GenKOL: 模块化生成 AI 框架用于可扩展虚拟 KOL 生成

Tan-Hiep To [1,2] , Duy-Khang Nguyen [1,2] , Tam V. Nguyen [3], Minh-Triet Tran [1,2], Trung-Nghia Le [1,2]*
[1]University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University - Ho Chi Minh, Ho Chi Minh City, Vietnam
[3]University of Dayton, Ohio, US

摘要—关键意见领袖（KOL）在现代营销中通过塑造消费者认知和增强品牌信誉发挥着至关重要的作用。然而，与人类 KOL 合作通常涉及高昂的成本和物流挑战。为了解决这个问题，我们提出了 GenKOL，这是一个交互式系统，使营销专业人员能够高效地使用生成 AI 创建高质量的虚拟 KOL 图像。GenKOL 允许用户通过一个集成多项 AI 功能（包括服装生成、妆容转移、背景合成和头发编辑）的直观界面动态组合促销视觉效果。这些功能被实现为模块化、可互换的服务，可以在本地机器或云中灵活部署。这种模块化架构确保了在各种用例和计算环境中的适应性。我们的系统可以显著简化品牌内容的生产，通过可扩展的虚拟 KOL 创建降低成本并加速营销工作流程。

***Index Terms*—模块化架构，生成式 AI，图像生成**

I. 介绍

关键意见领袖（KOLs）是在特定领域和社区中有影响力的个人，特别是在营销方面。与 KOLs 合作可以提升品牌的声誉，并强烈塑造消费者对其产品和服务的看法 [1]。然而，这样的合作也带来了显著的挑战。与高知名度的 KOLs 合作通常需要大量的财务投资，对营销预算造成巨大压力 [1]。此外，制作和管理必要的内容和视觉材料也需要大量时间和精力。这些挑战凸显了寻找既保持参与质量又减少资源消耗的成本效益替代方案的需求。

生成式人工智能的快速发展提供了一个有前景的解决方案：虚拟关键意见领袖（KOL）。通过利用深度学习模型，组织能够全面或部分自动化营销材料（如图像、视频、合成音频以及高度可定制化的数字人物）的生成，以吸引他们的目标受众 [2]。这种方法不仅减少了成本和生产时间，还提供了前所未有的创意灵活性和叙事适应性。相比之下，传统的图像编辑软件需要大量的专业知识和长时间的培训，限制了非专业用户的使用。这些局限进一步推动了能够使高质量内容创作大众化的系统的发展。

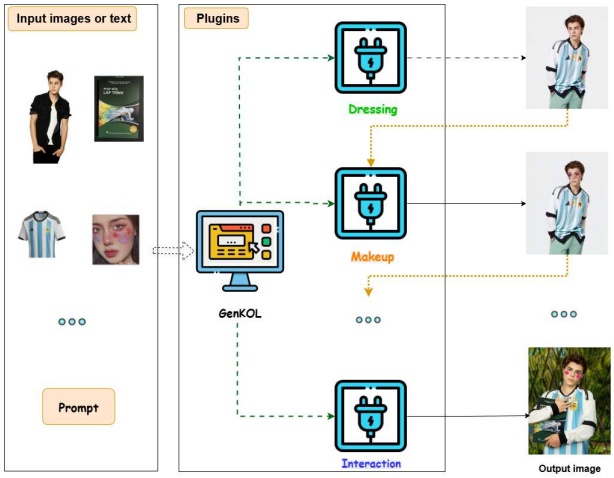

图 1. 所提出的 GenKOL 系统的流程。

尽管技术取得了进步，现有的 AI 模型仍然难以生成符合用户期望的高质量图像。这些模型可能非常复杂，并且可能无法准确反映用户的意图，这限制了它们灵活组合图像的能力。多个深度学习模型，特别是生成式 AI 模型 [2] 的整合，通常需要大量的硬件和软件资源，从而阻碍了用户的访问性并降低了灵活性和可扩展性。因此，创建一个能够使用基于插件的资源配置方法简单地集成各种 AI 模型的灵活系统至关重要且具有战略优势。开发一款用户友好、成本效益高且时间效率高的综合应用是至关重要的，同时还能产生高质量的结果。这将为产品广告提供有效的解决方案，并广泛适用于商业、营销和内容创作等领域的多种应用场景。

*Corresponding author. Email: ltnghia@fit.hcmus.edu.vn

在这篇论文中，我们介绍了 GenKOL，一个用于创建虚拟 KOL 的深度学习系统，提供了可扩展和有效的营销及内容制作解决方案。GenKOL 允许高效地管理资源并支持可扩展、模块化的流程，使得无需大量修改即可简单添加新的人工智能功能。特别地，我们引入了一个插件驱动框架，将服装生成 [3], [4]、化妆转移 [5] 和背景合成 [6], [7] 等任务模块化。每个组件作为独立的可部署服务运行，在异构环境中运作，包括本地设备和云端平台。这种设计使根据用户需求或计算资源灵活地更新、更换或扩展人工智能服务成为可能。通过标准化接口抽象模型功能，该框架提高了重用性，加速了开发，并确保了跨平台兼容性。图 1 说明了系统的工作流程，在 ACM 多媒体 2025 会议上进行了展示 [8]。

我们的主要贡献如下：

- 我们引入了一个生成式 AI 框架，该框架能够在保持语义一致性和独特视觉特征的同时，无缝地将基础身份图像与多个风格参考进行整合。
- 我们设计了一个模块化且可扩展的系统架构，便于集成新的算法和服务。此设计增强了对不断发展的技术的适应性，并使用户能够根据不同的应用特定需求定制生成管道。

II. 相关工作

图像生成已成为人工智能领域的核心议题，这一进展得益于深度学习的进步。早期的方法主要由生成对抗网络（GANs）[9] 引领，其引入了一种能够产生逼真、高质量图像的对抗训练框架。基于 GAN 的方法已成功应用于多种任务，包括图像到图像转换、草图转图像合成、条件生成和文本引导生成、视频合成、全景渲染以及场景图生成。

近期，扩散模型 [10] 作为强大的替代方案出现，提供了改进的训练稳定性、样本多样性和可控性。这些模型通过逐步用高斯噪声污染训练数据（前向过程）并学习反转这种污染（反向过程）来生成图像。通过对数据似然性的变分下界进行优化，基于扩散的方法实现了精确控制和始终如一的高质量输出。Ho 等人 [10] 以及 Rombach 等人 [11] 的奠基性贡献建立了扩散模型的可扩展性和视觉保真度。后续研究扩展了它们的应用范围：Tumanyan 等人 [12] 探索了特征级别控制，Baranchuk 等人 [13] 调查了条件采样，而 Xu 等人 [14] 推进了语义图像合成。基于这些发展，我们的工作利用最先进的生成模型 [2] 来创建表达性强、可定制且高质量的虚拟 KOL 图像。

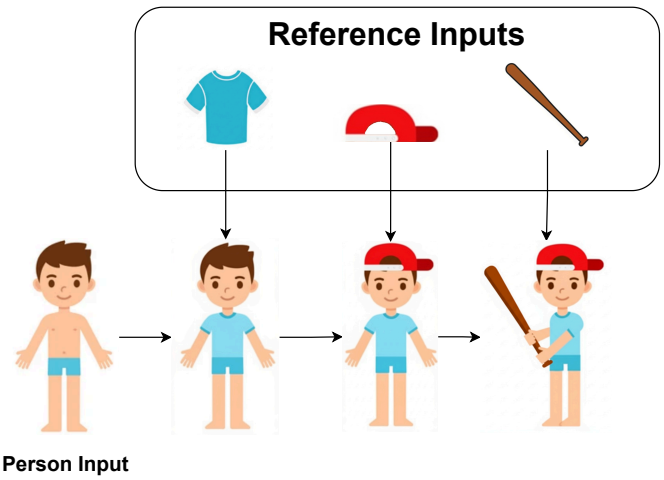

图 2. 多步骤图像生成管道的说明。

另一个新兴方向是多步骤图像生成，它将合成分解为一系列细化操作。与一次性生成图像不同，多步骤流水线允许迭代编辑，实现精细个性化。这种范式特别适用于需要用户控制的应用程序，如头像创建和交互式视觉编辑（图 2）。现代系统例如 ControlNet [15] 和 PhotoMaker [16] 通过支持分阶段变换来体现这一方法——调整人体姿势、编辑背景或修改面部属性。虽然多步骤流水线通过显示中间输出增强了可解释性和用户控制，但也带来了诸如误差累积、复杂度增加以及模块间需要标准化接口等挑战。

在 GenKOL 中，**多步生成是一项核心设计原则**。每个服务作为一个独立的生成单元，既消耗参考输入也消耗中间输出以产生更精细的结果供后续阶段使用。这种模块化结构与更加灵活、用户中心化的生成系统的大趋势相一致，同时解决了可扩展性、适应性和易用性的关键限制。

III. 提议的 GenKOL 系统

A. 概述

GenKOL 利用了一个灵活且模块化的 AI 插件架构，如图 4 所示。这使得用户能够从输入图像和详细的文本提示中创建具有丰富背景的虚拟 KOL 视觉效果，并保持强大的上下文一致性。因此，它优于传统的、松散集成的管道系统。该系统的模块化设计支持可扩展部署、资源的有效利用以及各种生成模型的无缝集成。通过将专业服务链接成定制的工作流，用户可以确保数据流畅传输，加速实验进程，并持续产出高质量且符合背景的输出。

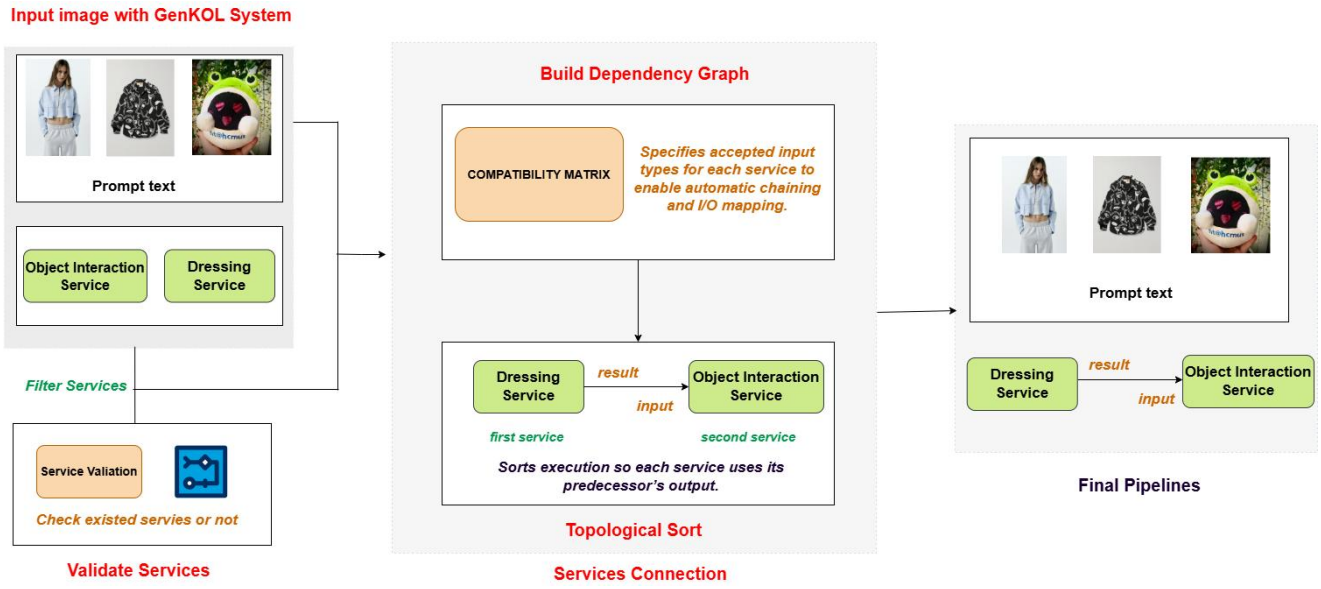

图 3. 兼容性和智能输入-输出映射用于管道生成。

系统的一个基本原则是在模块化 AI 服务之间建立清晰的执行路径。诸如更换服装、虚拟化妆、背景编辑和对象互动等任务以逻辑顺序相连。为了简化执行，GenKOL 采用了一个自动化编排过程来确定服务之间的适当顺序和连接。具体来说，应用了拓扑排序将服务组织为有向无环图（DAG），确保没有服务在其前置条件满足之前执行。一个兼容性矩阵进一步验证潜在的连接，防止不兼容的服务配对并避免执行失败。一旦验证完成，每个服务会自动从先前的输出中分配所需的输入，从而支持灵活且可靠的流程构建。

图 3 展示了在 GenKOL 管道中验证兼容性和集成服务的综合工作流程。其智能映射系统自动化了服务之间的数据传输，最小化人工干预并减少配置错误。这种自动化允许用户轻松地从搜索查询直接创建复杂的工作流，同时保持灵活性、技术正确性以及主动错误管理，最终增强模块化、可扩展性和动态部署。

面部姿态、形状和外观的变化是处理真实人类或 KOL 图像时面临的一个关键挑战，这可能导致合成过程中出现错位。为了解决这一问题，GenKOL 整合了一个预训练的面部特征点检测模型，该模型可以识别出 68 个关键面部特征点。通过建立一个标准化的初始姿态，系统显著提高了生成的人脸在流水线中的连贯性和对齐度。随后，合成的脸部会被无缝地重新集成到其原始上下文中，确保稳定性和视觉质量。图 5 提供了 GenKOL 工作流程中使用的预对齐面部输入示例。

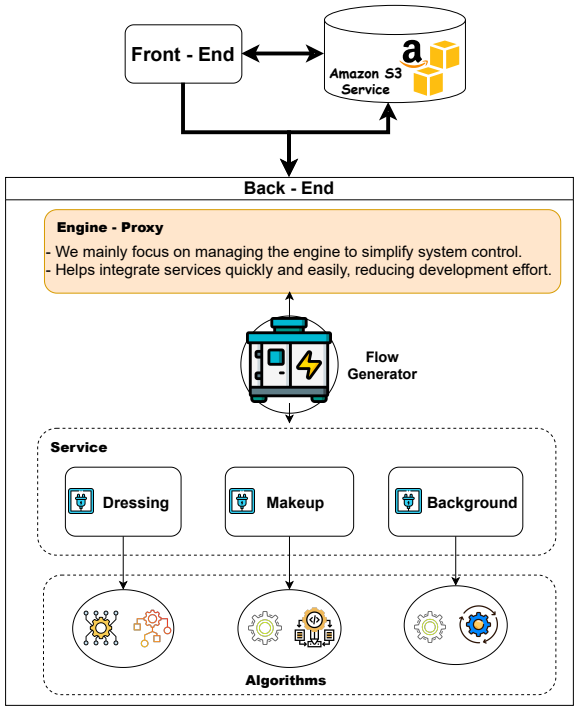

图 4. 提议的模块化人工智能服务架构。

B. 系统架构

我们提出了一种模块化、基于插件的架构，简化了将 AI 模型作为独立服务进行集成的过程（图 4）。每个

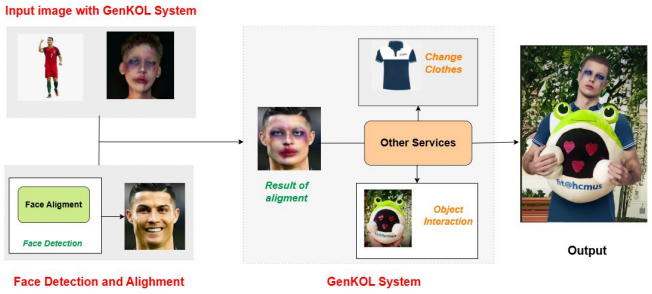

图 5. 面部检测和对齐程序的概述，使用预训练的关键点模型以确保在执行 GenKOL 的生成服务之前进行姿态归一化。

服务，例如图像编辑、文本转图像生成或背景替换，都被封装为具有标准化接口的独立模块，实现无缝通信和顺序执行。这种设计允许服务以最小的努力安装、更换或移除，支持快速实验并减少资源开销。通过仅加载必要的组件，系统优化了资源使用并提高了可扩展性，同时对多种算法版本的支持使用户能够根据应用需求平衡速度、准确性和视觉质量。该架构由四个关键组件组成：引擎、流程生成器、服务和算法。

发动机。引擎充当代理层，通过统一的接口标准化 AI 服务之间的通信。所有服务，无论是本地部署还是云部署（例如 AWS、Google Cloud），都必须符合此接口才能在系统中注册。这种即插即用的设计不仅简化了集成和维护，还能够动态重新配置工作流程而不影响整体功能。这种灵活性降低了用户的操作难度并加快了新功能的部署。

流生成器。流生成器从注册的服务中组合可执行的管道，允许用户构建特定任务的工作流。它支持 AI 驱动处理链的迭代细化，使得可以将诸如衣物转移和妆容应用等服务整合成连贯的管道。通过自动化管道构建，流生成器减少了配置错误，最大限度地减少了人工干预，并为开发人员和最终用户节省了时间。

服务。服务层管理算法的注册和执行上下文，无论是作为本地机器学习模型实现还是远程推理 API。此设计使用户能够精细控制资源分配，基于可用内存、GPU 容量或可扩展性要求做出部署决策。因此，该系统可以适应多种计算环境，从小型个人设备到大型商业基础设施，使其具有成本效益和高效性。

算法。每个服务（例如虚拟试衣、妆容转移或场景编辑）都可以使用多个可互换的算法。这些算法与部署无关，可以在异构环境中执行，包括本地设备、私有服务器和基于云的平台。这种灵活性确保架构能够适应各种用例，从个人内容创作到企业规模的营销活动。

C. 模块化扩展性

GenKOL 采用基于插件的架构，便于新模型的无缝集成。要纳入一个新模型，用户首先需要将其注册到 Engine 中，Engine 起着所有服务集中控制器和代理的作用。这一步注册使得该模型能够在整个系统中被一致地引用、调度和调用。

注册完成后，用户在算法模块中实现相应的服务逻辑。这包括定义标准化的输入和输出接口以及指定针对模型任务（例如服装合成、背景替换或面部属性编辑）定制的执行方法。完成这些步骤后，该模型将被封装为插件，允许其动态插入到任何用户定义的处理管道中。

引擎通过调用具有正确输入的适当插件服务来编排执行。为了保证跨相互依赖的服务之间的正确执行顺序，维护了一个依赖矩阵来表示插件之间的关系。这种结构化的方法允许系统在构建管道时自动确定有效的执行序列，从而最小化手动协调并减少错误传播。

插件集成工作流的一个示例，包括模型注册、方法绑定和服务映射，在图 6 中进行了说明。这种可扩展性确保了 GenKOL 可以轻松适应生成式 AI 的进步，通过支持快速整合新兴模型而不扰乱现有工作流程。

IV. 实验

A. 实验设置

由于当前没有应用程序提供与 GenKOL 相同的全部功能范围，因此无法进行一对一的直接比较。大多数工具专注于特定功能；例如，KlingAI[1] 专注于面部编辑和化妆效果，而 Fitroom[2] 则便于虚拟试衣但不包括面部特征。Maybelline[3] 允许尝试化妆品而不整合衣物或全图生成，而 TRYO[4] 提供基于 AR 的试穿但缺乏 AI 驱动的生成能力。相比之下，GenKOL 集成了这些功能，允许直观高效地创建可定制的虚拟关键意见领袖（KOL）。表 I 显示了各种主要工具与 GenKOL 的功能对比。

在缺乏全面基准的情况下，我们通过用户满意度研究来评估 GenKOL，以获得公正的反馈并评估 GenKOL 作为图像生成智能系统的有效性。实验在一个配备 40GB NVIDIA GPU 的 Linux 服务器上执行，以便在评估阶段促进并行图像生成。

[1]https://www.klingai.com/global/
[2]https://fitroom.app
[3]https://www.maybelline.com/virtual-makeover-makeup-tools
[4]https://apps.apple.com/us/app/tryo-virtual-try-on-ar-app/id1640247631

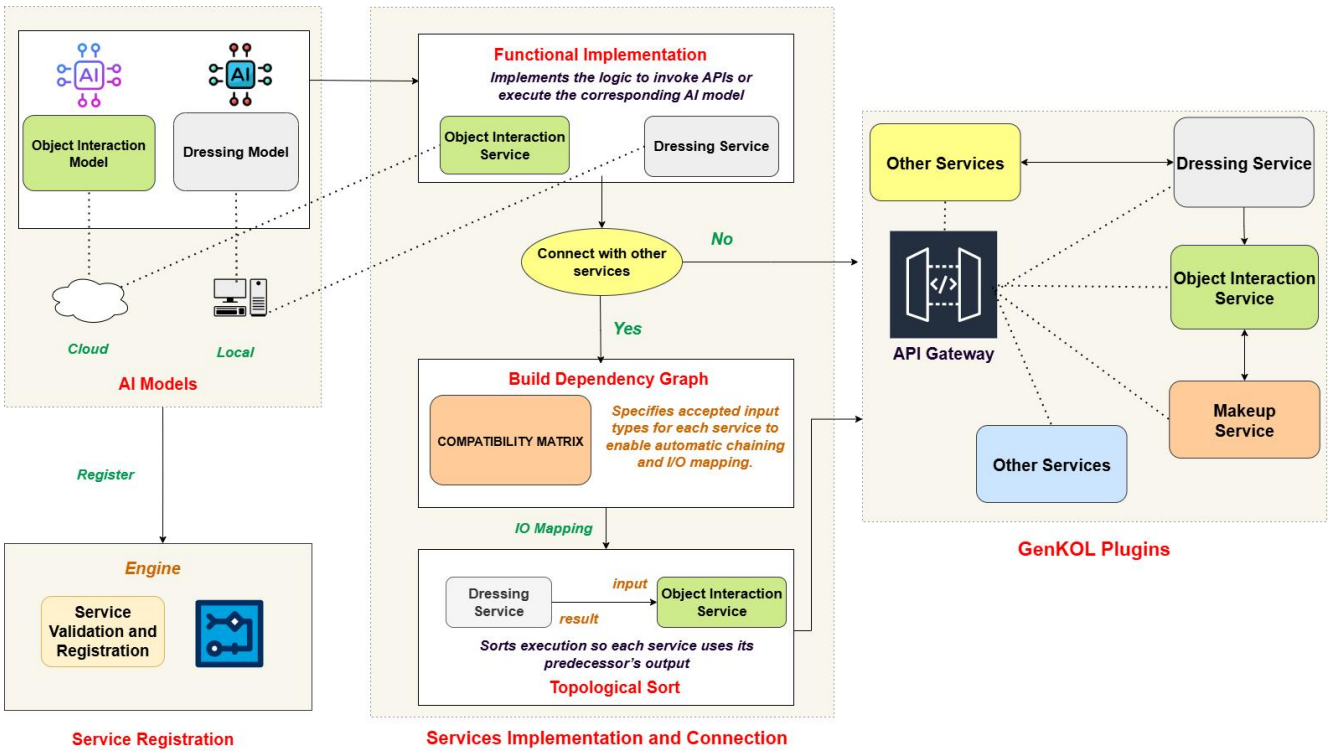

图 6. GenKOL 中基于插件的集成工作流概述。

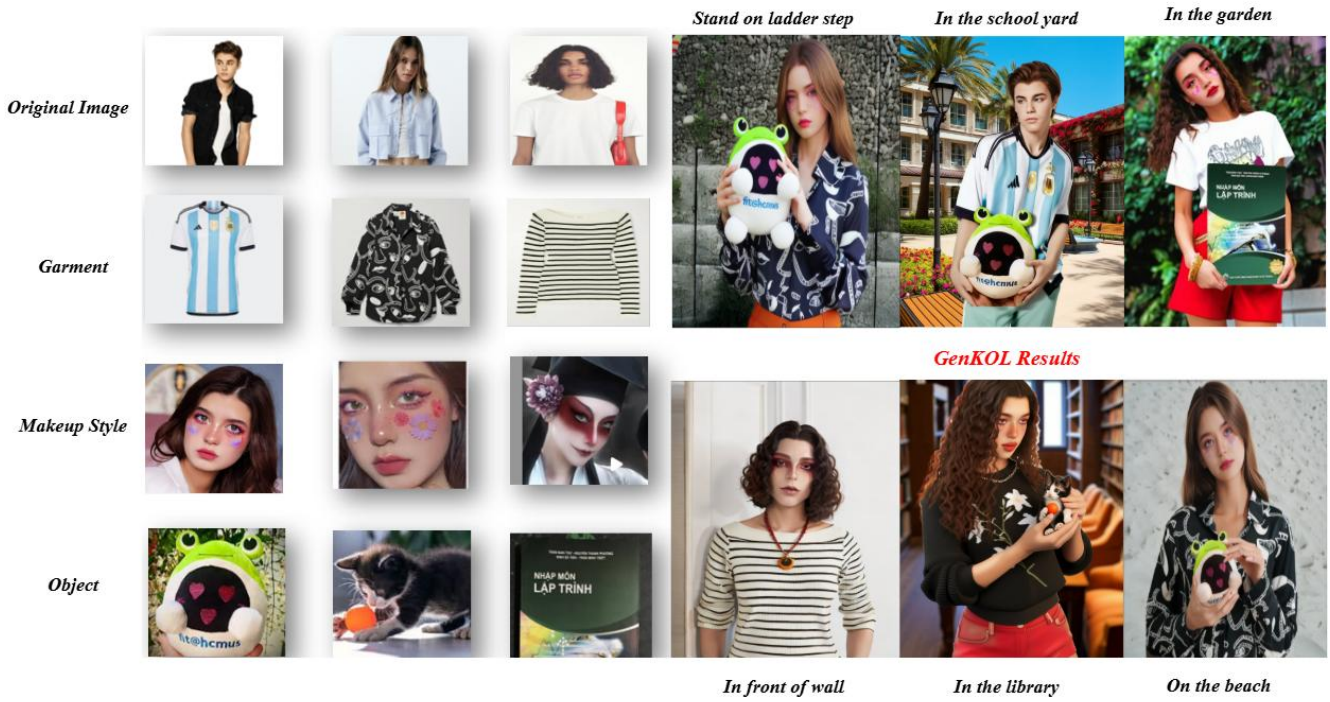

图 7. 生成结果示例的 GenKOL。给定一张原始图像（顶部行，左）和每个属性对应提示（服装、化妆风格和互动对象），我们的 GenKOL 系统（最右侧列）合成将所有元素（包括指定背景）无缝结合的真实虚拟 KOL。

图 8. GenKOL（我们的）、Gemini-2.0 和 ChatGPT-4o 在不同场景和视觉提示下的图像输出定性比较。

表 I
选定工具和应用程序中可用功能的调查与 GenKOL 相比。

工具	虚拟试穿	化妆	背景 修改	对象交互
KlingAI	✗	✓	✓	✗
健身室	✓	✗	✗	✗
梅贝尔琳	✗	✓	✗	✗
尝试	✓	✓	✗	✗
生成 kol	✓	✓	✓	✓

表 II
在图像生成时间、感知质量和对象一致性方面对 GenKOL、Gemini-2.0 和 ChatGPT-4o 进行定性比较。

方法	平均生成时间（秒）	图像质量	一致性
Gemini-2.0	**30**	Neutral	Bad
ChatGPT-4o	600	**非常好**	Neutral
GenKOL	300	Good	**非常好**

B. 定性评估

1) 生成图像的展示： 我们的实验图像生成结果展示了 GenKOL 在制作高度逼真的广告视觉效果方面的能力。生成的产品图片展现了多种风格、性别、年龄和场景，提供了大量选项以满足任何营销活动的需求。图 7 展示了这些 KOL 图像的示例，它们拥有出色的品质和真实感。这一突破大大减少了产品和品牌设计所需的时间，为公司带来了显著的成本节约。因此，组织可以更好地将资源分配到产品开发和客户互动等领域，提高整体营销活动的效果。

2) 与 Gemini-2.0 和 ChatGPT-4o 的比较： 我们评估了 GenKOL 系统，将其与最先进的图像生成模型进行了对比测试，具体与 Gemini-2.0-Flash-Preview-Image-Generation 和 ChatGPT-4o 进行了受控比较。评估重点集中在三个关键标准上：图像生成时间、视觉质量和情境一致性（图 8）。实验结果显示 GenKOL 能够产生高质量的图像，并且在服装、化妆和环境互动方面具有强烈的一致性。虽然与 Gemini-2.0-Flash-Preview-Image-Generation 相比，GenKOL 需要更长的生成时间，但其输出包含更多细节并且情境连贯性更强。相比之下，GenKOL 的生成速度几乎是 ChatGPT-4o 的两倍，而视觉质量仅略低一些。在表 II 中提供了模型性能的详细比较，包括定量和定性指标。

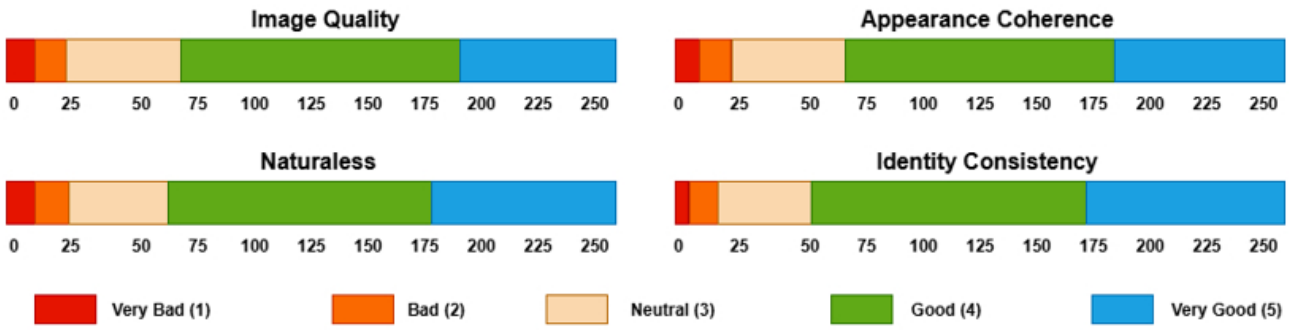

图 9. 生成图像用户研究中四个评估指标的评分分布。水平条形表示每个指标的汇总得分。

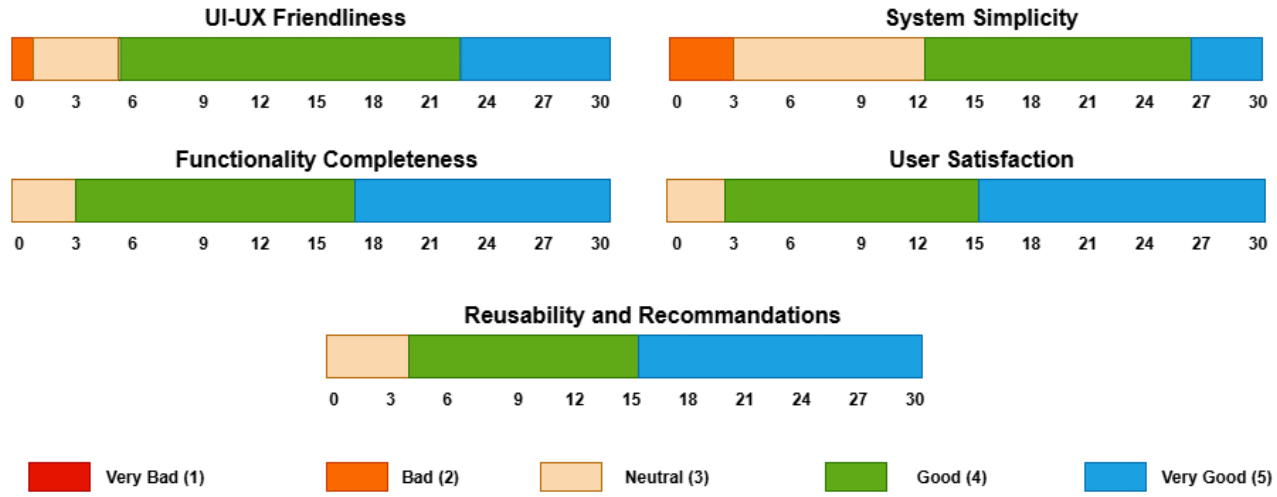

图 10. 用户研究评级分布针对提出的 GenKOL 系统，涵盖五个评估指标：界面友好性、系统简洁性、功能完整性、用户满意度和可重用性与推荐。

C. 用户研究

D. 生成图像的评估

为了评估图像生成的有效性，我们进行了一项全面的用户研究，重点关注视觉结果和生成过程中使用的提示。共有 254 名参与者被招募，年龄在 15 至 35 岁之间，来自软件开发、经济学、工程和技术教育等多个专业领域。目的是获得该系统实际可用性的见解、用户感知以及整体满意度。

我们编译了一个包含 200 张高质量虚拟 KOL 图像的数据集，并将评估组织成四个主题的 Google 表单，每个表单针对系统不同的功能方面：试穿、化妆应用、背景替换和对象交互。每个表单都提供了分步说明、示例和精心策划的图像集以供评估。

参与者根据逼真度、相关性和视觉质量等因素，以五点李克特量表（1=非常差，5=非常好）对生成的图像进行了评分。为了补充定量评估，在每个表格的末尾提供了开放式评论框，允许参与者提供定性反馈。

结果汇总如图 9 所示，表明用户满意度很高，大多数评分集中在 4 到 5 的范围内（良好至非常好）。反馈突显了 GenKOL 提供真实、美观且情境相关的视觉效果的能力。这些发现证实该系统在生成符合或超过用户对真实性、保真度、维度和审美质量期望的营销准备内容方面具有强大的潜力。

E. GenKOL 系统评估

除了评估视觉输出的质量外，我们还进行了一项用户体验（UX）研究来评价 GenKOL 平台的友好性、易用性和可访问性。为了确保广泛适用性，对来自不同专

业领域的 30 名参与者进行了开放调查，包括经济、工程和非 IT 领域。

为参与者设计了三种不同的生成流程，范围从简单的单服务转换到涉及多种服务的更复杂的流程，如化妆应用、服装转移和对象交互。为了促进这一过程，每个服务都提供了五个精选的样本输入图像，同时鼓励参与者使用来自在线来源的自选图像。这种双重方法确保了评估的一致性和探索系统在不同情境下适应性的灵活性。

参与者被指示探索 GenKOL 的具体功能，并在五点李克特量表上评估他们的体验（1 = 非常差，5 = 非常好）。评估涵盖了五个标准：UI/UX 友好性、系统简洁性、功能性完整性、用户满意度以及可重用性和推荐度。这种结构化设计使我们能够捕捉到界面的易访问性以及多步骤生成工作流的实际应用性。

如图 10 所示，结果表明评估一致为正面评价，大多数评分在所有五个标准中都落在 4-5 的范围内。这些发现证实了 GenKOL 提供了一个直观且用户友好的界面，并支持高质量、可适应的虚拟 KOL 生成，展示了对最终用户的强大可用性和易用性。

V. 结论与未来工作

在本文中，我们介绍了一种模块化、插件式的架构，旨在简化各种 AI 服务的集成，特别是 GenKOL。GenKOL 创造了虚拟 KOL 的真实视觉效果，并以其灵活性、可扩展性和用户友好的设计而著称，使其成为营销和电子商务领域理想的选择，用于驱动 AI 生成的视觉内容。它支持诸如虚拟着装和化妆应用等任务的动态工作流，显著减少了时间和资源投入以及生产成本，同时保证了高质量输出。

然而，GenKOL 面临的挑战是我们计划在未来更新中解决的问题。输出质量取决于各个插件的性能和兼容性，这促使我们开发一个自动化的插件验证系统。我们还旨在通过自适应重新排序任务来优化管道执行以提高性能。这些改进将巩固 GenKOL 作为下一代 AI 驱动视觉内容生成可靠平台的地位。

致谢

本研究由越南国家科技发展基金会（NAFOSTED）资助，资助编号为 102.05-2023.31。

参考文献

[1] Y. He, “The Influence of KOL (Key Opinion Leader) Marketing Model on the Consumption Behavior of Generation Z,” *Frontiers in Business, Economics and Management*, 12 2024.

[2] X. Du, N. Kolkin, G. Shakhnarovich, and A. Bhattad, “Generative models: What do they know? do they know things? let’s find out!” *arXiv preprint arXiv:2311.17137*, 2023.

[3] F. Shen, X. Jiang, X. He, H. Ye, C. Wang, X. Du, Z. Li, and J. Tang, “Imagdressing-v1: Customizable virtual dressing,” in *AAAI Conference on Artificial Intelligence*, 2025.

[4] K.-N. Nguyen-Ngoc, T.-T. Phan-Nguyen, K.-D. Le, T. V. Nguyen, M.-T. Tran, and T.-N. Le, “Dm-vton: Distilled mobile real-time virtual try-on,” in *IEEE international symposium on mixed and augmented reality adjunct (ISMAR-adjunct)*, 2023, pp. 695–700.

[5] Y. Zhang, L. Wei, Q. Zhang, Y. Song, J. Liu, H. Li, X. Tang, Y. Hu, and H. Zhao, “Stable-makeup: When real-world makeup transfer meets diffusion model,” *arXiv preprint arXiv:2403.07764*, 2024.

[6] A. E. Eshratifar, J. V. Soares, K. Thadani, S. Mishra, M. Kuznetsov, Y.-N. Ku, and P. De Juan, “Salient object-aware background generation using text-guided diffusion models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7489–7499.

[7] R. Islam and I. Ahmed, “Gemini-the most powerful llm: Myth or truth,” in *Information Communication Technologies Conference (ICTC)*, 2024.

[8] T.-H. To, D.-K. Nguyen, M.-T. Tran, and T.-N. Le, “Streamlining virtual kol generation through modular generative ai architecture,” in *ACM Multimedia*, 2025.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[10] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[12] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, “Plug-and-play diffusion features for text-driven image-to-image translation,” in *IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1921–1930.

[13] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, “Label-Efficient Semantic Segmentation with Diffusion Models,” in *International Conference on Learning Representations (ICLR)*, 2022.

[14] C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring Categorical Regularization for Domain Adaptive Object Detection,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.

[16] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, “Photomaker: Customizing realistic human photos via stacked id embedding,” in *IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8640–8650.