

基于扩散模型的个性化图像生成方法综述^{*}

何子健, 李冠彬

(中山大学 计算机学院, 广东 广州 510006)

通信作者: 李冠彬, E-mail: liguanbin@mail.sysu.edu.cn



摘 要: 随着深度学习技术和扩散模型的快速发展, 图像及视频生成模型展示了高质量、多样化的强大生成能力. 如何利用这些模型实现高效、精准的个性化生成成为当前研究的热点. 个性化图像生成方法能够通过结合文本描述和用户提供的特定概念或主体, 实现定制化图像的生成, 满足用户对个性化视觉内容的多样化需求. 综述基于扩散模型的个性化图像生成方法, 从生成目标的角度将现有方法分为单主体驱动生成和多概念组合生成两类, 前者聚焦于根据单一主体生成定制化图像, 重点研究如何精确捕捉和重建主体的视觉特征, 后者则专注于将多个概念或主体融合到同一图像中, 解决跨概念语义对齐和视觉一致性问题. 结合具体任务和应用场景, 对个性化生成代表性工作进行了详细分析. 此外, 比较和总结了常用的数据集、生成模型的评估方法和个性化生成方法间的性能对比, 进一步探讨了个性化生成方法在实际应用中面临的挑战及未来发展方向, 对研究趋势进行了展望. 旨在为相关领域的研究者提供全面的参考, 推动个性化生成方法的发展与创新.

关键词: 扩散模型; 个性化生成; 特征表达; 图像生成; 概念保持

中图法分类号: TP391

中文引用格式: 何子健, 李冠彬. 基于扩散模型的个性化图像生成方法综述. 软件学报, 2026, 37(4): 1854–1884. <http://www.jos.org.cn/1000-9825/7511.htm>

英文引用格式: He ZJ, Li GB. Review of Personalized Image Generation Methods Based on Diffusion Models. Ruan Jian Xue Bao/Journal of Software, 2026, 37(4): 1854–1884 (in Chinese). <http://www.jos.org.cn/1000-9825/7511.htm>

Review of Personalized Image Generation Methods Based on Diffusion Models

HE Zi-Jian, LI Guan-Bin

(School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

Abstract: With the rapid development of deep learning technologies and diffusion models, image and video generation models have displayed powerful capabilities to produce high-quality and diverse results. How to leverage these models for efficient and precise personalized generation has become a current research hotspot. Personalized image generation methods can combine text descriptions with specific concepts or subjects provided by users to enable the creation of customized images and meet the diverse needs of users for personalized visual content. This study reviews personalized image generation methods based on diffusion models, categorizing existing methods from the perspective of the generation target into single-subject driven generation and multi-concept combination generation. The former focuses on generating customized images according to individual subjects, emphasizing the accurate capture and reconstruction of the subjects' visual features. The latter focuses on merging multiple concepts or subjects into a single image, addressing challenges like semantic alignment across concepts and visual consistency. This study provides a detailed analysis of representative work in personalized generation by combining specific tasks and application scenarios. Additionally, this study compares and summarizes common datasets, evaluation methods of generation models, and performance comparisons between different personalized generation methods. It further discusses the challenges that personalized generation methods face in practical applications and the future development directions, and offers a prospect for the research trends. This study aims to provide comprehensive references for researchers in relevant fields, fostering the development and innovation of personalized generation methods.

^{*} 基金项目: 国家重点研发计划 (2024YFB3908503); 国家自然科学基金 (62322608); CCF 快手科研合作基金 (CCFKuaiShou 2024007)
收稿时间: 2025-04-17; 修改时间: 2025-06-16, 2025-07-21; 采用时间: 2025-08-12; jos 在线出版时间: 2025-11-26
CNKI 网络首发时间: 2025-11-27

Key words: diffusion model; personalized generation; feature representation; image generation; concept preservation

近年来,生成模型在图像生成领域取得了突破性进展,从早期的生成对抗网络 (generative adversarial network, GAN)^[1]到如今的扩散模型 (diffusion model, DM)^[2],这些模型展示了生成高质量、多样化图像的强大能力.特别是基于文本到图像的生成模型,通过大规模预训练和文本输入的灵活控制,能够生成与文本描述高度一致的高质量图像.这些技术进步不仅推动了图像生成领域的发展,还为条件生成、图像编辑、艺术创作等下游任务提供了强大的工具.其中,个性化图像生成作为生成模型的一个重要分支,其目标是根据用户提供的特定概念或主体(通常是1-5张图片)生成定制化的图像,满足用户对个性化视觉内容的多样化需求.早期研究主要采用 GAN Inversion 技术^[3-6],通过将用户提供的图像映射到 GAN 的隐空间,进而通过隐空间属性编辑实现个性化图像生成.然而,这类方法在多个关键维度存在明显局限:首先,训练稳定性方面存在模式崩溃问题,导致模型仅能复现训练数据而缺乏泛化能力,且生成器与判别器之间的对抗平衡难以维持;其次,生成质量方面,在纹理还原和细粒度细节(如毛发、复杂背景等)表现欠佳;最后,其有限的生成多样性和可控性也严重制约了实际应用潜力.这些固有缺陷使得 GAN Inversion 方法^[3-6]难以满足更广泛的应用需求.由于扩散模型在生成质量和多样性方面都取得了很大进步,GAN-based 的方法逐步被 Diffusion-based 的方法取代,并且研究者开始对扩散模型的条件可控生成进行更深入的研究.本文将重点综述基于扩散模型的个性化图像生成.

本文在公开的期刊以及会议论文中检索了自 2022 年起基于扩散模型的个性化图像生成相关的新方法,检索途径主要选用在 IEEE Xplore、ACM Digital Library、Google 学术、arXiv 等文献数据集和搜索引擎进行相关关键词和主题进行检索.结果如图 1 所示,自 2022 年起基于扩散模型的个性化生成相关论文方法增长迅速,至 2025 年 3 月已经达到 670 篇论文.其中,发表在 CCF 评级 A 类期刊或会议的文献共 326 篇,发表在其他非 CCF 评级 A 类期刊或会议的文献共 344 篇.这些数据都表明基于扩散模型的发展,个性化生成已经成为当下计算机视觉的研究热点,图 1 列出了上述 670 篇论文当中有代表性的 35 篇.

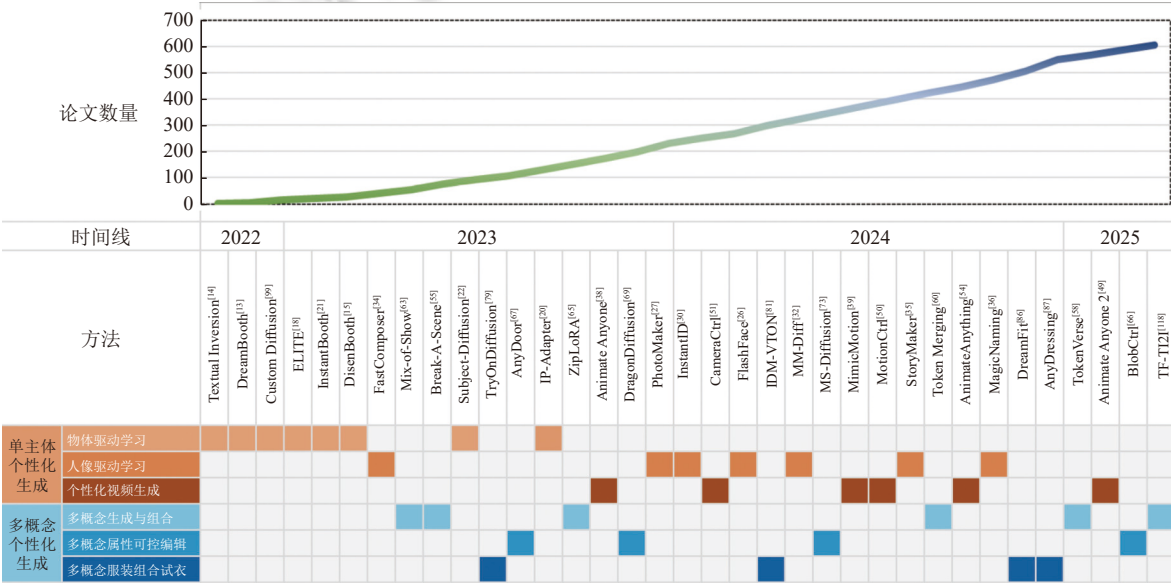


图 1 基于扩散模型的个性化图像生成论文自 2022 年起快速增长

个性化图像生成的核心任务是根据用户提供的少量图像(通常包含特定主体、风格或概念)和文本描述,生成既符合用户需求又保持视觉一致性的图像.这一任务在广告设计、虚拟现实、艺术创作等领域具有广泛的应用前景.例如,用户可以通过提供几张包含特定宠物、人物或艺术风格的图像,生成该主体在不同场景或风格下的定制化图像.现有的个性化图像生成方法可以从任务上分为两大类:单主体个性化生成和多概念组合生成.单主体个

个性化生成侧重于根据用户提供的单一主体 (如特定物体、人物或风格) 生成定制化图像, 重点在于如何精确捕捉和重建主体的视觉特征. 例如, 用户可以通过提供几张包含特定宠物的图像, 生成该宠物在不同场景下的图像. 而多概念组合生成则专注于将多个概念或主体融合到同一图像中, 解决跨概念的语义对齐和视觉一致性问题. 例如, 用户可以通过提供多个主体的图像, 生成这些主体在同一场景下的组合图像. 图 2 列出个性化生成方法的分类, 大致可以分为单主体个性化生成和多概念个性化生成. 目前研究有逐步从单主体到多概念的发展趋势. 基于这些研究成果, 企业发布了基于大数据训练的个性化内容生成模型, 以 OpenAI 发布的 ChatGPT4o、谷歌的 Gemini、快手的可灵以及字节跳动的豆包为代表, 这类模型展现出强大的多模态理解和生成能力, 能够根据用户输入的文本描述以及参考图像生成高质量、高保真度的个性化内容 (如图 3 所示). 这些大模型通常采用扩散模型或混合自回归架构, 通过海量多模态数据预训练获得强大的跨模态对齐能力. 相比传统工具, 它们不仅能实现更精细的风格控制和内容定制, 还能处理复杂的时空动态生成任务.

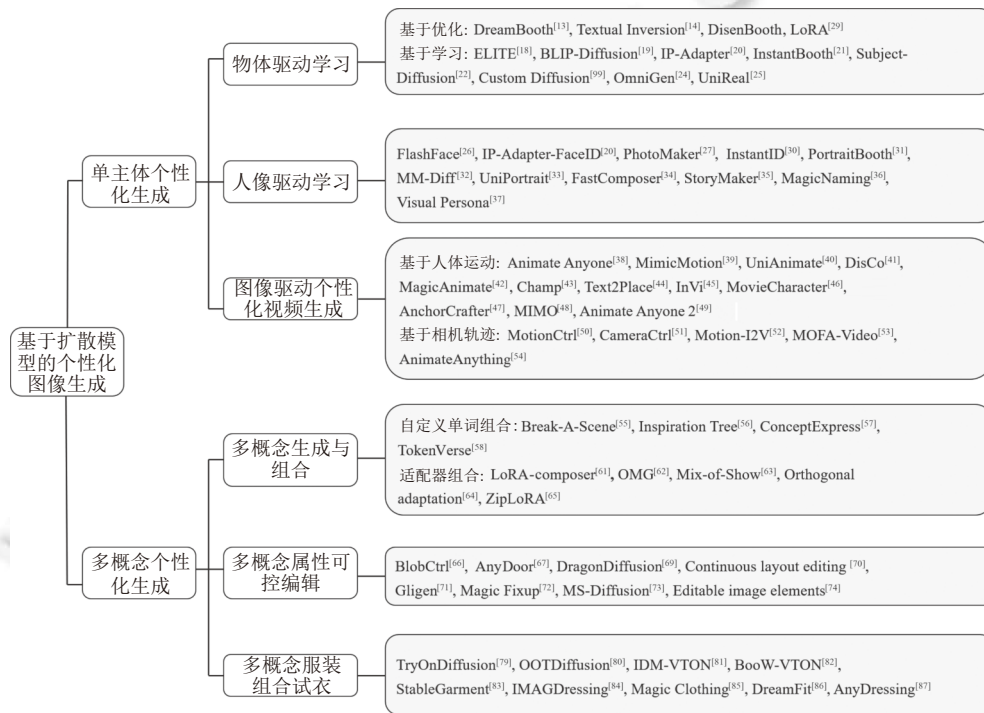


图 2 个性化图像生成方法分类

本文旨在系统性地综述基于扩散的个性化图像生成方法, 为研究者对该领域的研究提供全面的参考. 本文从个性化生成的相关任务出发, 从物体驱动学习、人像驱动学习以及图像驱动个性化视频生成共 3 个任务综述单主体个性化生成的方法; 并从多概念生成与组合、多概念属性可控编辑以及多概念服装组合试衣共 3 个任务综述多概念个性化生成. 这些任务的可视化如图 4 所示. 本文第 1 节给出扩散模型的数学背景以及个性化生成的定义, 并对个性化生成给出统一的框架, 将个性化生成过程标准化为概念反演和个性化生成两个主要阶段. 第 2 节归纳梳理单主体个性化生成的方法, 并从研究对象将这些方法分为物体驱动学习, 人像驱动学习以及图像驱动的个性化视频生成这 3 部分. 第 3 节则从单主体拓展到多概念, 归纳梳理多概念组合生成的方法, 包括多概念组合和生成, 多概念属性可控编辑和针对虚拟试衣的多概念组合. 第 4 节列出相关评估指标和方法, 总结常用的数据集, 并对个性化生成方法的实验结果进行分析. 第 5 节梳理当前基于扩散模型的个性化生成仍存在的问题和挑战, 并基于此展开对未来的研究方向. 最后第 6 节进行本文的总结.

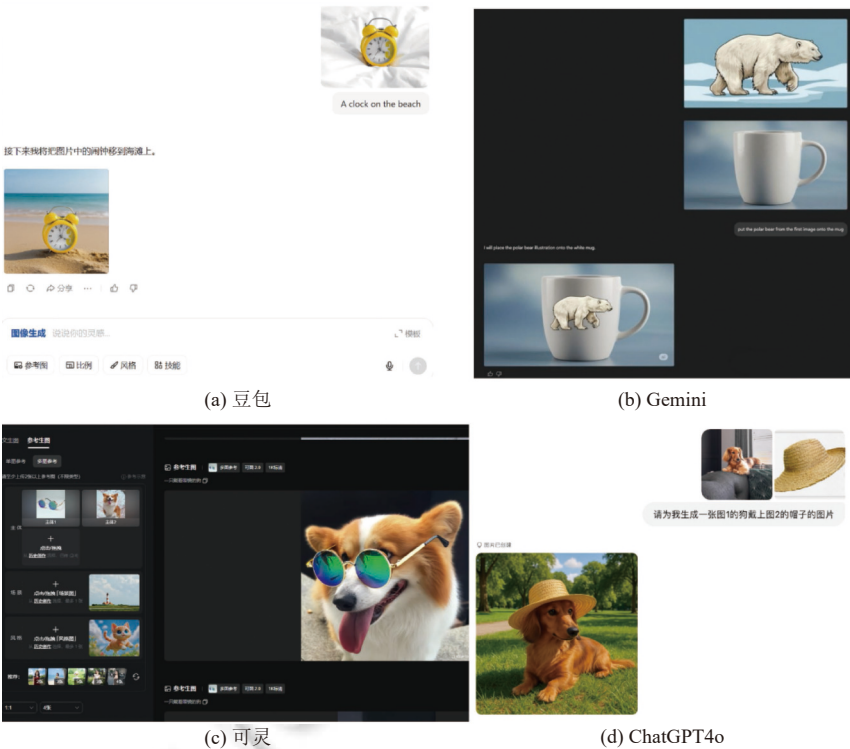


图 3 企业发布的个性化内容生成模型交互界面

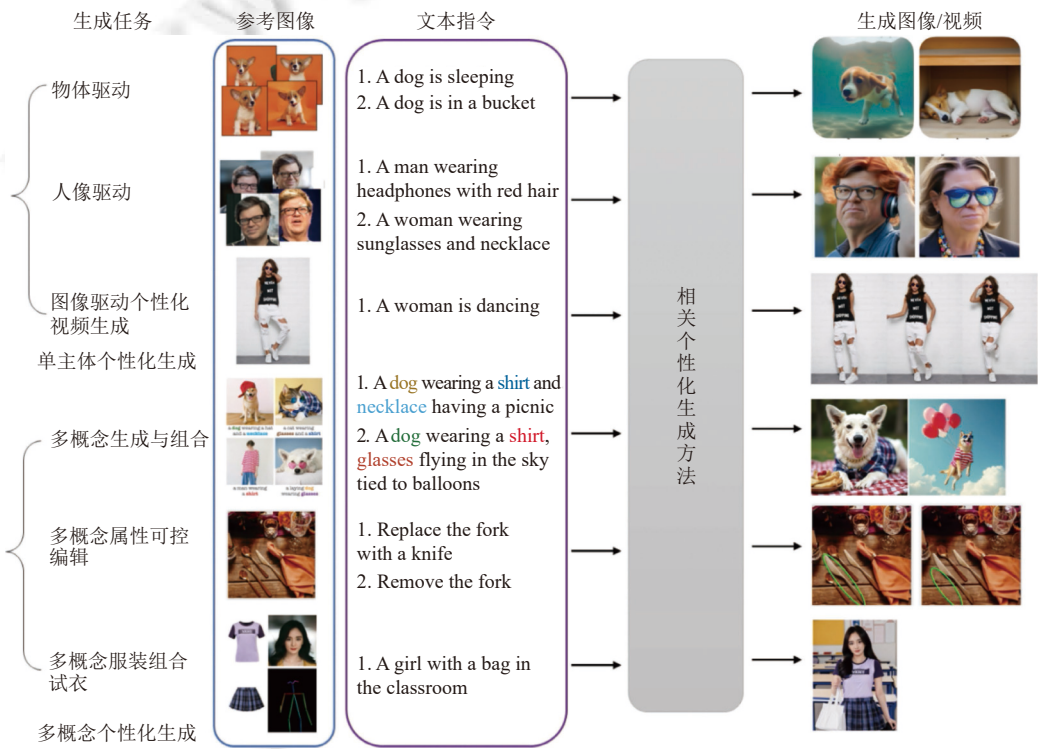


图 4 可视化基于扩散模型的个性化图像生成相关任务

1 问题定义

1.1 扩散模型基础

扩散模型的发展经历了从理论构建到大规模应用的完整演进过程. 早期的理论基础可追溯至 Sohl-Dickstein 等人^[7]提出的非平衡热力学框架, 将扩散过程与生成建模相结合. 2020 年 Ho 等人^[8]提出的去噪扩散概率模型 (DDPM) 通过固定方差和逐步去噪的策略显著提升了生成质量, 成为该领域的重要里程碑. 随后在 2021 年, 研究者们通过 Song 等人^[9]的 DDIM 加速采样、Rombach 等人^[2]的潜在扩散模型 (LDM) 降低计算成本, 大幅提升了模型的实用性和生成能力. 2022 年成为扩散模型爆发式发展的关键年份, 开源的 Stable Diffusion 推动了技术普及, 同年 OpenAI 发布的 DALL·E 2 和谷歌的 Imagen 展示了强大的多模态生成能力. 截至目前, 绝大部分基于扩散模型的图像生成依然采用潜在扩散模型的架构, DDIM 也是最常用的采样器之一.

具体来说, 扩散模型^[2]是一种通过迭代去噪过程去学习目标分布的生成模型. 扩散模型包含了一个马尔可夫链前向传播以及反向传播的过程. 在前向传播的过程中, 数据样本 x 经过 T 次加噪成噪声样本 z_T ; 而在反向传播过程中, 我们需要将噪声样本 z_T 恢复成原样本 x , 这一过程也是样本生成的过程. 扩散模型能够在反向传播过程中加入各种信号, 例如文本和图像, 从而去引导样本的生成. 一个条件可控的扩散模型 $\hat{x}_\theta$ 能够通过加权去噪的得分匹配损失函数:

$$E_{x,c,\varepsilon,t}[w_t\|\hat{x}_\theta(\alpha_t x + \sigma_t \varepsilon) - x\|_2^2] \quad (1)$$

其中, x 是目标样本, c 是条件输入, $\varepsilon \sim N(0, I)$ 表示高斯噪声. α_t 、 σ_t 以及 w_t 是扩散模型中与时间步数 t 相关的函数变量. 目标样本通过马尔可夫链前向传播被污染为 $z_T \equiv \alpha_T x + \sigma_T \varepsilon$, 而 $\hat{x}_\theta$ 则是噪声预测函数, 将 z_T 还原为原样本 x . 在推理生成阶段, 数据样本能够从高斯噪声 $z_T \sim N(0, I)$ 开始并通过 DDIM^[9]采样器进行去噪生成. 为减少训练和推理所需的计算资源并且尽量不降低图像的生成质量, 一般使用预训练的自编码器将数据样本 x 压缩到潜在空间中. 图 5 给出扩散模型去噪进行图像生成的过程.

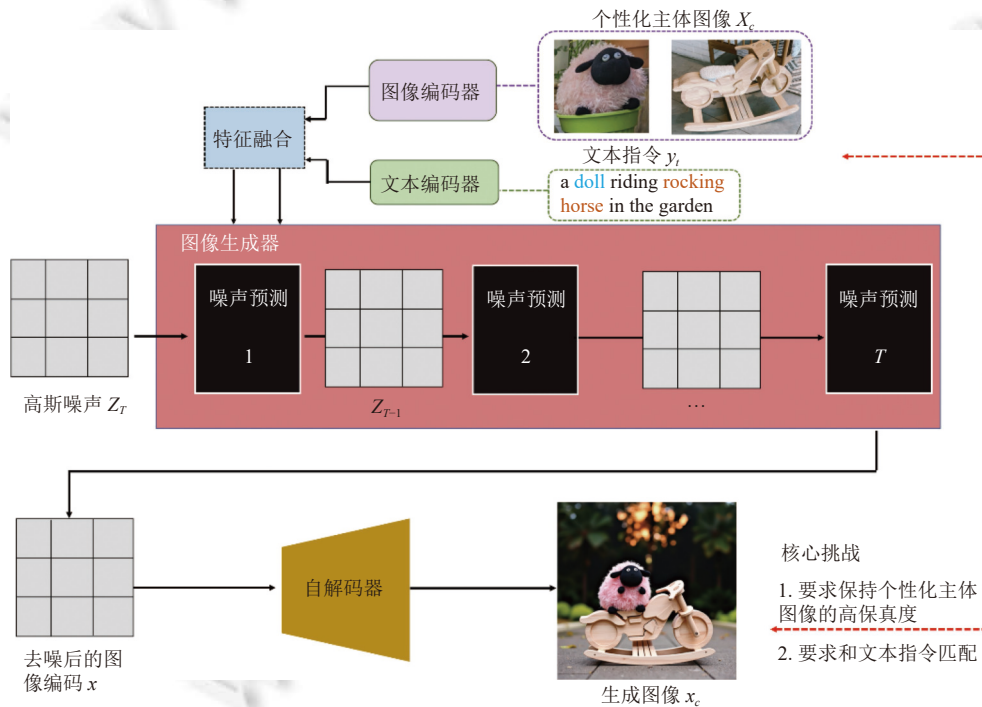


图 5 基于扩散模型的个性化图像生成过程及挑战

1.2 个性化图像生成任务

个性化图像生成的核心任务是生成包含用户指定概念 (例如特定主体、人脸或风格) 并符合给定文本指令的图像. 视觉概念由一组图像表示, 记为 X_c , 这些图像包含目标概念 y_c 和文本指令 y_t . 例如图 5 中 X_c 包含了两个概念主体玩偶以及木马, 这两个概念分别用文本“doll”和“rocking horse”表示, 文本指令是“a doll riding rocking horse in the garden”. 个性化图像生成的核心挑战在于两点, 一是要求保持个性化主体图像的高保真度, 二是满足多样化文本指令的精准对齐. 在此, 我们首先在不同生成模型的统一框架下定义个性化图像生成. 形式上, 生成模型 G 的生成过程可以描述为:

$$x = G(\varepsilon, c; \theta_G) \quad (2)$$

其中, θ_G 表示生成模型 G 的参数, $\varepsilon \sim N(0, I)$ 是从高斯分布中随机采样的噪声. $c = [c_0, \dots, c_N]$ 是一组可选条件, 例如文本或语义图. 给定概念图像 X_c 和目标文本 y_t , 个性化图像生成的目标是生成图像 x_c , 该图像既包含给定的视觉概念 y_c , 又与 y_t 所指示的描述保持一致:

$$x_c = G(z, c, y_t, X_c; \theta_G) \quad (3)$$

生成过程通常包括概念反演和个性化生成两个主要步骤. 首先在概念反演步骤中, 给定的概念图像 X_c 被投影到表征空间, 以获得条件 c_c :

$$c_c = \phi(X_c) \quad (4)$$

其中, $\phi(\cdot)$ 表示反演操作, 可以使用各种反演空间和反演方法来完成这一投影. 接着我们将 c_c 输入生成模型 G 会生成包含概念 y_c 的图像. 接着在个性化生成中, 概念条件 c_c 与目标文本 y_t 结合, 生成所需图像:

$$x_c = G(z, c, y_t, c_c; \theta_G) \quad (5)$$

本文后续将探讨 G 是扩散模型的情况, 基于扩散模型的个性化生成研究目前的发展.

2 单主体个性化生成

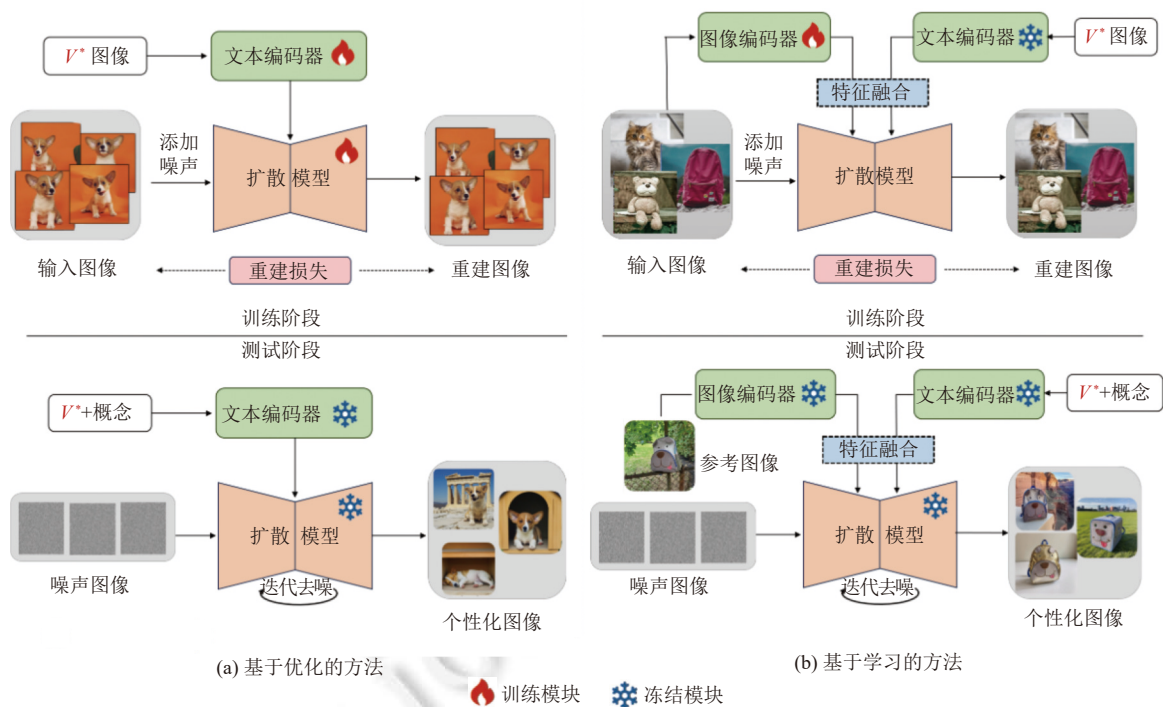
随着基于扩散模型的生成式方法的快速发展, 图像生成质量得到很大的提升^[2]. 研究者逐渐向可控生成的方向^[10-12]进行探索, 其中一个重要的课题就是保持主体的个性化生成. 本节先从单主体的个性化生成开始介绍, 包含物体驱动学习、人像驱动学习、图像驱动个性化视频生成这 3 部分.

2.1 物体驱动学习

物体驱动学习是个性化生成最常见的学习范式, 旨在通过用户提供的少量物体图像或描述, 生成或识别与该物体相关的内容. 与传统的生成或识别任务不同, 物体驱动学习强调对特定物体的精确建模和控制, 使得生成或识别的内容能够高度符合用户的需求. 这些方法大致上可以分为两类: 基于优化的方法和基于学习的方法. 这两类方法训练和推理的流程如图 6 所示.

2.1.1 基于优化的方法

基于优化的方法通过微调生成模型的参数, 将用户提供的主体图像映射到生成模型的表示空间中, 从而生成符合文本描述的图像. 个性化生成的一个关键问题是如何用文本表示特定的概念物体, 一个通用的做法为使用需优化学习的符号 V^* 代表需要学习的新概念. 该符号通过少量数据训练进行优化学习, 这个过程完成后在推理时能够和其他词自由组合进行个性化的生成. DreamBooth^[13]和 Textual Inversion^[14]是两种经典的主体驱动生成方法. DreamBooth 通过微调 U-Net 模型的参数, 将主体图像映射到一个特殊的提示词 V^* 上. 具体来说, DreamBooth 使用少量主体图像和对应的文本描述对生成模型进行微调, 使得模型能够在新生成的图像中保持主体的身份特征. 尽管 DreamBooth 在生成高质量图像方面表现出色, 但其计算成本较高, 且容易受到背景和姿态等无关因素的影响. Textual Inversion^[14]则通过微调提示词的嵌入表示来实现主体驱动生成. 与 DreamBooth 不同, Textual Inversion 不修改生成模型的参数, 而是通过学习一个新的提示词嵌入来表示主体. 这种方法在保持主体身份的同时, 减少了计算开销, 但在生成复杂场景或多主体图像时, 容易出现主体特征丢失或文本描述不一致的问题.



DisenBooth^[15]注意到了微调时主体相关信息和非主体相关信息在潜在嵌入空间中存在耦合,这使得: 1) 隐藏在耦合嵌入中与主体无关的信息主导生成过程,使得生成的图像严重依赖于无关信息,从而忽略了控制的文本描述; 2) 本应该主导生成的主体信息无法得到恰当保留,进而导致生成时图像中的主体发生畸变。为解决这些问题,DisenBooth 在去噪过程中对预训练的扩散模型进行微调,采用解耦的嵌入来分别保留主体身份信息并捕捉与身份无关的信息,并设计了新颖的弱去噪和对比嵌入辅助机制,以实现更高效的解耦。

2.1.2 基于学习的方法

基于学习的方法框架因其在测试阶段无需微调的特性,逐渐成为更具现实应用潜力的解决方案。这类方法的核心思想是通过引入额外的模块或预训练模型(如视觉编码器、适配器模块或分割掩码等),将主体图像直接映射到生成模型的表示空间中,从而避免微调网络的时间开销。如图 6(b) 所示,基于学习的方法会以参考图像和文本描述作为两部分独立输入,并通过在训练阶段大量数据样本重建的训练得到一个良好的图像编码器,同时还会设计一个合适的图像特征和文本特征融合模块,最终将特征注入扩散模型之中。

如何高效获取图像特征是学习方法的关键,一个通用的做法是借助预训练的跨模态模型例如 CLIP^[16]和 BLIP^[17]。ELITE^[18]和 BLIP-Diffusion^[19]是这个思想的两个典型方法。ELITE 通过预训练的视觉编码器 CLIP 提取主体多层次图像特征,然后将这些特征直接映射到文本空间,从而实现高效的主体驱动生成。

ELITE 使用全局和局部映射机制,分别捕捉主体的整体外观和细节特征,从而生成高质量的个性化图像。而 BLIP-Diffusion 设计了一个基于 BLIP-2 的多模态表示学习框架,这个框架主要包含 2 个组成部分:一个预训练的图像编码器,用于提取图像特征,而另一个是多模态编码器(Q-Former),用于图像文本对齐。该多模态编码器的输入是可学习的查找 token 和文本,输出与文本对齐的图像特征。这些查找 token 通过自注意力与文本进行交互以及通过交叉注意力与图像特征进行交互。然而,这两个方法对主体姿态和背景的鲁棒性较差,容易受到无关因素的影响,且依赖于预训练的视觉编码器,可能无法完全捕捉复杂主体的特征。

为突破预训练图像特征提取模型的能力限制, 后续的方法设计了基于参考图像的条件网络架构并将其融合到扩散模型生成的骨干网络之中. 例如 IP-Adapter^[20]训练了一个轻量级的解耦跨注意力模块, 其中图像特征和文本特征分别作为注意力的查询特征. IP-Adapter 无需像 ELITE 或 BLIP-Diffusion 设计可学习的空白 token. Instant-Booth^[21]则采用不同的策略, 通过将主体图像的特征直接注入生成模型的中间层以及引入一些适配器层来学习丰富的视觉特征表示, 能够在不修改模型参数的情况下生成个性化图像. 进一步地, Subject-Diffusion^[22]融合了可学习空白 token 和基于参考图像的条件网络架构做法, 综合两者的优势提高了个性化生成的效果. 此外, 针对主体身份无关因素不可避免会对生成结果的影响, Jia 等人^[23]通过引入分割掩码或用户标注的掩码来排除背景对生成结果的影响. 这种设计使得模型在生成过程中专注于主体特征, 提高生成图像的准确性, 且无需微调生成模型, 计算效率较高.

最近还有研究者采用 diffusion Transformer (DiT) 网络结构实现基于学习的个性化生成^[24,25]. DiT 网络架构比起 U-Net 的一大优势是无须再额外训练一个图像编码器, 仅需把图像 token 输入到 DiT 中, 然后微调预训练好的 DiT 网络或者采用 LoRA 等轻量级模型进行适配. OmniGen^[24]将 Flux (一个预训练 DiT 模型) 应用到物体的个性化生成, 大幅增强了个性化生成的性能. UniReal^[25]引入了图像索引嵌入机制, 将多张图像与其对应的提示词项进行关联映射, 将框架拓展到多主体的个性化生成.

2.2 人像驱动学习

身份保持的人像生成是主体驱动生成的一个重要分支, 在虚拟形象设计、影视制作、广告创意及社交媒体内容生成等实际应用场景中展现出重要价值. 这类方法的核心挑战在于保持生成图像中人物身份特征的一致性. 如图 7 所示, 参考人像也可以被视为一个特殊主体, 其典型处理流程包括: 首先通过人脸编码器提取参考人像的关键面部特征, 经特征投影后输入图像编码器; 随后将文本特征与图像特征融合并注入扩散模型, 最终实现个性化人像生成. FlashFace^[26]、IP-Adapter-FaceID^[20]和 PhotoMaker^[27]等方法是人像驱动学习的代表方法. 这些方法通常利用 ArcFace^[28]提取的 ID 嵌入作为条件, 确保生成图像中面部的高保真度. ArcFace 是一种基于深度学习的面部识别模型, 能够提取高维特征向量来表征面部身份信息. 通过将这些特征向量作为生成模型的输入条件, 可以在生成图像中保留目标人物的身份特征. IP-Adapter-FaceID^[20]是一种基于 LoRA^[29]的 AI 绘画工具, 支持多张参考图像和多种特征提取, 能够根据人脸输入通过快速微调网络, 从而根据提示生成身份一致的图像. FlashFace^[26]将人脸身份编码通过 ArcFace 编码为一系列特征图, 这使得模型能关注更多的脸部细节 (例如疤痕、文身和脸部形状). FlashFace 还设计了一种解耦集成策略, 在文本和图像引导生成过程中取得一个良好的平衡, 缓解人脸和文本提示间的冲突. PhotoMaker^[27]则通过将任意数量的输入 ID 图像编码为堆叠 ID 嵌入, 保留了身份信息, 从而生成高质量的人像图像. 这些方法在保持身份一致性的同时, 各具特色地解决了人像生成中的关键问题.

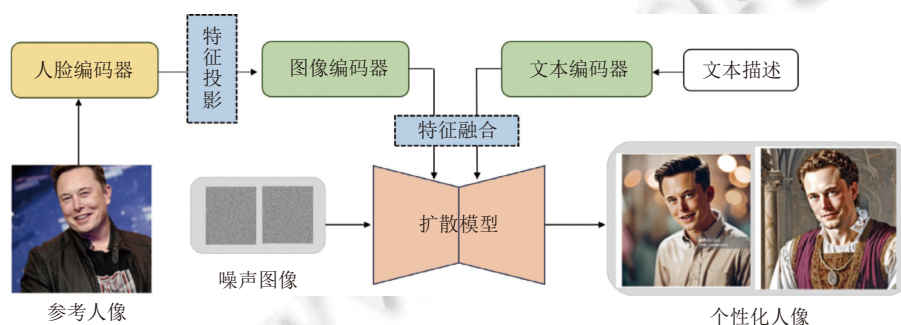


图 7 人像驱动学习示意图

在面部特征控制方面, 除使用 ArcFace 引入面部特征外, 现有研究已发展出多种创新方法. IdentityNet^[30]提出的 IdentityNet 通过引入 5 个面部关键点 (包括眼睛、鼻子和嘴巴等关键部位的位置信息) 来精确控制面部结构, 该方法不仅实现了优异的面部相似性, 还能在复杂场景下保持稳定的身份一致性. PortraitBooth^[31]利用一个无需

微调的人脸识别模型,将该模型对人脸的特征嵌入作为条件进行个性化图像生成,这种引入的动态身份保留策略进一步确保了生成图像与原始图像身份的高度相似性。

此外,一些研究还专注于多 ID 图像生成,即在同一张图像中生成多个不同身份的人物。例如,MM-Diff^[32]通过在训练阶段对不同主题的交叉注意力图施加约束,实现了无需预定义输入的多 ID 图像生成。UniPortrait^[33]则采用 ID 路由模块来统一多 ID 定制,从而避免身份混合问题。FastComposer^[34]是一种无需微调的高效多主体生成方法,通过结合图像编码器提取的主题嵌入和通用文本条件,仅需单次前向传播即可完成个性化图像生成,显著降低了计算成本。StoryMaker^[35]方法在这一领域进一步扩展了应用范围。它不仅专注于保留面部特征,还确保生成图像中服装、发型和身体的一致性。对于多角色生成任务,StoryMaker 引入了位置感知重采样器和注意力损失机制,以解决多角色混合问题。这些创新使得 StoryMaker 在生成复杂场景和多角色图像时表现出色。MagicNaming^[36]提出了一种无需微调 U-Net 的解决方案。通过引入 NN 空间, MagicNaming^[36]能够在不损害模型生成能力的情况下,兼容任何 SDXL 模型变体。这一创新不仅保留了模型的原始生成能力,还提升了技术的可扩展性和实用性。

包含细致人像的全身生成最近也受到了研究者的关注。Nam 等人^[37]提出了一个人类全身个性化生成的基座模型。该模型致力于解决两个关键问题,一是生成符合文本指令的全身图像,包含脸部表情、姿态、动作等的描述;二是全身人像的一致性,包括人脸、服装和饰品等。为解决这 2 个问题, Nam 等人提出了一个数据构造流程并构造了一个 580k 人像样本对的 Visual Persona-500K 数据集,并提出一种基于预训练文本到图像扩散模型的 Transformer 编码器-解码器架构。该架构首先将输入图像解耦为不同身体区域,将这些区域编码为局部外观特征后,独立映射为稠密身份嵌入向量,进而引导扩散模型生成定制化图像。

目前身份保留的个性化图像生成技术在面部保真度、多 ID 生成和语义一致性方面取得了显著进展。然而,如何在保持身份一致性的同时,进一步提升生成图像的多样性和质量,仍然是未来研究的重要方向。此外,探索无需微调的高效方法,以及提升技术的通用性和可扩展性,也将成为该领域的关键挑战。随着深度学习技术和多模态学习方法的不断发展,身份保留的个性化图像生成技术有望在更多实际场景中得到广泛应用,为人工智能生成内容领域带来更多创新和突破。

2.3 图像驱动个性化视频生成

图像驱动个性化视频生成的核心目标是通过结合参考图像和一系列运动信号,合成具有高度真实感和一致性的动画视频序列。基于文本生成图像的扩散模型,我们可以通过简单地在 U-Net 每个模块增加一个时序注意力层实现从图像生成到视频生成的拓展。个性化视频生成最重要的一个应用是人物图像的视频生成。Animate Anyone^[38]使用动作序列引导参考人体静态图像的运动。它基于 Stable Diffusion^[2]设计的框架(如图 8 所示),对 2D U-Net 加入时序注意力模块,将图像生成拓展到参考图像的个性化视频生成。该方法的网络架构采用双 U-Net 设计,通过多层次特征融合实现外观一致性和运动控制。具体而言,参考 U-Net 首先提取参考图像的多尺度特征,随后在每个模块中通过创新的空间注意力机制与主 U-Net 的生成特征进行融合。该空间注意力机制的操作流程包括 3 个步骤:1) 将参考图像特征沿时间维度复制 t 次;2) 在通道维度上与生成图像特征进行拼接;3) 通过自注意力计算实现特征交互。这种设计不仅保留了参考图像的细节特征,还确保了时序上的连贯性。结合姿态引导器的运动控制,后续的研究如 MimicMotion^[39]和 UniAnimate^[40]则进一步改进了基础框架。MimicMotion^[39]利用预训练的视频扩散能力,设计了 PoseNet 来注入骨架信息;UniAnimate^[40]则通过 Mamba 时间建模技术,在时间维度上堆叠参考图像,以实现更高效的角色动画生成。此外,一些研究探索了不同的运动控制信号。例如,DisCo^[41]和 MagicAnimate^[42]使用 DensePose 作为人体表示,而 Champ^[43]则采用 3D 参数化人体模型 SMPL 来整合多模态信息,包括深度、法线和语义信号。

使用动作序列信号等去引导个性化视频生成虽然取得了一定的成功,但也存在着很大的局限性。在当前的跨身份动画 workflows 中,运动信号通常是从不同的视频中提取的,而个性化内容则来源于静态图像。这种分离的处理方式导致生成的角色动画在空间关系和交互性方面存在明显不足。具体而言,角色与其环境之间的空间关系往往缺乏真实感,人与物体之间的内在交互也被破坏。例如,一个角色在复杂场景中的行走或与物体的互动可能显得生硬

或不自然. 因此, 大多数现有方法仅限于生成简单的动作 (如单独的手势或舞蹈), 而无法充分捕捉角色与其环境之间复杂的空间和交互关系. 这些局限性严重限制了角色动画技术的进一步发展, 尤其是在需要高度真实感和复杂交互的场景中. 因此, 一些工作开始研究利用扩散模型生成与场景或交互物体上下文一致的图像或视频, 通过插入或修复技术将人物融入给定场景中, 增强个性化视频的真实性. 例如, Text2Place^[44]通过学习语义掩码并使用文本引导来定位区域, 从而将人物放置在背景场景中. InVi^[45]则通过图像修复和扩展注意力机制生成帧, 实现物体插入. 此外, 一些研究专注于角色与场景或物体的交互动画. 例如, MovieCharacter^[46]将动画角色结果合成到去除人物的视频序列中, 而 AnchorCrafter^[47]则专注于人-物体交互建模, 通过同时感知交互外观特征和注入交互运动模式, 实现了高质量的产品推广视频生成. MIMO^[48]通过空间分解扩散方法, 将视频分解为人类、背景和遮挡物, 并基于 3D 深度重新组合这些元素以生成角色视频. Animate Anyone 2^[49]设计了一种框架, 从源视频中捕捉环境表示作为条件输入, 使模型能够以端到端的方式学习角色与环境之间的内在关系. Animate Anyone 2 通过移除角色区域来构建环境表示, 并生成与环境上下文一致的角色填充这些区域. 为了进一步提升生成效果, 它开发了一种形状无关的掩码策略, 更好地表示角色与其上下文场景之间的边界关系. 这种策略不仅能够有效缓解形状泄漏问题, 还能实现角色与场景的高效整合.

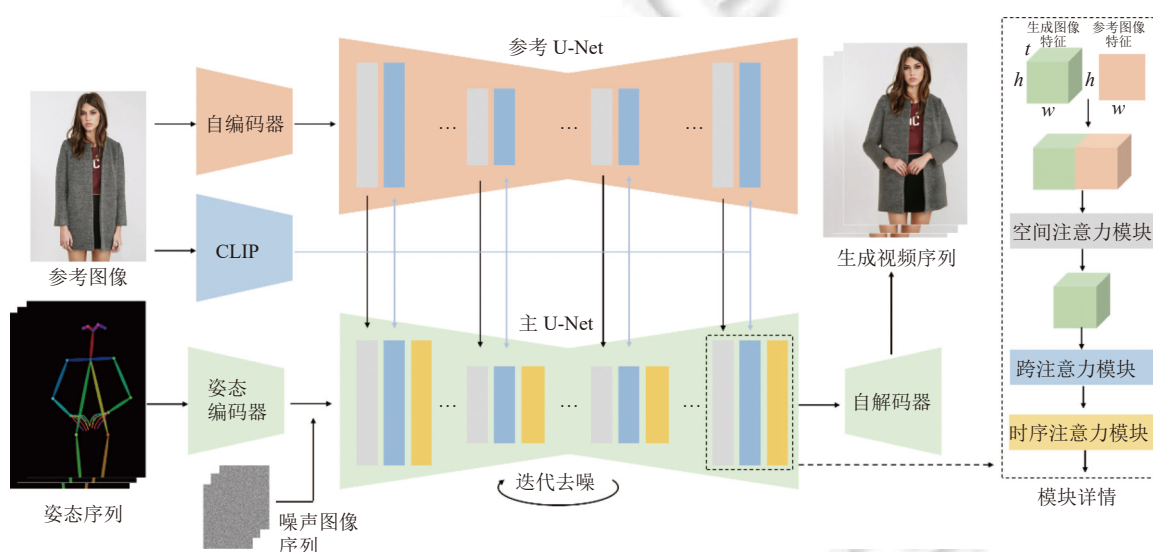


图 8 图像驱动视频生成示意图

除了个性化人物视频生成外, 通过相机轨迹操控动态视频生成是另一种个性化视频生成的形式. MotionCtrl^[50]和 CameraCtrl^[51]是这方面经典的工作, 它们通过文本输入生成控制轨迹的语义向量, 并最终基于此引导参考图像的运动并生成具有特定相机运动效果的视频. 除文本描述外, Motion-I2V^[52]设计了一种支持基于图像引导的视频生成方法. 它允许用户通过拖拽注释来调整物体的运动, 例如调整眼睛的方向或指示腿部的运动. 然而, Motion-I2V^[52]的功能较为有限, 仅支持小幅度的物体运动, 无法实现复杂的相机运动效果 (如环绕拍摄或镜头推拉). MOFA-Video^[53]是另一种能够实现像素级运动控制的方法. 它支持对输入图像中每个局部区域的运动方向进行精确控制, 从而实现小范围的相机运动. 然而, 为了确保动态视频的全局一致性, MOFA-Video 要求用户为输入图像中可能移动的每个局部区域指定运动方向, 确保了高精度的控制和动态视频的全局一致性. AnimateAnything^[54]通过多尺度控制特征融合网络将不同条件 (如相机轨迹、文本提示和用户注释) 统一转化为逐帧光流, 并结合光流作为运动先验来指导视频生成; 同时, 引入基于频率的稳定化模块, 确保视频频域一致性, 有效减少大尺度运动引起的闪烁问题.

综上所述, 个性化视频生成技术在角色动画与环境交互方面取得了显著进展, 但仍面临生成质量、复杂交

互处理和模型适应性等挑战. 未来的研究需要进一步探索更高效的框架和更精细的交互建模方法, 以推动该技术在更广泛的实际应用中的发展. 例如, 如何在不损害模型生成能力的前提下实现更复杂的角色和环境交互, 以及如何提升生成视频的多样性和真实感, 都是值得深入研究的方向. 随着深度学习技术和多模态学习方法的不断发展, 个性化视频生成技术有望在更多实际场景中得到广泛应用, 为人工智能生成内容 (AIGC) 领域带来更多创新和突破.

3 多概念个性化生成

视觉世界是一个由丰富视觉概念组成的复杂图景, 涵盖了多样的物体、姿态、光照条件、材质和纹理, 这些元素通常以复杂的方式交织在一起. 文本到图像模型展现了学习和表达这种复杂性的卓越能力, 生成的图像不仅能够捕捉单个概念, 还能将它们无缝整合到多样且连贯的场景中. 基于此, 研究者逐步从单概念的个性化生成迈向多概念的个性化生成. 本节将分多概念生成与组合、多概念属性可控编辑以及多概念服装组合试衣这 3 部分进行介绍.

3.1 多概念生成与组合

如后文图 9 所示, 多概念的生成与组合目前大概可以分为两类, 第 1 种是自定义单词组合方法, 旨在从单张或少量图像中学习多个概念, 在学习的过程中对这些概念进行语义解耦, 使自定义的概念 token 能够在新的组合中重复使用. 第 2 种是适配器组合方法, 考虑直接组合多个学习好特定概念的适配器模型, 在组合的过程中实现概念间的调制和优化. 先谈第 1 种, Break-A-Scene^[55]采用 DreamBooth 微调方法, 组合概念时使用用户提供的空间掩码来隔离图像中的概念. 然而, 因为它无法解耦对象的姿态和外观, 这种方法对空间掩码的依赖存在显著局限性. Inspiration Tree^[56]和 ConceptExpress^[57]通过从单张图像中联合学习多个词嵌入来提取多个概念, 每个词嵌入代表一个不同的概念. 然而, 这些方法实现的解耦是不可预测的, 无法控制图像的哪些方面被分离为独立概念, 这限制了方法的灵活性和可用性, 使其在生成复杂和多样化组合时效果不佳. TokenVerse^[58]针对基于 DiT^[59]的文本到图像模型进行研究. 它观察到 DiT 的调制空间具有语义性, 能够实现对复杂概念的局部化控制. 基于这一发现, TokenVerse 设计了一种基于优化的框架, 该框架以图像和文本描述为输入, 并为每个词在调制空间中找到一个独特的方向. 这些方向随后可用于生成新图像, 将学习到的概念以所需的配置组合在一起. 对于概念的组合问题, token merging (ToMe)^[60]发现词语 token 经过编码后天然地存在语义耦合并且 token 存在语义可加性. ToMe 手动聚合相关 token 成混合概念, 使得多个混合概念不容易互相干扰, 大大增强了多概念组合的稳定性.

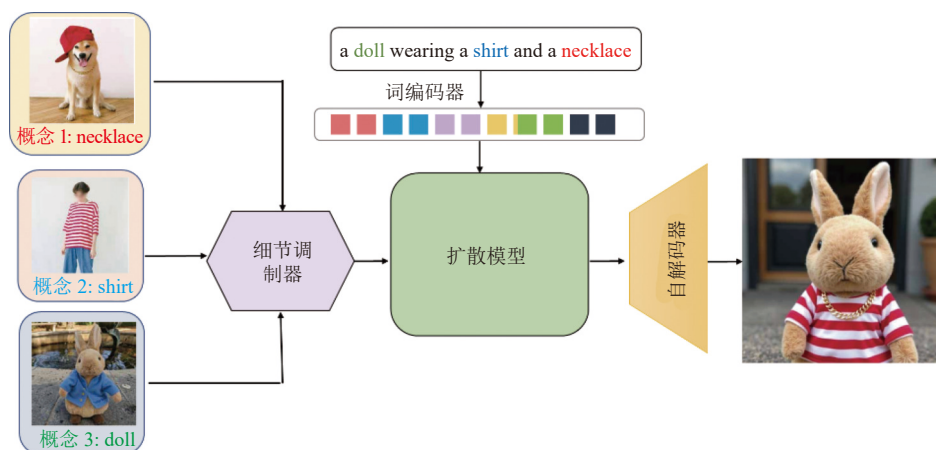
第 2 类方法旨在将来自不同图像的概念组合到单张生成图像中. 概念组合方法通常利用 LoRA^[29]来学习新概念 (通常每张图像一个概念), 并提出合并多个 LoRA 以同时运行的技术. 一些方法如 LoRA-composer^[61]和 OMG^[62], 通过结合空间条件 (例如空间掩码) 并为图像的特定区域分配局部提示来实现这一点. 然而, 这限制了自然重叠概念的组合能力, 例如一个人佩戴的项链, 这无法通过掩码区别划分. 其他方法 (Gu 等人^[63]、Po 等人^[64]、Shah 等人^[65]) 则融合多个 LoRA, 使它们能够协同工作. 其中 Mix-of-Show^[63]定义利用多个概念 LoRA 来联合支持多个定制概念的场景为去中心化概念定制. Mix-of-Show 采用嵌入分解 LoRA (ED-LoRA) 进行单模型调优, 并通过梯度融合在中心节点上保留单一概念的领域本质, 理论上支持无限的概念融合. Po 等人^[64]提出了正交适应方法, 旨在鼓励在微调过程中彼此独立的定制模型具有正交的残差权重. 这确保了在推理过程中, 定制模型能够以最小干扰进行叠加. Shah 等人^[65]则提出了 ZipLoRA, 专注于简单有效地组合单独的内容主体和风格概念.

多概念生成与组合的研究在近年来取得了显著进展, 但仍面临诸多挑战. 单图像多概念解耦学习方法目前仍不够成熟, 但在导致生成多样化和复杂组合时仍需进一步提升灵活性和可控性. 而多模型概念组合优化方法通过 LoRA 技术和梯度融合等手段, 显著提升了概念组合的效率和生成图像的质量, 但在处理自然重叠概念和复杂交互场景时仍有改进空间.

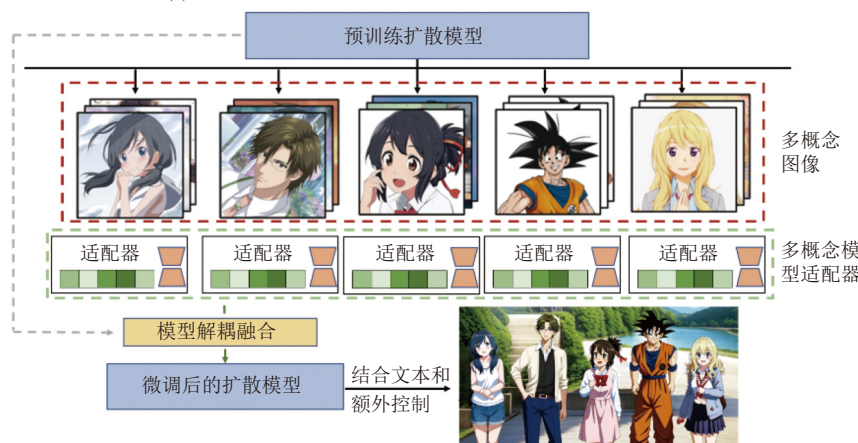
3.2 多概念属性可控编辑

元素级别的图像处理长期以来一直是数字艺术的目标, 现实生活中人们采用 Adobe Photoshop 这样的编辑工具进行视觉元素的精确操控. 除了多概念的图像生成外, 研究者进一步地对图像这些元素进行更深入的讨论, 包括

插入、去除、移动等, 希望能够用自动化工具实现像 Adobe Photoshop 那样对多元素的细粒度控制. 图 10 列出主体移动, 多主体和场景混合以及图像重绘这 3 种可控编辑操作^[66]. ControlNet^[10]和 IP-Adapter^[20]最早为这样基于扩散模型的生成式模型设计了可控生成的范式. 它们将条件图像 (例如姿态、布局图像) 作为额外输入, 并将它们的特征注入主 U-Net 中. 但这个工作不支持交互式、多轮、基于元素的操作 (例如, 组合、调整大小、排列), 这些操作对于创意工作流程至关重要. 面临的挑战包括: 1) 视觉元素的解耦和表示, 2) 连续布局控制, 3) 保持外观和身份, 4) 维持视觉特征, 5) 缺乏用于端到端训练的大规模配对训练数据.



(a) 自定义单词组合的代表方法 TokenVerse^[58] 的流程



(b) 适配器组合的代表方法 Mix-of-Show^[63] 的流程

图 9 多概念生成与组合示意图

AnyDoor^[67]首先研究了概念插入的问题, 如何既准确且无缝地将目标物体放置到场景图像的期望位置中. 具体而言, AnyDoor 通过将目标物体作为模板, 重新生成场景图像中由框/掩码标记的局部区域. 为实现这一点, AnyDoor 使用 DINO^[68]作为 ID 提取器来提取概念物体的关键 ID 特征, 并使用了一个类似 ControlNet 的结构来获取物体的细节图作为补充. DragonDiffusion^[69]则探讨了物体移动编辑问题, 在扩散模型上实现了拖拽式的图像元素编辑操作. 该方法将图像编辑操作重新定义为预训练扩散模型中特征对应关系的优化过程, 其核心在于构建与编辑目标对齐的能量函数, 并通过梯度引导实现精确控制. 具体而言, 该方法提出了两项关键技术: 1) 基于特征对应关系的梯度优化机制, 将用户拖拽操作转化为潜在空间的特征调整; 2) 视觉交叉注意力策略, 利用记忆库设计保持编辑区域与原始图像的一致性. 得益于这些高效的设计, 所有内容编辑和一致性操作都来自原图像特征对应关系, 无需额外的模型微调或附加模块.

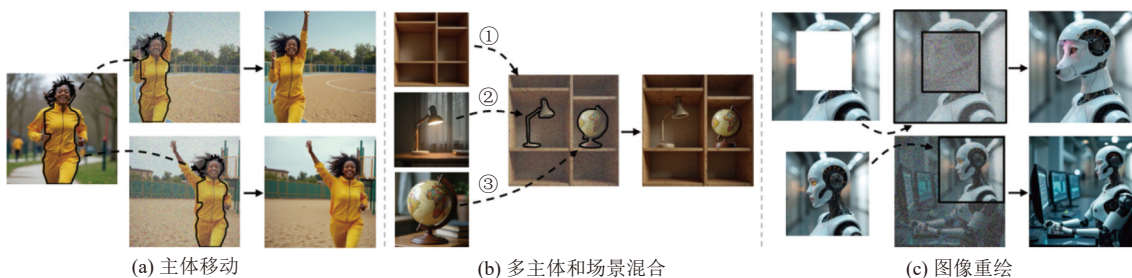


图 10 常见的多概念属性可控编辑

仅用图像特征进行操作往往缺乏足够的精度,一些工作开始借助布局进行更稳定的图像编辑. Zhang 等人^[70]提出了首个用于单张图像布局编辑的框架,实现对单张图像进行连续编辑的同时并保持图像概念的外观属性. 该研究通过两个关键技术突破解决了这一挑战: 首先,开发了掩码文本反演方法,将图像中的多个对象概念解耦并嵌入到独立的文本标记中,有效保持了各对象的视觉属性;其次,设计了一种无需训练的优化策略,通过对预训练扩散模型施加布局控制约束,使生成图像既能忠实呈现学习到的视觉概念,又能精确匹配用户指定的新布局. 这种端到端的解决方案不仅实现了高质量的布局编辑效果,还保持了原始图像的核心视觉特征. Gligen^[71]可以根据文本描述和定位指令生成图像. 定位指令提供有关图像的附加信息,例如边界框、深度图、语义地图等. 为了防止知识遗忘, Gligen^[71]冻结基座模型权重,并添加新的可训练门控 Transformer 层,以接收新的定位输入. 在训练期间, Gligen 使用门控机制逐渐将新的定位信息融合到预训练模型中. 这种设计使得生成灵活,并提高了质量和可控性. Magic Fixup^[72]提出了一种包含变换与协调两阶段处理的流程框架,能够从原始图像中迁移精细细节特征,并保持各组成部分的视觉一致性. 但由于其训练依赖于视频数据,在实际图像编辑应用时可能会导致性能下降. MS-Diffusion^[73]提出了基于布局引导的多概念零样本图像个性化框架,该框架整合了多主体的兼容性、零样本学习能力的融入、布局引导的提供以及基础模型参数的保留. MS-Diffusion 将布局 token 与特征重采样器相结合以保持主题之间的细节保真度,并改进了交叉注意力机制,使其能够适应多主体输入,确保每个主体条件作用于特定区域. 该多主体交叉注意力机制在保持文本控制的同时,协调了主题之间的和谐组合.

上面的工作大多只探讨了多概念属性编辑的部分操作,为进一步设计多概念属性统一的视觉理解框架,研究者们从元素建模进行设计,增强编辑的可控性. 元素级别视觉建模的本质在于位置、语义和身份的灵活解耦和表示. BlobCtrl^[66]设计了 Blob 作为视觉基元来实现这一点. 这种 Blob 是一个概率二维高斯分布,几何上表现为一个椭圆. Blob 参数能够精确指定位置、大小和方向,而高斯平滑性确保了和谐且连续的布局控制. 对于视觉概念保持, BlobCtrl 使用可微分斑点溅射结合变分自编码器特征来保持外观. 借助这种特殊的 Blob 基元设计, BlobCtrl 实现了高效精细的元素级别的操作,包括图像中概念主体的插入、移动、去除、替换等. 扩散模型的高维噪声输入空间并不天然适合图像反演或空间编辑. Mu 等人^[74]提出了一种图像表示方法,利用扩散模型促进输入图像的空间编辑. 具体地,他们考虑学习将输入编码为能够忠实重建输入图像的“图像元素”. 这些元素可以由用户直观地编辑,并通过扩散模型解码为真实的图像.

3.3 多概念服装组合试衣

虚拟试衣作为多概念个性化生成的重要子任务,旨在生成穿着指定服装且保持人物身份一致性的图像,在电子商务和创意设计领域具有广泛的应用前景. 这个任务的示意图如图 11 所示,服装提取 U-Net 网络用于提取多个服装的特征,然后将这些特征以注意力的方式注入生成图像的主 U-Net 中. 同时为保持人物的一致性,服装试衣的方法会将人物身体进行遮挡,将仅保留人脸信息的图像以及姿态图作为条件输入. 一些方法还会设计相应的文本辅助最终的图像生成.

早期的虚拟试衣方法^[75-77]采用二阶段网络. 第 1 阶段是学习服装图像到生成图像的像素级别的变换,这一步被称为“变形”;第 2 阶段则是将变形后的服装“融合”到人物之中. 该方法缺乏将服装概念深入挖掘的过程,单纯

学习像素变换, 因此应用场景十分有限, 也容易受到遮挡和环境的影响. 后续 Stable VITON^[78]、TryOnDiffusion^[79]、OOTDiffusion^[80]、IDM-VTON^[81]等工作采用基于扩散模型并且一步直接生成的设计. 为增强虚拟试衣效果, 它们还使用了双 U-Net 架构, 即除主 U-Net 去生成图像外, 额外引入服装提取 U-Net, 通过交叉注意力机制将提取的服装特征动态注入生成过程. 相较于传统方法, 该技术方案在保持服装图案、材质等细粒度特征方面展现出明显优势, 为虚拟试衣的实际应用提供了更可靠的技术支持. BooW-VTON^[82]通过建立高质量伪掩码数据实现了无需掩码的虚拟试衣, 这种无需掩码的设计避免了空间信息的损失, 使得服装的个性化生成更加鲁棒以及更好地应用于复杂的姿态和遮挡. 图 11 给出了基于扩散模型和双 U-Net 架构的多概念服装组合试衣示意图.

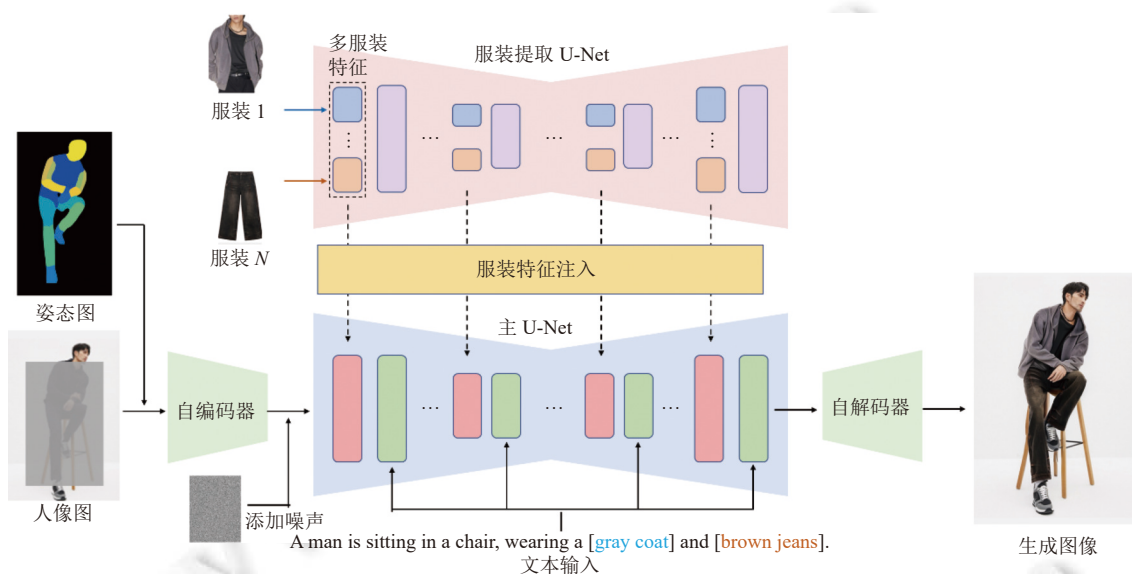


图 11 多概念服装组合试衣示意图

然而, 这些方法仍然将虚拟试衣视为基于模板的编辑任务, 需要在原有人物图像上进行编辑. 此时编辑受到人物模板的影响, 无法自由进行概念属性组合, 在实际应用场景中灵活性不足. 一些方法通过尽可能保留文生图基座模型的生成能力实现更自由的虚拟试衣. StableGarment^[83]和 IMAGDressing^[84]重新定义了虚拟试穿的任务, 其目标是基于给定衣服和人物以及其他可选条件 (如姿态、场景、表情等) 灵活地生成可控的人物图像. StableGarment 和 IMAGDressing 借助服装 U-Net 去提取细粒度的服装特征的同时, 引入一个混合注意力的模块去将这些特征在扩散模型去噪过程中注入, 实现了能人物服装多属性自由组合的虚拟试衣. Magic Clothing^[85]设计了一个联合 classifier-free 引导去平衡服装特征以及文本输入, 尽可能保持文生图基座模型的生成能力. DreamFit^[86]则提出一个轻量级的服装编码器, 使用可训练的 LoRA^[29]层代替平行 U-Net 去提取服装特征, 轻量级服装特征提取的设计同样使得不容易损害原生文生图基座的能力. 尽管如此, 这些方法仅适用于单件服装, 缺乏对多条件的支持, 限制了自由搭配多种服装的能力.

为实现服装概念的灵活组合生成, AnyDressing^[87]提出了多服装虚拟试衣任务框架, 支持根据文本提示或控制条件生成穿着任意服装组合的人物图像. 该任务面临以下挑战: 1) 服装保真度: 防止多件服装之间的混淆, 同时保留每件服装的复杂纹理; 2) 文本-图像一致性: 最小化多件服装对无关区域的影响, 确保生成图像与文本提示的一致性; 3) 插件兼容性: 实现与社区控制插件的无缝集成. 为解决上述问题, AnyDressing 设计了一种能够根据任意组合的服装和个性化文本提示定制人物图像的新方法. AnyDressing 依然由服装提取 U-Net 以及主 U-Net 两个核心网络组成, 记为 GarmentsNet 和 DressingNet. GarmentsNet 利用核心的服装特定特征提取器 (GFE) 模块, 通过共

享 U-Net 架构中的并行自注意力层提取多件服装的细节特征,并结合 LoRA 机制减少参数增加,避免服装混淆并确保网络效率. DressingNet 采用着装注意力 (DA) 机制,将多件服装特征无缝集成到去噪过程中,并通过实例级服装定位 (IGL) 学习策略,确保每件服装仅影响其对应区域,从而提高生成图像与文本提示的一致性.此外,AnyDressing 还设计了服装增强纹理学习 (GTL) 策略,通过感知特征和高频信息的约束,增强对服装细节的监督.得益于这些设计,AnyDressing 可以作为插件模块,轻松集成到任何基于扩散模型的社区控制扩展中,从而提升合成图像的多样性和可控性.

多概念服装组合试衣是多概念个性化生成的一个重要应用,服装试衣的自由度和服装一致的保真度是该任务核心要求.随着个性化生成的发展,相信服装这种特殊的概念最终也能更可控地引导用户需要的试衣图像生成.

4 评估方法

本节对上述介绍的个性化生成的 6 个任务分别进行方法对比.第 4.1 节先介绍个性化生成的常用数据集,第 4.2 节介绍定量评估中常用的指标,第 4.3 节则给出具体的方法定量和定性的对比结果.

4.1 数据集

下面是个性化生成的常用数据集.

- DreamBench^[13]: 一个主要的评价个性化物体生成的公开数据集,包含 30 个对象,涵盖背包、动物、汽车和玩具等类别.对于每个对象,纳入了 25 条提示词,以确保进行多样化的评估,这些提示词涵盖了各种情境、配饰和属性.

- DreamBench++^[88]: 是对 DreamBench 数据集的扩展,包含 150 张图像和 1350 条提示词,为评估模型性能提供了一个更广泛的评价框架.

- 互联网人物数据集: StoryMaker^[35]使用的互联网收集的人物图像,包含 30 万单人物图像和 20 万的双人物图像,图像说明使用 CogVLM^[89]生成,主要用于人像个性化生成的评估.由于人像数据的敏感性,目前该数据集暂未开源.

- PPR10K 数据集^[90]: 一个包含多样视角的真实人像图像数据集,包含 11 161 张高质量的人像照片.这些照片被分为 1 681 个场景,每个场景包含多张照片.

- TikTok benchmark^[91]: 一个由 TikTok 短视频平台下载的人物舞蹈数据集,总共包含 300 个舞蹈视频,每个视频大约是 10–15 s.该数据集可以用于图像驱动个性化视频生成.

- CelebA^[92]: 一个大型人脸属性数据集,包含 202 599 张名人的面部图像,涵盖 10 177 个不同的身份,主要用于计算机视觉中与人脸相关的研究和任务.数据图像大小为 178×218,每个图像附有 40 条人脸属性标注.由于 CelebA 数据集分辨率较低且背景简单,难以衡量个性化生成的多样性,一般用于人像驱动学习方法的训练但不用于评估.

- VITON-HD^[93]: 一个经典的高分辨率虚拟试衣数据集,包含 13 679 对服装与人物上身着衣图像,主要用于虚拟试衣算法的训练和测试.

- Dressing-Pair^[87]: 由 AnyDressing 提出的三元组 (上身衣服,下身衣服和人物图像) 服装组合数据集.该数据集包含了从公开数据集 Dresscode 提取的 26 114 个三元组对以及从互联网收集的私有的 37 065 个三元组对.

- BlobBench^[66]: 由 BlobCtrl^[66]提出的综合性的用于评估元素级生成和能力的数据集. BlobBench 包含 100 张图像,来自不同场景的真实世界图像和人工智能生成的图像,如室内和室外场景、动物和风景等,以确保评估的公平性和有效性.这些图像均匀地包含多种元素级操作,并且标注了椭圆参数、前景掩模和详细的文字描述.

4.2 评估指标

定量评估使用的评估指标一般是利用生成图像的像素、深度特征并借助某种算法或者预训练模型对生成结果量化评估的.下面列举了常见的个性化生成的评估指标,包括图像以及视频.

- Frechet inception distance (FID)^[94]: 使用深度网络 Inception-v3^[95]提取特征,计算生成图像的特征分布与真实

图像的特征分布距离. FID 可以衡量生成图像的质量, 分数越低代表生成图像的分布越接近真实图像, 说明质量越好.

- Frechet video inception distance (FVD)^[96]: 一种用于评估视频生成质量的指标, 类似于图像生成领域中常用的 FID. FVD 使用预训练的 3D 卷积神经网络 (如 I3D 模型) 提取视频的特征表示, 并计算生成视频特征分布与真实视频特征分布之间的 Frechet 距离 (即两个多元高斯分布之间的差异). FVD 值越低, 表明生成视频的质量越高, 与真实视频的分布越接近.

- 峰值信噪比 (PSNR): 评估图像失真程度, 值越大代表失真程度越低, 生成图像的质量越高.

- 结构相似性差异 (SSIM)^[97]: 评估生成图像和真实图像之间的结构相似性, 数值越大代表生成图像与真实图像越相似.

- 学习感知图像块相似性 (LPIPS)^[98]: 使用深度特征评估生成图像与真实图像之间的相似度. 数值越低表示两张图像越相似, 反之代表生成图像的多样性越高.

- BLIP-VQA score^[17]: 基于 BLIP 模型的视觉问答 (VQA) 任务的评估指标. BLIP 是一种先进的视觉-语言预训练模型, 能够同时理解图像内容和自然语言问题, 并生成准确的文本回答. BLIP-VQA score 通过评估模型在 VQA 任务中的表现来衡量生成图像与文本问题之间的对齐程度. 其优势在于它不仅评估图像的视觉质量, 还评估图像与文本语义的一致性. 值越高代表生成图片和文本语义越一致.

- CLIP-T^[16]: 一种基于 CLIP 模型的评估指标, 用于衡量生成图像与文本描述之间的语义一致性. CLIP 是由 OpenAI 提出的一种多模态预训练模型, 能够同时理解图像和文本, 并在共享的嵌入空间中对齐两者的语义表示. CLIP-T 的优势在于它能够捕捉图像和文本之间的高层次语义关联, 而不仅是低层次的视觉特征匹配. CLIP-T 得分的范围通常在 $[-1, 1]$ 之间, 得分越高, 表明生成图像与文本描述的语义一致性越好.

- CLIP-I^[16]: 一种基于 CLIP 模型的评估指标, 用于衡量生成图像与概念图像之间的语义一致性. 得分越高表明它们之间越相似.

- CLIP-AS^[16]: 一个基于 CLIP 模型相似度计算的图像美学评分指标, 范围通常在 0–10 分, 分数越高代表图像越符合人类主流审美标准 (如构图、色彩、风格等). 该指标通过对比图像与预设美学文本提示 (如“高质量照片”) 的匹配度生成.

- DINO-I^[68]: 一种基于 DINO 模型的评估指标, 与 CLIP-I 类似, 用于衡量生成图像与概念图像之间的语义一致性. 得分越高表明它们之间越相似.

4.3 方法对比

本节将对个性化生成方法进行定量和定性评估. 对于物体驱动学习方法, 表 1 列出了 Textual Inversion^[14]、DreamBooth^[13]、Custom Diffusion^[99]、ELITE^[18]、BLIP-Diffusion^[19]、IP-Adapter^[20]、Subject-Diffusion^[22]、OmniGen^[24]和 UniReal^[25]这 9 个方法在 DreamBench 数据集上的定量对比结果. 可视化结果对比在图 12 中给出. 基于学习的方法需要大规模图像数据对图像编码器进行预训练, 应用时无须对给定图片进行网络微调, 因而速度较快, 对于分辨率为 512×512 的图像生成, 这些方法在 A100 显卡上耗时基本小于 5 s. 然而, 该类方法对于纹理复杂的物体存在一定的保真度不足, 例如图 12 中第 1 行展示 Textual Inversion^[14]、DreamBooth^[13]、Custom Diffusion^[99]这 3 个方法对汽水罐的编辑结果. 这主要是由于难以学习一个特定的单词向量对这些复杂的物体进行描述. 而基于优化的方法则需要根据给定图片进行网络微调, 本文采用常用的学习率为 $1\text{E}-5$, 批次大小为 8, 训练迭代次数为 200 次, 训练设备为 2 张 A100 显卡的训练设置进行对比. 这些方法的效果很大程度上取决于图像编码器的训练以及图像特征的注入方式, 如图 12 所示, IP-Adapter^[20]这种轻量级设计无法保存足够的细节, 基于 DiT 网络的 OmniGen^[24]和 UniReal^[25]注入参考图像 token 方式在文本指令匹配和细节保持两方面都取得了较好的效果. 在模型大小和显存消耗方面, 前 7 个方法是基于 Stable Diffusion 上进行微调, 模型参数量在 1B 左右, 推理显存消耗一般在 6–8 GB 之间, OmniGen^[24]和 UniReal^[25]则是基于 DiT 模型, 模型大小在 12B 左右, 推理显存消耗在 40 GB 左右. Custom Diffusion^[99]训练的参数仅占用 75 MB 左右的显存, 在显存占用和训练速度在基于优化的方法取得领先.

表 1 物体驱动学习方法在 DreamBench 数据集上定量对比

方法	类型	网络架构	DINO-I↑	CLIP-I↑	CLIP-T↑	推理时间	推理显存消耗 (GB)
真实匹配图像	—	—	0.774	0.885	—	—	—
Textual Inversion ^[14]	基于优化	U-Net	0.569	0.780	0.255	19.2 min	6
DreamBooth ^[13]	基于优化	U-Net	0.668	0.803	0.305	22.3 min	6
Custom Diffusion ^[99]	基于优化	U-Net	0.643	0.790	0.305	6.1 min	5
ELITE ^[18]	基于学习	U-Net	0.621	0.771	0.293	3.2 s	6
BLIP-Diffusion ^[19]	基于学习	U-Net	0.594	0.779	0.300	4.1 s	8
IP-Adapter ^[20]	基于学习	U-Net	0.667	0.813	0.289	3.3 s	7
Subject-Diffusion ^[22]	基于学习	U-Net	0.711	0.787	0.293	4.3 s	8
OmniGen ^[24]	基于学习	DiT	0.693	0.810	0.320	6.2 s	40
UniReal ^[25]	基于学习	DiT	0.702	0.806	0.326	6.5 s	40

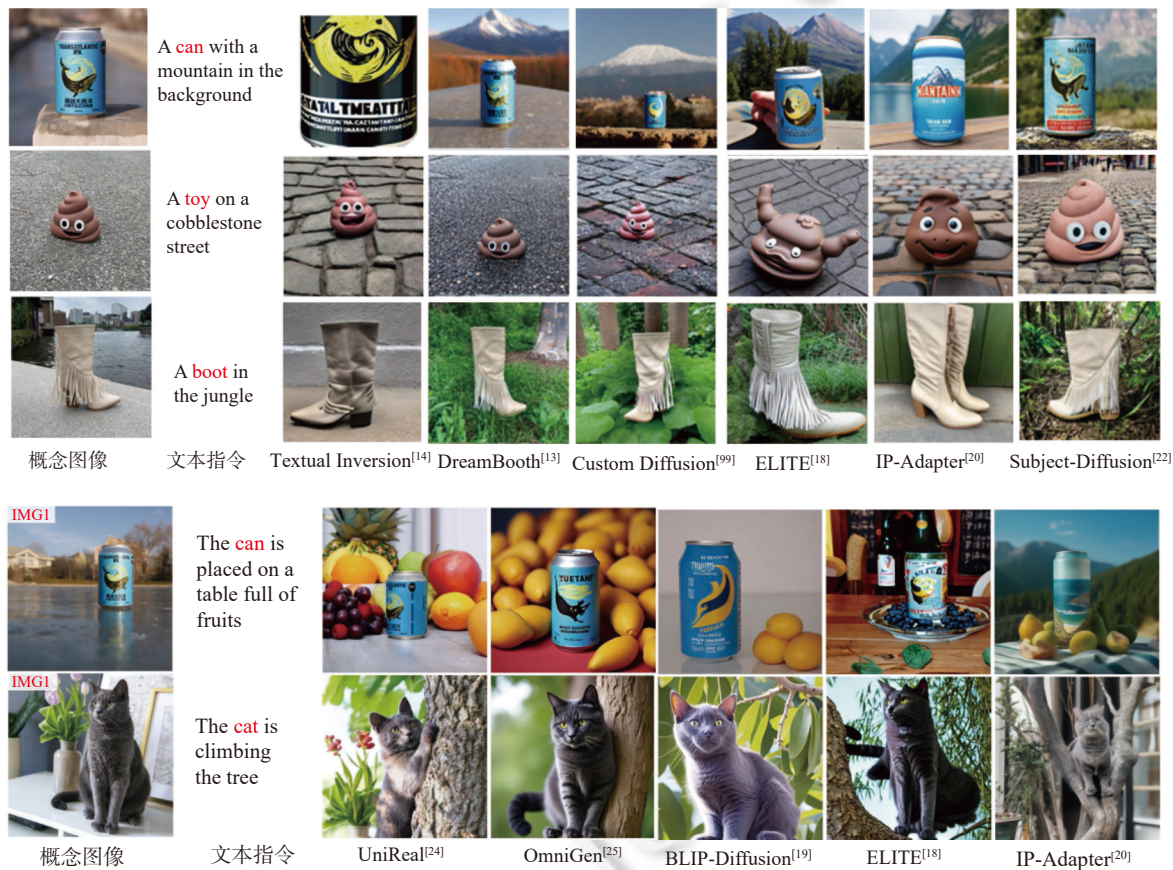


图 12 物体驱动学习方法在 DreamBench 数据集上定性对比

对于人像驱动学习方法, 基于 StoryMaker^[35]和 Visual Persona^[37], 表 2 和图 13 列出 MM-Diff^[32]、PhotoMaker-V2^[27]、IP-Adapter-FaceID^[20]、InstantID^[30]、StoryMaker^[35]和 Visual Persona^[37]在互联网人物数据集和 PPR10K 数据集上的定量和定性对比结果. 从图 13 中可以看到, 这些方法都能在一定程度上保持住 ID, 而 StoryMaker^[35]通过对人脸和角色的分别建模, 在半身人像生成上性能取得领先. 而全身人像生成上, 除 Visual Persona 外的方法都缺乏对全身的建模, 因而身体的生成并不好. 人像驱动学习方法均为基于学习的方法, 推理速度在 A100 显卡上小于 5 s. 这些方法都是基于 Stable Diffusion 上进行微调, 模型参数量在 1B 左右. StoryMaker 需同时提取人脸和图像特征, 推理的显存占用较高, 需 14 GB.

表 2 人像驱动学习方法定量对比

类别	方法	是否支持多人	Face Sim↑	CLIP-I↑	CLIP-T↑	推理显存消耗 (GB)
互联网人物数据集 半身人像效果	MM-Diff ^[32]	是	0.607	0.198	0.784	10
	PhotoMaker-V2 ^[27]	否	0.618	0.237	0.753	10
	IP-Adapter-FaceID ^[20]	否	0.666	0.229	0.687	7
	InstantID ^[30]	否	0.739	0.233	0.725	6
	StoryMaker ^[35]	是	0.673	0.217	0.795	14
PPR10K数据集 全身人像效果	IP-Adapter-FaceID ^[20]	否	—	0.186	0.749	7
	InstantID ^[30]	否	—	0.170	0.712	6
	PhotoMaker-V2 ^[27]	否	—	0.203	0.764	10
	IP-Adapter ^[20]	否	—	0.504	0.738	7
	StoryMaker ^[35]	是	—	0.680	0.677	14
	Visual Persona ^[37]	否	—	0.730	0.667	9

(a) 在互联网人物数据集上
半身人像效果的对比



(b) 在PPR10K数据集上
全身人像效果的对比



图 13 人像驱动学习方法定性对比

对于图像驱动视频生成方法,表 3 列出了 DisCo^[41]、MagicAnimate^[42]、Animate Anyone^[38]、Champ^[43]、UniAnimate^[40]和 Animate Anyone 2^[49]共 6 个方法在 TikTok benchmark 的定量对比结果. 图像驱动的视频生成需要较大计算资源,在不插帧优化的情况下,上述 6 个方法在 32 帧的视频生成至少需要约 64 GB 的显存,参数量在 1.2B 左右,同时在 A100 显卡上运算约为 4 min 左右.

表 3 图像驱动视频生成方法在 TikTok 数据集上定量对比

方法	SSIM↑	PSNR (dB)↑	LPIPS↑	FVD↓
DisCo ^[41]	0.668	29.03	0.292	292.80
MagicAnimate ^[42]	0.714	29.16	0.239	179.07
Animate Anyone ^[38]	0.718	29.56	0.285	171.90
Champ ^[43]	0.802	29.91	0.234	160.82
UniAnimate ^[40]	0.811	30.77	0.231	148.06
Animate Anyone 2 ^[49]	0.812	30.82	0.223	144.65

对于多概念生成与组合方法,由于目前缺乏合适的定量指标去衡量生成的图片是否符合多概念条件,一般采用定性对比的方法.图 14 列出 TokenVerse^[58]中 ConceptExpress^[57]、Break-A-Scene^[55]、DreamBooth^[13]、OMG^[62]和 TokenVerse^[58]在 3 组概念图像上的对比,其中“ ”表示使用多概念联合训练策略.基于适配器组合的方法 OMG^[62]对于输入图像中多概念的隔离依然存在挑战,而自定义单词组合方法 ConceptExpress^[57]和 Break-A-Scene^[55]则在细节保持方面存在缺陷,例如图 14 中第 1 行的猫以及第 3 行的帽子.目前基于 DiT^[59]模型的方法 TokenVerse 取得了最好的效果.在模型大小和显存消耗方面,前 5 个方法都是基于 Stable Diffusion 进行微调,模型大小为 1B 左右,显存消耗在 8–10 GB 之间,而 TokenVerse 采用 DiT 模型,模型大小为 12B 左右,显存消耗增至 42 GB.



图 14 多概念生成与组合方法对比

对于多概念属性可控编辑方法,表 4 和图 15 列出基于 BlobCtrl^[66]中 AnyDoor^[67]、Magic Fixup^[72]、Gligen^[71]和 BlobCtrl^[66]在 BlobBench 的定量和定性对比结果.如图 15 所示,基于布局引导的图像生成方法 Gligen^[71]无法重现输入场景,Magic Fixup^[72]则在移动和改变尺寸的任务上缺乏对空白的修补,AnyDoor^[67]和 BlobCtrl^[66]都取得了不错的效果.在模型大小和显存消耗方面,AnyDoor^[67]和 Magic Fixup^[72]的参数量在 1B 左右,推理显存消耗在 8–10 GB. Gligen 由于需要引入定位指令,参数量增至 1.1B 左右,显存消耗增至 11 GB 左右.而 BlobCtrl^[66]的参数

量在 1.5B 左右,推理显存消耗在 17 GB 左右. 相对应地, 由于 BlobCtrl^[66]能够实现高效精细的元素级别的操作, 目前取得了最好的编辑效果.

表 4 多概念属性可控编辑方法定量对比

方法	组合		移动		替换		改变尺寸		推理显存消耗 (GB)
	CLIP↑	DINO↑	CLIP↑	DINO↑	CLIP↑	DINO↑	CLIP↑	DINO↑	
AnyDoor ^[67]	0.867	0.812	0.854	0.817	0.817	0.802	0.833	0.837	10
Gligen ^[71]	0.707	0.578	0.712	0.624	0.684	0.606	0.782	0.694	12
Magic Fixup ^[72]	0.805	0.786	0.846	0.824	0.842	0.801	0.837	0.852	8
BlobCtrl ^[66]	0.883	0.869	0.889	0.878	0.862	0.860	0.865	0.891	17

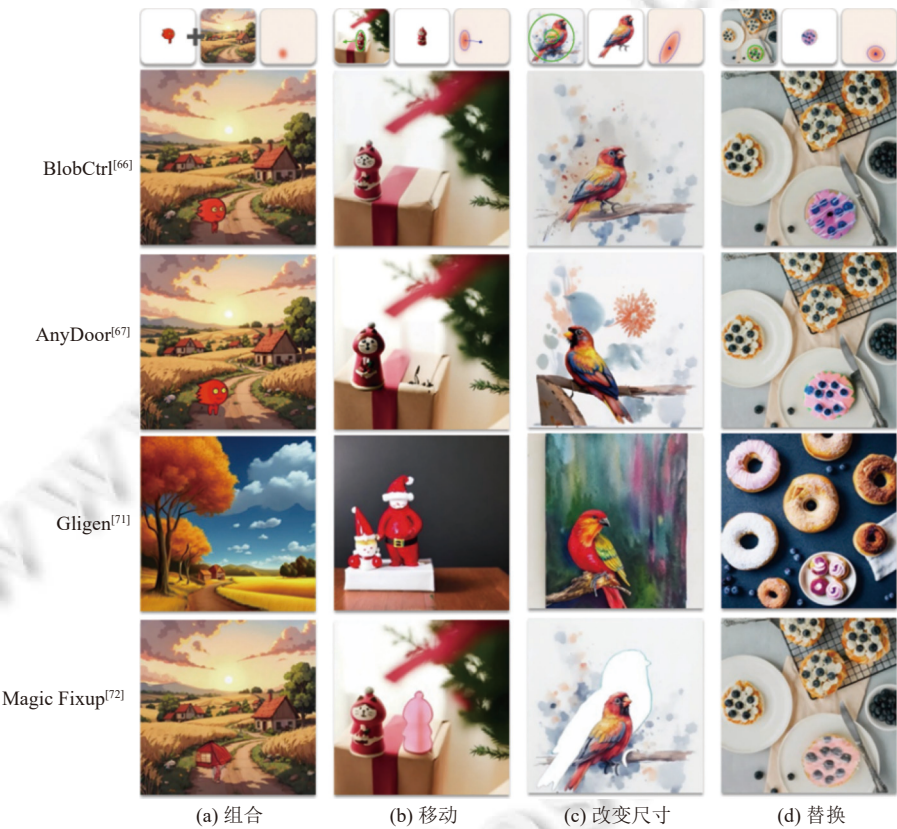


图 15 多概念属性可控编辑方法在 BlobBench 数据集上定性对比

对于虚拟试衣方法, 表 5 和表 6 分别对基于模板的方法和自由生成两种类别的算法进行定量对比. 基于模板的方法包括 HR-VTON^[75]、LaDI-VTON^[76]、DCI-VTON^[77]、Stable VITON^[78]、OOTDiffusion^[80]、IDM-VTON^[81]和 BooW-VTON^[82]. 自由生成的方法包括 StableGarment^[83]、Magic Clothing^[85]、IMAGDressing^[84]和 AnyDressing^[87]. 图 16 给出 BooW-VTON^[82]对于基于模板的虚拟试衣方法的定性对比, 而图 17 则给出了 AnyDressing^[87]中对自由生成该类方法的定性对比. 对于图 16 展示的结果对比, HR-VTON 是基于 GAN 的方法, 虽然生成速度很快 (<200 ms), 但是生成质量较差. 其他 8 个方法均为基于扩散模型的算法, LaDI-VTON 和 DCI-VTON 采用跨注意力输入衣服特征, 生成质量有所提高但泛化性能依然不足 (如图 16 第 3 行), 生成速度在 3 s 左右, 模型大小约为 1B, 显存消耗为 7–8 GB. 而 OOTDiffusion、IDM-VTON、BooW-VTON、Stable VITON 均使用双 U-Net, 衣服细节特征保留得

更好, 但相应推理时间增加至 5 s 左右, 模型大小增至 1.5B 左右, 显存消耗增至 16 GB 左右. 对于图 17 展示的对
比结果, Magic Clothing、StableGarment、IMAGDressing 和 AnyDressing 都采用双 U-Net 架构, 这些方法对于最常
见的上装都起到一定的保持效果. 在自由组合方面, AnyDressing 最为自然, 其他方法在多衣服组合都存在缺陷, 上
下装出现不协调的情况. 上述的讨论是基于 768×512 分辨率的.

表 5 基于模板生成的虚拟试衣方法在 VITON 数据集上定量对比

方法	SSIM↑	LPIPS↑	FID↓	推理显存消耗 (GB)
HR-VITON ^[75]	0.883	0.115	9.70	5
LaDI-VTON ^[76]	0.872	0.156	8.85	7
DCI-VTON ^[77]	0.856	0.166	8.73	8
Stable VITON ^[78]	0.885	0.133	6.52	10
OOTDiffusion ^[80]	0.862	0.124	6.03	16
IDM-VTON ^[81]	0.870	0.102	6.29	16
BooW-VTON ^[82]	0.869	0.093	5.89	17

表 6 基于自由生成的虚拟试衣方法定量对比

方法	数据集	CLIP-T↑	CLIP-I↑	CLIP-AS↑	推理显存消耗 (GB)
StableGarment ^[83]	VITON-HD	0.285	0.583	5.781	8
	Dressing-Pair	0.284	0.556	5.735	
Magic Clothing ^[85]	VITON-HD	0.288	0.640	5.703	16
	Dressing-Pair	0.266	0.583	5.540	
IMAGDressing ^[84]	VITON-HD	0.202	0.734	5.077	16
	Dressing-Pair	0.242	0.614	5.291	
AnyDressing ^[87]	VITON-HD	0.289	0.741	5.881	16
	Dressing-Pair	0.296	0.734	5.874	



图 16 基于模板生成的虚拟试衣方法在 VITON-HD 数据集上定性对比



图 17 自由生成的虚拟试衣方法定性对比

5 挑战和展望

5.1 挑战

(1) 过拟合

因参考训练图像数量有限, 当前个性化概念合成系统仍然面临过拟合这一关键挑战. 该过拟合问题表现为两种形式: 1) 主体对象可编辑性丧失. 个性化模型生成的图像往往简单复制参考图像中的主体对象, 例如总是以相同姿势呈现猫咪; 2) 无关语义混入. 参考图像在训练时模型混入了与主体无关的信息 (如背景或不相关物体), 导致输出中会生成当前语境无关的元素. 为探究其根源, Compositional inversion^[100]研究发现, 相比预训练词汇形成的中心分布, 学习到的词嵌入位于分布外区域. 另一研究^[101]也发现可学习词嵌入显著偏离初始嵌入分布. 此外, 还有文献表明^[23,100,102]中, 特定修饰符会压制其他 token, 导致其他语义特征缺失.

针对该问题, 目前已提出多种解决方案, 例如排除冗余背景、注意力操控、可学习参数正则化和数据增强等. 然而该问题尚未完全解决, 特别是当主体对象具有非刚性外观^[100], 或上下文提示词与参考图像中无关元素存在语义关联时^[103]. 很明显, 解决个性化生成中的过拟合问题, 不仅是一个技术挑战, 而且是确保这些提出的方法以及系统在多样且动态的现实环境中得以实际部署和扩展的必要条件. 因此, 迫切需要一种有效的策略和可靠的评估指标, 以便在实际应用中实现更广泛的应用并获得更高的用户满意度.

(2) 在主体保真度与生成自由性方面的权衡

个性化内容合成的终极目标是构建能够同时实现两种关键能力的系统: 一方面需要以高保真度呈现特定主体, 另一方面要能灵活响应多样化文本或图像等提示. 然而, 这两种需求本质上存在显著矛盾. 首先主体保真度要求模型通过最小化重建损失来精确捕捉和复现参考主体的细微特征, 包括材质纹理、微观结构等细节特征. 这种精确复现通常需要模型具备强大的特征记忆能力. 而文本对齐则要求系统根据文本或参考图像等描述灵活调整主

体属性,包括改变姿态、表情、环境或艺术风格等.这种创造性调整往往需要打破对参考图像的机械复制.在物体驱动学习方面,基于优化的方法这个问题尤为突出,因为这些大多仅在少量样本上微调基座模型,这种做法在训练时间较长的情况下很容易产生过拟合损坏原有基座模型的生成能力.如果较少训练时长,则模型可能欠拟合,导致主体保真度低. *Perfusion*^[104]方法通过正则化注意力投影机制,在特征空间构建双重约束.但这种方法在复杂场景下容易导致特征混淆. *Cao* 等人^[105]设计了双路解耦架构,将条件引导过程分离为独立的保真度通道和文本适应通道.虽然缓解了冲突,但增加了模型复杂度,且难以处理跨模态的细粒度交互.

(3) 多概念组合生成冲突

当前个性化生成方法主要集中于特定领域,如物体主体、人脸、服装或风格等方向的定制化生成,但在实际应用中,用户往往期望能够在一张图像中融合多个定制化概念.当多个概念进行组合生成时,部分概念很可能会压制其他概念的生成,导致最后生成的结果没办法包含用户所需的全部概念.在适配器组合方法中,训练时间较长的适配器获得一个相对基座模型更大的偏移量,后续组合适配器时,偏移量较大的一个容易压制其他适配器并使其失效.而对于自定义 token 组合方法,虽然 token 间的学习是独立的,但是这些自定义 token 的语义很可能存在冲突,例如相似动物或同类别物体同时组合时,这种冲突会使得这些 token 对应的语义区域相互影响^[58].如何解耦适配器或者自定义 token 的学习仍然是个性化生成的一大挑战.这些显著的语义冲突和组合灵活性不足的问题使得用户无法实现自由创作.

(4) 缺乏面向个性化生成的审慎推理机制

在某些以内容质量优先于时间效率的个性化场景中,比如数字广告领域,广告商通常每天只为每个用户投放最合适的几条广告.受大语言模型推理方面取得巨大成功的启发,审慎推理会注重事先进行的广泛逻辑推理和上下文推理,从而能够对用户偏好进行全面分析,以推动更有效的内容个性化^[106].然而,目前的个性化生成缺乏多轮优化的做法或是用户反馈的过程,只能采用大量生成的办法覆盖用户需求,但这种做法生成效率和质量较为低下.

(5) 评估体系不足

尽管个性化生成技术日益流行,但该领域仍缺乏标准化的测试数据集和能够全面评估不同方法性能的指标体系.现有的评估方式存在明显缺陷,首先是过度依赖 CLIP 相似度.当前广泛采用的视觉保真度评估方法主要基于 CLIP 图像相似度计算,但这种方法存在严重偏差:1) 当模型对参考图像过拟合时,会虚高评估分数;2) 无法区分简单复制和语义级保真;3) 对细粒度属性变化的敏感度不足.除 CLIP 相似度外,其他评估维度也呈现单一化:1) 缺乏对主体可编辑性的量化标准;2) 忽视生成多样性等关键指标;3) 对多模态对齐能力的评估不完善.

(6) 伦理隐忧

个性化生成技术在推动创意表达和数字内容创新的同时,也伴随着深刻的伦理隐忧和滥用风险,尤其是在人像生成领域,技术的双刃剑效应尤为突出.深度伪造技术的泛滥使得伪造身份、制造虚假新闻或恶意诽谤成为可能,不仅威胁个人隐私与名誉,还可能被用于政治操纵、金融诈骗等系统性危害,动摇社会信任基础.生成技术的高真实度模糊了真实与虚构的边界,导致公众对数字内容的信任危机,而现有法律体系在肖像权、数据来源合法性等方面的滞后性进一步加剧了责任认定的困难.一些文章开始对个人照片隐私保护进行研究, *Anti-DreamBooth*^[107]设计了一个防御系统对抗 LoRA^[29]和 DreamBooth^[13]等个性化生成.这个防御系统通过对图像增加微妙的噪声扰动显著降低了 LoRA^[29]和 DreamBooth^[13]的个性化生成质量.对于基于编码器的人像生成方法,例如 InstantID^[30]、IP-Adapter^[20]和 PhotoMaker^[27], IDProtector^[108]提出了一种对抗性噪声编码器进行人像保护.它的方法是在单次前向传播的过程中为人像照片添加难以察觉的对抗噪声.尽管如此,目前隐私保护的研究仍然不足,未来研究需要构建动态演进的保护框架,既能适应生成模型的快速迭代,又能满足不同应用场景下隐私保护与可用性的双重需求.

5.2 研究趋势

(1) 防止过拟合

过拟合一直是困扰少样本训练的问题,为突破现有局限,可以从下面 3 个维度进行创新:1) 动态特征门控机制:开发能自动识别任务需求(保真/创意)的智能开关,在特征提取层实现自适应处理.2) 层次化训练策略:分阶段

优化模型能力, 先建立强大的特征重建基础, 再培养创造性适应能力. 3) 多模态对比学习: 构建联合嵌入空间, 使模型能自然理解文本修改与视觉特征间的映射关系.

这些创新方向未来有望推动个性化内容合成系统突破现有瓶颈, 在个性化创作的众多领域实现更广泛的应用. 当前研究虽已取得进展, 但在处理复杂语义修改, 同时需要模型保真的创意能力时仍面临挑战 (例如“保持人物身份的同时改变年龄特征”), 这将成为未来研究的一个热点.

(2) 通用的个性化生成

当前个性化生成方法主要集中于特定领域, 如前文提及物体主体、人脸、服装的定制化生成. 然而在实际应用中, 用户往往需要将多个定制化概念融合至单张图像中. 虽然已有部分研究^[58,62,65,73,109], 但其生成能力仍存在明显局限. 因此, 亟需发展支持通用类别的个性化生成方法, 以提供更灵活多样的生成能力. 除了现有的对象级别控制外, 还需要实现属性级别、部件级别更细粒度的个性化控制, 使用户能够精确调控生成图像的各个细节特征. 这种细粒度的控制能力将大幅提升个性化生成系统的实用价值, 满足专业创作场景中对细节把控的严苛要求. 要实现这一目标, 需要突破现有方法在跨概念兼容性和组合稳定性方面的技术瓶颈, 开发新型的特征解耦与重组机制, 这也是未来通用人工智能的发展方向.

(3) 视频和三维方向的个性化生成

近年来, 视频与三维内容生成技术发展迅猛, 一系列研究方法^[110-118]已开始探索个性化技术在这些领域的应用. 这些技术为内容创作开辟了新途径, 同时也带来了新的技术挑战. 在个性化视频生成方面, 需要解决跨帧外观与结构一致性的维护问题, 其复杂度远超单帧图像生成; 而在三维内容生成领域, 则需确保多视角一致性, 这就要求生成模型必须具备准确理解和表现三维结构的能力. 未来研究的一大趋势相信会是着力推动个性化生成技术从静态图像向动态内容和空间复杂内容的拓展, 这不仅丰富现有创作手段, 也将为虚拟现实、影视特效等行业带来新的技术突破. 值得注意的是, 动态内容的时序连贯性与三维场景的空间一致性将成为技术突破的关键所在, 需要开发新型的神经网络架构和训练范式来应对这些挑战.

(4) 生成与检索的协同演进

传统个性化内容分发系统主要依赖于基于检索的方法 (如推荐系统), 但现有内容库往往难以完全满足用户的个性化需求. 将个性化生成技术与检索式方法相结合, 有望构建更强大的内容分发体系. 这种融合架构既能利用检索系统对用户偏好的精准把握, 又能通过生成技术突破内容库的局限性, 实现“检索-生成”的闭环优化. 具体而言, 系统可以先通过检索确定用户的内容偏好模式, 再基于这些模式动态生成符合用户期待的新内容; 同时, 生成内容的质量反馈又可反哺检索模型的优化. 这种协同机制不仅能显著提升内容分发的精准度, 还能创造传统检索系统无法提供的新型个性化体验, 为数字内容产业带来范式变革. 要实现这种协同效应, 需要解决生成内容的质量控制、风格一致性保持以及实时响应等技术挑战.

(5) 更多模态的个性化提示

除文本和图像外, 个性化图像生成的提示模态正朝着多元化方向发展, 涵盖语音、脑电信号及触觉等新兴引导提示. 在语音到图像生成领域, Sung-Bin 等人^[119]通过跨模态潜在空间对齐方法, 利用视觉信息增强音频特征表达, 有效缓解了视听模态间的语义鸿沟问题. 进一步地, Petermann 等人^[120]提出基于视觉-语言模型的推理框架, 通过检索高质量音频-图像配对数据构建可扩展的声学-视觉关联体系. 其生成的模型支持隐式语义混合与插值, 例如仅输入火焰燃烧声与马匹奔跑声, 即可生成“烈焰战场中奔腾的骏马”这一复杂场景, 证明了音频模态对图像生成的细粒度控制能力. 在脑电信号 (EEG) 引导生成方向, 研究者关注如何通过解码脑信号中的视觉编码信息推动脑机接口应用. 代表性工作“Guess what I think”^[121]将扩散模型引入该领域, 基于 ControlNet 轻量化框架实现 EEG 信号对潜在扩散模型 (LDM) 的调控, 为高精度脑控图像生成提供了可行方案. 触觉引导生成的研究尚处探索阶段, 但已显现潜力. Yang 等人^[122]开发的触觉-视觉双向转换模型, 实现了触觉信号与图像的端到端合成, 并在多项跨模态任务中验证了其有效性, 为触觉在交互式生成中的应用奠定了基础.

(6) 自回归与非自回归范式的统一

近年来, 生成模型的发展趋势逐渐向自回归与非自回归方法的统一方向演进. 传统自回归模型 (如 GPT) 逐元

素生成数据,虽可控性强但效率较低;而非自回归方法(如扩散模型)并行生成,速度更快但可能牺牲一致性.最新研究如 Transfusion^[123]和 Show-O^[124]尝试结合两者优势,使用自回归-扩散混合建模,同时处理理解和生成任务.为适应任务的不同输入数据和变化,它们采用文本编码器和图像编码器将文本和图像各自编码为离散 token,并进一步提出了统一的提示策略.它们将这些 token 作为输入处理,最终得到了图像和文本统一生成的效果.未来,生成模型可能进一步融合不同范式,在可控性、速度和生成能力上实现更优平衡.

6 总 结

本文系统综述了基于扩散模型的个性化图像生成方法,通过相关文献的统计得知,个性化生成技术已成为计算机视觉领域的研究热点,其核心目标是通过结合用户提供的特定概念与文本描述,实现高质量、可控的定制化图像生成.本文首先从理论框架层面,明确了扩散模型的数学基础与个性化生成的任务定义,提出将生成过程标准化为概念反演和个性化生成两阶段的统一范式.其次,针对当前研究从单主体向多概念发展的趋势,本文分别对两大方向展开深度分析:1)单主体生成方面,本文介绍了物体驱动学习、人像驱动学习以及图像驱动个性化视频生成这3个子任务,现有方法通过设计概念反演和特征注入技术,显著提升了主体视觉特征的保真度(如物体、人像等细节重建);2)多概念生成任务方面,本文从多概念生成与组合、多概念属性可控编辑和多概念服装组合试衣这3方面进行综述,这些任务聚焦于概念的语义对齐与视觉一致性,通过分层控制、注意力机制优化、概念解耦等手段解决了多主体融合的冲突问题.本文进一步总结了常用数据集和评估指标,并对比了不同方法的生成质量,为后续研究提供了基准参考.最后,本文整理当前个性化生成方法仍存在的挑战,并基于此给出了未来可能的研究方向.本文的综述为研究者提供了系统性的技术梳理与问题分析,有望推动个性化生成技术在艺术创作、虚拟试衣等场景的落地,并促进生成模型领域的可持续发展.

References

- [1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.
- [2] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10674–10685. [doi: [10.1109/CVPR52688.2022.01042](https://doi.org/10.1109/CVPR52688.2022.01042)]
- [3] Zhu JY, Krähenbühl P, Shechtman E, Efros AA. Generative visual manipulation on the natural image manifold. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 597–613. [doi: [10.1007/978-3-319-46454-1_36](https://doi.org/10.1007/978-3-319-46454-1_36)]
- [4] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D. Encoding in style: A StyleGAN encoder for image-to-image translation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2287–2296. [doi: [10.1109/CVPR46437.2021.00232](https://doi.org/10.1109/CVPR46437.2021.00232)]
- [5] Tov O, Alaluf Y, Nitzan Y, Patashnik O, Cohen-Or D. Designing an encoder for StyleGAN image manipulation. ACM Trans. on Graphics, 2021, 40(4): 133. [doi: [10.1145/3450626.3459838](https://doi.org/10.1145/3450626.3459838)]
- [6] Liu M, Wei YX, Wu XH, Zuo WM, Zhang L. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. Science China Information Sciences, 2023, 66(5): 151101. [doi: [10.1007/s11432-022-3679-0](https://doi.org/10.1007/s11432-022-3679-0)]
- [7] Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: PMLR, 2015. 2256–2265.
- [8] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 574.
- [9] Song J, Meng CL, Ermon S. Denoising diffusion implicit models. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–20.
- [10] Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 3813–3824. [doi: [10.1109/ICCV51070.2023.00355](https://doi.org/10.1109/ICCV51070.2023.00355)]
- [11] Zuo R, Hu HX, Deng XM, Ma CX, Wang HA. Survey on deep learning methods for freehand-sketch-based visual content generation. Ruan Jian Xue Bao/Journal of Software, 2024, 35(7): 3497–3530 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/7053.htm> [doi: [10.13328/j.cnki.jos.007053](https://doi.org/10.13328/j.cnki.jos.007053)]

- [12] Yu K, Bin Y, Zheng ZQ, Yang Y. Text-to-image generation with conditional semantic augmentation. *Ruan Jian Xue Bao/Journal of Software*, 2024, 35(5): 2150–2164 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/7024.htm> [doi: 10.13328/j.cnki.jos.007024]
- [13] Ruiz N, Li YZ, Jampani V, Pritch Y, Rubinstein M, Aberman K. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Vancouver: IEEE, 2023. 22500–22510. [doi: 10.1109/CVPR52729.2023.02155]
- [14] Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D. An image is worth one word: Personalizing text-to-image generation using textual inversion. In: *Proc. of the 11th Int'l Conf. on Learning Representations*. Kigali: OpenReview.net, 2023. 1–31.
- [15] Chen H, Zhang YP, Wu SM, Wang X, Duan XG, Zhou YW, Zhu WW. DisenBooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In: *Proc. of the 12th Int'l Conf. on Learning Representations*. Vienna: OpenReview.net, 2024. 1–23.
- [16] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: *Proc. of the 38th Int'l Conf. on Machine Learning*. Kigali: PMLR, 2021. 8748–8763.
- [17] Li JN, Li DX, Xiong CM, Hoi S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: *Proc. of the 39th Int'l Conf. on Machine Learning*. Baltimore: PMLR, 2022. 12888–12900.
- [18] Wei YX, Zhang YB, Ji ZL, Bai JF, Zhang L, Zuo WM. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: *Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision*. Paris: IEEE, 2023. 15897–15907. [doi: 10.1109/ICCV51070.2023.01461]
- [19] Li DX, Li JN, Hoi SCH. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In: *Proc. of the 37th Int'l Conf. on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2023. 1312.
- [20] Ye H, Zhang J, Liu SB, Han X, Yang W. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv*: 2308.06721, 2023.
- [21] Shi J, Xiong W, Lin Z, Jung HJ. InstantBooth: Personalized text-to-image generation without test-time finetuning. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 8543–8552. [doi: 10.1109/CVPR52733.2024.00816]
- [22] Ma J, Liang JH, Chen C, Lu HN. Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In: *Proc. of the 2024 ACM SIGGRAPH Conf. Denver*: ACM, 2024. 25. [doi: 10.1145/3641519.3657469]
- [23] Jia XH, Zhao Y, Chan KCK, Li YD, Zhang H, Gong BQ, Hou TB, Wang HS, Su YC. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv*:2304.02642, 2023.
- [24] Xiao ST, Wang YZ, Zhou JJ, Yuan HY, Xing XR, Yan RR, Liu Z. OmniGen: Unified image generation. In: *Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2025. 13294–13304. [doi: 10.1109/CVPR52734.2025.01241]
- [25] Chen X, Zhang ZF, Zhang H, Zhou YQ, Kim SY, Liu Q, Li YJ, Zhang JM, Zhao NX, Wang YL, Ding H, Lin Z, Zhao HS. UniReal: Universal image generation and editing via learning real-world dynamics. In: *Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2025. 12501–12511. [doi: 10.1109/CVPR52734.2025.01166]
- [26] Zhang SL, Huang LH, Chen X, Zhang YF, Wu ZF, Feng YT, Wang W, Shen YJ, Liu Y, Luo P. FlashFace: Human image personalization with high-fidelity identity preservation. *arXiv*:2403.17008, 2024.
- [27] Li Z, Cao MD, Wang XT, Qi Z, Cheng MM, Shan Y. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 8640–8650. [doi: 10.1109/CVPR52733.2024.00825]
- [28] Deng JK, Guo J, Xue NN, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. In: *Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019. 4685–4694. [doi: 10.1109/CVPR.2019.00482]
- [29] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: *Proc. of the 10th Int'l Conf. on Learning Representations*. OpenReview.net, 2022. 1–13.
- [30] Wang QX, Bai X, Wang HF, Qin ZK, Chen A, Li HX, Tang X, Hu Y. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv*:2401.07519, 2024.
- [31] Peng X, Zhu JW, Jiang BY, Tai Y, Luo DH, Zhang J, Lin W, Jin TS, Wang CJ, Ji RR. PortraitBooth: A versatile portrait model for fast identity-preserved personalization. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 27070–27080. [doi: 10.1109/CVPR52733.2024.02557]

- [32] Wei ZC, Su QK, Qin L, Wang WZ. MM-Diff: High-fidelity image personalization via multi-modal condition integration. arXiv:2403.15059, 2024.
- [33] He JJ, Geng YF, Bo LF. UniPortrait: A unified framework for identity-preserving single-and multi-human image personalization. arXiv:2408.05939, 2024.
- [34] Xiao GX, Yin TW, Freeman WT, Durand F, Han S. FastComposer: Tuning-free multi-subject image generation with localized attention. *Int'l Journal of Computer Vision*, 2025, 133(3): 1175–1194. [doi: [10.1007/s11263-024-02227-z](https://doi.org/10.1007/s11263-024-02227-z)]
- [35] Zhou ZG, Li J, Li HX, Chen N, Tang X. StoryMaker: Towards holistic consistent characters in text-to-image generation. arXiv:2409.12576, 2024.
- [36] Zhao J, Zheng HL, Wang CY, Lan L, Huang WR, Tang YH. MagicNaming: Consistent identity generation by finding a “name space” in T2I diffusion models. In: *Proc. of the 39th AAAI Conf. on Artificial Intelligence*. Philadelphia: AAAI Press, 2024. 10439–10447. [doi: [10.1609/aaai.v39i10.33133](https://doi.org/10.1609/aaai.v39i10.33133)]
- [37] Nam J, Son S, Xu Z, Shi J, Liu DF, Liu F, Kim K, Zhou Y. Visual Persona: Foundation model for full-body human customization. In: *Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2025. 18630–18641. [doi: [10.1109/CVPR52734.2025.01736](https://doi.org/10.1109/CVPR52734.2025.01736)]
- [38] Li H. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 8153–8163. [doi: [10.1109/CVPR52733.2024.00779](https://doi.org/10.1109/CVPR52733.2024.00779)]
- [39] Zhang Y, Gu JX, Wang LW, Wang H, Cheng JQ, Zhu YF, Zou FY. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv:2406.19680, 2024.
- [40] Wang X, Zhang SW, Gao CX, Wang JY, Zhou XQ, Zhang YY, Yan LX, Sang N. UniAnimate: Taming unified video diffusion models for consistent human image animation. arXiv:2406.01188, 2024.
- [41] Wang T, Li LJ, Lin K, Zhai YH, Lin CC, Yang ZY, Zhang HW, Liu ZC, Wang LJ. DisCo: Disentangled control for realistic human dance generation. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 9326–9336. [doi: [10.1109/CVPR52733.2024.00891](https://doi.org/10.1109/CVPR52733.2024.00891)]
- [42] Xu ZC, Zhang JF, Liew JH, Yan HS, Liu JW, Zhang CX, Feng JS, Shou MZ. MagicAnimate: Temporally consistent human image animation using diffusion model. In: *Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2024. 1481–1490. [doi: [10.1109/CVPR52733.2024.00147](https://doi.org/10.1109/CVPR52733.2024.00147)]
- [43] Zhu SH, Chen JL, Dai ZZ, Dong ZL, Xu YH, Cao X, Yao Y, Zhu H, Zhu SY. Champ: Controllable and consistent human image animation with 3D parametric guidance. In: *Proc. of the 18th European Conf. on Computer Vision*. Milan: Springer, 2024. 145–162. [doi: [10.1007/978-3-031-73001-6_9](https://doi.org/10.1007/978-3-031-73001-6_9)]
- [44] Parihar R, Gupta H, Vs S, Babu RV. Text2Place: Affordance-aware text guided human placement. In: *Proc. of the 18th European Conf. on Computer Vision*. Milan: Springer, 2024. 57–77. [doi: [10.1007/978-3-031-72646-0_4](https://doi.org/10.1007/978-3-031-72646-0_4)]
- [45] Saini N, Bodla N, Shrivastava A, Ravichandran A, Zhang X, Shrivastava A, Singh B. InVi: Object insertion in videos using off-the-shelf diffusion models. arXiv:2407.10958, 2024.
- [46] Qiu D, Chen ZY, Wang R, Fan MY, Yu CQ, Huang JS, Wen X. MovieCharacter: A tuning-free framework for controllable character video synthesis. arXiv:2410.20974, 2024.
- [47] Xu ZY, Huang ZY, Cao J, Zhang Y, Cun XD, Shuai Q, Wang YC, Bao LC, Li JT, Tang F. AnchorCrafter: Animate cyber-anchors selling your products via human-object interacting video generation. arXiv:2411.17383, 2024.
- [48] Men YF, Yao Y, Cui MM, Bo LF. MIMO: Controllable character video synthesis with spatial decomposed modeling. In: *Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2025. 21181–21191. [doi: [10.1109/CVPR52734.2025.01973](https://doi.org/10.1109/CVPR52734.2025.01973)]
- [49] Hu L, Wang GY, Shen Z, Gao X, Meng DC, Zhuo L, Zhang P, Zhang B, Bo LF. Animate Anyone 2: High-fidelity character image animation with environment affordance. arXiv:2502.06145, 2025.
- [50] Wang ZX, Yuan ZY, Wang XT, Li YW, Chen TS, Xia MH, Luo P, Shan Y. MotionCtrl: A unified and flexible motion controller for video generation. In: *Proc. of the 2024 ACM SIGGRAPH Conf.* Denver: ACM, 2024. 114. [doi: [10.1145/3641519.3657518](https://doi.org/10.1145/3641519.3657518)]
- [51] He H, Xu YH, Guo YW, Wetzstein G, Dai B, Li HS, Yang CY. CameraCtrl: Enabling camera control for text-to-video generation. arXiv:2404.02101, 2024.
- [52] Shi XY, Huang ZY, Wang FY, Bian WK, Li DS, Zhang Y, Zhang MY, Cheung KC, See S, Qin HW, Dai JF, Li HS. Motion-I2V: Consistent and controllable image-to-video generation with explicit motion modeling. In: *Proc. of the 2024 ACM SIGGRAPH Conf.* Denver: ACM, 2024. 111. [doi: [10.1145/3641519.3657497](https://doi.org/10.1145/3641519.3657497)]
- [53] Niu MY, Cun XD, Wang XT, Zhang Y, Shan Y, Zheng YQ. MOFA-Video: Controllable image animation via generative motion field

- adaptions in frozen image-to-video diffusion model. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 111–128. [doi: [10.1007/978-3-031-72655-2_7](https://doi.org/10.1007/978-3-031-72655-2_7)]
- [54] Dai ZZ, Zhang ZH, Yao Y, Qiu BX, Zhu SY, Qin L, Wang WZ. AnimateAnything: Fine-grained open domain image animation with motion guidance. arXiv:2311.12886, 2023.
- [55] Avrahami O, Aberman K, Fried O, Cohen-Or D, Lischinski D. Break-A-Scene: Extracting multiple concepts from a single image. In: Proc. of the 2024 SIGGRAPH Asia Conf. Sydney: ACM, 2023. 96. [doi: [10.1145/3610548.3618154](https://doi.org/10.1145/3610548.3618154)]
- [56] Vinker Y, Voynov A, Cohen-Or D, Shamir A. Concept decomposition for visual exploration and inspiration. ACM Trans. on Graphics, 2023, 42(6): 1–13. [doi: [10.1145/3618315](https://doi.org/10.1145/3618315)]
- [57] Hao SZ, Han K, Lv ZY, Zhao SH, Wong KYK. ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 215–233. [doi: [10.1007/978-3-031-73202-7_13](https://doi.org/10.1007/978-3-031-73202-7_13)]
- [58] Garibi D, Yadin S, Paiss R, Tov O, Zada S, Ephrat A, Michaeli T, Mosseri I, Dekel T. TokenVerse: Versatile multi-concept personalization in token modulation space. ACM Trans. on Graphics, 2025, 44(4): 41.
- [59] Peebles W, Xie SN. Scalable diffusion models with Transformers. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 4172–4182. [doi: [10.1109/ICCV51070.2023.00387](https://doi.org/10.1109/ICCV51070.2023.00387)]
- [60] Hu TH, Li LX, van de Weijer J, Gao HC, Shahbaz Khan F, Yang J, Cheng MM, Wang K, Wang YX. Token merging for training-free semantic binding in text-to-image synthesis. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 4372.
- [61] Yang Y, Wang W, Peng L, Song CT, Chen Y, Li HJ, Yang XL, Lu QL, Cai D, Wu BX, Liu W. LoRA-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. arXiv:2403.11627, 2024.
- [62] Kong Z, Zhang Y, Yang TY, Wang T, Zhang KH, Wu BZ, Chen GY, Liu W, Luo WH. OMG: Occlusion-friendly personalized multi-concept generation in diffusion models. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 253–270. [doi: [10.1007/978-3-031-72751-1_15](https://doi.org/10.1007/978-3-031-72751-1_15)]
- [63] Gu YC, Wang XT, Wu JZ, Shi YJ, Chen YP, Fan ZH, Xiao WY, Zhao R, Chang SN, Wu WJ, Ge YX, Shan Y, Shou MZ. Mix-of-Show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In: Proc. of the 37th Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 699.
- [64] Po R, Yang GD, Aberman K, Wetzstein G. Orthogonal adaptation for modular customization of diffusion models. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 7964–7973. [doi: [10.1109/CVPR52733.2024.00761](https://doi.org/10.1109/CVPR52733.2024.00761)]
- [65] Shah V, Ruiz N, Cole F, Lu E, Lazebnik S, Li YZ, Jampani V. ZipLoRA: Any subject in any style by effectively merging LoRAs. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 422–438. [doi: [10.1007/978-3-031-73232-4_24](https://doi.org/10.1007/978-3-031-73232-4_24)]
- [66] Li YW, Li LG, Zhang ZY, Li XY, Wang GZ, Li HX, Cun XD, Shan Y, Zou YX. BlobCtrl: A unified and flexible framework for element-level image generation and editing. arXiv:2503.13434, 2025.
- [67] Chen X, Huang LH, Liu Y, Shen YJ, Zhao DL, Zhao HS. AnyDoor: Zero-shot object-level image customization. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 6593–6602. [doi: [10.1109/CVPR52733.2024.00630](https://doi.org/10.1109/CVPR52733.2024.00630)]
- [68] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A. Emerging properties in self-supervised vision Transformers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9630–9640. [doi: [10.1109/ICCV48922.2021.00951](https://doi.org/10.1109/ICCV48922.2021.00951)]
- [69] Mou C, Wang XT, Song JC, Shan Y, Zhang J. DragonDiffusion: Enabling drag-style manipulation on diffusion models. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: OpenReview.net, 2024. 1–12.
- [70] Zhang ZY, Huang ZT, Liao J. Continuous layout editing of single images with diffusion models. Computer Graphics Forum, 2023, 42(7): e14966. [doi: [10.1111/cgf.14966](https://doi.org/10.1111/cgf.14966)]
- [71] Li YH, Liu HT, Wu QY, Mu FZ, Yang JW, Gao JF, Li CY, Lee YJ. Gligen: Open-set grounded text-to-image generation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22511–22521. [doi: [10.1109/CVPR52729.2023.02156](https://doi.org/10.1109/CVPR52729.2023.02156)]
- [72] Alzayer H, Xia ZH, Zhang X, Shechtman E, Huang JB, Gharbi M. Magic Fixup: Streamlining photo editing by watching dynamic videos. arXiv:2403.13044, 2024.
- [73] Wang XR, Fu SM, Huang QH, He WG, Jiang H. MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 1–29.

- [74] Mu JT, Gharbi M, Zhang R, Shechtman E, Vasconcelos N, Wang XL, Park T. Editable image elements for controllable synthesis. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 39–56. [doi: [10.1007/978-3-031-72627-9_3](https://doi.org/10.1007/978-3-031-72627-9_3)]
- [75] Lee S, Gu G, Park S, Choi S, Choo J. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: Proc. of the 17th European Conf. on Computer Vision. Milan: Springer, 2022. 204–219. [doi: [10.1007/978-3-031-19790-1_13](https://doi.org/10.1007/978-3-031-19790-1_13)]
- [76] Morelli D, Baldtrati A, Cartella G, Cornia M, Bertini M, Cucchiara R. LaDI-VTON: Latent diffusion textual-inversion enhanced virtual try-on. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 8580–8589. [doi: [10.1145/3581783.3612137](https://doi.org/10.1145/3581783.3612137)]
- [77] Gou JH, Sun SY, Zhang JF, Si JL, Qian C, Zhang LQ. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 7599–7607. [doi: [10.1145/3581783.3612255](https://doi.org/10.1145/3581783.3612255)]
- [78] Kim J, Gu G, Park M, Park S, Choo J. Stable VITON: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 8176–8185. [doi: [10.1109/CVPR52733.2024.00781](https://doi.org/10.1109/CVPR52733.2024.00781)]
- [79] Zhu LY, Yang DW, Zhu T, Reda F, Chan W, Saharia C, Norouzi M, Kemelmacher-Shlizerman I. TryOnDiffusion: A tale of two UNets. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 4606–4615. [doi: [10.1109/CVPR52729.2023.00447](https://doi.org/10.1109/CVPR52729.2023.00447)]
- [80] Xu YH, Gu T, Chen WF, Chen A. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 8996–9004. [doi: [10.1609/aaai.v39i9.32973](https://doi.org/10.1609/aaai.v39i9.32973)]
- [81] Choi Y, Kwak S, Lee K, Choi H, Shin J. Improving diffusion models for authentic virtual try-on in the wild. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 206–235. [doi: [10.1007/978-3-031-73016-0_13](https://doi.org/10.1007/978-3-031-73016-0_13)]
- [82] Zhang XP, Song D, Zhan PX, Chang TY, Zeng JH, Chen QG, Luo WH, Liu AA. BooW-VTON: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 26399–26408. [doi: [10.1109/CVPR52734.2025.02458](https://doi.org/10.1109/CVPR52734.2025.02458)]
- [83] Wang R, Guo HL, Liu JM, Li HX, Zhao HB, Tang X, Hu Y, Tang H, Li PP. StableGarment: Garment-centric generation via stable diffusion. arXiv:2403.10783, 2024.
- [84] Shen F, Jiang X, He X, Ye H, Wang C, Du XY, Li ZC, Tang JH. IMAGDressing-v1: Customizable virtual dressing. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 6795–6804. [doi: [10.1609/aaai.v39i7.32729](https://doi.org/10.1609/aaai.v39i7.32729)]
- [85] Chen WF, Gu T, Xu YH, Chen A. Magic Clothing: Controllable garment-driven image synthesis. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 6939–6948. [doi: [10.1145/3664647.3680691](https://doi.org/10.1145/3664647.3680691)]
- [86] Lin ET, Zhang XJ, Zhao FW, Luo YX, Dong X, Zeng L, Liang XD. DreamFit: Garment-centric human generation via a lightweight anything-dressing encoder. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 23723–23733. [doi: [10.1109/CVPR52734.2025.02209](https://doi.org/10.1109/CVPR52734.2025.02209)]
- [87] Li XH, Sun QC, Zhang PZ, Ye FL, Liao ZC, Feng WQ, Zhao ST, He Q. AnyDressing: Customizable multi-garment virtual dressing via latent diffusion models. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 5218–5226. [doi: [10.1609/aaai.v39i5.32554](https://doi.org/10.1609/aaai.v39i5.32554)]
- [88] Peng Y, Cui YX, Tang HM, Qi ZK, Dong RP, Bai J, Han CR, Ge Z, Zhang XY, Xia ST. DreamBench++: A human-aligned benchmark for personalized image generation. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 1–23.
- [89] Wang WH, Lv QS, Yu WM, Hong WY, Qi J, Wang Y, Ji JH, Yang ZY, Zhao L, Song XX, Xu JZ, Chen KP, Xu B, Li JZ, Dong YX, Ding M, Tang J. CogVLM: Visual expert for pretrained language models. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 3860.
- [90] Liang J, Zeng H, Cui MM, Xie XS, Zhang L. PPR10K: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 653–661. [doi: [10.1109/CVPR46437.2021.00071](https://doi.org/10.1109/CVPR46437.2021.00071)]
- [91] Jafarian Y, Park HS. Self-supervised 3D representation learning of dressed humans from social media videos. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 45(7): 8969–8983. [doi: [10.1109/TPAMI.2022.3231558](https://doi.org/10.1109/TPAMI.2022.3231558)]
- [92] Liu ZW, Luo P, Wang XG, Tang XO. Deep learning face attributes in the wild. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 3730–3738. [doi: [10.1109/ICCV.2015.425](https://doi.org/10.1109/ICCV.2015.425)]
- [93] Choi S, Park S, Lee M, Choo J. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 14126–14135. [doi: [10.1109/CVPR46437.2021.01391](https://doi.org/10.1109/CVPR46437.2021.01391)]

- [94] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. of the 31st Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.
- [95] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: [10.1109/CVPR.2016.308](https://doi.org/10.1109/CVPR.2016.308)]
- [96] Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717, 2018.
- [97] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612. [doi: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861)]
- [98] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: [10.1109/CVPR.2018.00068](https://doi.org/10.1109/CVPR.2018.00068)]
- [99] Kumari N, Zhang BL, Zhang R, Shechtman E, Zhu JY. Multi-concept customization of text-to-image diffusion. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 1931–1941. [doi: [10.1109/CVPR52729.2023.00192](https://doi.org/10.1109/CVPR52729.2023.00192)]
- [100] Zhang XL, Wei XY, Wu JL, Zhang TY, Zhang ZX, Lei Z, Li Q. Compositional inversion for stable diffusion models. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 7350–7358. [doi: [10.1609/aaai.v38i7.28565](https://doi.org/10.1609/aaai.v38i7.28565)]
- [101] Pang LY, Yin J, Xie HR, Wang QP, Li Q, Mao XD. Cross initialization for personalized text-to-image generation. arXiv:2312.15905, 2023.
- [102] Hao SZ, Han K, Zhao SH, Wong KYK. ViCo: Plug-and-play visual condition for personalized text-to-image generation. arXiv:2306.00971, 2023.
- [103] Zhang XL, Zhang WY, Wei XY, Wu JL, Zhang ZX, Lei Z, Li Q. Generative active learning for image synthesis personalization. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 10669–10677. [doi: [10.1145/3664647.3680773](https://doi.org/10.1145/3664647.3680773)]
- [104] Tewel Y, Gal R, Chechik G, Atzmon Y. Key-locked rank one editing for text-to-image personalization. In: Proc. of the 2023 ACM SIGGRAPH Conf. Los Angeles: ACM, 2023. 12. [doi: [10.1145/3588432.3591506](https://doi.org/10.1145/3588432.3591506)]
- [105] Cao P, Yang L, Zhou F, Huang TR, Song Q. Concept-centric personalization with large-scale diffusion priors. arXiv:2312.08195, 2023.
- [106] Fang Y, Wang WJ, Zhang Y, Zhu FB, Wang QF, Feng FL, He XN. Reason4Rec: Large language models for recommendation with deliberative user preference alignment. arXiv:2502.02061, 2025.
- [107] van Le T, Phung H, Nguyen TH, Dao Q, Tran NN, Tran A. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 2116–2127. [doi: [10.1109/ICCV51070.2023.00202](https://doi.org/10.1109/ICCV51070.2023.00202)]
- [108] Song YR, Yang P, Ci H, Shou MZ. IDProtector: An adversarial noise encoder to protect against ID-preserving image generation. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 3019–3028. [doi: [10.1109/CVPR52734.2025.00287](https://doi.org/10.1109/CVPR52734.2025.00287)]
- [109] Jones M, Wang SY, Kumari N, Bau D, Zhu JY. Customizing text-to-image models with a single image pair. In: Proc. of the 2024 SIGGRAPH Asia Conf. Papers. Tokyo: ACM, 2024. 6. [doi: [10.1145/3680528.3687642](https://doi.org/10.1145/3680528.3687642)]
- [110] She D, Liu MS, Pang JX, Wang J, Yang Z, He WG, Zhang GH, Wang Y, Huang QH, Tang HB, Yu YL, Fu SM. CustomVideoX: 3D reference attention driven dynamic adaptation for zero-shot customized video diffusion Transformers. arXiv:2502.06527, 2025.
- [111] Zhang YM, Xing ZN, Zeng YH, Fang YQ, Chen K. PIA: Your personalized image animator via plug-and-play modules in text-to-image models. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 7747–7756. [doi: [10.1109/CVPR52733.2024.00740](https://doi.org/10.1109/CVPR52733.2024.00740)]
- [112] Liang F, Ma HY, He ZC, Hou TB, Hou J, Li KP, Dai XL, Juefei-Xu F, Azadi S, Sinha A, Zhang PZ, Vajda P, Marculescu D. Movie Weaver: Tuning-free multi-concept video personalization with anchored prompts. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 13146–13156. [doi: [10.1109/CVPR52734.2025.01227](https://doi.org/10.1109/CVPR52734.2025.01227)]
- [113] Li XM, Jia X, Wang QH, Diao HW, Ge MM, Li PX, He Y, Lu HC. MoTrans: Customized motion transfer with text-driven video diffusion models. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 3421–3430. [doi: [10.1145/3664647.3680718](https://doi.org/10.1145/3664647.3680718)]
- [114] Li HJ, Qiu HN, Zhang SW, Wang X, Wei YJ, Li ZK, Zhang YY, Wu BX, Cai D. PersonalVideo: High ID-fidelity video customization without dynamic and semantic degradation. arXiv:2411.17048, 2024.
- [115] Wang Z, Li AX, Zhu LT, Guo Y, Dou Q, Li ZG. CustomVideo: Customizing text-to-video generation with multiple subjects.

- arXiv:2401.09962, 2024.
- [116] Raj A, Kaza S, Poole B, Niemeyer M, Ruiz N, Mildenhall B, Zada S, Aberman K, Rubinstein M, Barron J, Li YZ, Jampani V. DreamBooth3D: Subject-driven text-to-3D generation. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 2349–2359. [doi: [10.1109/ICCV51070.2023.00223](https://doi.org/10.1109/ICCV51070.2023.00223)]
 - [117] Ouyang YC, Chai WH, Ye JY, Tao DP, Zhan YB, Wang GA. Chasing consistency in text-to-3D generation from a single image. arXiv:2309.03599, 2023.
 - [118] Hsiao TF, Ruan BK, Wu YL, Lin TL, Shuai HH. TF-TI2I: Training-free text-and-image-to-image generation via multi-modal implicit-context learning in text-to-image models. arXiv:2503.15283, 2025.
 - [119] Sung-Bin K, Senocak A, Ha H, Owens A, Oh TH. Sound to visual scene generation by audio-to-visual latent alignment. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6430–6440. [doi: [10.1109/CVPR52729.2023.00622](https://doi.org/10.1109/CVPR52729.2023.00622)]
 - [120] Petermann D, Kalayeh MM. Seeing sound: Assembling sounds from visuals for audio-to-image generation. arXiv:2501.05413, 2025.
 - [121] Lopez E, Sigillo L, Colonnese F, Panella M, Communiello D. Guess what I think: Streamlined EEG-to-image generation with latent diffusion models. In: Proc. of the 2025 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025. 1–5. [doi: [10.1109/ICASSP49660.2025.10890059](https://doi.org/10.1109/ICASSP49660.2025.10890059)]
 - [122] Yang FY, Zhang JC, Owens A. Generating visual scenes from touch. In: Proc. of the 2023 IEEE/CVF Int'l Conf. on Computer Vision. Paris: IEEE, 2023. 22013–22023. [doi: [10.1109/ICCV51070.2023.02017](https://doi.org/10.1109/ICCV51070.2023.02017)]
 - [123] Zhou CT, Yu LL, Babu A, Tirumala K, Yasunaga M, Shamis L, Kahn J, Ma XZ, Zettlemoyer L, Levy O. Transfusion: Predict the next token and diffuse images with one multi-modal model. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 1–24.
 - [124] Xie JH, Mao WJ, Bai ZC, Zhang DJ, Wang WH, Lin KQ, Gu YC, Chen ZJ, Yang ZH, Shou MZ. Show-O: One single Transformer to unify multimodal understanding and generation. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025. 1–25.

附中文参考文献

- [11] 左然, 胡皓翔, 邓小明, 马翠霞, 王宏安. 基于手绘草图的视觉内容生成深度学习方法综述. 软件学报, 2024, 35(7): 3497–3530. <http://www.jos.org.cn/1000-9825/7053.htm> [doi: [10.13328/j.cnki.jos.007053](https://doi.org/10.13328/j.cnki.jos.007053)]
- [12] 余凯, 宾斌, 郑自强, 杨阳. 基于条件语义增强的文本到图像生成. 软件学报, 2024, 35(5): 2150–2164. <http://www.jos.org.cn/1000-9825/7024.htm> [doi: [10.13328/j.cnki.jos.007024](https://doi.org/10.13328/j.cnki.jos.007024)]

作者简介

何子健, 博士, 主要研究领域为计算机视觉, 图像视频生成, 多模态融合.

李冠彬, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为视觉内容理解与生成.