GenFlow: 交互式模块化图像生成系统

Duc-Hung Nguyen*[1, 2], Huu-Phuc Huynh*[1, 2], Minh-Triet Tran[1, 2], and Trung-Nghia Le**[1, 2]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

摘要—生成艺术解锁了无限的创意可能性，但由于需要高级架构概念和计算工作流程的技术专业知识，其全部潜力尚未得到充分发挥。为弥合这一差距，我们提出了 GenFlow，这是一个新型模块化框架，使所有技能水平的用户都能轻松精确地生成图像。该框架配备了基于节点的编辑器以实现无缝定制，并采用由自然语言处理驱动的智能助手，将工作流程创建的复杂性转化为直观易用的体验。通过自动化部署过程并减少技术障碍，我们的框架使得尖端的生成艺术工具变得触手可及。一项用户研究表明，GenFlow 能够优化工作流、缩短任务完成时间并通过其直观界面和自适应功能增强用户理解。这些结果将 GenFlow 定位为一个具有突破性的解决方案，在生成艺术领域重新定义了访问性和效率。

***Index Terms*—信息检索，网络搜索，稳定扩散，基于节点的界面。**

I. 介绍

生成艺术由于机器学习和基于视觉的模型的进步而人气飙升。从编辑和广告到电影行业，由人工智能驱动的图像生成工具正在改变创意工作流程 [1]。流行的平台如 Stable Diffusion WebUI、Midjourney 和 DALL-E 3 提供了强大的功能。然而，它们有限的界面可能对需要更高定制的任务构成挑战，比如风格迁移、图像构图或应用高级滤镜和效果 [2]。

诸如 ComfyUI[1] 等工具已经出现，以通过基于节点的可视化编程提供增强的灵活性。在此类可视化编程中，源代码以图形式呈现，称为工作流，其中每个节点（即执行特定任务的代码单元）是一个顶点，每条连接（即从一个节点到另一个节点的数据流）是一条边。尽管这些工具提供了强大的定制选项，但也引入了陡峭的学习曲线，这需要像 IPAdapter [3] 或 ControlNet [4] 这样的大量机器学习技术。因此，许多用户，特别是新手或没有技术背景的用户，无法充分利用这些高级工具的潜力，导致可访问性和易用性 [5] 方面存在差距。

为了填补这一空白，我们引入了 GenFlow，这是一个新颖的模块化框架，旨在使高级图像生成工具更加易于访问，同时保持高度定制化的特性 [6]。GenFlow 集成了两个核心组件：编辑器是一个用户友好的基于节点的界面，用于创建和管理工作流程，适用于具有不同技术水平的用户；流助理是由自然语言处理驱动的智能助手，通过允许用户以自然语言描述所需任务来简化工作流程的发现和部署过程 [7]。此外，GenFlow 还集成了自动模型检索、本地数据库以加快工作流程搜索以及智能网络探索等功能，以提高效率和可用性。通过消除技术障碍，该系统使用户能够轻松探索生成艺术工具，促进创意与实验 [8]。

我们通过一项用户研究来评估 GenFlow 的有效性，并强调它重新定义生成艺术可访问性的潜力。在这项用户研究中，参与者称赞 GenFlow 简化了工作流程选择过程，并提供了一个易于使用但功能强大的图像生成工具。

我们的贡献如下：

- 我们介绍 GenFlow，一个用户友好的可定制图像生成系统，适合初学者和专家。
- 我们开发了 Flow Pilot，一个基于自然语言的工作流程发现的智能助手。
- 用户研究证明了所提出的系统的有效性和可用性。

II. 相关工作

A. 基于节点的工作流系统

基于节点的可视化编程环境在多个领域已被证明是有效的，包括 3D 建模（Blender 的节点编辑器）、游戏开发（Unreal Engine 的蓝图）和音频处理（Max/MSP）[9] [10]。这些系统允许用户通过连接代表不同操作的节

*Both authors contributed equally to this research.

**Corresponding author. *e-mail: ltnghia@fit.hcmus.edu.vn*

[1]https://www.comfy.org/

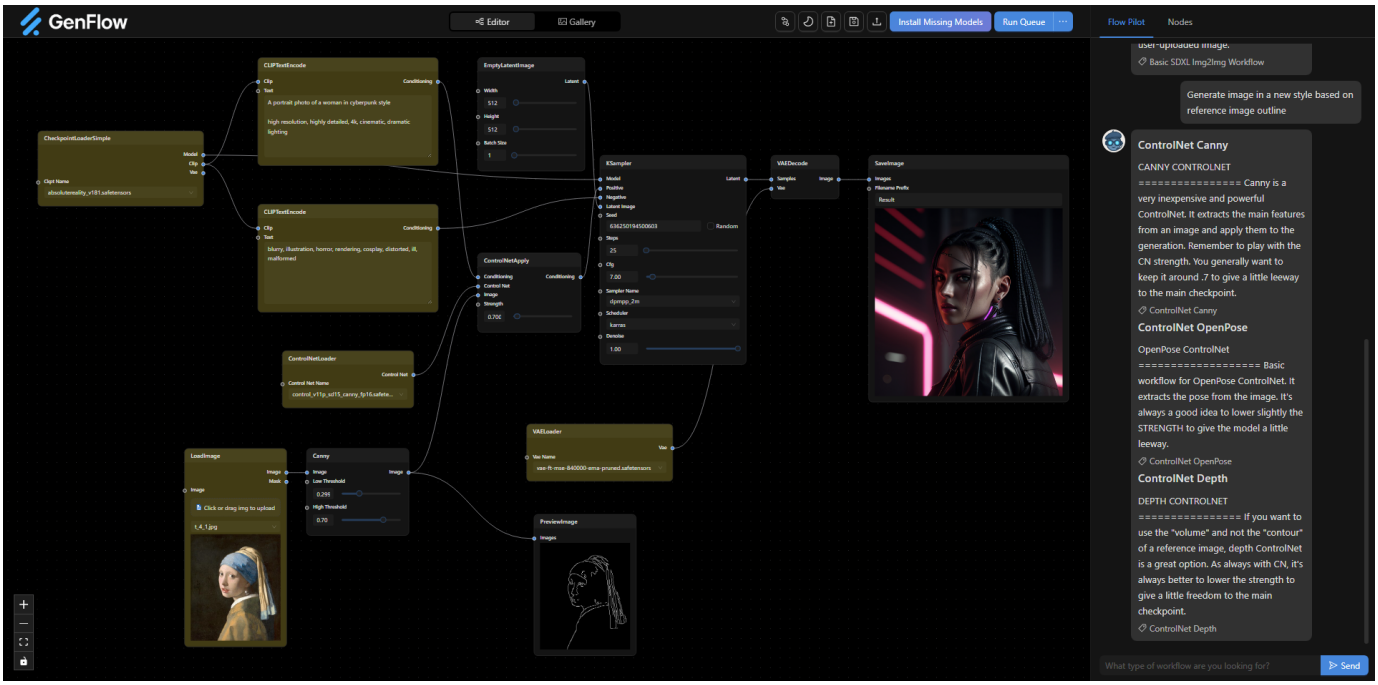

图 1: GenFlow 接口，包括两个关键组件：Canvas 用于高效的工作流创建和管理，以及 Flow Pilot 用于智能工作流搜索。

点来创建复杂的流程。这种方法提供了高度的灵活性和控制力，但也可能对新用户造成陡峭的学习曲线 [11]。ComfyUI 是 Stable Diffusion [12] 的一个基于节点界面的显著例子。虽然提供了强大的自定义选项，但它需要对 Stable Diffusion 管道有相当的技术知识，这使得它对于普通用户来说不太容易访问。

B. 检索增强生成（RAG）

RAG 结合了信息检索和生成模型的优点。通过将其与外部知识来源 [13] [14] 进行关联，提高了生成文本的质量和事实准确性。RAG 已成功应用于各种自然语言处理任务中，如问答 [15] 和对话生成 [16]。在我们的工作中，我们通过根据用户提供的描述检索相关工作流程，将 RAG 方法适应到人工智能艺术生成领域，弥合了自然语言用户输入与结构化工作流执行 [17] 之间的差距。

C. 网络搜索

自主网络代理领域随着大型语言模型（LLMs）的出现取得了显著进展，这些模型能够实现数字平台上从网页导航到与图形用户界面（GUIs）交互等复杂任务的自动化。例如，多模态代理如 WebVoyager 利用结合视觉-语言理解来识别页面上的可交互元素，尽管它们有时无法完全把握这些元素背后的语义意图 [18]。在此基础上，OmniParser 系统进一步通过解析 UI 组件来推断其功能，但它们可能仍然会忽略某些在动态或深度嵌套布局中的可交互小部件 [19]。为了补充这些建模进展，引入了诸如 VisualWebArena 这样的基准套件来评估代理在现实网络场景下的性能，而 SeeClick 等努力则致力于改进元素检测和定位以确保在网络浏览任务中选择更可靠的行动方案 [20], [21]。综上所述，这些贡献为 Web Suffer Agent 奠定了基础，后者旨在将多模态识别、语义 UI 解析和稳健的基准测试整合到一个统一框架中，实现更加有效且通用的网络自动化。

GenFlow 通过引入自然语言界面扩展了先前基于节点的系统，降低了非专家用户的使用门槛，同时保留了高级创作者所需的功能性。与仅专注于视觉编程或文本交互的现有工具不同，GenFlow 通过 Flow Pilot 结合了两者，实现了直观的工作流程发现和定制。此外，多代理网络探索引擎增强了现有的网络自动化工作，特别针对图像生成任务进行了检索优化。这种统一的方法使得 GenFlow 成为一种更易于访问且智能的生成式图像创作解决方案。

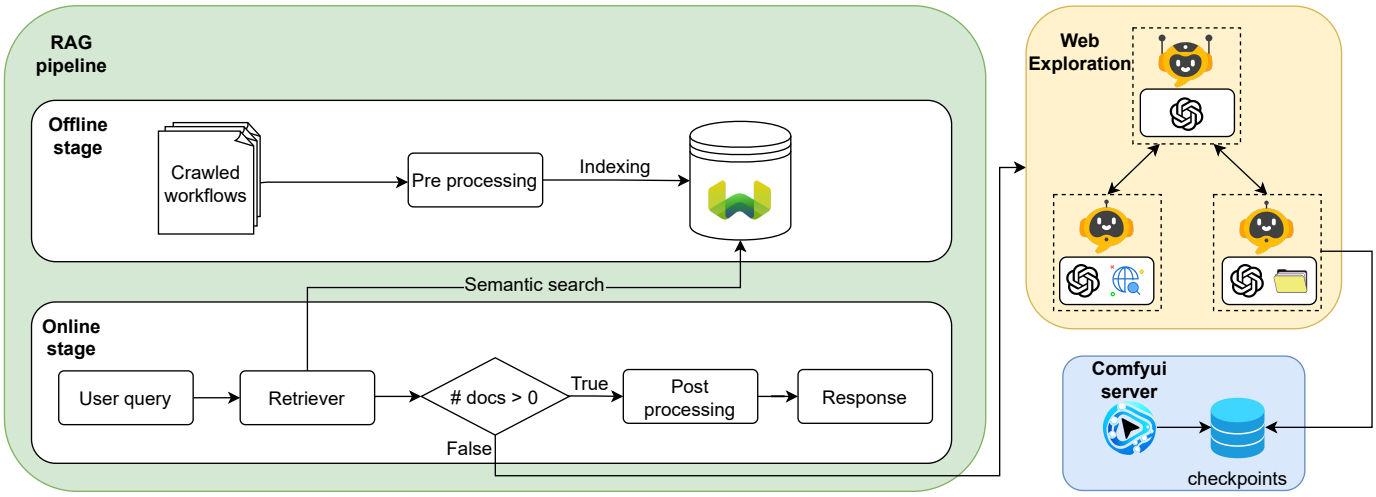

图 2: 该框架通过两个不同的阶段运行：**(a) 离线阶段**- 爬取的工作流被预处理以移除噪声（例如，标签、链接）并索引到 Weaviate 向量数据库中。**(b) 在线阶段**- 用户查询触发语义搜索；如果找到相关结果，则进行后处理并返回。如果没有检索到合适的工作流，则激活分层多代理网络探索系统。一个主管代理协调两个工人：其中一个在网络上搜索相关工作流和模型检查点，另一个下载模型检查点并将它们存储在 ComfyUI 正常运行所需正确的文件夹结构中。

III. 建议的 GenFlow 系统

A. 用户界面

GenFlow 是一个交互式模块化框架，简化了图像生成工作流的创建和发现。它将基于节点的可视化编辑器与自然语言助手相结合，使所有技能水平的用户都能以最少的技术努力构建复杂的管道。与其他优先考虑灵活性而牺牲易用性的现有工具不同，GenFlow 降低了入门门槛，同时保留了高级自定义功能。这些目标通过一个集成可视化编程和智能工作流检索的统一界面得以实现。该界面由两个主要组件组成，如图 1 所示：

画布: 画布提供了一个可视化的编程空间，用户可以拖放节点来构建工作流程。每个节点代表一个任务（例如，文本到图像转换、图像增强），用户可以定义参数、上传媒体并建立连接。侧边栏包括一个“节点”标签页用于浏览可用操作和一个“作品集”标签页用于保存和管理生成的输出和工作流程。这种设置支持快速迭代，并通过清晰的可视化结构鼓励探索。

流量舵手: 该助手使用户能够以自然语言描述任务，例如“使用这张草图生成肖像”。系统将输入与相关的工作流程进行语义匹配，按相关性和受欢迎程度对它们进行排名，并显示描述供审查。用户可以直接在画布上部署选定的工作流程，大大加快了发现和定制图像生成任务的过程。

B. 系统工作流程

1) 概览: 我们的系统核心由稳定扩散引擎驱动，负责处理图像生成管道的执行。我们在 ComfyUI 后端的基础上管理这些管道，并通过 ComfyUI Manager 扩展其功能，自动支持自定义节点集成。如图 2 所示，ComfyUI Manager 从本地数据库中检索安装元数据（例如 URL 和保存路径）来验证并安装所需的模型和节点。如果找不到必要的资产，则控制权返回到 ComfyUI，然后它调用 Web Exploration 模块搜索并下载缺失的组件。一旦检索到这些资源，它们将被存储在数据库中，并立即在 ComfyUI 环境中可用。

2) RAG 管道流程: 为了确保嵌入和检索的高质量输入，我们通过去除噪音（如 URL、HTML 标签、电子邮件地址、社交媒体引用、表情符号、标点符号、停用词和特殊字符）来预处理描述字段。这将产生干净且语义上有意义的文本，适合下游分析。使用可配置模型生成嵌入，默认使用 GoogleAI 的“models/embedding-001”，并将这些嵌入存储在 Weaviate 中，这是一个针对语义搜索和元数据过滤优化的向量数据库。

用户查询以相同的方式进行处理，并通过相似度阈值与存储条目匹配，以确保相关性。检索到的工作流然后根据受欢迎程度指标（如点赞数量）进行排名。一个大型语言模型（LLM）进一步细化和验证顶部结果，为用户提供包含名称、描述和元数据的精选工作流选项，

Algorithm 1 结合交互元素

输入： W (WebVoyager 元素), O (OmniParser 元素), τ (IoU 阈值)
输出： C (组合元素列表)
初始化： $C \leftarrow [\,]$, $U \leftarrow \emptyset$
for all $w \in W$ **do**
    $o^* \leftarrow \arg\max_{o \in O \setminus U} \text{IoU}(w.\text{bbox}, o.\text{bbox})$
    if o^* exists and $\text{IoU}(w.\text{bbox}, o^*.\text{bbox}) \geq \tau$ **then**
        将 w 和 o^* 的合并属性添加到 C; $U \leftarrow U \cup \{\text{index of } o^*\}$
    else
        添加 w 到 C
    end if
end for
for all $o \in O \setminus U$ **do**
    将 o 加到 C
end for
return C

以便支持高效发现和部署。

3) 网络探索引擎: 诸如 WorkflowComfyUI、OpenArt 和 Civitai 之类的平台随着生成式 AI 的兴起而变得流行，促进了用户分享和探索图像生成与编辑工作流程的社区。

这些平台提供了宝贵的创意和技术知识库，使用户能够发现并适应各种 ComfyUI 工作流程。

尽管它们有价值，但在这些平台上定位特定任务的工作流程仍然具有挑战性。用户必须经常浏览不同的界面，理解平台特定的惯例，并手动审查众多选项以识别合适的工作流程。这一过程耗时且为创意探索引入了不必要的摩擦。为了克服这一障碍，我们介绍了一个多智能体网络探索引擎（图 2），该引擎根据用户意图自动化发现与任务相关的工作流程。

a) 网络受苦代理: Web Suffer Agent 设计用于通过网络抓取、API 调用和优化的搜索查询自动收集来自在线平台的相关工作流程。传统网页代理通常难以准确识别和与视觉元素互动，原因在于感知能力有限及语义理解不足。为克服这些限制，最近的研究提出了提高代理解析和解释复杂网页界面能力的方法。

WebVoyager [18] 提出了使用标记集（SoM）提示 [22] 来从网页中提取交互元素，增强了检测和交互的准确性。然而，这种启发式方法无法推断出元素的功能性，需要代理依赖于容易出错的基于视觉的推理。OmniParser [19] 通过从图标动作对中学习元素功能来解决这个问题，但它可能仍然会错过复杂界面中的某些可交互元素。通过结合 WebVoyager 的元素检测和 OmniParser 的功能性推断，我们的代理获得了视觉基础和语义上下文。这种整合在算法 1 中进行了说明，使网络交互更加准确和全面，标志着人工智能系统向稳健且可扩展的网页探索迈出了重要一步。

b) 文件受损代理: 文件处理代理在多代理管道中扮演关键角色，通过检索、解析和组织工作流数据。它处理现有文件以及由网络下载代理下载的文件，确保收集所有相关信息。该代理提取工作流结构、参数和依赖关系，将非结构化数据转换为结构化格式，以便进行下游分析和集成。

c) 多智能体系统: 管理跨越多样化任务的工作流发现对单一代理来说是困难的。为了解决这个问题，我们采用了一种监督多代理架构，其中编排代理将任务委托给专门的工作代理，例如网络搜索代理和文件处理代理 [23]。每个代理维护自己的内存以跟踪过去的行动和交互，使系统能够保留上下文、适应新任务并确保准确有效的探索。这种设计与 Magentic-One 框架 [24] 相一致，该框架展示了由主要编排者协调专门代理解决复杂问题 [25], [26] 的有效性。

IV. 工作流程效率与传统方法的比较

没有我们的系统，用户需要在网络上搜索或向聊天机器人求助以根据参考图像和条件文本生成图像。这个过程包括寻找合适的技术、搜索仓库、查阅文档以及从 GitHub 或 Hugging Face 等平台上下载代码和模型。平均而言，这大约需要 12 分钟 47 秒。

相比之下，使用我们的系统，用户可以直接将任务输入 FlowPilot，它会立即检索相关的流程。这种简化的做法将所需时间缩短到仅 3 分钟 13 秒。

为了评估使用本地数据库的影响，我们测量了检索提示 “给我将图像转换为图像的基本工作流程” 的时间。安装缺失节点和模型所需的时间取决于缺失组件的数量及其大小等因素。例如，如表 I 所示，“ThinkDiffusion—Img2Img” 通过网络检索的工作流只需要一个缺失的节点和一个模型（ThinkDiffusionXL.safetensors 大小为

表 I: 以毫秒为单位比较使用本地数据库和不使用本地数据库的操作。

	本地数据库	无本地数据库
Search Workflow	3,871	152,008
Install Missing Nodes	2,755	228,182
Install Missing Models	7,255	935,878,301
总计	**13,881**	**936,258,491**

表 II: 任务详情和难度。

标识符	任务	难度
T1	Image Super-Resolution	Easy
T2	Text-to-Image Generation	Moderate
T3	Generate Image with Referenced Face	Advanced
T4	Generate Image with Referenced Outline	Advanced
T5	Virtual Try-On	Advanced

6.94 GB)。这突显了利用本地数据库显著减少了检索时间（以毫秒计），为用户提供更快更高效的服务。

V. 用户研究

我们进行了一项用户研究，以评估 GenFlow 在帮助具有心盲症的个体可视化抽象概念和生成个性化艺术作品方面的有效性。

A. 设置

1) 参与者: 我们邀请了 13 名参与者，从 P1 到 P13，来自一所大学，包括 3 名研究人员和 10 名 STEM 学生，来体验我们的系统。所有参与者具备从中级到熟练的编程技能。其中，38%是人工智能和计算机视觉方面的专家，38%具有中级知识，24%是初学者。

2) 任务: 在用户研究之前，我们进行了一次小组访谈以了解参与者典型的任务。一位参与者强调了图像增强和生成这类任务的耗时性质，这些任务通常需要多种工具和手动调整。另一位参与者指出从参考输入（如面部图像或草图）生成逼真图像的挑战，这需要专业知识和技术调试。基于这些见解，我们设计了五项任务以简化这些过程，具体见表 II。

3) 装置与程序: 每次研究会话都在我们的实验室进行，参与者在我们的观察下使用提供的笔记本电脑完成分配的任务。每个会话包括 30 分钟时间段内的四个部分。介绍之后，参与者跟随视频中的分步说明（10 分钟），并可以自由提问以澄清疑问。接下来，参与者完成了表 II 中详述的五项研究任务。鼓励参与者大声思考。在每次会话结束时，我们通过调查收集反馈。

B. 结果

1) 任务完成: 所有参与者均成功完成了全部五个任务，任务 1 至任务 5 的平均完成时间分别为 144.23 秒、205.07 秒、166.61 秒、193.31 秒和 662.08 秒。

提示语简洁，通常少于 25 个单词，并根据任务复杂度不同而风格各异。对于简单任务（T1：图像超分辨率），参与者使用直接命令如 “提高分辨率” [P13] 或 “增强图像质量” [P6]。中等难度的任务（T2：文本到图像生成）鼓励更具创造性的输入，例如 “生成一辆超级跑车的图像” [P5]。高级任务则需要更高的具体性。在 T3（参考人脸生成）中，提示包括 “创建一张与这个参考匹配的人脸图像” [P8]。在 T4（参考轮廓生成）中，示例为 “通过填充这个轮廓来生成图像” [P12]。T5（虚拟试穿）涉及精确指令如 “改变模特儿的衣物” [P7]。

参与者还使用了目标驱动的措辞，例如，“我想在这个中世纪背景下创作一张肖像” [P1] 和 “我想调整这个人的衣服” [P2]。这种提示风格的变化反映了参与者适应不同任务难度的能力。对于类似 T1 这样的简单任务，简单的命令就足够了，而对于像 T3、T4 和 T5 这样的高级任务，则需要更复杂且描述性的提示。这一进展突显了参与者根据任务复杂性调整其提示策略的能力。

2) 见解: 参与者反思了他们使用该应用程序的经验，经常将其与 Photoshop、ChatGPT 和手动模型管道等熟悉的工具进行比较。他们的反馈揭示了几点关键观察：

- **效率提升**：参与者，尤其是那些有编程经验的人，称赞该应用程序简化工作流程的能力。P1 表示，“在一个视频文件系统中，我还没见过能做这种事的”，强调了其独特的功能。P9 认为它相比自托管设置 “节省时间”，而 P13 则将其描述为比手动方案 “更快”。P4 强调不需要搜索模型的好处，称这是一个重要的时间节约。
- **用户界面**：用户界面因支持快速迭代而受到称赞。P2 观察到 “这里的迭代由于减少了对复杂提示的需求，通过 UI 变得快得多”。拖放界面因其易用性脱颖而出，P6 表示，“只需拖放”，使得非编码人员也能轻松使用。P10 欣赏系统的透明度和控制

表 III: 用户研究的结果。第一行显示输入图像和文本提示（如适用），第二行展示相应的生成输出。

T1: 图像超分辨率	T2: 文本到图像	T3: 带参考人脸的生成	T4: 带参考大纲的生成部分	T5: 虚拟试穿
	–			 
–	a photo of a hyper sport car	a photo of a man walking down a city sidewalk, wearing a blue shirt and beige pants, high resolution	a photo of a female knight in medieval times, high resolution, highly detailed, 4k, cinematic, dramatic lighting	–
				

表 IV: 系统方面参与者评分。

方面	描述	平均评分
Usefulness	Effectiveness in reducing task completion time.	4.3 / 5
Friendly Interface	Moderate user-friendliness with potential for improved intuitiveness.	3.8 / 5
Ease of Use	Straightforward and accessible navigation and usability.	4.2 / 5
Performance	Consistent and reliable processing speed and responsiveness.	4.1 / 5
Overall Satisfaction	General approval of functionality and performance.	4.0 / 5

力，表示，“它显示了整个工作流程并让我们进行配置”，提供了比生成式 AI 工具更多的控制。

参与者广泛称赞该工具的自动化功能、可视化工作流程设计以及定制的便捷性。这些特性使用户能够减少手动操作，无缝切换任务，并专注于创意或更高层次的目标。工作流的可视化表示帮助用户，尤其是新手，更好地理解复杂过程并自信地进行实验。支持多种模型和灵活的工作流配置增强了工具的适应性和长期价值。用户研究中参与者创建的例子如表 III 所示。

3) 评分: 该系统通过 13 名参与者的调查进行了评估，重点关注五个关键维度：有用性、界面友好度、易用性、性能和总体满意度。平均评分来自参与者反馈。如表 IV 所示，参与者认为该系统在支持任务完成方面有效，特别是通过减少执行时间和保持高性能。尽管总体满意度和易用性较高，但反馈表明，改进用户界面可以进一步提高可用性，并使系统对更广泛的用户群体更具可访问性。

4) 讨论: 用户研究证实了 GenFlow 在减少任务完成时间方面以及提高用户满意度方面的有效性。参与者一致称赞其直观的界面和简化的流程，强化了我们设计理念的价值。通过利用基于节点的架构，GenFlow 建立在熟悉的视觉编程范式之上，同时提供了足够的灵活性以适应人工智能技术的快速发展。此外，Flow Pilot 的智能检索系统依托于自然语言处理和检索增强生成领

域的最新进展，使 GenFlow 成为人工智能艺术领域的一个前瞻性工具。

然而，参与者反馈也揭示了系统可以改进的关键领域，以更好地支持多样化的用户需求。一些用户希望获得更精确的工作流程建议，表明需要增强 Flow Pilot 的自然语言理解和多模态查询（例如草图或参考图像）的支持。另一些建议共享和协作精化工序将促进学习和创新，突显了社区驱动库的价值。此外，对基于节点界面的不同舒适度水平表明，适应性 UI 功能，如个性化布局或基于使用情况的推荐，可以提高初学者的易用性。

这些发现指出了使 GenFlow 更具情境感知能力、协作性和适应性的重要性，确保其在生成式 AI 工具变得越来越复杂的情况下，仍能为广泛用户提供持续有效的服务。

VI. 结论

本文介绍了 GenFlow，一个新颖的交互式模块化框架，它弥合了 AI 艺术生成中用户友好性和强大定制能力之间的差距。我们的工作在使高级工具更广泛地普及方面迈出了重要一步，使得这些工具不仅对广大受众易于使用，同时也未牺牲专家用户所需的深度。

未来的工作将集中在通过改进的语言理解和多模态输入（如草图和图像）来增强 GenFlow 的工作流程检索，以及开发一个协作存储库用于共享和精化工作流程。这些方向旨在促进社区驱动的创新并使系统适应不同用户的专长水平。

致谢

此项研究得到了越南国家大学胡志明市分校信息技术学院的研究基金支持。

参考文献

[1] A. Chatterjee, “Art in an age of artificial intelligence,” *Frontiers in Psychology*, vol. 13, p. 1024449, 2022.

[2] T. Brooks, A. Holynski, and A. A. Efros, “Instructpix2pix: Learning to follow image editing instructions,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 392–18 402.

[3] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” *arXiv preprint arXiv:2308.06721*, 2023.

[4] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[5] C. Zhang and J. Liu, “Content based deep learning image retrieval: A survey,” in *Proceedings of the 2023 9th International Conference on Communication and Information Processing*, ser. ICCIP ’23. New York, NY, USA: Association for Computing Machinery, 2024, p. 158 – 163. [Online]. Available: https://doi.org/10.1145/3638884.3638908

[6] R. Wingström, J. Hautala, and R. Lundman, “Redefining creativity in the era of ai? perspectives of computer scientists and new media artists,” *Creativity Research Journal*, vol. 36, no. 2, pp. 177–193, 2024.

[7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[8] S. R. Dubey, “A decade survey of content based image retrieval using deep learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2022.

[9] T. Angert, M. Suzara, J. Han, C. Pondoc, and H. Subramonyam, “Spellburst: A node-based interface for exploratory creative coding with natural language prompts,” in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[10] J. Bieniek, M. Rahouti, and D. C. Verma, “Generative ai in multimodal user interfaces: Trends, challenges, and cross-platform adaptability,” 2024. [Online]. Available: https://arxiv.org/abs/2411.10234

[11] M. Noone and A. Mooney, “Visual and textual programming languages: a systematic review of the literature,” *Journal of Computers in Education*, vol. 5, pp. 149–174, 2018.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, vol. 2, 2023.

[15] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, and B. Hooi, “G-retriever: Retrieval-augmented generation for textual graph understanding and question answering,” *arXiv preprint arXiv:2402.07630*, 2024.

[16] H. Wang, W. Huang, Y. Deng, R. Wang, Z. Wang, Y. Wang, F. Mi, J. Z. Pan, and K.-F. Wong, “Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems,” *arXiv preprint arXiv:2401.13256*, 2024.

[17] I. M. Hameed, S. H. Abdulhussain, and B. M. M. and, “Content-based image retrieval: A review of recent trends,” *Cogent Engineering*, vol. 8, no. 1, p. 1927469, 2021. [Online]. Available: https://doi.org/10.1080/23311916.2021.1927469

[18] H. He, W. Yao, K. Ma, W. Yu, Y. Dai, H. Zhang, Z. Lan, and D. Yu, "Webvoyager: Building an end-to-end web agent with large multimodal models," *arXiv preprint arXiv:2401.13919*, 2024.

[19] Y. Lu, J. Yang, Y. Shen, and A. Awadallah, "Omniparser for pure vision based gui agent," *arXiv preprint arXiv:2408.00203*, 2024.

[20] J. Y. Koh, R. Lo, L. Jang, V. Duvvur, M. Lim, P.-Y. Huang, G. Neubig, S. Zhou, R. Salakhutdinov, and D. Fried, "VisualWebArena: Evaluating multimodal agents on realistic visual web tasks," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 881–905. [Online]. Available: https://aclanthology.org/2024.acl-long.50/

[21] K. Cheng, Q. Sun, Y. Chu, F. Xu, L. YanTao, J. Zhang, and Z. Wu, "SeeClick: Harnessing GUI grounding for advanced visual GUI agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9313–9332. [Online]. Available: https://aclanthology.org/2024.acl-long.505/

[22] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.

[23] P. Delias, A. Doulamis, and N. Matsatsinis, "What agents can do in workflow management systems," *Artificial Intelligence Review*, vol. 35, pp. 155–189, 2011.

[24] A. Fourney, G. Bansal, H. Mozannar, C. Tan, E. Salinas, E. E. Zhu, F. Niedtner, G. Proebsting, G. Bassman, J. Gerrits, J. Alber, P. Chang, R. Loynd, R. West, V. Dibia, A. Awadallah, E. Kamar, R. Hosn, and S. Amershi, "Magentic-one: A generalist multi-agent system for solving complex tasks," Microsoft, Tech. Rep. MSR-TR-2024-47, November 2024. [Online]. Available: https://www.microsoft.com/en-us/research/publication/magentic-one-a-generalist-multi-agent-system-for-solving-complex-tasks/

[25] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. Wiley Publishing, 2009.

[26] Y. Li, H. Wang, and Z. Zhang, "Multi-agent based workflow management systems design," *Advanced Science Letters*, vol. 6, no. 1, pp. 727–731, 2012.