程这一设计目标方面的有效性。研究旨在深入了解用户如何感知系统的各项功能并与之交互，同时识别潜在的障碍。用户评估分为 3 个部分：比较评估、自由形式使用观察和半结构式访谈。对于比较评估，参与者同时使用 Promptify 和常用的开源界面 Automatic1111。随后，他们自由使用 Promptify 创建图像。最后，我们开展半结构化访谈，以了解他们使用 Promptify 的体验。Promptify 和 Automatic1111 均安装在一台配备 Nvidia Titan RTX GPU 的台式计算机上。我们使用 Stable-Diffusion-v1-5 检查点，并将所有超参数设为相同值，包括 50 个去噪步数、7.5 的 CFG 引导尺度和 Euler A 噪声采样。

7.1 参与者

我们共招募了 15 名参与者开展研究，并完成了 14 项用户研究（平均年龄=24.2，标准差=1.7）。在其中一项研究中，一名职业视觉艺术家尝试生成埃贡·席勒 [3] 风格的图像。然而，这些图像屡次被我们设置的不适宜工作场所（NSFW）内容过滤器拦截。该参与者最终决定退出研究。在完成研究的 14 名参与者中，没有人具备丰富的 Stable Diffusion 使用经验。9 名参与者表示此前没有使用文本到图像模型的经验，另外 5 名参与者仅使用过 Stable Diffusion 或 DALL·E 几次。此外，没有参与者表示会定期使用文本到图像模型，或将其用于专业工作。研究约耗时 1.5 小时，参与者获得了 30 加元的补偿。

7.2 方法论

我们在安静的实验室环境中开展了这项研究。参与者首先听取了关于研究和各系统的简要介绍。随后，他们测试了每个系统并填写相应的问卷。在参与者完成比较评估和自由形式使用观察后，我们进行了半结构化访谈。研究期间，我们鼓励参与者说出自己的想法并提出任何问题。

7.2.1 比较评估。我们采用被试内方法，比较参与者使用 Promptify 和 Automatic1111 [11] 的体验。我们请参与者使用 Automatic1111 中专为文本生成图像功能设计的 txt2img 选项卡。我们对系统使用顺序进行了平衡：一半参与者先使用 Promptify，另一半先使用 Automatic111。为帮助参与者寻找灵感，我们建议他们先选择自己最喜欢的动物，以及一位自己喜欢的艺术家或一种视觉艺术风格。这两项选择将作为起始主题内容和风格，用于两个界面。我们承认，这种方法可能会限制创作自由，但它为所有参与者提供了共同的起点，从而确保比较更加可控。参与者也可以根据需要添加其他细节。下面概述每种条件下的操作步骤。这些步骤以受控方式重现了我们在形成性研究中发现的提示工程迭代过程。

步骤 1：根据自己最喜欢的动物和艺术风格撰写初始提示。步骤 2：根据初始提示生成 10 张图像。步骤 3：查看生成图像，并选出最喜欢和最不喜欢的图像。步骤 4：优化提示，以强化期望的特征并抑制不想要的特征。步骤 5：再重复两次优化过程。步骤 6：使用最喜欢的提示生成 50 张图像。步骤 7：查看生成图像，并从 50 张图像中选出最喜欢的一张。

7.2.2 自由使用观察。完成比较性评估后，我们让参与者自由使用 Promptify 的功能来创建图像。

[3]埃贡·席勒（维基百科）：https://en.wikipedia.org/wiki/Egon_Schiele

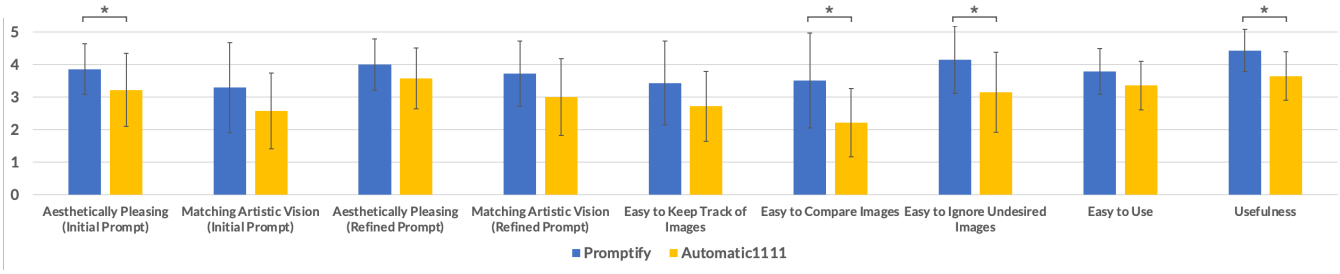

图 7：比较评估中的主观评分结果。Promptify 获得的平均分数一致高于基线工具。误差线表示标准差，星号突出显示 $p < 0.05$ 的结果。

7.2.3 半结构式访谈。最后，我们开展了时长为 10-20 分钟的半结构式访谈，以了解用户使用 *Promptify* 的体验，包括各项功能的有效性以及他们遇到的任何困难。

7.3 主观评分

我们设计了主观评分，以在比较评估的每种条件之后收集用户反馈。所有评分均采用 5 点李克特量表（1—非常不同意，5—非常同意），任务负荷问题除外，这些问题采用 7 点 NASA-TLX量表。 [20]

审美品质。 我们询问参与者是否认为他们根据初始提示和优化后的提示生成的图像具有审美吸引力。该问题与步骤 1-5 相关。

符合艺术愿景。 我们询问参与者创作的图像是否符合他们的艺术愿景。该问题同样与步骤 1-5 相关，并要求分别针对初始提示和优化后的提示进行回答。

追踪和检查生成图像。 参与者评价了自己追踪和检查所生成的大量图像的体验。他们评估了收藏夹管理、图像对比以及忽略不喜欢的图像的难易程度。这与步骤 5–7 有关。

任务负荷。 我们请参与者评价使用各系统生成图像时的心理需求、努力程度和挫折感。

有用性和易用性。 参与者评估了各系统的有用性和易用性。对于 Promptify，我们还请他们对各项功能分别进行评分。

7.4 比较评估结果

如图 7 和图 8 所示，Promptify 在所有评分项上的平均分均一致较高，证明了我们系统的有效性。进一步采用 Wilcoxon 符号秩检验进行分析后发现，与基线工具相比，我们的系统在支持文生图生成工作流程方面显著更有用（$Z = 11.0$，$p < .05$）。此外，用户认为，使用他们的初始提示词生成的图像（图 10）在美感上显著优于使用基线工具生成的图像（$Z =$

2.5，$p < .05$）。值得注意的是，使用 Promptify 撰写的初始提示词明显更长（平均=34.43 个词，标准差=8.84），相比之下

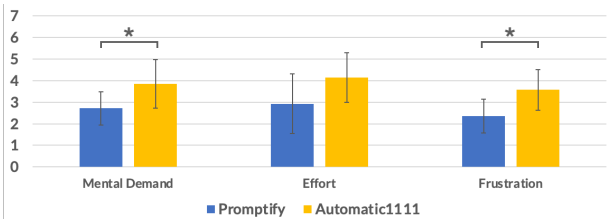

图8：任务负荷评分结果：心理需求、努力程度、挫折感。误差条表示标准差，星号标示 $p < 0.05$ 的结果。

使用 Automatic1111 撰写的提示词（平均=8.6 个词，标准差=6.31），在使用 NLTK [12] 移除停用词后仍是如此。这表明，Promptify 提供的提示词建议能够帮助用户撰写更详细的提示词，从而在首次尝试时生成更具吸引力的视觉效果。使用 Promptify 时，用户的心理需求显著降低（$Z = 9.5$，$p < .05$），挫折感也更少（$Z = 11.0$，$p < .05$）。参与者还发现，在使用 Promptfiy 搜索大型图像集合中的心仪图像时，比较和对照图像（$Z = 13.0$，$p < .05$）以及忽略不喜欢的图像（$Z = 13.5$，$p < .05$）都更加容易。

7.5 用户对 Promptify 功能的反馈

下面讨论与 Promptify 功能相关的用户反馈。这些反馈基于问卷和访谈数据，以及在比较评估和自由使用环节中进行的观察。图9展示了用户对 Promptify 功能的主观评分。

7.5.1 主题构思。 参与者认为 Promptify 的主题建议引擎有用且易用，有用性评分平均为 4.43（标准差=0.76），易用性评分为 4.21（标准差=0.70）。参与者指出，主题构思帮助他们为图像补充合适的细节，而这些细节是他们独自构思时不会想到的。P9 表示：“我发现 *[Promptify]* 一开始让我更轻松地生成第一张图像。我觉得在 *[自动1111]*模式下，我可能找不到合适的词语来准确描述脑海中的想法。” P1 尝试为“云钓鱼”这个抽象想法生成主题扩展。她觉得，“在这个非常抽象的例子中，*[Promptify]* 帮助我完善想法，并顺利开展了一次很棒的探索，完全没有挫折感。” 另一方面，参与者们.....

已经有具体想法的参与者认为主题扩展不太有用。例如，P11 想要生成一幅绵羊烘焙派的图像，并指出，由于自己已经有了明确目标，主题内容建议对于生成该图像而言相关性较低。

7.5.2 风格扩展。 风格扩展被认为易于使用，尤其被评为我们最有用的功能，受访者对其有用性的平均评分为 4.57（标准差=0.65），易用性评分为 4.56（标准差=0.51）。参与者强调，风格扩展帮助他们快速实现理想的风格。例如，在对比研究中，P1想创作一幅猫的印象派画作，但觉得自己用 Automatic1111 生成的图像“看起来太有印象派风格了——颜色有点过于饱和”。她在半结构化访谈中进一步评论道“我刚开始使用 *Automatic1111* 时总会卡住……我不知道该输入哪些提示词，才能得到我脑海中想要的效果”，并表示 Promptify 提供了丰富的风格建议，让她更轻松地实现了理想的风格。此外，P5认为，风格扩展会补充一些特定的关键短语，这些短语对于增强 Stable Diffusion生成的图像尤其有用。她指出“我不会想到要加上‘高度精细’、‘艺术站热门’或‘数字绘画’这样的标签。”其他参与者则担心，风格扩展功能可能会推荐他们不熟悉的艺术家姓名，从而妨碍他们理解提示词。正如P10所说，“如果你是艺术家，而且非常了解各种风格，可能就会更好地使用它。”

7.5.3 图像布局与聚类。 图像布局与聚类的平均有用性评分为 4.29（标准差=0.61），易用性评分为 4.36（标准差=0.63）。参与者表示，该功能通过呈现特定主题，并帮助他们比较以往提示和经过优化的提示所生成的图像，协助他们管理大批量的生成图像。一些参与者利用拖放功能，将自己最喜欢的图像分组；另一些则使用基于相似性的聚类和定位功能来识别图像中的主题。用户提到的最简单用例，是能够忽略他们不喜欢的整个图像簇。例如，P2 正在生成涂鸦图像

猎豹，并注意到有些聚类中完全没有猎豹，还评论道：“你很快就几乎可以排除一个群组。”其他参与者指出，图像布局帮助他们思考今后编写提示词时可以做出的不同选择。P4 生成了 NBA压哨球的图像，他指出，图像布局与聚类功能有助于了解可能生成的不同类型图像。它还帮助他决定“是想要多个球员，还是想要人们扣篮或投篮；需要多近的特写；以及是否想要大量彩纸屑。能够快速浏览这些选项，感觉非常有用。”一位参与者（P10）对聚类功能颇为不满，并评论道：“聚类似乎是基于一些非常表面的特征；如果有人告诉你它就是按照这些特征进行聚类的，你可能会认同。”这表明，基于 CLIP 嵌入进行聚类可能存在局限，而且通过 CLIP 嵌入确定的相似性可能与用户的感知不一致。

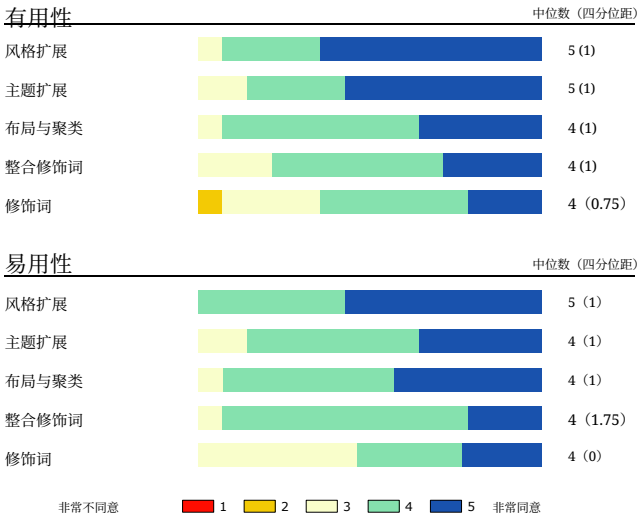

图9：Promptify 功能的主观评分结果显示，参与者普遍认为这些功能易于使用且有用。各行按平均有用性评分排序（中位数、四分位距）。

7.5.4 CLIP Interrogator 提供的提示词修饰语。 大多数参与者认为 CLIP Interrogator 提供的提示词修饰语很有价值，有用性评分为 3.79（标准差=0.89），且易于使用，评分为 3.86（标准差=0.86）。风格修饰词帮助参与者强化了自己喜欢但无法描述的风格。例如，P1 在生成印象派猫咪图像时，她最喜欢的图像都带有建议的修饰语“by Tsuguharu Foujita”。在后续生成中，她使用该关键词强化自己喜欢的城市背景风格，并评论道：“选择几个 *[修饰词]*帮助我锁定了自己喜欢的风格。”。P8 在以莫奈风格生成雪豹图像时，其第一批图像中最喜欢的一张描绘了以独特艺术风格绘制的树木。作为优化，他们选择在提示中添加“by Boris Kustodiev”，以生成更多相同风格的图像（图5），并评论说新图像“更加相似了。毫无疑问，是树木的风格。”我们观察到包括上述例子在内的7位参与者都进行了类似的互动，这表明该功能能够有效促进提示词的迭代优化，使生成内容更集中于用户期望的内容。然而，一些参与者表示不确定如何使用修饰词，因为他们不熟悉所显示的艺术家姓名。例如，P12 解释道：“困难在于，我之前并不认识很多艺术家。所以它给出了艺术家的名字，我可以搜索并查阅相关资料，但仅凭记忆，它无法立即向我提供信息。”CLIP Interrogator 的修饰词为用户缩小偏好的风格范围提供了有用工具。然而，如果不了解相关艺术家的身份，一些参与者会发现难以利用这些修饰词进行提示词优化。这可能导致该功能在 Promptify 的所有功能中获得了最低的平均有用性评分和易用性评分。



图 10：研究参与者使用 Promptify（上排）和 Automatic1111（下排）生成的第一批图像示例。

8 讨论与未来工作

8.1 促进文本生成图像工作流程

Promptify 旨在支持文本到图像工作流，包括提示词编写、生成和审阅图像，以及优化初始提示词。用户研究结果表明，与用户自行生成的图像相比，Promptify 能够有效推荐生成美学效果更佳图像的提示词。图像布局与聚类用户界面有助于检查和比较生成图像。

尽管许多参与者认为 CLIP Interrogator [37] 所推荐的提示词修饰语很有用，但我们未观察到具有统计显著性的结果，无法据此得出它们能够实现更好的提示词优化。该工具提供两类关键词：1）用于描述主题内容的图像描述；2）用于描述风格的艺术家姓名。在实验过程中，我们发现图像描述往往不太有用，因为它们可能较为笼统，并且会与现有提示中的主题内容重叠。艺术家姓名有助于生成特定风格的艺术作品，但对于不熟悉这些艺术家的用户而言，可能较难使用。未来的研究可以利用专注于特定区域或物体（而非整幅图像）的图像描述模型，来改进从生成图像中提取修饰词的建议，从而避免生成笼统的图像描述。此外，结合更强大的视觉语言模型（如 GPT-4 [35]）可能会生成更具针对性的风格修饰词，并扩充 CLIP Interrogator 中预先整理的艺术家列表。Promptify 做出了全面的系统贡献，未来还可以无缝整合各个系统组件的高级模块。

8.2 使用大语言模型自动生成提示词建议

Promptify 使用大语言模型辅助编写文本到图像生成的提示。我们方法的一个独特之处在于，用户可以通过语言指令引导主题构思。用户发现，这项功能有助于拓展思路并探索新颖的主体。我们的风格扩展功能利用大语言模型的少样本提示和社区数据，推荐模仿社区用语的风格关键词 [31, 47]。虽然我们目前的方法使用特定的提示词格式，但我们也认识到，其他格式同样可能有效。不过，我们强调少样本提示方法具有灵活性：只需提供几条采用新格式的示例提示词，就能轻松适配不同的提示词格式。少样本提示的另一种替代方法是对大语言模型进行微调 [5]。然而，微调大语言模型成本高昂，而且需要仔细整理微调数据集。我们的初步探索发现，微调往往会导致大语言模型遗忘有助于推荐艺术家和风格的既有知识（灾难性遗忘 [16]），并对所用数据过度拟合

用于微调的数据，例如，总是推荐“Greg Rutkowski”——一位在 SD 社区中颇受欢迎的数字艺术家。另一个未来方向是采用思维链提示 [52] 等技术，引导大语言模型为推荐的提示提供逻辑推理。这可以提高透明度，帮助用户理解为何推荐某些关键词。例如，大语言模型可以给出如下解释：“推荐加入宫崎骏，以体现吉卜力工作室相关风格，因为他是该工作室的联合创始人。”最后，尽管我们使用的 SD 模型允许用户输入负面提示词，以引导图像生成避开不希望出现的特征，但我们目前的大语言模型引擎不会推荐负面提示词。未来研究可以探讨这一方向。

8.3 模型随机性与单个关键词的有效性

使用基于扩散的模型生成图像的过程会受到随机性的影响，这是因为模型使用了随机噪声。这种随机性可能导致同一提示生成各种各样的图像。此外，预期的关键词并不总能在生成的图像中体现出来。例如，要求生成一头在草地上的奶牛，可能只会得到呈现草地的图像；要求生成一只梵高风格的鸡（如用户研究中的 P6 所期望的），结果可能会是一幅文森特·梵高的自画像，而不是一幅具有艺术家风格的鸡的肖像。为缓解这一问题，Promptify 会对图像进行语义组织，并允许用户移除不想要的图像。随着提示词变长，一些参与者发现，很难让生成的图像准确呈现具体细节。当提示建议包含多个细节，而用户预期这些细节都会出现在生成的图像中时，这种情况尤为常见。然而，我们发现，随着提示词变长，Stable Diffusion 模型往往会忽略某些短语，导致生成的图像缺少提示词中指定的某些细节。例如，在图 10 中，P4-top 的图像不太像梵高的风格，而 P12-top 的图像没有呈现提示词中描述的兔子。我们推测，这正是 Promptify 在匹配艺术愿景方面未能显著优于 Automatic1111 的主要原因。此外，理解单个关键词及其在提示中的位置如何影响图像生成过程，仍是一项重大挑战。我们初步开展了消融分析：从提示中移除单个关键词，并使用 CLIP 嵌入计算其与原始提示的句子相似度；但分析未能提供有力证据来说明单个关键词的影响。未来的研究亟须进一步探索应对模型随机性的方法，并调查单个短语的影响，以帮助用户创建有效的提示词。

8.4 将 Promptify 与现有工具进行对比

文本到图像模型的流行推动了各种配套工具的发展 [1, 2, 5, 11]。我们将 Promptify 与最受欢迎的开源工具之一 Automatic1111 进行了比较，展示了 Promptify 的显著优势。现有的一些工具与 Promptify 的部分功能相似。例如，Magic Prompt [5] 和 Promptgen [2] 都能够为文本到图像生成推荐提示词。然而，Promptify 提供了更为先进的提示词建议引擎，能够分别针对主题内容和风格提供建议。此外，Promptify 还独创性地允许用户通过大语言模型引导主题构思和风格扩展。除此之外，Promptify 还推出了一种基于 CLIP 嵌入的新颖图像布局与聚类界面，让用户能够浏览和查看使用 Stable Diffusion 生成的图像。

近期机器学习领域的研究开始开发先进技术，以支持文本到图像生成中的图像编辑 [21, 59]。例如，Prompt-to-Prompt [21] 利用从扩散模型中提取的注意力图，通过修改先前的提示来实现保留布局的图像编辑。我们的系统则着眼于一个不同但互补的目标，即帮助新手用户发现他们原本可能不会想到的有效关键词。

在当前研究中，我们重点探索提示词编写，并固定了所有其他可能影响文本到图像生成的重要参数，例如随机种子、噪声调度器和 CFG 引导尺度。此外，SD 模型还有其他变体，可接受多种用户输入来引导图像生成，例如图生图 [43]、图像修复 [43] 和 ControlNet [59]。未来的工作可以将这些参数和模型集成到 Promptify 中，从而实现更广泛的图像生成能力。

9 结论

我们介绍了 Promptify，这是一个交互式系统，支持用户交互式地探索并创建用于文本生成图像的文本提示词。Promptify 利用大语言模型，分别针对主题内容和风格提供提示词建议。用户可以通过自然语言指令引导主题内容建议。风格关键词则通过少样本提示生成：利用在线社区中的文本生成图像提示词示例向大语言模型提供提示。Promptify 还提供布局与聚类界面，帮助用户高效地整理、审阅和比较图像。此外，Promptify 还会为每张生成的图像推荐相关关键词，用户可以利用这些关键词来优化初始提示词。我们对 14 名参与者开展的用户研究表明，与广泛使用的文本生成图像基线工具相比，Promptify 的实用性显著更高，能够帮助用户首次尝试就创作出视觉效果更出色的图像，同时大幅降低认知负荷。

参考文献

[1] 2022. *Automatic1111 扩展*。 https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Extensions [2] 2022. *Promptgen*。 https://github.com/AUTOMATIC1111/stable-diffusion-webui- promptgen [3] 2023. *CLIP*。 https://github.com/openai/CLIP [4] 2023. *Lexica*。 https://lexica.art/ [5] 2023. *MagicPrompt-Stable-Diffusion*。 https://huggingface.co/spaces/Gustavosta/MagicPrompt-Stable-Diffusion [6] 2023. *Midjourney*。 https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F [7] 2023. *OpenAI API*。 https://openai.com/blog/openai-api [8] 2023. *React Flow*。 https://reactflow.dev/ [9] 2023. *SentenceTransformers*。 https://www.sbert.net/ [10] Maneesh Agrawala。2023。不可预测的黑箱是糟糕的界面。 https: // magrawala.substack.com/p/unpredictable-black-boxes-are-terrible [11] automatic1111。2022。Stable Diffusion Web UI。https://github.com/AUTOMATIC1111/stable-diffusion-webui。[12] Steven Bird、Ewan Klein 和 Edward Loper。2009。使用 *Python* 进行自然语言处理：借助自然语言工具包分析文本。" O’Reilly Media, Inc."。

[13] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020。语言模型是少样本学习者。arXiv:2005。14165 [cs.CL][14] Siddhartha Chaudhuri、Evangelos Kalogerakis、Stephen Giguere 和 Thomas Funkhouser。2013。Attribit：利用语义属性进行内容创作。载于第26届年度ACM用户界面软件与技术研讨会论文集（苏格兰圣安德鲁斯，英国）（UIST '13）。美国计算机协会，纽约，纽约州，美国，193–202。https://doi.org/10.1145/ 2501988.2502008[15] cmdr2。2022。Easy Diffusion。https://github.com/cmdr2/stable-diffusion-ui。[16] Danny Driess、Fei Xia、Mehdi S. M. Sajjadi、Corey Lynch、Aakanksha Chowdh- ery、Brian Ichter、Ayzaan Wahid、Jonathan Tompson、Quan Vuong、Tianhe Yu、Wenlong Huang、Yevgen Chebotar、Pierre Sermanet、Daniel Duckworth、Sergey Levine、Vincent Vanhoucke、Karol Hausman、Marc Toussaint、Klaus Greff、Andy Zeng、Igor Mordatch 和 Pete Florence。2023。PaLM-E：具身多模态语言模型。载于arXiv预印本 arXiv:2303.03378。[17] Noyan Evirgen 和 Xiang 'Anthony' Chen。2022。GANzilla：生成对抗网络中的用户驱动方向发现。载于第35届年度ACM用户界面软件与技术研讨会。ACM。https: //doi.org/ 10 1145 3526113 3545638 。/。Noyan Evirgen 和 Xiang'Anthony' Chen。2023。[18] GANravel：生成对抗网络中的用户驱动方向解耦。arXiv预印本 arXiv:2302.00079（2023）。[19] Brendan J. Frey 和 Delbert Dueck。2007。通过数据点间传递消息进行聚类。《科学》315，5814（2007），972–976。https://doi.org/10.1126/science。 1136800 10 1126 1136800 [20] 1988 arXiv:https://www.science.org/doi/pdf/。/science。Sandra G Hart 和 LE Staveland。。NASA-TLX（任务负荷指数）的开发：实证与理论研究结果。[W]：PA Hancock、N. Meshkati（编）：人的心理工作负荷。[21] Amir Hertz、Ron Mokady、Jay Tenenbaum、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022。利用交叉注意力控制进行提示到提示图像编辑。（2022）。[22] Hyung-Kwon Ko、Gwanmo Park、Hyeon Jeon、Jaemin Jo、Juho Kim 和 Jinwook Seo。2022。面向视觉艺术家创作的大规模文本到图像生成模型。arXiv abs/2210.08477（2022）。[23] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022。BLIP：用于统一视觉语言理解与生成的自举式语言 - 图像预训练。arXiv:2201。12086 [cs.CV][24] Jiachang Liu、Dinghan Shen、Yizhe Zhang、Bill Dolan、Lawrence Carin 和 Weizhu Chen。2021。GPT-3 的上下文示例具备哪些优点？arXiv: 2101 06804 [ ] 。cs.CL Vivian Liu 和 Lydia B Chilton。2022。文本到图像生成模型提示工程设计指南。载于2022年CHI计算机系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。美国计算机协会，纽约，纽约州，美国，文章384，23页。https://doi.org/ 10 1145 3491102 3501825 。/。Vivian Liu、Han Qiao 和 Lydia Chilton。2022。[26] Opal：用于新闻插图的多模态图像生成。载于第35届年度ACM用户界面软件与技术研讨会论文集（美国俄勒冈州本德）（UIST '22）。美国计算机协会，纽约，纽约州，美国，文章73，17页。https://doi.org/ 10 1145 3526113 3545621 。/。Vivian Liu、Jo Vermeulen、George Fitzmaurice 和 Justin Matejka。2022。3DALL-E：在 3 2210 11603 [ ] D设计工作流中集成文本到图像人工智能。arXiv：。cs.HC Elman Mansimov、Emilio Parisotto、Jimmy Lei Ba 和 Ruslan Salakhutdinov。2016。利用注意力根据描述生成图像。arXiv:1511。02793 [cs.LG][29] Joe Marks、Brad Andalman、Paul A Beardsley、William Freeman、Sarah Gibson、Jessica Hodgins、Thomas Kang、Brian Mirtich、Hanspeter Pfister、Wheeler Ruml 等。1997。Design Galleries：一种设置计算机图形学和动画参数的通用方法。载于第24届年度计算机图形学与交互技术会议论文集。ACM出版社/Addison-Wesley出版公司，389–400。[30] Justin Matejka、Michael Glueck、Erin Bradner、Ali Hashemi、Tovi Grossman 和 George Fitzmaurice。2018。Dream Lens：大型生成式设计数据集的探索与可视化。载于2018年CHI计算机系统中的人为因素会议论文集（加拿大魁北克省蒙特利尔）（CHI '18）。美国计算机协会，纽约，纽约州，美国，1–12。https://doi.org/ 10 1145 3173574 3173943 。/。Justin Matejka、Wei Li、Tovi Grossman 和 George Fitzmaurice。2009。CommunityCommands：软件应用程序的命令推荐。载于第22届年度ACM用户界面软件与技术研讨会论文集（加拿大不列颠哥伦比亚省维多利亚）（UIST '09）。美国计算机协会，纽约，纽约州，美国，193–202。https://doi.org/10.1145/1622176.1622214

[32] Anh Nguyen、Jeff Clune、Yoshua Bengio、Alexey Dosovitskiy 和 Jason Yosinski。2017年。《即插即用生成网络：潜在空间中的条件迭代图像生成》。arXiv: 1612 00005 [ ] 。cs.CV nolan dev。. GANInterface。[33] 2019 https://github.com/nolan-dev/GANInterface。[34] 2023 4 2303 08774 [ ] OpenAI。。GPT-技术报告。[35] 2023 4 2303 08774 [arXiv:。cs.CL OpenAI。。GPT-技术报告。arXiv:。cs.CL] [36] 2022 Jonas Oppenlaender。。文本到图像生成的提示词修饰符分类。arXiv: 2204 13988 [ ] 。[37] 2022 cs.MM pharmapsychotic。。clip-interrogator。https://github.com/pharmapsychotic/ clip-interrogator。[38] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021年。通过自然语言监督学习可迁移视觉模型。arXiv: 2103 00020 [ ] 。cs.CV Aditya Ramesh、[39] Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。使用CLIP潜变量的层次化文本条件图像生成。arXiv: 2204 06125 [ ] 。cs.CV Aditya Ramesh、[40] Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Rad- ford、Mark Chen 和 Ilya Sutskever。2021年。零样本文本到图像生成。arXiv:2102.12092 [cs.CV][41] Scott Reed、Zeynep Akata、Santosh Mohan、Samuel Tenka、Bernt Schiele 和 Honglak Lee。2016年。学习绘制什么与在哪里绘制。arXiv:1610.02454 [cs.CV][42] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。2016年。生成对抗式文本到图像合成。载于《第33届国际机器学习会议论文集》（《机器学习研究论文集》，第48卷），Maria Florina Balcan 和 Kilian Q. Weinberger（编）。PMLR，美国纽约州纽约市，1060–1069。https://proceedings.mlr.press/v48/reed16.html[43] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2021年。使用潜在扩散模型的高分辨率图像合成。arXiv: 2112 10752 [ ] 。cs.CV Elvis Saravia。。提示工程指南。https://github.com/dair-ai/Prompt- Engineering-Guide（2022年12月）。[44] 2022 [45] Ana Serrano、Diego Gutierrez、Karol Myszkowski、Hans-Peter Seidel 和 Belen Masia。2016年。材质外观的直观控制空间。《ACM图形学汇刊》35（6），文章186（2016年12月），12页。https://doi.org/10.1145/2980179。2980242[46] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。2015年。使用非平衡热力学进行深度无监督学习。载于《第32届国际机器学习会议论文集》（《机器学习研究论文集》，第37卷），Francis Bach 和 David Blei（编）。PMLR，法国里尔，2256–2265。https://proceedings.mlr.press/v37/sohl-dickstein15.html [47] John Swales。2014年。话语共同体的概念。Wardle 和 Downs（2014），215–28。[48] Michael Terry 和 Elizabeth D. Mynatt。2002 年。侧视图：面向开放式任务的持久按需预览。载于《第 15 届 ACM用户界面软件与技术年度研讨会论文集》（法国巴黎）（UIST '02）。美国计算机协会，美国纽约州纽约市，71–80。https://doi.org/ 10 1145 571985 571996 。/。Laurens Van der Maaten 和 Geoffrey Hinton[49]。2008年。使用t-SNE可视化数据。《机器学习研究期刊》9，11（2008）。[50] Yunlong Wang、Shuyuan Shen 和 Brian Y Lim。2023年。RePrompt：自动提示编辑，将人工智能生成艺术优化为精确表达。arXiv预印本 arXiv:2302.09466（2023）。[51] Zijie J. Wang、Evan Montoya、David Munechika、Haoyang Yang、Benjamin Hoover 和 Duen Horng Chau。2022年。DiffusionDB：面向文本到图像生成模型的大规模提示词图库数据集。arXiv:2210.*14896 [cs]*（2022）。https://arxiv. org/abs/2210.14896[52] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 Denny Zhou。2023年。思维链提示在大语言模型中引发推理。arXiv:2201.11903 [cs.CL][53] Yuxin Wen、Neel Jain、John Kirchenbauer、Micah Goldblum、Jonas Geiping 和 Tom Goldstein。2023年。硬提示轻松实现：用于提示调优与发现的基于梯度的离散优化。arXiv: 2302 03668 [ ] 。cs.LG Sam [54] 2022 Witteveen 和 Martin Andrews。。研究扩散模型中的提示工程。arXiv: 2211 15462 [ ] 。cs.CV Mehmet Ersin Yumer、Paul [55] Asente、Radomir Mech 和 Levent Burak Kara。2015年。使用自编码器网络进行程序化建模。载于《第28届ACM用户界面软件与技术年度研讨会论文集》（美国北卡罗来纳州夏洛特）（UIST '15）。美国计算机协会，美国纽约州纽约市， 109 118 10 1145 2807442 2807448 –。https://doi.org/。/。Mehmet Ersin Yumer、Siddhartha Chaudhuri、Jessica K. Hodgins 和 Levent Bu- rak Kara。2015年。使用变形控制柄进行语义形状编辑。《ACM图形学汇刊》34（4），文章86（2015年7月），12页。https://doi.org/10.1145/2766908[57] Loutfouz Zaman、Wolfgang Stuerzlinger、Christian Neugebauer、Rob Woodbury、Maher Elkhaldi、Naghmi Shireen 和 Michael Terry。2015年。GEM-NI：用于创建和管理生成式设计备选方案的系统。载于《第33届ACM计算系统中的人为因素会议论文集》（首尔，[56]

大韩民国）（*CHI '15*）。美国计算机协会，美国纽约州纽约市，1201–1210。https://doi.org/10.1145/2702123.2702398[58] Han Zhang、Tao Xu、Hongsheng Li、Shaoting Zhang、Xiaogang Wang、Xiaolei Huang 和 Dimitris Metaxas。2017年。StackGAN：使用堆叠生成对抗网络将文本合成为逼真照片。arXiv:1612.03242 [cs.CV]

[59] Lvmin Zhang 和 Maneesh Agrawala。2023年。为文本生成图像扩散模型添加条件控制。arXiv: 2302 05543 [ ] 。cs.CV

[60] Tony Z. Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。2021年。在使用前进行校准：提升语言模型的少样本性能。arXiv:2102.09690 [cs.CL]