

北京郵電大學

本科 毕业 设计（论文）



题目： 基于 LLM 的交互式多模态图像编辑系统的设计与搭建

姓 名 罗彬慈

学 院 人工智能学院

专 业 智能科学与技术

班 级 2020219107

学 号 2020212053

指导教师 李佩佩

2024 年 5 月

北京邮电大学

本科毕业设计（论文）诚信声明

本人声明所呈交的毕业设计（论文），题目《基于 LLM 的交互式多模态图像编辑系统的设计与搭建》是本人在指导教师的指导下，独立进行研究工作所取得的成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 罗彬慈 日期：_____ 2024年5月19日

关于论文使用授权的说明

本人完全了解并同意北京邮电大学有关保留、使用学位论文的规定，即：北京邮电大学拥有以下关于学位论文的无偿使用权，具体包括：学校有权保留并向国家有关部门或机构送交学位论文，有权允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，有权允许采用影印、缩印或其它复制手段保存。汇编学位论文，将学位论文的全部或部分内容编入有关数据库进行检索。（保密的学位论文在解密后遵守此规定）

本人签名：_____ 罗彬慈 日期：_____ 2024年5月19日

导师签名：_____ 李佩佩 日期：_____ 2024年5月19日

北京邮电大学

本科生毕业设计（论文）成绩评定表

学生姓名	罗彬慈			学院	人工智能	
学号	2020212053	专业	智能科学与技术		班级	202 021 910 7
论文题目	(中文) 基于 LLM 的交互式多模态图像编辑系统的 设计与搭建 (英文) Design and Construction of Interactive Multimodal Image Editing System Based on LLM					
指导教师	李佩佩	指导教师职称	副教授		指导教师单位	人工智能学院
中期检查小组评分	(满分 30 分)： 26		中期检查小组长签字: 李佩佩		日期: 2024 年 3 月 21 日	
指导教师评分	指导教师成绩评定标准					
	评价内容	具体要求	分值	评分		
	调研论证	能独立查阅理解文献和从事相关调研; 正确翻译外文资料; 有分析、综合各种信息、获取新知识及拓展更新知识的能力和自学能力。	3	3		
	方案设计	能独立提出符合需求的可行性研究方案、实验方案、设计方案，独立进行实验（如安装、调试、操作）和研究方案论证。能合理评估系统成本，理解局限性，并考虑社会、健康、安全、法律、文化以及环境等因素的影响。	4	4		
	能力水平	能综合运用所学知识和技能分析解决毕设过程中遇到的实际问题; 能正确处理实验数据; 能对课题进行理论分析, 获得有价值的结论。	3	3		
	学习态度	认真、勤奋、努力、诚实、严格遵守纪律, 按期饱满完成规定的任务。	3	3		
	设计(论文)水平	文题相符、综述简练完整, 有见解; 立论正确, 论述充分, 结论严谨合理; 实验正确, 分析处理科学; 文字通顺, 技术用语准确, 符合工程管理规范; 设计(论文)有理论价值和应用价值。	5	4		
	文本规范	装订顺序正确, 字体字号等与基本规范相符, 符号统一, 编号齐全, 图表完备、整洁、正确。	2	2		
指导教师评分合计(满分 20 分)			19			

	评语：该毕业设计在内容丰富性和独创性上表现出色，尤其是在实际应用方面，学生成功地将理论知识与实际问题结合，展示了扎实的工程能力和出色的逻辑思维、系统分析能力。该毕业设计工作量饱满，实现效果优秀，论文逻辑清晰，格式规范，很好地达到了预期的研究目标和要求。		
	指导教师签字: 李佩佩 日期: 2024年5月17日		
复议	<input type="checkbox"/> 是 <input type="checkbox"/> 否 复议评分合计: 复议人签字: 复议日期: 复议有权限修改指导教师评分, 选择复议后指导教师评分将由复议评分替换		

本科生毕业设计（论文）答辩成绩评定标准

答辩小组成绩评定	评价内容	具体要求	分值	评分
	选题	符合专业培养目标, 符合社会实际、结合工程实际, 难易适度, 体现新颖性、综合性。	5	4
	设计(论文)质量水平	全面完成任务书中规定的各项要求, 文题相符, 工作量饱满, 写作规范, 达到综合训练标准和毕业要求, 有理论成果和应用价值, 并考虑社会、安全、环境等因素。	20	16
	答辩准备	准备充分: 有简洁、清晰、美观的演示文稿; 准时到场。	5	4
	内容陈述	语言表达简洁、流利、清楚、准确, 重点突出, 逻辑性强, 概念清楚, 论点正确; 实验方法科学, 分析归纳合理; 结论严谨; 对毕业设计(论文)的内容掌握透彻。	15	12
	回答问题	回答问题准确、有深度、有理论根据、基本概念清晰。	5	4
	答辩小组评分合计(满分50分)		40	

意见: 学生对毕业完成的工作做了详细地讲解, 回答问题正确, 思路明确。论文完成的工作以及论文的撰写较好地符合了毕业设计评价指标的要求, 通过答辩。

答辩小组组长签字: 高欣 2024年5月20日

答辩小组成员: 于平 王一帆 刘洋

学院意见	同意
	最终成绩: 百分制 85 ; 五分制 良好
	院长签章: 范新波 学院盖章: 人工智能学院 2024年5月30日



基于 LLM 的交互式多模态图像编辑系统的设计与搭建

摘要

随着深度学习技术在图像处理领域和文本生成领域的迅速发展，多模态交互系统的构建已成为研究的热点。本文介绍了一个基于最新图像生成模型和大语言模型的交互式多模态图像编辑系统的设计与搭建，系统利用了 Stable Diffusion、DALL-E 等图像生成模型和 ChatGPT 系列、ChatGLM2-6B 等大语言模型，通过图形用户界面（GUI）、中间件（middleware）进行图像生成模型和大语言模型的整合，实现了一个既直观又高效的基于文本交互的多模态图像编辑系统。用户可以通过简单的文本指令控制图像编辑过程，系统能够自动解析这些指令并修改图像。同时，本项目通过自动化脚本使用特定图像数据集、图像分割模型和 GPT3.5 Turbo 构建了适用于特定任务的大语言模型微调数据集，使用了 LoRA 方法对 ChatGLM2-6B 模型进行微调，并在微调后在特定任务中获得了近似于 GPT3.5 Turbo 的性能表现。本项目评估了系统在实际应用中的表现，结果显示该系统能够有效地提高图像编辑效率和用户交互体验。

关键词 图像编辑 大语言模型 多模态

Design and Construction of Interactive Multimodal Image Editing System Based on LLM

ABSTRACT

With the rapid advancement of deep learning technologies in the fields of image processing and text generation, the construction of multimodal interaction systems has become a focal point of research. This paper presents the design and construction of an interactive multimodal image editing system based on the latest image generation models and large language models. The system utilizes image generation models such as Stable Diffusion and DALL-E, alongside large language models like the ChatGPT series and ChatGLM2-6B. Through a graphical user interface (GUI) and middleware, it integrates these models to create an intuitive and efficient text-based multimodal image editing system. Users can control the image editing process through simple text commands, which the system automatically parses and applies to modify images. Additionally, this project has constructed a fine-tuning dataset for specific tasks using automated scripts with specific image datasets, image segmentation models, and GPT-3.5 Turbo. The LoRA method was used to fine-tune the ChatGLM2-6B model, achieving performance close to that of GPT-3.5 Turbo in specific tasks. The system's performance in practical applications has been evaluated, showing that it can effectively improve image editing efficiency and enhance user interaction experience.

KEY WORDS Image Editing Large Language Models Multimodal

目 录

第一章 绪论	1
1.1 项目背景与意义	1
1.1.1 项目背景	1
1.1.2 项目意义	1
1.2 国内外研究现状	1
1.2.1 图像编辑	2
1.2.2 大语言模型微调	2
1.2.3 多模态图像编辑方法	3
1.3 项目内容及创新点	4
1.3.1 项目内容	4
1.3.2 项目创新点	5
1.4 论文结构	5
第二章 相关技术研究	6
2.1 扩散模型	6
2.2 基于扩散模型的可控图像生成	7
2.3 大语言模型	7
2.4 系统开发工具	8
2.5 本章小结	9
第三章 基于 LLM 的交互式多模态图像编辑系统的需求分析	10
3.1 系统业务与用户角色分析	10
3.2 系统功能需求分析	10
3.3 本章小结	10
第四章 基于 LLM 的交互式多模态图像编辑系统的设计与实现	11
4.1 GUI 的构建	11
4.1.1 图像自动遮罩与优化	12
4.1.1.1 图像自动遮罩	13
4.1.1.2 图像遮罩性能优化	13
4.1.1.3 对自动生成的遮罩进行优化	14
4.1.2 多模态	15
4.1.2.1 JSON 指令生成	15
4.1.2.2 JSON 指令校验	16
4.1.2.3 多指令处理	16

4.1.2.4 图像模型请求参数生成	16
4.1.3 图像修改建议	17
4.2 Middleware 的构建	17
4.2.1 对多个平台的 API 进行配置和整合	17
4.2.2 使用 Beego 框架提供 API 服务	18
4.3 LLM 的微调	18
4.3.1 LLM 微调数据集生成与性能评估方法	18
4.3.1.1 微调数据集生成	18
4.3.1.2 LLM 指令生成任务性能评估方法	19
4.3.2 ChatGLM2-6B 针对指令生成任务的微调	20
4.3.3 各个 LLM 在本任务下的性能评估	20
4.4 Stable Diffusion 及扩展的使用	22
4.4.1 Stable Diffusion API 的使用	23
4.4.2 ControlNet 的使用及效果	24
4.4.3 Roop 的使用及效果	24
4.5 本章小结	25
第五章 系统实现效果与使用	26
5.1 系统实现效果	26
5.2 系统使用方法	26
5.2.1 系统部署	26
5.2.2 GUI 使用说明	28
5.3 本章小结	29
第六章 项目管理与维护	30
6.1 代码管理	30
6.2 自动化测试	30
6.3 持续集成与持续部署	31
6.4 本章小结	32
第七章 总结及未来展望	33
7.1 总结	33
7.2 未来展望	34
参考文献	
致 谢	
附 录	
外 文 资 料	
外 文 译 文	

任 务 书

开 题 报 告

中 期 检 查 表

教师指导毕业设计(论文)记录表

第一章 绪论

1.1 项目背景与意义

1.1.1 项目背景

随着技术的迅速发展，图像生成编辑在媒体娱乐、数字营销等领域等多个行业中发挥越来越重要的作用，然而传统的图像编辑技术在交互性和生成图像质量上仍面临许多挑战。传统图像编辑工具往往依赖于专业的技术知识和复杂的操作界面，交互性通常较差，不能很好地根据用户的具体需求进行灵活调整和响应，对于普通用户来说门槛较高，用户需要花费大量时间学习如何使用这些工具，限制了工具的普及性和易用性。

为了改善图像编辑的交互性和灵活性，人们正在利用深度学习技术来改善图像编辑系统。目前的研究显示扩散模型已经在图像生成领域具有巨大的潜力，其能够学习大量图像数据，自动提取复杂的特征，并生成高质量的图像。若将其运用于图像编辑，或许可以改善图像编辑的质量。对于交互性，若将最新的 GPT4Turbo 等性能优异的大语言模型融入到图像编辑系统中，或许可以进一步提升系统的交互性，实现更加自然和直观的交互。

通过将图像生成模型与大语言模型结合，可以创建一个更加灵活且易用性强的图像编辑系统，该系统不仅能够提供更加直观的编辑界面，降低用户的操作难度，还能根据用户的描述自动生成或修改图像，极大地提升生成图像的质量和编辑效率。用户可以通过简单的语言指令，如“增加图片亮度”或“改变背景为海滩”，直接与编辑系统交互，系统能够理解这些指令并即时作出响应。

通过整合深度学习和语言模型技术，我们有望构建出一个全新的交互式图像编辑系统，其不仅能够提供更高质量的图像编辑结果，更能够提供给用户更加自然的交互方式，为该领域的专业人士和普通用户都带好的图像编辑体验。

1.1.2 项目意义

通过融合先进的图像生成技术与大语言模型，本项目希望能搭建一个基于 LLM 的交互式图像编辑系统，提供更为易用、精准且高效的图像编辑工具，提升媒体娱乐、数字营销以及智能医疗等多个行业的图像处理能力。通过实现更加智能化和用户友好型的编辑系统，为广大用户带来前所未有的图像编辑体验。这样的研究与开发，为图像编辑技术的未来提供了一种可能性和一条新的探索和实践路径。

1.2 国内外研究现状

Multimodal image synthesis and editing: A survey^[1] 这篇论文对多模态图像合成和编辑的进展进行了综述，讨论如何有效地结合多种模态信息来创造和编辑图像。该论文提出了一种分类体系，根据数据模态和模型架构进行分类，并详细介绍了多模态图像合成

和编辑的各种方法。该论文还讨论了基准数据集、评价指标以及当前研究中的挑战，并提出了未来的研究方向，强调在图像合成和编辑任务中引入跨模态指导的重要性和潜力。

1.2.1 图像编辑

ImageBART^[2] 是一种基于自回归变换的图像生成和编辑模型，灵感来源于自然语言处理中的 BART(Bidirectional and Auto-Regressive Transformers)^[3]，其结合了自回归和编码-解码架构，用于高效地生成和编辑图像。EditGAN^[4] 提出了一种高精度的语义图像编辑方法，允许用户通过修改图像的详细部分分割掩模来编辑图像，其基于生成对抗网络(GAN)构建，只需要少量标记样本即可进行训练，实现了高效的编辑。Generating images from captions with attention^[5] 介绍了一个能从自然语言描述中生成图像的模型，这个模型在关注描述中的相关词汇的同时利用深度循环注意力编写器(DRAW)迭代地在画布上绘制图像。Object-based image editing^[6] 介绍了基于对象的图像编辑技术，允许用户直接在对象层面而不是像素层面进行编辑。Faceshop: Deep sketch-based face image editing^[7] 提供了一种利用深度学习来进行面部图像编辑的方法。Image-based modeling and photo editing^[8] 探讨了基于图像的建模与照片编辑技术，特别强调了分层编辑和实体分离技术的应用。Invertible conditional gans for image editing^[9] 探索了可逆 Conditional GANs 在图像编辑中的应用，这种模型结合了编码器和 cGAN，可对真实图像进行精确修改。In-domain gan inversion for real image editing^[10] 研究了在领域GAN反转技术在真实图像编辑中的应用，通过GAN学习特定图像域的编辑操作。Poisson image editing^[11] 探讨了泊松图像编辑技术，该技术能够无缝地将图像区域融合和编辑，并提供高质量的编辑效果。Image editing in the contour domain^[12] 提出了一种在轮廓域进行图像编辑的方法，这种方法直接在轮廓层面而不是像素层面进行操作。Diffusion maps for edge-aware image editing^[13] 使用扩散映射技术进行边缘感知图像编辑以在保持边缘清晰的同时平滑颜色和纹理。

1.2.2 大语言模型微调

Parameter-efficient fine-tuning of large-scale pre-trained language models^[14] 介绍了Delta-tuning 技术，它在不改变预训练大模型(PLMs)架构的基础上，通过微调少量参数来适应新任务。其主要包括以下方法：增加式、指定式和重新参数化方法，以及在变换器层添加适配器模块的实际应用，这些方法比全参数微调更节省计算资源，提高了模型的适应性和效率。How fine can fine-tuning be? learning efficient language models^[15] 分析了微调过程的效率，对在保持性能的同时减少计算成本的方法进行探索。# InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models^[16] 使用指令标签分析监督式微调过程，以提高大型语言模型在特定任务上的性能。Scaling federated learning for fine-tuning of large language models^[17] 应用联邦学习方法来微调大型语言模型，以提高模型在多个数据源上的通用性和隐私保护。Longlora: Efficient fine-tuning of

long-context large language models^[18] 针对长上下文的大型语言模型进行高效微调，以提升模型对长距离依赖信息的处理能力。Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively^[19] 通过自适应优化子网络来有效微调预训练语言模型，以提高模型的灵活性和适应新任务的能力。Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models^[20] 提出一种用于大型语言模型参数高效微调的适配器家族，以减少模型调整过程中的资源消耗。Language models are few-shot learners^[21] 探讨语言模型在小样本学习中的表现，并分析其微调和适应新任务的能力。Fine-tuning language models with just forward passes^[22] 提出一种仅通过前向传播进行微调的新方法，以简化语言模型的调整过程。

1.2.3 多模态图像编辑方法

PixelTone: A Multimodal Interface for Image Editing^[23] 介绍了一种结合语音和直接操作的多模态照片编辑界面，旨在简化移动设备上的图像编辑任务，通过自然语言和草图来定位图像中的具体更改区域。研究还开发了一个定制的自然语言解释器，将用户的语言指令映射到具体的图像处理操作上。通过用户研究，验证了接口的有效性，展示了其在简化编辑流程和提高用户互动体验方面的潜力。DiffusionCLIP^[24] 是一种结合了扩散模型和 CLIP^[25] 模型的新型图像编辑方法。这种方法利用了扩散模型的生成能力和 CLIP 的语义理解能力，以实现通过文本描述来精确控制图像的内容和属性的编辑。Language-based image editing with recurrent attentive models^[26] 提出一种基于语言的图像编辑方法，通过递归注意模型允许用户用自然语言描述来指导图像编辑过程。Imagic: Text-based real image editing with diffusion models^[27] 利用扩散模型和文本描述来进行真实图像的编辑，并允许细粒度控制和高度个性化的编辑。Sequential attention GAN for interactive image editing^[28] 介绍了一个用于交互式图像编辑的顺序注意力 GAN 模型，允许用户通过多轮对话逐步指导图像编辑。Towards automatic image editing: Learning to see another you^[29] 研究了自动图像编辑的可能性，通过机器学习方法让系统能够根据用户的需求生成或修改图像。Instructpix2pix: Learning to follow image editing instructions^[30] 使用 pix2pix 模型学习遵循图像编辑指令，使得模型能够根据文字描述自动进行图像编辑。Blended diffusion for text-driven editing of natural images^[31] 使用自然语言界面和数据驱动的图像生成技术进行文本驱动的图像编辑。Tigan: Text-based interactive image generation and manipulation^[32] 提出了一种基于文本的交互式图像生成和操作框架。Shape-aware text-driven layered video editing^[33] 对基于形状感知的文本驱动分层视频编辑方法进行探索。Prompt tuning inversion for text-driven image editing using diffusion models^[34] 使用扩散模型，为基于文本的图像编辑提供一种新的反演方法。Lightweight text-driven image editing with disentangled content and attributes^[35] 提供了通过解耦内容和属性以进行轻量级的基于文本驱动的图像编辑的方法。

1.3 项目内容及创新点

1.3.1 项目内容

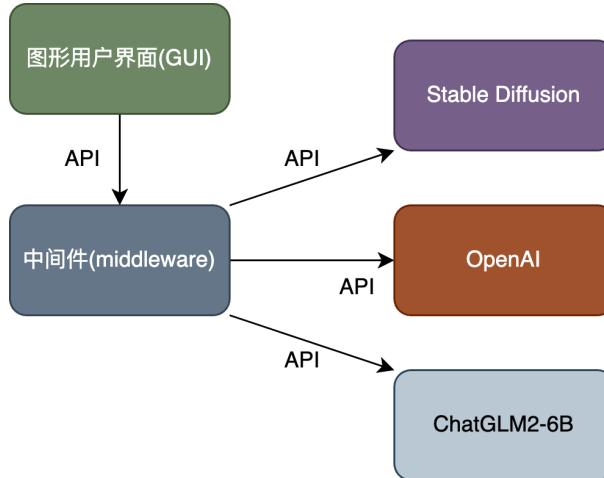


图 1-1 项目整体结构

该项目是一个多模态交互式图像编辑系统，它主要实现了 GUI、Middleware、以及对 Stable Diffusion 和 ChatGLM2-6B 模型的修改与适配。在整体结构上，GUI、Middleware、以及模型修改之间的相互关系和数据流向见图 1-1。

项目基于 Stable Diffusion 的开源项目 stable-diffusion-webui 进行了扩展，增强了其功能。系统可以调用不同的 Stable Diffusion 模型并结合多个扩展后的功能来对图像进行精细的修改和调整。这一功能的实现大大增强了系统对图像处理的灵活性和多样性。

项目还包括了对 ChatGLM2-6B 模型的微调。通过利用专门为本项目的需求生成的微调数据集进行微调，进一步提升了语言模型处理特定任务的能力和准确性，且能够利用 API 调用这些经过微调的 ChatGLM2-6B 模型。

通过调用 OpenAI 的 API，本项目实现了多项功能：利用 GPT4V 生成关于图像修改的建议，使用 GPT3.5Turbo 来辅助生成用于微调大语言模型的数据集以及图像修改指令，以及在 Stable Diffusion 模型不可用时，使用 DALL-E2 作为替代模型来进行图像修改。这些功能为基于 LLM 的交互式图像编辑系统提供了强大的工具。

在 GUI 方面，本系统主要通过 Python 语言实现，构建了一个直观且用户友好的交互界面。该界面简洁易用，通过 API 调用向 Middleware 发出请求，有效减轻了用户在执行计算密集型任务时对硬件资源的需求，从而降低了用户侧的使用门槛。

Middleware 部分使用 Golang 语言构建了一个后端服务。这个服务不仅接入了 Stable Diffusion、ChatGLM2-6B、OpenAI 的 API，还将这些 API 进行了有效的整合，向 GUI 提供了一致风格的 API 接口。这种设计不仅提升了 GUI 调用多方 API 的便利性，也通过统一的配置和管理，极大地增强了系统的可维护性和稳定性。

通过整合深度学习和大语言模型技术，本项目不仅能够提供更高质量的图像编辑结果，还能为用户提供更加自然的交互方式，极大地提升了图像编辑的效率和体验。此外，

系统设计中还考虑了扩展性和未来技术的整合，预留了接口以适应未来可能的升级，以适应快速发展的技术需求。

1.3.2 项目创新点

本项目的方案设计充分考虑了技术实现的可行性与用户操作的便捷性，力求在满足复杂功能需求的同时，保证系统的易用性和稳定性。本项目将复杂的模型调用参数抽象为用户可理解的模版与设置，建立了一个基于 JSON 的指令机制打通了大语言模型与图像生成模型，同时对多种主流大语言模型和图像编辑模型进行适配预留有接口，可随着大语言模型和图像生成模型技术的发展快速迭代。通过系统化的设计和技术的整合，本项目实现了一个高效且用户友好的多模态交互式图像编辑系统。

1.4 论文结构

本论文共分为七个章节，每个章节的具体内容如下：

第一章，绪论。本章主要论述了基于 LLM 的交互式多模态图像编辑系统的项目背景与意义，总结了相关的国内外研究现状，简述了项目内容并列举了项目创新点。

第二章，相关技术研究。本章主要介绍了本项目所使用到的相关技术，包括扩散模型、基于扩散模型的可控图像生成、大语言模型、系统开发工具。

第三章，基于 LLM 的交互式多模态图像编辑系统的需求分析。本章主要对系统业务与用户角色进行分析，结合本项目所使用技术的特点和传统图像编辑系统的痛点，分析了系统功能需求。

第四章，基于 LLM 的交互式多模态图像编辑系统的设计与实现。本章主要论述系统的设计与实现方式，包括 GUI 与 Middleware 的构建、多模态的实现、LLM 的微调、StableDiffusion 及扩展的使用。

第五章，系统实现效果与使用。本章主要展示了系统实现效果，并对系统使用方法进行说明。

第六章，项目管理与维护。本章主要论述了项目所使用的管理与维护方法，包括代码管理、自动化测试、持续集成与持续部署。

第七章，总结及未来展望。本章对全文和开发工作进行总结与归纳，并提出基于 LLM 的交互式多模态图像编辑系统的未来展望。

第二章 相关技术研究

2.1 扩散模型

扩散模型是近年来发展起来的一种新型生成模型，与传统的生成对抗网络（GANs）和变分自编码器（VAEs）相比，扩散模型在生成图像的质量和多样性方面展现出了卓越的性能，其基本原理是模拟从高质量数据分布到高熵噪声分布的逐步转变过程，然后再逆向这一过程以生成新的数据。扩散模型的关键优势在于其生成的图像质量较高且自然，而且在训练过程中相对稳定，不易出现生成对抗网络中常见的模式崩溃问题，其基本原理如下：

前向过程（Forward Process）也称为扩散过程逐步将原始数据 x_0 加入高斯噪声，最终转化为纯噪声 x_t 。此过程可以表示为算法 1：

算法 1 扩散模型的前向过程

```

1: 初始化  $x_0 \sim q(x_0)$ 
2: for  $t = 1$  到  $T$  do
3:    $x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I})$ 
4: end for
```

反向过程（Reverse Process）是前向过程的逆过程，其能通过通过训练的参数化模型 $\epsilon_\theta(x_t, t)$ 从纯噪声状态 x_t 逐步重构出原始数据 x_0 。此过程可以表示为算法 2：

算法 2 扩散模型的反向过程

```

1: 初始化  $x_T \sim \mathcal{N}(0, \mathbf{I})$ 
2: for  $t = T$  到  $1$  do
3:   如果  $t > 1$ ，则采样  $z \sim \mathcal{N}(0, \mathbf{I})$ ，否则  $z = 0$ 
4:    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$ 
5: end for
6: return  $x_0$ 
```

扩散模型的优化最小化真实噪声和模型估计噪声之间的差异，其过程可以表示为算法 3：

算法 3 扩散模型的优化过程

```

1: repeat
2:   采样  $x_0 \sim q(x_0)$ ,  $t \sim \text{Uniform}\{1, \dots, T\}$ ,  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
3:    $\tilde{x}_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ 
4:   对  $\nabla_\theta \|\epsilon - \epsilon_\theta(\tilde{x}_t, t)\|^2$  进行梯度下降
5: until 收敛
```

Stable Diffusion^[36] 是一种基于扩散模型的深度学习图像生成模型，它能够根据文本描述生成高质量的图像。该模型采用了条件生成技术，允许用户通过文本指令来引导噪

声逆向还原从而生成高质量的图片，在艺术创作、媒体娱乐、广告和数字营销等多个领域具有广泛的应用。

由于本项目对于图像生成模型的要求较高且需求复杂，为了便于结合 Stable Diffusion 模型和其他前沿研究成果及开源社区项目，本项目在构建 Stable Diffusion 模块时以开源项目 stable-diffusion-webui¹为基础，结合 sd-webui-controlnet 和 sd-webui-roop 等扩展，通过 API 为 Middleware 提供服务。当 Stable Diffusion 模型不可用时，系统则会调用 OpenAI 的 DALL-E 2 模型。DALL-E 2 是一个先进的图像生成模型，可以根据用户提供的文本描述生成详细、高质量的图像，其核心优势在于其创造力和多样性，能够在遵循描述的同时，创造出独特和富有创意的视觉内容。

2.2 基于扩散模型的可控图像生成

ControlNet^[37] 是一种新的神经网络架构，用于在大型预训练的文本到图像扩散模型中添加空间条件控制。ControlNet 的核心是利用预训练模型的深层和健壮的编码层作为强大的支撑，学习多种条件控制。该网络通过零初始化的卷积层 (zero convolutions) 连接，这些层从零开始逐步增长参数，确保训练初期不会引入有害的噪声。通过使用 ControlNet，用户可以更精细地控制图像生成过程，使生成的图像更贴近用户的具体需求，尤其是在空间布局和细节表达上。这种方法能够有效地减少试错循环，提高图像生成的效率和质量。其与 Stable Diffusion 结合的方式如图 2-1 所示（图片摘自论文 Adding Conditional Control to Text-to-Image Diffusion Models^[37]）。

2.3 大语言模型

近些年大语言模型发展异常迅猛，其通过学习大量文本数据，能够出色地完成生成文本、回答问题、翻译语言等任务。随着算力的提升和语料的增加，大语言模型已经取得了显著的进步，并在多个应用场景中展现出了强大的能力。目前大语言模型主要使用 Transformers^[38] 架构，其能够有效处理长距离依赖问题。GPT (Generative Pre-trained Transformer)、BERT (Bidirectional Encoder Representations from Transformers) 等主流大语言模型，通过大规模的语料进行预训练，对语言的深层次结构和语义的理解能力有了显著的提升，在应用方面展现出广泛的适用性。在大语言模型中，最著名的便是 OpenAI 的 GPT 系列：GPT-3.5 Turbo 是 OpenAI 开发的一款先进的自然语言处理模型，属于 GPT-3 系列的增强版本，在处理大量文本和生成文本方面表现出色；GPT-4^[39] 是 GPT-3 的后续版本，代表了目前最新一代的大语言模型技术，在模型结构和训练数据量上进行了大幅扩展，能够更准确地理解和生成复杂的文本，在理解上下文、维持一致性以及生成更自然的语言方面具有显著优势；GPT-4V 是 GPT-4 的一个特殊版本，优化了图像标注、视觉问答等视觉任务，结合文本和视觉处理能力，能够更好地理解和生成与

¹<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

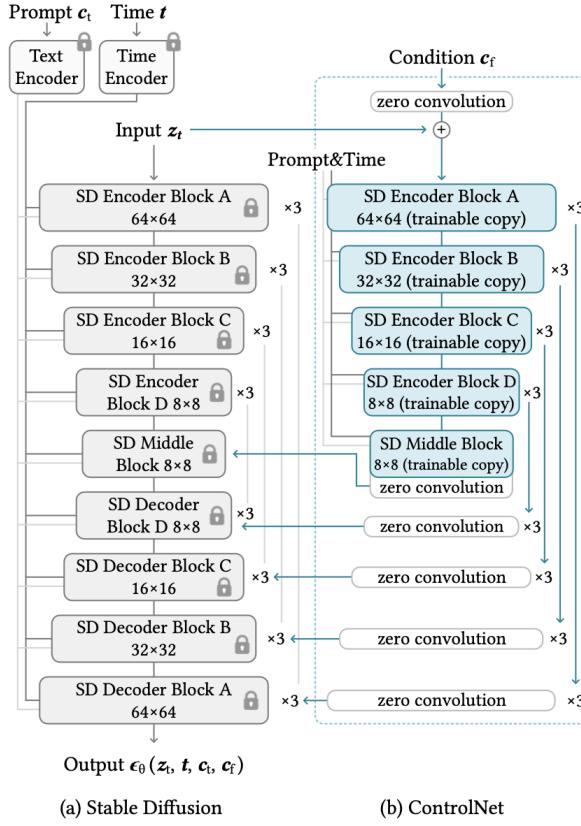


图 2-1 ControlNet 与 Stable Diffusion 结合

图像相关的文本内容。

ChatGLM2-6B 是由清华大学开发的第二代开源双语(中英)对话模型, 基于 ChatGLM-6B 进行迭代并提升了性能, 这款模型经过大规模预训练, 实现了显著的性能提升, 并在多个数据集上表现出色。ChatGLM2-6B 支持更长的对话上下文, 并提高了推理速度和降低了显存占用, 使得即使在资源有限的环境下也能有效运行。本项目使用 ChatGLM2-6B 模型, 结合开源项目 LLaMA-Factory¹, 利用本项目提供的数据自动生成功能所生成的数据集, 使用 LoRA^[40] 方法对模型进行微调以在本项目所需的任务中获得更佳的表现。微调后的模型通过 fastapi 提供 API 服务。

LoRA 通过在每个 Transformer 层中注入可训练的低秩矩阵对权重进行更新, 而不是更新整个权重矩阵。这种方法不仅减少了模型的存储和计算成本, 还保持了与完全微调相当的模型性能。其算法可概括为算法 4。

2.4 系统开发工具

Gradio 是一个开源的 Python 库, 旨在简化为机器学习模型创建自定义用户界面的过程。本项目使用 Gradio 框架构建了 GUI 组件以提供简单易用的交互界面。

beego 是一个使用 Go 语言开发的开源 Web 框架, 它支持快速开发各种应用程序,

¹<https://github.com/hiyouga/LLaMA-Factory>

算法 4 LoRA: 低秩适应大型语言模型

输入: 预训练模型权重 $W_0 \in \mathbb{R}^{d \times k}$, 输入 x , 低秩 r

输出: 适应后的输出

- 1: 初始化低秩矩阵 $A \in \mathbb{R}^{d \times r}$ 和 $B \in \mathbb{R}^{r \times k}$
 - 2: 设定更新权重 $\Delta W = BA$
 - 3: 冻结预训练权重 W_0
 - 4: **for** 每一层 Transformer **do**
 - 5: 使用 $W = W_0 + \Delta W$ 进行前向传播
 - 6: 优化 A 和 B 来最小化损失函数
 - 7: **end for**
 - 8: 输出调整后的模型结果
-

包括 API、Web 应用和后端服务。该框架设计灵感来源于 Tornado、Sinatra 和 Flask，结合了 Go 的接口 (interface) 和结构体嵌入 (struct embedding) 等特性，提供了高效的性能和简便的操作。本项目使用 beego 框架构建了 Middleware 组件以提供多个 API 服务，以便 GUI 可以高效地使用各种模型和工具。

2.5 本章小结

在本章中，我们深入探讨了与多模态图像编辑系统相关的关键技术，首先详细介绍了扩散模型的工作原理及其在图像生成中的应用，接着讨论了大语言模型的相关信息与 LoRA 微调技术，最后概述了系统开发中使用的主要工具和方法。本章不仅为理解系统的技术基础提供了理论支撑，也为后续章节中系统设计和实现的详细讨论奠定了基础。

第三章 基于 LLM 的交互式多模态图像编辑系统的需求分析

3.1 系统业务与用户角色分析

本项目的目标是构建一个基于 LLM 的交互式多模态图像编辑系统，结合大语言模型和图像生成技术，提供一个友好的图形用户界面，支持用户通过自然语言指令进行复杂的图像编辑操作，使非技术用户也能轻松使用复杂的图像编辑功能。系统需能够理解和解析用户的自然语言指令，将其转换为具体的图像编辑任务，并利用最新的图像生成模型如 Stable Diffusion 根据由用户的指令生成的相关参数修改图像。

本项目的目标用户群体包括设计师、营销专业人员、教育工作者和普通消费者。设计师可以利用系统快速实现创意构思，改善设计流程的效率。营销专业人员可以使用该系统快速调整广告图像，以适应不同的市场需求。教育工作者可以使用此系统来创建或修改教学资料中的图像，使教学内容更加生动有趣。普通消费者则可以利用系统提供的易于使用的平台，探索个人图像编辑和创意表达。

3.2 系统功能需求分析

项目旨在整合 Stable Diffusion 和大语言模型等最新的深度学习模型以实现通过自然语言指令实现高质量的图像编辑。图形用户界面 (GUI) 设计应简洁直观，包含预览功能，提升用户体验。中间件 (Middleware) 需高效处理 GUI 请求，并向模型 API 转发，确保系统稳定性和响应速度。系统应能高效处理复杂的图像编辑任务，并考虑多种使用场景和用户需求，以保证广泛应用和良好用户体验。

3.3 本章小结

本章主要对基于 LLM 的交互式多模态图像编辑系统进行了需求分析。项目旨在结合大语言模型和图像生成技术提供一个高效、直观的图像编辑系统，使用户能够轻松使用复杂的图像编辑功能。项目的目标用户群体包括设计师、营销专业人员、教育工作者和普通消费者，需支持用户通过自然语言指令利用图像生成模型进行高质量的图像编辑操作。系统需要整合多个深度学习模型、设计直观的 GUI 界面、实现高效处理请求的中间件，并考虑多种使用场景和用户需求，以提供良好的用户体验。

第四章 基于 LLM 的交互式多模态图像编辑系统的设计与实现

在本项目中，图形用户界面（GUI）作为用户与系统交互的前端界面扮演了至关重要的角色，不仅需要具备直观操作的特性，还应支持复杂的自定义图像处理功能，因此本项目将简洁且直观考虑为 GUI 设计中的重点，通过图形化元素如按钮、图标和菜单等，允许用户以简单的点击操作来进行交互。GUI 还集成了一些预览功能，用户可以查看遮罩效果，有效提升用户体验和操作的精确性。

Middleware 负责处理来自 GUI 的请求，在本系统中起到了桥梁的作用。Middleware 采用了高效的 Golang 语言结合 beego 框架进行构建以保证系统的响应速度和稳定性，将来自 GUI 的请求进行处理并向对应的部署在云平台或 OpenAI 的模型调用 API 转发。它整合了包括 Stable Diffusion、ChatGLM2-6B 及 OpenAI 提供的 API 等多种服务，通过提供统一风格的 API 接口，极大地简化了 GUI 与模型之间的交互复杂度，不仅提高了开发效率，也便于系统的后期维护和升级。

图像编辑和文本交互方面所用到的图像生成模型和大语言模型是本项目的核心技术。项目中采用了最新的扩散模型技术——Stable Diffusion，其通过学习大量的图像数据，能够生成高质量的图像内容。项目还引入了大语言模型（如 GPT 系列和 ChatGLM2-6B）来处理和理解用户的自然语言指令，实现更加智能的图像编辑功能。用户可以使用简单的语言描述如“将背景更换为海滩”来进行图像编辑，系统能够将用户输入自动解析为特定指令并进行相应的图像编辑。

通过整合 GUI、Middleware、图像生成模型和大语言模型，本项目能够接受用户的文本输入并进行相应的复杂图像编辑。系统的设计考虑到了多种使用场景和用户的不同需求，确保了广泛的应用性和良好的用户体验。

4.1 GUI 的构建

图形用户界面是现代软件项目中不可或缺的组成部分，它极大地提升了应用程序的可访问性和用户体验。GUI 通过可视化元素如按钮、图标和菜单等，允许用户以直观的方式与系统进行交互，简化了操作过程并降低了用户的使用门槛。本项目所构建的 GUI 为用户提供了一个交互方式简单且功能强大的图形用户界面，通过图形化的方式展示信息和选择，使得用户能够通过简单的点击或触摸来执行命令或更改设置。GUI 虽然承担计算任务较少，但却是承载本项目结构与逻辑的关键部分。通过使用符合规则的指令作为中枢，GUI 打通了大语言模型和图像生成模型之间的壁垒，使基于 LLM 的创新交互式图像编辑系统成为可能。GUI 的模块构成如表 4-1：

用户首先上传需要修改的图片，然后可在 Chat 模块中选择不同的大语言模型进行交互并得到相应的指令，最后在 Operation Board 模块中选择指令执行或一键全部执行。如果对自动生成的遮罩不满意，可在 Edit Image 中对遮罩进行修改。

在 Auto 模块中，用户可通过选择多张图片批量生成满足微调大语言模型微调所需的数据。其会循环地从给定的图片集中随机选择图片继续分割，将分割后的结果和特定

表 4-1 GUI 模块

模块	描述
BaseImage	接受上传的原始图片并预览
EditedImage	预览修改后的图片
Operation Board	执行指令
Settings	对系统进行设置
Chat	与大语言模型交互的聊天界面
Edit Image	对图像进行自定义遮罩和换脸等操作
Auto	执行自动化操作
Manual	系统使用说明

的 prompt 通过 GPT3.5Turbo 生成对应的修改建议，再将分割的结果、生成的建议通过 GPT3.5Turbo 生成指令。

4.1.1 图像自动遮罩与优化

由于本项目需要提供对图像进行部分修改的功能，所以需要在使用图像生成模型进行图像编辑时需要提供一个遮罩以明确需要修改的部分和不需要修改的部分。为了自动生成符合要求的遮罩，本项目借助图像分割和大语言模型的辅助，可通过两种方式生成自动遮罩：基于关键词对自动生成遮罩和基于已给出的点自动填充生成遮罩。两种方法都会首先使用图像分割模型对图像进行分割（如图 4-1），然后根据给出的要求对相应的部分进行遮罩生成原始的遮罩。受制于图像分割模型在边缘上的表现并不理想，需要对特定的分割部分进行处理以提高遮罩的质量，因此最后会通过本项目设计的优化算法生成最终的遮罩。



图 4-1 图像分割结果：(a)原始图像，(b)分割结果

4.1.1.1 图像自动遮罩

基于关键词自动生成遮罩的方法会根据关键词和图像分割结果生成自动原始的遮罩，该功能会遍历每个给出的关键词，若关键词与分割结果之一吻合，则会对相应的分割区域进行遮罩，生成原始的遮罩（如图 4-2）。

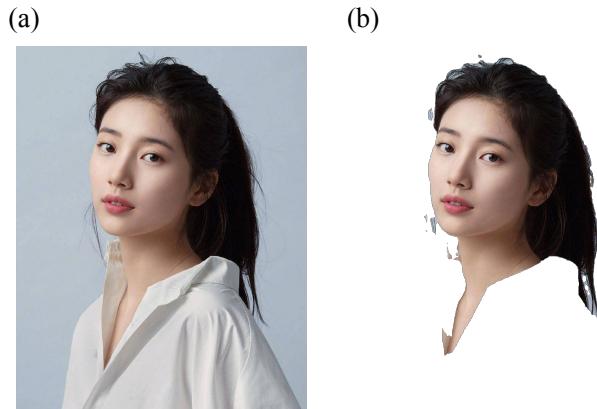


图 4-2 关键词自动生成遮罩结果：(a)原始图像，(b)keywords=[’Background’, ’Upper-clothes’, ’Dress’, ’Right-arm’] 得到的遮罩

基于已给出的点自动填充生成遮罩的方法会根据在图片中标记的点和图像分割结果生成自动原始的遮罩，该功能会遍历每个给出的点，将该点所在的部分全部进行遮罩，最后生成原始的遮罩（如图 4-3）。



图 4-3 基于已给出的点自动填充生成遮罩：(a)标记后的图像，(b)生成的遮罩

4.1.1.2 图像遮罩性能优化

LRU(Least Recently Used)缓存是一种常用的缓存淘汰算法，用于在有限的缓存空间中管理数据。它的核心思想是优先淘汰最长时间未被使用的数据项。functools.lru_cache 是 python 标准库中 functools 模块提供的一个装饰器，它实现了 LRU 缓存机制。该装饰器可以非常方便地被添加到任何想要进行缓存的参数可哈希的函数上，自动地保存最近

执行的函数调用结果并在后续相同的调用中直接返回缓存的结果，避免重复计算的开销。

由于在本项目中图像遮罩存在一张图片多次调用的特点，本项目使用了 LRU 缓存实现性能优化。由于 python 中 `PIL.Image.Image` 对象不可哈希，缓存分割结果时将图像转为 base64 字符串进行映射。

4.1.1.3 对自动生成的遮罩进行优化

由于分割模型性能的限制，生成的原始遮罩可能在某些细节上表现不佳而影响图像编辑模型的结果，因此设计了一个算法对自动生成的遮罩进行优化。算法 5 可以根据配置文件的设置，对特定的未被遮罩的部分在遮罩的边缘进行收缩。

算法 5 遮罩优化算法

输入： 原始遮罩 *OriginMask*，图像分割结果 *SegmentResult*，配置文件 *Config*

输出： 优化后的遮罩 *OptimizedMask*

- 1: 获取遮罩与非遮罩的描边得到像素 *EdgePixels*
 - 2: 从配置文件和图像分割结果获取 *ConfigPixels*
 - 3: 仅保留出现在 *EdgePixels* 中的 *ConfigPixels*
 - 4: **for** *pixel* in *ConfigPixels* **do**
 - 5: Apply MinFilter Kernel(in *Config*) in *OriginMask*[*pixel*]
 - 6: **end for**
 - 7: 得到优化后的遮罩 *OptimizedMask*
-

由于该算法仅会对遮罩边缘上的像素进行卷积且在设计时充分考虑到了内存中像素的存储顺序的原因，虽然需要复杂的处理过程，但经过多次的迭代后算法的时间复杂度降低至 $O(mnr)$ (m,n 表示图片的长宽， r 表示设置的优化强度)。算法实现的效果如图 4-4 所示，可见在发丝附近遮罩的质量得到了明显的改善。

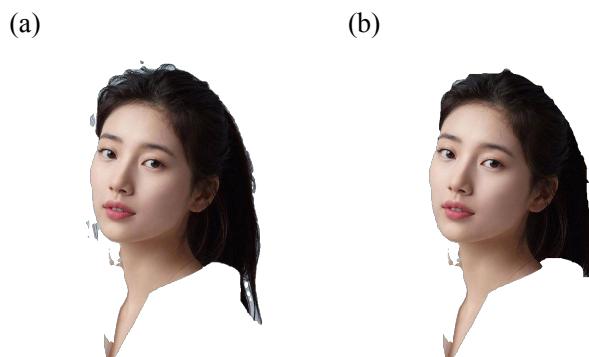


图 4-4 遮罩优化结果：(a)原始遮罩，(b)算法优化后的遮罩

4.1.2 多模态

如何打通大语言模型和图像生成模型是本项目的关键。本项目通过特定的 prompt 和图像分割结果，使用大语言模型生成 JSON 格式的指令并校验，并支持多轮对话。用户可有选择性地执行生成的指令或执行全部指令。系统首先会按照给定的规则对指令进行预处理和排序，然后通过指令生成请求参数来调用图像生成模型。多模态任务的实现方式如图 4-5。

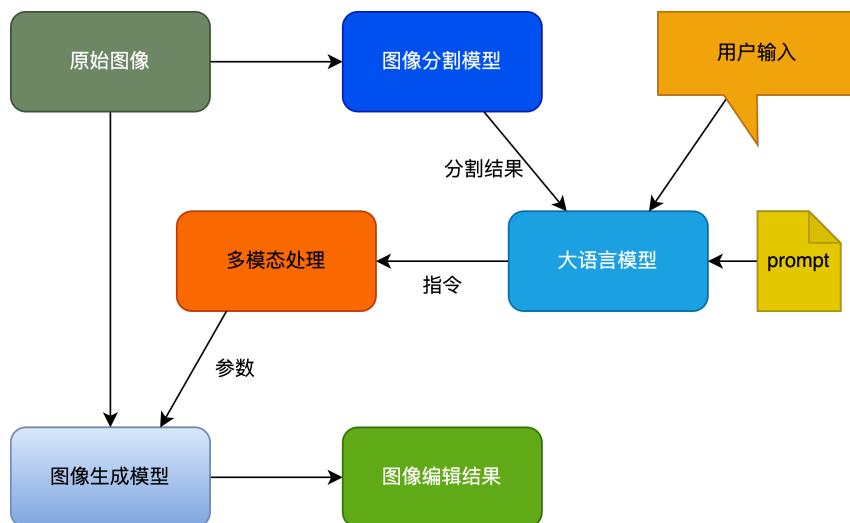


图 4-5 多模态实现方式

4.1.2.1 JSON 指令生成

JSON (JavaScript Object Notation) 是一种轻量级的数据交换格式，易于人阅读和编写，同时也易于机器解析和生成。它基于 JavaScript 的一个子集，但独立于语言，被广泛应用于许多编程语言中。JSON 主要用于网络应用间浏览器与服务器之间的数据传输。在 JSON 中，数据以键值对的形式存在，可以表示数组、布尔值、数字、对象或字符串。由于其简洁性和易于交互的特性，JSON 已成为 Web 应用中数据交换的主流技术。由于 JSON 应用范围广且大语言模型 JSON 处理能力较强，本项目采用此格式来承载大语言模型和图像生成模型的联系。

通过特定的 prompt 和图像分割结果以及用户输入的修改意图，本项目可使用 GPT3.5Turbo、GPT4Turbo、微调后的 ChatGLM2-6B 生成 JSON 指令。例：当图像分割结果为 Background, Hair, Upper-clothes, Dress, Face, Right-arm，用户输入为“将背景更换为蓝天白云，将衣服更改为白色的 T-shirt”时，生成的 JSON 指令如代码 4-1 所示。

代码 4-1 生成的指令

```

1  [
2    {
3      "command" : "change",
4      "paras" : [ ["Background", "Upper-clothes"] , "blue_sky, white_T-shirt"]

```

```

5      }
6  ]

```

4.1.2.2 JSON 指令校验

由于大语言模型生成指令不稳定，需要对生成指令的合规性进行校验。校验规则存储为一个 JSON 文件，以修改非面部和面部的指令为例，指令校验的算法如算法 6 所示，其中校验规则如代码附-1 所示。

算法 6 JSON 指令校验算法

输入：待校验的指令 *JsonCommands*、校验规则 *Rules*

输出：校验后的指令 *ValidJsonCommands*

```

1: for Command in JsonCommands do
2:   if Command satisfy Rules then
3:     Add Command to ValidJsonCommands
4:   end if
5: end for

```

4.1.2.3 多指令处理

由于一次执行可能会涉及到多个指令，会遇到指令重复、指令优先性等问题，所以会对需要执行的指令进行合并与排序。指令合并的算法如下：

算法 7 多指令处理算法

输入：待处理的指令 *OriginCommands*、指令合并规则 *Rules*、指令优先性 *Priority*

输出：处理后的指令 *ProcessedCommands*

```

1: for Command in OriginCommands do
2:   if Same Command type already in ProcessedCommands then
3:     Combine Command to the same one in ValidJsonCommands use Rules
4:   end if
5: end for
6: 根据 Priority 对 ValidJsonCommands 进行排序

```

4.1.2.4 图像模型请求参数生成

图像模型请求参数生成较为复杂，对于某个参数，其可能来源于 GUI 中可修改的设置，可能来源于指令，可能来源于模版，否则设置为默认参数。由于每个参数来说，其来源的优先性可能不一致，因此设计了算法 8 来生成图像模型请求参数。

算法 8 图像模型请求参数生成算法

输入：指令 *Command*、设置 *Settings*、默认参数 *Default*、参数来源优先性 *PriorityRules*

输出：图像模型请求参数 *Parameters*

```

1: for ParaKey in Parameters do
2:   Get Template from Command
3:   if ParaKey found in Command or Settings or Template then
4:     Choose the highest priority source use PriorityRules[ParaKey]
5:   else
6:     Set this parameter to Default[ParaKey]
7:   end if
8: end for

```

4.1.3 图像修改建议

本项目提供了根据图像自动生成图像建议的功能。由于传统的大语言模型只能接受文本输入，因此本项目采用了 GPT4V 来自动生成图像修改建议。

GPT-4V 是由 OpenAI 开发的多模态大型语言模型，是 GPT 系列基础模型的第四代。该模型具有视觉能力，可以将图片作为输入，进行各种任务，例如描述图片中的幽默、总结截图文本、回答包含图表的考试题目等。

用户通过 GUI 界面的 Advise 按键，可以生成建议并将其转换为指令。

4.2 Middleware 的构建

在本项目中，Middleware 作为核心组件，通过整合来自不同平台的 API，为 GUI 提供了统一且易于接入的 API 服务。其整合了多个图像生成和大语言语言模型，通过统一的接口设计，使得 GUI 能够方便地调用所需的功能。此外，Middleware 采用了 Golang 语言和 Beego 框架，不仅保证了 API 服务的高并发处理能力和稳定性，还通过模块化的设计提高了系统的可维护性和可扩展性。主要 API 服务包括 Stable Diffusion 和 DALL-E2 模型、图像分割模型，以及多种大语言模型。这样的架构设计不仅优化了开发效率，也确保了系统的稳定运行和长期发展。

4.2.1 对多个平台的 API 进行配置和整合

Middleware 通过整合不同平台如图像生成模型、语言模型等的 API，使得 GUI 开发者可以通过一个统一的接口调用多种功能。这种整合不仅包括 API 的聚合，还涉及统一 API 调用的风格和路由规范，不仅保证了 API 服务的高稳定性和可靠性，还便于日后的维护和扩展。例如，无论是调用 ChatGLM2-6B 模型还是多种 GPT 模型，GUI 都能通过

相同的结构化请求方式访问不同的服务。

4.2.2 使用 Beego 框架提供 API 服务

本项目使用 beego 框架构建了 Middleware 组件提供了多个 API 服务，以便 GUI 可以高效地使用各种模型和工具。*PostSDTxt2Img* 和 *PostSDImg2Img* 是通过 Stable Diffusion 模型来生成或修改图像的 API，这使得用户能够通过简单的 API 调用，进行复杂的图像生成和编辑操作。*GetLoras* 这个 API 用于获取 Stable Diffusion 模型可用的 LoRA 模型列表。*PostDALLE2Edit* 利用 DALL-E2 模型修改图片，这进一步扩展了图像处理的能力，在 Stable Diffusion 模型不可用时作为替代。*PostHuggingFaceImgSegment* API 通过可在部署在本地的分割模型不可用时通过调用 Hugging Face 上的图像分割模型 API 来实现图像分割。在文本处理方面，*PostGPT3Dot5Turbo*、*PostGPT4Turbo* 和 *PostGPT4V* 等 API 利用 OpenAI 的不同版本 GPT 模型来处理指令理解和生成任务，*PostChatGLM2_6B* 可调用微调后的 ChatGLM2-6B 模型来进行指令生成。

4.3 LLM 的微调

大语言模型 (LLM) 的微调是一种在预训练模型基础上通过特定数据集进一步训练的过程，用于优化模型在特定任务或场景中的表现。特别是对于参数量较小的 LLM，微调不仅可以提升其性能，还能增强其针对具体任务的适应性和泛化能力。微调的一个主要优势是性能提升，即使是较小的模型，通过针对性的微调，可以在特定任务上实现甚至超过大型模型未经微调时的性能，可见微调能够根据具体需求调整模型的行为，使其更加专注于特定的输出目标。微调技术提供了一种有效途径，通过少量的定制化数据提升模型的应用性能，特别是在参数量较小的模型中，这种优势尤为显著。

原始的 ChatGLM2-6B 模型已经通过大规模数据预训练，具备了强大的语言理解和生成能力，但受制于参数量较小，其在本项目需要高精准度的指令生成任务上表现不佳，因此本项目通过自动生成并进行筛选的高质量指令生成微调数据集对 ChatGLM2-6B 进行微调以进一步提升模型在指令生成任务的表现。

本项目利用开源项目 LLaMA-Factory¹和本项目中自动生成的指令生成任务数据集对 ChatGLM2-6B 模型进行微调。

4.3.1 LLM 微调数据集生成与性能评估方法

4.3.1.1 微调数据集生成

在大语言模型 (LLM) 的微调过程中，数据集的功能和作用至关重要。微调数据集不仅提供了模型训练所需的具体数据，还直接影响了模型微调后的性能和适应性。对于参数量较小的模型，高质量的微调数据集尤其重要，因为这些模型通常缺乏足够的参数

¹<https://github.com/hiyouga/LLaMA-Factory>

量来从大规模数据中学习复杂的特征。通过高质量的微调数据集，参数量较小的大语言模型可以有效地提高这些模型的学习效率和最终性能。构造高质量的微调数据集涉及两个关键步骤：数据收集和数据预处理。数据收集需要确保获得足够多的、具有代表性的数据，这些数据能够覆盖模型在实际应用中可能遇到的各种情况。数据预处理是将收集到的原始数据转换成模型可以直接处理的格式，这可能包括数据清洗、特征提取、标签编码等。

由于没有适用于本任务的开源数据集，本项目尝试建立一个自动化工作流程，通过利用 ATR Dataset¹ 调用多个模型和一定的校验规则来生成所需的数据并整合为一个数据集。数据集生成的流程如图 4-6。同时，在系统使用过程中产生的数据可设置是否保存，这些数据也会在运行数据集生成脚本添加到数据集中。

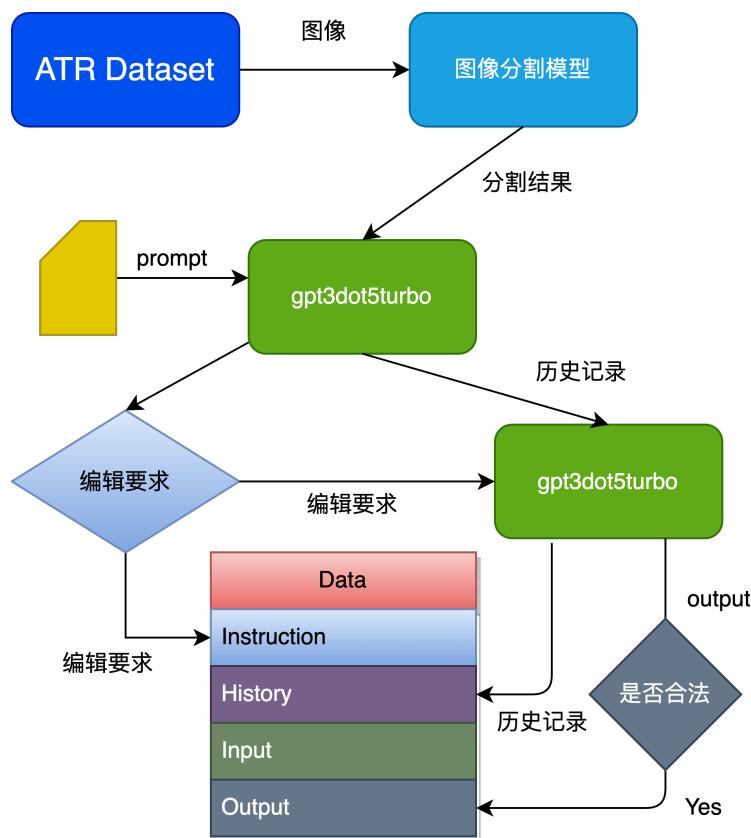


图 4-6 数据集自动生成流程

通过建立的自动化工作流程，本项目创建了一个包含 3000 个样本的数据集，数据集中的每一项数据都经过合法性校验以保证数据集的质量。数据集的格式满足 LLaMA-Factory 的要求并可使用此数据集对 ChatGLM2-6B 进行微调。

4.3.1.2 LLM 指令生成任务性能评估方法

性能评估是大语言模型（LLM）开发和应用的关键环节，尤其是在模型的实用性和可靠性方面。有效的评估方法可以帮助研究者和开发者了解模型在特定任务上的表现，

¹<https://github.com/lemondan/HumanParsing-Dataset>

从而进行进一步的优化和调整。

由于需要一种直观的性能评估方法来对不同的大语言模型在指令生成任务上的性能进行评估，本项目采用生成指令的合法率对不同的大语言模型在指令生成任务上的性能进行评估。

4.3.2 ChatGLM2-6B 针对指令生成任务的微调

LoRA: Low-Rank Adaptation of Large Language Models^[40] 这篇论文提出了一种新颖且高效的 LoRA 微调方法，用于微调大型预训练语言模型以适应特定任务。传统的微调方法往往需要重新训练模型的所有参数，而全参数训练的方法在模型参数规模庞大时需要巨大的量。算法主要通过在 Transformer 模型的每一层中注入低秩的分解矩阵来更新权重，而不改动预训练的权重，其具体算法描述见算法 9。

算法 9 LoRA: 大型语言模型的低秩适配

输入： 预训练权重矩阵 $W_0 \in \mathbb{R}^{d \times d}$, 输入 x , 秩 r

输出： 调整后的输出 h

- 1: 初始化低秩矩阵 $A \in \mathbb{R}^{r \times d}$ 和 $B \in \mathbb{R}^{d \times r}$
 - 2: 冻结预训练权重 W_0
 - 3: **for** 对于每个训练步骤: **do**
 - 4: 计算更新矩阵 $\Delta W = BA$
 - 5: 应用更新: $W = W_0 + \Delta W$
 - 6: 计算输出: $h = Wx$
 - 7: **end for**
 - 8: 优化参数 A 和 B 以最小化损失函数
-

本项目使用生成的指令生成任务数据集结合 LoRA 方法，通过开源项目 LLaMA-Factory¹对 ChatGLM2-6B 模型进行微调。LoRA 微调在大语言模型上的训练损失随训练步数变化的情况如图 4-7 所示。从图中可以看出，最初损失值很高，但随着训练步数的增加，损失值迅速下降，特别是在前 50 步之内下降最为显著。在经过约 50 步之后，损失下降的速度开始放缓，但仍然持续下降，表明模型继续从训练数据中学习。在约 200 步之后，损失曲线趋于平缓，说明模型已经接近收敛，额外的训练步骤在减少损失方面的效果变得有限。图中还展示了一个平滑处理的损失曲线，更清晰地显示了训练过程的整体趋势，而不是每一步的波动。

4.3.3 各个 LLM 在本任务下的性能评估

结合本项目的 LLM 指令生成任务性能评估方法，本项目对不同的大语言模型进行了指令生成任务性能评估。各个模型的性能表现如表 4-2 和图 4-8 所示。

¹<https://github.com/hiyouga/LLaMA-Factory>

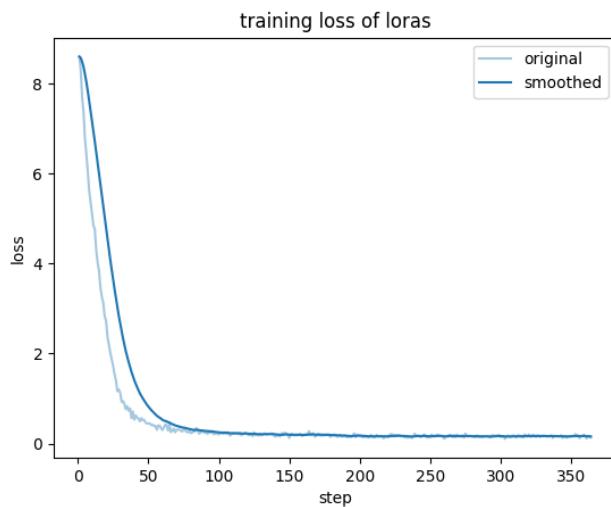


图 4-7 step-loss 图

表 4-2 LLM 指令生成性能

模型	测试样本数	合格率
GPT3.5Turbo	100	97%
GPT4Turbo	100	99%
ChatGLM2-6B(Origin)	100	6%
ChatGLM2-6B(LoRA trained 50 steps)	1000	19.1%
ChatGLM2-6B(LoRA trained 60 steps)	1000	43.7%
ChatGLM2-6B(LoRA trained 80 steps)	1000	77.8%
ChatGLM2-6B(LoRA trained 100 steps)	1000	88.1%
ChatGLM2-6B(LoRA trained 150 steps)	1000	93.9%
ChatGLM2-6B(LoRA trained 200 steps)	1000	95.8%
ChatGLM2-6B(LoRA trained 250 steps)	1000	95.2%
ChatGLM2-6B(LoRA trained 300 steps)	1000	96.7%

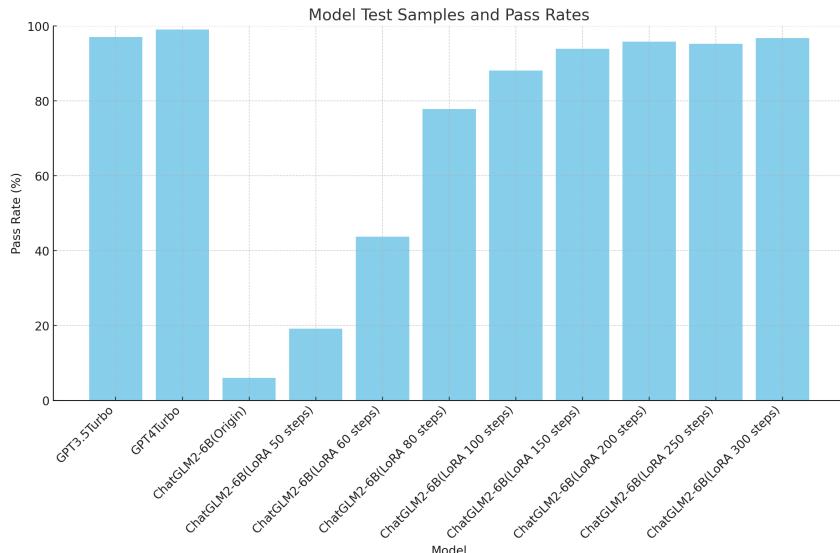


图 4-8 LLM 指令生成性能

可观察到原始的 ChatGLM2-6B 模型在未经过微调时，在指令生成任务中的合格率仅为 6%。这一低下的性能表现暴露了 ChatGLM2-6B 模型在没有针对性训练的情况下难以完成需要高精准度指令生成任务。通过使用 LoRA 方法进行微调，模型性能随着微调步骤的增加显著提升。当微调步骤增至 80 步时，合格率提升至 77.8%；而在经过 200 步的微调后，合格率达到 95.8%，并在 300 步训练后稳定在 96% 左右。与 ChatGLM2-6B 相比，GPT3.5Turbo 和 GPT4Turbo 在无需额外训练的情况下即可达到分别为 97% 和 99% 的高合格率。

4.4 Stable Diffusion 及扩展的使用

Stable Diffusion 是一种基于深度学习的图像生成模型，能够根据文本描述生成高质量的图像。该模型利用条件生成技术，允许用户通过简单的文本指令来引导图像的生成过程。这种技术在艺术创作、媒体娱乐、广告以及数字营销等多个领域显示了广泛的应用潜力。

除了 Stable Diffusion 的基础功能外，本项目还集成了多个扩展，如 ControlNet 和 roop。ControlNet 允许用户对生成的图像进行更精确的控制，通过这种方式 ControlNet 增强了图像的细节质量和一致性，使生成的图像更贴近用户的具体需求。roop 专注于提供面部替换功能，这一功能特别适用于需要在图像中修改人物面部特征的场景。roop 通过简单的操作接口，使得用户能够轻松地将一张人脸图像替换到另一张脸上，而不需要复杂的图像处理知识背景。

sd-webui-controlnet¹ 使用了 ControlNet 的原理，旨在增强原有 Stable Diffusion 模型的图像生成控制能力。通过集成一个额外的控制网络（ControlNet），允许用户精确指导图像的具体内容，显著提升了生成图像的细节质量和一致性。

¹<https://github.com/Mikubill/sd-webui-controlnet>

sd-webui-roop¹基于 DeepFake^[41]，允许用户在图片中进行面部替换，简化了面部交换的过程，无需训练特定的模型，大大降低了使用复杂度。

系统还提供了丰富的参数调整选项，如调整噪声去除强度。这些功能为用户提供了广泛的自定义空间，可以根据具体的应用场景调整生成效果，以满足不同的视觉表达需求。Stable Diffusion 模型及其扩展的结合使用，为本项目的图像编辑方面提供了强大的功能和更高的灵活性。

4.4.1 Stable Diffusion API 的使用

本项目通过将开源项目 stable-diffusion-webui²部署在揽睿星舟机器学习平台³，通过 API 来调用 Stable Diffusion 模型，主要使用的参数见表 4-3。

表 4-3 主要使用的 Stable Diffusion webui img2img API 参数

参数	描述	形式
prompt	输出图像的期望修改或主题	str
negative_prompt	生成图像中应避免的内容	str(base64 Image)
mask	选择性编辑或生成的区域的图像遮罩	str
inpainting_fill	修补时的填充方法	int
inpaint_full_res	是否在图像的全分辨率下应用修补	bool
inpaint_full_res_padding	使用全分辨率时修补区域周围的填充	int
inpainting_mask_invert	是否反转修补遮罩	bool
mask_blur	遮罩边缘的模糊量	float
denoising_strength	去噪强度	float
sampler_index	采样算法	str
seed	初始化随机数生成器的值	int
steps	生成图像时的步骤数	int
width	输出图像的宽度	int
height	输出图像的高度	float
cfg_scale	输入提示的权重	float
restore_faces	是否修复生成图像中的面部	bool
alwayson_scripts	插件参数	dict

¹<https://github.com/s0md3v/sd-webui-roop>

²<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

³<https://www.lanrui-ai.com>

4.4.2 ControlNet 的使用及效果

sd-webui-controlnet¹是一个用于 stable-diffusion-webui 的扩展，允许用户通过添加额外的条件来控制扩散模型的行为，从而增强生成图像的精确度和控制性。这一扩展可以实时添加到原始的 **Stable Diffusion** 模型中，不需要进行合并处理。本项目使用该插件以保持原始图像的主要轮廓和布局不受改变，其实现的效果如图 4-9。



图 4-9 ControlNet 效果：(a)原始图像，(b)未使用 ControlNet，(c)使用 ControlNet

4.4.3 Roop 的使用及效果

roop²是一个用于 stable-diffusion-webui 的扩展，提供面部替换的功能。这个扩展需要一张原始图像和目标图像，并能将原始图像的面部替换到目标图像上。其实现的效果如图 4-10。



图 4-10 roop 效果：(a)原始图像，(b)目标图像，(c)结果

¹<https://github.com/Mikubill/sd-webui-controlnet>

²<https://github.com/s0md3v/sd-webui-roop>

4.5 本章小结

本章中详细介绍了基于大语言模型（LLM）的交互式多模态图像编辑系统的设计与实现。本章首先阐述了图形用户界面（GUI）的构建过程，接着描述了中间件的开发过程，还探讨了 LLM 的微调技术及其在本系统中的应用，以及如何利用 Stable Diffusion 模型和其他技术来增强图像编辑的功能。这些元素共同构成了一个高效、直观的图像编辑系统。

第五章 系统实现效果与使用

系统实现的核心在于其高度集成的图形用户界面 (GUI)，该界面不仅简洁易用，还充分支持复杂的图像处理操作。GUI 设计考虑了用户的操作习惯，提供了直观的图像上传、编辑和预览功能。用户可以通过简单的点击和拖拽操作，实现图像的上传和编辑。GUI 还集成了多个图像生成模型和大语言模型并提供控制选项，如模型选择、参数调整等，使用户能够根据具体需求调整图像处理流程。

5.1 系统实现效果

系统实现了一个简单易用的 GUI，可通过特定端口进行访问。GUI 的总体效果如图 5-1 所示。

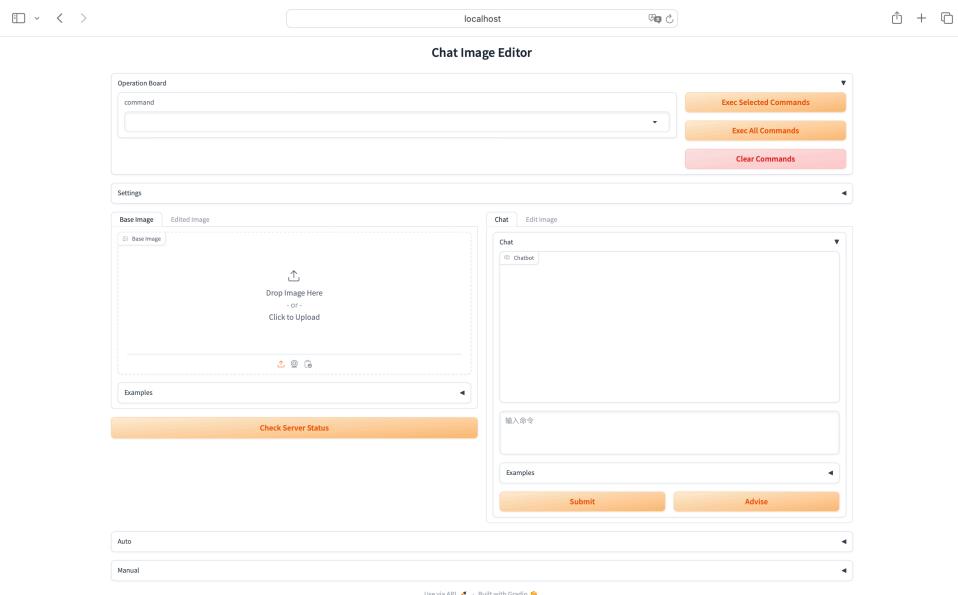


图 5-1 GUI 总体效果

可使用 GUI 的 Auto 模块进行 LLM 微调数据生成和 LLM 指令生成任务性能测试，其实现效果如图 5-2 所示。

可使用 GUI 的 Manual 模块查看 GUI 使用方法，其实现效果如图 5-3 所示。

使用 GUI 进行基于 LLM 的交互式图像编辑时，系统主要模块效果如图 5-4 所示。

5.2 系统使用方法

5.2.1 系统部署

本项目提供了 Docker、Kubernetes 等多种常用的项目部署方式以适应不同的用户需求和操作环境。如果选择通过 Docker 进行部署，用户可以根据需求选择不同的镜像。通过运行命令 `docker run --name multimodal -p 27777:80 binciluo/multimodal:latest`，可以在



图 5-2 Auto 模块

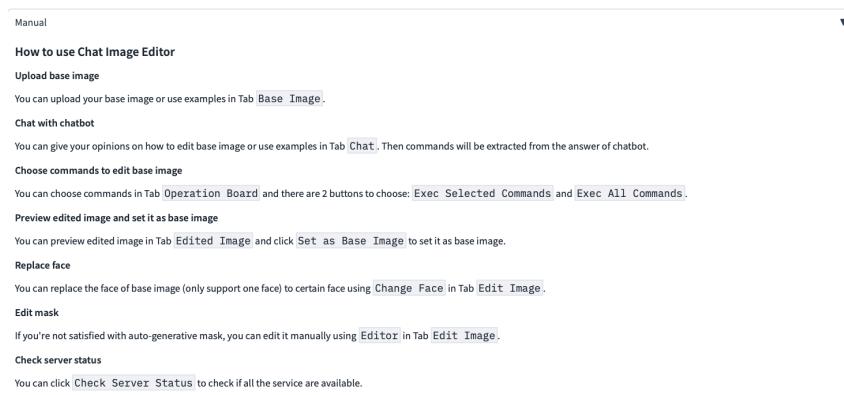


图 5-3 Manual 模块

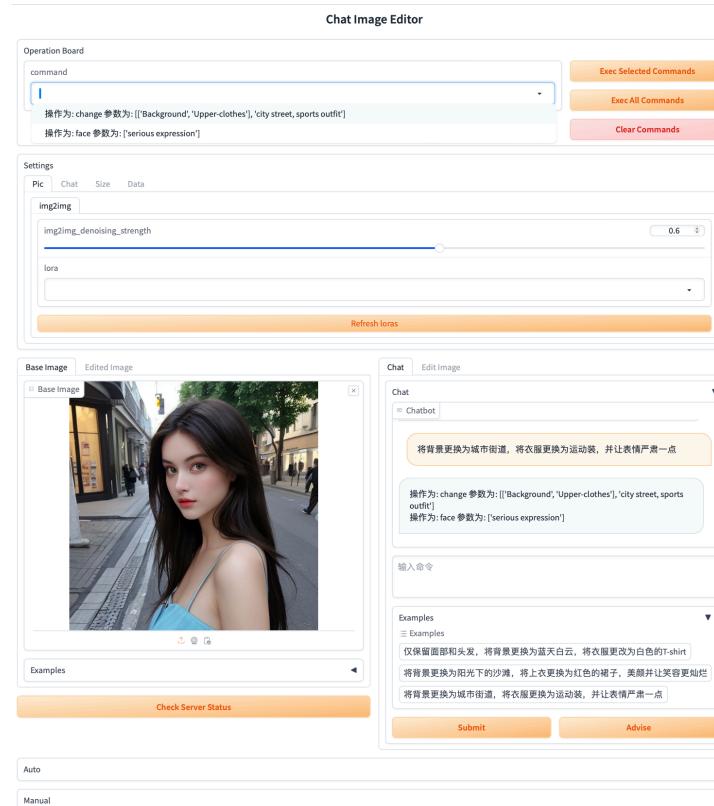


图 5-4 交互式图像编辑

本地部署图像分割模型。如果希望通过 Hugging Face API 使用图像分割功能以在性能受限的设备上运行，则可以使用命令 `docker run -name multimodal -p 27777:80 bincluo/multimodal:mini_latest`。部署完成后，用户可通过访问本地地址 `127.0.0.1:27777` 来使用服务。对于 Kubernetes 部署，用户需要先切换到包含 Kubernetes 配置文件的目录，使用 `kubectl apply -f pod.yaml` 命令部署服务（对于使用 Arm 架构的用户，则需使用 `kubectl apply -f pod_arm.yaml`）。非 Linux 用户需运行 `kubectl port-forward mm-service-pod -n default 27777:27777` 以访问服务。服务可通过 `127.0.0.1:27777` 地址访问。

项目还支持本地直接运行。用户需要先安装 Golang 和 Beego 框架，然后安装 Python 所需的依赖包。首先安装 Golang，接着通过命令 `go install github.com/beego/bee/v2@latest` 安装 Beego，然后运行 `pip install -r gradio_web/requirements.txt` 安装 Python 依赖。最终，通过运行脚本 `runner.sh` 启动服务，或者设置环境变量 `SEG_MODEL_ENV='local'` 并运行 `runner_local.sh` 脚本以在本地使用图像分割模型。

5.2.2 GUI 使用说明

1. 上传基础图像：您可以上传您的基础图像，或者使用在 Base Image 中的 Examples 提供的示例图像。图像上传后，系统会自动地进行图像分割任务并在 Chat 界面显示分割结果。
2. 与 LLM 对话您可以就如何编辑基础图像提出您的看法，或者使用在 Chat 中的 Examples 提供的示例。LLM 的回复会被识别是否含有指令，若有指令存在则会对指令进行提取并自动添加至 Operation Board 中。
3. 选择指令来修改基础图像您可以在 Operation Board 中选择单个或多个指令，并有两个按钮可供选择：Exec Selected Commands 和 Exec All Commands。Exec Selected Commands 会执行选中的指令，而 Exec All Commands 会执行所有指令。系统会对这些指令进行合并与排序等预处理，然后生成对应的参数向图像编辑模型发起请求。
4. 预览编辑后的图像并设置为基础图像您可以在 Edited Image 标签中预览编辑后的图像，并点击 Set as Base Image 将其设置为新的基础图像。
5. 替换面部您可以使用 Edit Image 标签中的 Change Face 功能，将基础图像的面部（仅支持单一面部）替换为特定面部。
6. 编辑遮罩如果您对自动生成的遮罩不满意，可以在 Edit Image 标签的 Editor 中手动编辑，然后重新运行之前的指令。
7. 检查服务器状态您可以点击 Check Server Status 来确认所有服务是否可用。系统会在大约三秒后弹出提示框显示哪些服务目前不可用。
8. 自动生成 LLM 微调数据您可在 Auto 模块中的 Gen Data 选择图片，输入 OpenAI 的 API Key 并指定生成数量和线程数以生成 LLM 微调数据。
9. 对 LLM 指令生成合格率进行测试您可在 Auto 模块中的 Test LLM 选择测试集并指定测试样本数量和线程数以对 LLM 指令生成合格率进行测试。

5.3 本章小结

在本章中，我们展示了基于 LLM 的交互式多模态图像编辑系统的实际运行效果和用户使用方法，提供了详细的用户操作指南，确保用户能够充分利用系统功能进行图像编辑。

第六章 项目管理与维护

6.1 代码管理

Git 是一个开源的分布式版本控制系统，其允许多个开发者在共同的代码基础上工作，同时能够追踪和记录所有文件的历史变更，兼具高性能与灵活性。Git 能处理从小到大的项目，让开发者能够在本地机器上工作，并保持代码的多个版本，以便在不同的分支上进行试验和开发新功能而不影响主代码库。

GitHub¹是一个基于 Git 的代码托管和协作平台，其为开发者提供了一个强大而便捷的环境来管理代码和协作，不仅能够追踪和记录代码的变更历史以确保代码的完整性和回溯能力，还能通过分支管理支持多线程的工作流以允许多个开发者同时推进不同的功能。GitHub 的 pull 请求机制促进了团队成员之间的代码审查和讨论，不仅提高了代码质量也加强了团队协作能力。GitHub 的集成系统支持持续集成和持续部署流程，与各种开发工具链的无缝连接，使得项目管理更加高效。由于其承载了众多开源项目，GitHub 成为开发者的一个展示和交流的平台，促进了知识共享和技术交流。

为了便于进行代码管理和版本控制，本项目在 GitHub 上创建了一个仓库²，结合 GitHub 的其他功能，将其发展成了功能完整、文档详细的开源项目。

6.2 自动化测试

自动化测试是一种利用软件来控制执行测试的过程，其能自动比较实际的运行结果与预期结果来验证被测软件功能是否符合预期。自动化测试的主要功能是提高测试的效率和覆盖率，其可以快速地执行大量的测试用例，并且可以反复运行这些测试，在软件开发的早期发现缺陷，从而减少修复缺陷的成本并确保软件在开发迭代中未引入错误。自动化测试还可以将测试人员从繁琐的手动测试工作中解脱出来，使他们有更多时间专注于更复杂的测试任务。在持续集成和持续部署（CI/CD）的实践中，自动化测试是不可或缺的一环，它提高了软件交付的速度和质量，是现代软件开发流程中的关键组成部分。

GitHub Actions 是 GitHub 提供的一个持续集成与持续部署（CI/CD）平台，允许用户在代码仓库中直接自动化、自定义地执行特定的软件开发工作流程相关任务。用户可定义一系列的事件和操作，当如代码推送、合并请求等指定事件发生时，GitHub Actions 会自动运行对应的构建代码、运行测试、部署到生产环境等任务。GitHub Actions 使得开发者无需离开 GitHub 环境就能自动化处理软件的构建、测试和部署过程，从而大幅提升开发的效率。它支持多种操作系统，提供了大量现成的 Actions 供用户使用，并且允许创建私有的、自定义的 Actions。作为 CI/CD 的解决方案，GitHub Actions 简化了开

¹<https://github.com>

²<https://github.com/BinciLuo/multimodal-service>

发流程，加快了从编写代码到部署产品的过程，同时还提高了软件的质量和交付的可靠性。

本项目使用 GitHub Actions 对代码中的部分模块进行自动化测试以保证代码的正确性和项目的稳定性。当有新的 pull 请求对 main 或 dev 分支进行更新时，自动化测试工作将在最新版本的 Ubuntu 运行环境上执行。其首先会使用 actions/checkout@v3 获取最新的仓库代码，利用 actions/setup-python@v3 来设置 Python 3.10 版本的 Python 环境。当设置好 Python 环境后，需要安装测试所需的依赖。在 gradio_web 目录下首先升级 pip，然后安装本项目自动化测试所需的代码检查和测试框架 flake8 和 pytest，如果存在 requirements.txt 文件，还会安装该文件中列出的依赖。最后，流程将继续在 gradio_web 目录下执行名为 test_utils.py 的测试脚本。

6.3 持续集成与持续部署

持续集成 (Continuous Integration, CI) 和持续部署 (Continuous Deployment, CD) 是现代软件开发流程中的关键部分。CI 的核心是自动化地将代码变合并到特定分支时自动运行构建和测试流程，这样可以迅速发现并解决集成错误，提高代码质量，缩短反馈周期。CD 扩展了 CI 的概念，不仅包括自动化测试，还包括自动化部署过程，确保经过测试的代码可以被自动地部署到指定环境中。这使得产品能够快速迭代，缩短从开发到产品投放市场的时间，同时减少了部署过程中的人为错误，提升了软件交付的速度和安全性。CI/CD 通过自动化的流程减少了手动工作以使开发团队更加专注于功能开发和创新。

本项目使用 GitHub Actions 进行 CI 和 CD 流程。除了测试外，本项目还有几个关键的 CI/CD 流程，主要包括部署将项目部署到 Azure Web 应用和 Docker 镜像构建。

名为“Build and deploy container app to Azure Web App - gradio-app”的 Action 在代码推送到 main 分支时手动触发，自动化了在 Azure Web App 上的部署过程。其首先构建 gradio_web 的 Docker 镜像，并将其推送到 DockerHub，随后将镜像部署到 Azure 的生产环境，从而实现高效和一致的应用发布。名为“Build and deploy container app to Azure Web App - middleware-app”的 Action 以相近的方式实现了 Middleware 在 Azure Web App 上的自动化部署。

名为“Docker Image CI”的 Action 主要用于构建和推送 Docker 镜像到 DockerHub。当代码被推送到 main 分支时，此工作流程触发并执行以下操作：使用最新的 Ubuntu 环境，首先通过 GitHub Secrets 进行 Docker 登录，然后分别从 docker/Dockerfile 和 docker/DockerfileMini 两个文件构建构建标准镜像和更小的 Mini 镜像，其区别为标准镜像使用本地的模型进行图像分割而 Mini 模型使用 HuggingFace¹的 API 进行图像分割。这些镜像将在构建完成后被标记并推送到 DockerHub 账户下，确保最新的容器镜像版本可供部署和分发。此自动化流程加快了软件的交付速度，保证了镜像的最新状态和可用性。

¹<https://huggingface.co>

名为“Docker Image CI for ARM64”的 Action 通过相同的方式实现了适用于 arm64 架构的镜像的构建与发布。

6.4 本章小结

在本章中，我们探讨了项目管理与维护的关键策略和实践，确保了基于 LLM 的交互式多模态图像编辑系统的高效开发和长期可持续性发展。本章详细介绍了项目的代码管理方式、自动化测试以及持续集成与部署的实施方法。这些管理和维护策略不仅提高了开发过程的效率，也确保了系统在实际部署后的稳定性和可靠性，为未来可能的功能扩展和技术升级打下了坚实的基础。

第七章 总结及未来展望

7.1 总结

本系统的研发始于当前图像编辑工具的局限性，这些工具往往需要用户具备专业知识和技能，且操作复杂，难以满足非专业用户的需求。为了解决这些问题，本项目通过打通最新的大语言模型和图像生成模型，开发了一个既强大又用户友好的图像编辑系统。系统的核心在于其能够通过简单的与大语言模型聊天的方式来自动化地生成图像编辑的指令从而调用图像生成模型来执行复杂的图像编辑任务。从系统架构层面，打通大语言模型和图像生成模型需要两个主要组件的协同工作：一个直观的图形用户界面（GUI）和一个功能强大的中间件（Middleware）。

图形用户界面（GUI）的设计充分考虑了易用性和高效性，使得即使是没有图像编辑经验的用户也能够轻松使用。通过各个简洁明了的模块，用户可以执行包括文本交互、图像自动遮罩、更换面部、参数设置等多种任务。GUI 既为用户提供了足够简单易用的自动化操作，也能让用户对如遮罩生成等高精度要求的操作进行手动的调整。GUI 不仅提高了操作的直观性，还通过使用中间件（middleware）访问一些对资源需求较大的模型，极大地降低了对用户设备性能的需求。

项目的中间件（Middleware）部分是整个系统的枢纽。它整合了多个不同平台、不同请求格式、不同返回格式的 API，包括图像生成模型如 Stable Diffusion、DALL-E2 和大语言模型如 ChatGLM2-6B、GPT-3.5 Turbo 和 GPT-4 Turbo。中间件（Middleware）的设计保证了这些模型的高效协同工作，支持了 GUI 所需的从文本交互到图像编辑的一系列高级功能。此外，中间件（Middleware）还处理所有后端逻辑，包括一对多且互相隔离的服务、以及用户请求的响应。

系统创新地引入了 LoRA 进行大语言模型的微调，通过训练专门的 LoRA 模型，原本在本任务下表现较差的 ChatGLM2-6B 模型能够理解复杂的用户输入并精准地将其转化为指令，然后自动对指令进行抽取并执行高质量的图像编辑任务。微调数据集的自动生成是其中的一个重要环节，系统采用了结合多个模型输出和验证规则的自动化工作流程。在已有的图像数据集下，利用现有的图像分割模型和文本生成模型，自动地生成图像编辑的要求、指令，并进行数据合格校验，生成了大量高质量的训练数据。系统设计了自动化的测试流程对各个大语言模型的指令生成能力进行评估，并对 ChatGLM2-6B 的多个微调模型和 GPT3.5 turbo、GPT4 Turbo 进行评估，以便对不同的模型的指令生成能力进行量化。从结果中可观察到 GPT3.5 turbo、GPT4 Turbo 在不进行微调的情况下就已经具有很强的能力，而参数量较小、初始能力极弱的 ChatGLM2-6B 模型在微调后在指令生成上也能达到接近 GPT3.5 turbo 的能力。

7.2 未来展望

随着最近几年 GPU 和 TPU 等专用硬件的发展，深度学习的训练和推理速度得到了极大的提升，深度学习在图像和语音识别、自然语言处理和自动驾驶等许多领域都取得了显著进展。在文本生成和图像生成任务上，深度学习技术已从理论探索逐步过渡到实际应用，图像生成技术已经实现了从简单的图像生成到复杂的场景重构的跨越，在这一领域，生成对抗网络（GANs）和最新的扩散模型都已经能够生成高质量的图像内容。而 OpenAI 的 GPT 系列和清华大学的 ChatGLM2-6B 等大语言模型也已经能够理解并生成复杂的自然语言文本。在这些新技术的加持下，通过一定方法对大语言模型和图像生成模型进行链接，从而使基于 LLM 的交互式图像编辑系统成为可能。本项目尝试了基于特定指令链接大语言模型和图像生成模型的方法，并搭建了基础框架且得到了优异的效果。

大语言模型和图像生成技术的结合带来了许多机遇，但也存在不少技术和伦理方面的挑战。如何确保生成的内容的准确性和适当性，防止生成有偏见或不当信息的风险，数据隐私和安全成为重点关注的问题，必须妥善处理，确保内容合规性和用户信息受到保护的同时不妨碍技术的有效应用。

随着深度学习技术的继续发展，未来的交互式技术必将更加智能和高效。大语言模型和图像生成技术的结合不仅能提升用户在图像编辑领域的体验，还将开启全新的应用，为创新和改进现有服务提供强大动力。

参考文献

- [1] Zhan Fangneng, Yu Yingchen, Wu Rongliang et al. Multimodal image synthesis and editing: A survey [J]. arXiv preprint arXiv:2112.13592. 2022.
- [2] Esser Patrick, Rombach Robin, Blattmann Andreas et al. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis [J]. Advances in Neural Information Processing Systems. 34. 2021: 3518–3532.
- [3] Lewis Mike, Liu Yinhuan, Goyal Naman et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [J]. arXiv preprint arXiv:1910.13461. 2019.
- [4] Ling Huan, Kreis Karsten, Li Daiqing et al. Editgan: High-precision semantic image editing [J]. Advances in Neural Information Processing Systems. 34. 2021: 16331–16345.
- [5] Mansimov Elman, Parisotto Emilio, Ba Jimmy Lei et al. Generating images from captions with attention [J]. arXiv preprint arXiv:1511.02793. 2015.
- [6] Barrett William A, Cheney Alan S. Object-based image editing [C]. In 29th Annual Conference on Computer Graphics and Interactive Techniques. 2002 : 777–784.
- [7] Portenier Tiziano, Hu Qiyang, Szabo Attila et al. Faceshop: Deep sketch-based face image editing [J]. arXiv preprint arXiv:1804.08972. 2018.
- [8] Oh Byong Mok, Chen Max, Dorsey Julie et al. Image-based modeling and photo editing [C]. In 28th Annual Conference on Computer Graphics and Interactive Techniques. 2001 : 433–442.
- [9] Perarnau Guim, Van De Weijer Joost, Raducanu Bogdan et al. Invertible conditional gans for image editing [J]. arXiv preprint arXiv:1611.06355. 2016.
- [10] Zhu Jiapeng, Shen Yujun, Zhao Deli et al. In-domain gan inversion for real image editing [C]. In European Conference on Computer Vision. 2020 : 592–608.
- [11] Di Martino J Matías, Facciolo Gabriele, Meinhardt-Llopis Enric. Poisson image editing [J]. Image Processing On Line. 6. 2016: 300–325.
- [12] Elder James H, Goldberg Richard M. Image editing in the contour domain [C]. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1998 : 374–381.
- [13] Farbman Zeev, Fattal Raanan, Lischinski Dani. Diffusion maps for edge-aware image editing [J]. ACM Transactions on Graphics (TOG). 29 (6). 2010: 1–10.
- [14] Ding Ning, Qin Yujia, Yang Guang et al. Parameter-efficient fine-tuning of large-scale pre-trained language models [J]. Nature Machine Intelligence. 5 (3). 2023: 220–235.
- [15] Radiya-Dixit Evani, Wang Xin. How fine can fine-tuning be? learning efficient language models [C]. In International Conference on Artificial Intelligence and Statistics. 2020 : 2435–2443.
- [16] Lu Keming, Yuan Hongyi, Yuan Zheng et al. # InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models [C]. In The Twelfth International Conference on Learning Representations. 2023 .
- [17] Hilmkil Agrin, Callh Sebastian, Barbieri Matteo et al. Scaling federated learning for fine-tuning of large language models [C]. In International Conference on Applications of Natural Language to Information

- Systems. 2021 : 15–23.
- [18] Chen Yukang, Qian Shengju, Tang Haotian et al. Longlora: Efficient fine-tuning of long-context large language models [J]. arXiv preprint arXiv:2309.12307. 2023.
- [19] Zhang Haojie, Li Ge, Li Jia et al. Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively [J]. Advances in Neural Information Processing Systems. 35. 2022: 21442–21454.
- [20] Hu Zhiqiang, Wang Lei, Lan Yihuai et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models [J]. arXiv preprint arXiv:2304.01933. 2023.
- [21] Brown Tom, Mann Benjamin, Ryder Nick et al. Language models are few-shot learners [J]. Advances in Neural Information Processing Systems. 33. 2020: 1877–1901.
- [22] Malladi Sadhika, Gao Tianyu, Nichani Eshaan et al. Fine-tuning language models with just forward passes [J]. Advances in Neural Information Processing Systems. 36. 2024.
- [23] Laput Gierad P, Dontcheva Mira, Wilensky Gregg et al. Pixeltone: A multimodal interface for image editing [C]. In SIGCHI Conference on Human Factors in Computing Systems. 2013 : 2185–2194.
- [24] Kim Gwanghyun, Kwon Taesung, Ye Jong Chul. Diffusionclip: Text-guided diffusion models for robust image manipulation [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 : 2426–2435.
- [25] Radford Alec, Kim Jong Wook, Hallacy Chris et al. Learning transferable visual models from natural language supervision [C]. In International Conference on Machine Learning. 2021 : 8748–8763.
- [26] Chen Jianbo, Shen Yelong, Gao Jianfeng et al. Language-based image editing with recurrent attentive models [C]. In IEEE Conference on Computer Vision and Pattern Recognition. 2018 : 8721–8729.
- [27] Kawar Bahjat, Zada Shiran, Lang Oran et al. Imagic: Text-based real image editing with diffusion models [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 : 6007–6017.
- [28] Cheng Yu, Gan Zhe, Li Yitong et al. Sequential attention GAN for interactive image editing [C]. In 28th ACM International Conference on Multimedia. 2020 : 4383–4391.
- [29] Ghodrati Amir, Jia Xu, Pedersoli Marco et al. Towards automatic image editing: Learning to see another you [J]. arXiv preprint arXiv:1511.08446. 2015.
- [30] Brooks Tim, Holynski Aleksander, Efros Alexei A. Instructpix2pix: Learning to follow image editing instructions [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 : 18392–18402.
- [31] Avrahami Omri, Lischinski Dani, Fried Ohad. Blended diffusion for text-driven editing of natural images [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 : 18208–18218.
- [32] Zhou Yufan, Zhang Ruiyi, Gu Jiuxiang et al. Tigan: Text-based interactive image generation and manipulation [C]. In AAAI Conference on Artificial Intelligence. 2022 : 3580–3588.
- [33] Lee Yao-Chih, Jang Ji-Ze Genevieve, Chen Yi-Ting et al. Shape-aware text-driven layered video editing [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023 : 14317–14326.
- [34] Dong Wenkai, Xue Song, Duan Xiaoyue et al. Prompt tuning inversion for text-driven image editing using diffusion models [C]. In IEEE/CVF International Conference on Computer Vision. 2023 : 7430–7440.
- [35] Li Bo, Lin Xiao, Liu Bin et al. Lightweight text-driven image editing with disentangled content and

- attributes [J]. IEEE Transactions on Multimedia. 2023.
- [36] Rombach Robin, Blattmann Andreas, Lorenz Dominik et al. High-resolution image synthesis with latent diffusion models [C]. In IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 : 10684–10695.
- [37] Zhang Lvmin, Rao Anyi, Agrawala Maneesh. Adding conditional control to text-to-image diffusion models [C]. In IEEE/CVF International Conference on Computer Vision. 2023 : 3836–3847.
- [38] Vaswani Ashish, Shazeer Noam, Parmar Niki et al. Attention is all you need [J]. Advances in Neural Information Processing Systems. 30. 2017.
- [39] Achiam Josh, Adler Steven, Agarwal Sandhini et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774. 2023.
- [40] Hu Edward J, Shen Yelong, Wallis Phillip et al. Lora: Low-rank adaptation of large language models [J]. arXiv preprint arXiv:2106.09685. 2021.
- [41] Van Huynh Nguyen, Hoang Dinh Thai, Nguyen Diep N et al. DeepFake: Deep dueling-based deception strategy to defeat reactive jammers [J]. IEEE Transactions on Wireless Communications. 20 (10). 2021: 6898–6914.

致 谢

我对李佩佩老师在本论文的研究和写作过程中的悉心指导和无私帮助深感感激，她严谨的学术态度和深厚的专业知识不断激励着我。同时，感谢所有参与讨论和提供宝贵意见的同学和朋友们，感谢北京邮电大学提供的研究平台和资源，这篇论文的完成离不开每一位给予我支持和帮助的人。

附录

代码附-1 指令校验规则

```
1  {
2      "face": {
3          "paras_type": [
4              "<class\u00b7'str'\u00b7"
5          ],
6          "paras_enum": null,
7          "paras_min": null,
8          "paras_max": null,
9          "combine": true,
10         "priority": 1
11     },
12     "change": {
13         "paras_type": [
14             "<class\u00b7'list'\u00b7",
15             "<class\u00b7'str'\u00b7"
16         ],
17         "paras_enum": null,
18         "paras_min": null,
19         "paras_max": null,
20         "combine": true,
21         "priority": 1
22     }
23 }
```

外 文 资 料

Adding Conditional Control to Text-to-Image Diffusion Models

L vmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University

{l vmin, anyirao, maneesh}@cs.stanford.edu



Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), etc., to control the image generation of large pretrained diffusion models. The default results use the prompt “a high-quality, detailed, and professional image”. Users can optionally give prompts like the “chef in kitchen”.

Abstract

We present *ControlNet*, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. *ControlNet* locks the production-ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional controls. The neural architecture is connected with “zero convolutions” (zero-initialized convolution layers) that progressively grow the parameters from zero and ensure that no harmful noise could affect the finetuning. We test various conditioning controls, e.g., edges, depth, segmentation, human pose, etc., with *Stable Diffusion*, using single or multiple conditions, with or without prompts. We show that the training of *ControlNets* is robust with small ($<50k$) and large ($>1m$) datasets. Extensive results show that *ControlNet* may facilitate wider applications to control image diffusion models.

1. Introduction

Many of us have experienced flashes of visual inspiration that we wish to capture in a unique image. With the advent of text-to-image diffusion models [54, 62, 72], we can now create visually stunning images by typing in a text prompt. Yet, text-to-image models are limited in the control they provide over the spatial composition of the image; precisely expressing complex layouts, poses, shapes and forms can be difficult via text prompts alone. Generating an image that accurately matches our mental imagery often requires numerous trial-and-error cycles of editing a prompt, inspecting the resulting images and then re-editing the prompt.

Can we enable finer grained spatial control by letting users provide additional images that directly specify their desired image composition? In computer vision and machine learning, these additional images (e.g., edge maps, human pose skeletons, segmentation maps, depth, normals, etc.) are often treated as conditioning on the image generation process. Image-to-image translation models [34, 98] learn

the mapping from conditioning images to target images. The research community has also taken steps to control text-to-image models with spatial masks [6, 20], image editing instructions [10], personalization via finetuning [21, 75], *etc.* While a few problems (*e.g.*, generating image variations, inpainting) can be resolved with training-free techniques like constraining the denoising diffusion process or editing attention layer activations, a wider variety of problems like depth-to-image, pose-to-image, *etc.*, require end-to-end learning and data-driven solutions.

Learning conditional controls for large text-to-image diffusion models in an end-to-end way is challenging. The amount of training data for a specific condition may be significantly smaller than the data available for general text-to-image training. For instance, the largest datasets for various specific problems (*e.g.*, object shape/normal, human pose extraction, *etc.*) are usually about 100K in size, which is 50,000 times smaller than the LAION-5B [79] dataset that was used to train Stable Diffusion [82]. The direct finetuning or continued training of a large pretrained model with limited data may cause overfitting and catastrophic forgetting [31, 75]. Researchers have shown that such forgetting can be alleviated by restricting the number or rank of trainable parameters [14, 25, 31, 92]. For our problem, designing deeper or more customized neural architectures might be necessary for handling in-the-wild conditioning images with complex shapes and diverse high-level semantics.

This paper presents ControlNet, an end-to-end neural network architecture that learns conditional controls for large pretrained text-to-image diffusion models (Stable Diffusion in our implementation). ControlNet preserves the quality and capabilities of the large model by locking its parameters, and also making a *trainable copy* of its encoding layers. This architecture treats the large pretrained model as a strong backbone for learning diverse conditional controls. The trainable copy and the original, locked model are connected with *zero convolution* layers, with weights initialized to zeros so that they progressively grow during the training. This architecture ensures that harmful noise is not added to the deep features of the large diffusion model at the beginning of training, and protects the large-scale pretrained backbone in the trainable copy from being damaged by such noise.

Our experiments show that ControlNet can control Stable Diffusion with various conditioning inputs, including Canny edges, Hough lines, user scribbles, human key points, segmentation maps, shape normals, depths, *etc.* (Figure 1). We test our approach using a single conditioning image, with or without text prompts, and we demonstrate how our approach supports the composition of multiple conditions. Additionally, we report that the training of ControlNet is robust and scalable on datasets of different sizes, and that for some tasks like depth-to-image conditioning, training ControlNets on a single NVIDIA RTX 3090Ti GPU can achieve

results competitive with industrial models trained on large computation clusters. Finally, we conduct ablative studies to investigate the contribution of each component of our model, and compare our models to several strong conditional image generation baselines with user studies.

In summary, (1) we propose ControlNet, a neural network architecture that can add spatially localized input conditions to a pretrained text-to-image diffusion model via efficient finetuning, (2) we present pretrained ControlNets to control Stable Diffusion, conditioned on Canny edges, Hough lines, user scribbles, human key points, segmentation maps, shape normals, depths, and cartoon line drawings, and (3) we validate the method with ablative experiments comparing to several alternative architectures, and conduct user studies focused on several previous baselines across different tasks.

2. Related Work

2.1. Finetuning Neural Networks

One way to finetune a neural network is to directly continue training it with the additional training data. But this approach can lead to overfitting, mode collapse, and catastrophic forgetting. Extensive research has focused on developing finetuning strategies that avoid such issues.

HyperNetwork is an approach that originated in the Natural Language Processing (NLP) community [25], with the aim of training a small recurrent neural network to influence the weights of a larger one. It has been applied to image generation with generative adversarial networks (GANs) [4, 18]. Heathen *et al.* [26] and Kurumuz [43] implement HyperNetworks for Stable Diffusion [72] to change the artistic style of its output images.

Adapter methods are widely used in NLP for customizing a pretrained transformer model to other tasks by embedding new module layers into it [30, 84]. In computer vision, adapters are used for incremental learning [74] and domain adaptation [70]. This technique is often used with CLIP [66] for transferring pretrained backbone models to different tasks [23, 66, 85, 94]. More recently, adapters have yielded successful results in vision transformers [49, 50] and ViT-Adapter [14]. In concurrent work with ours, T2I-Adapter [56] adapts Stable Diffusion to external conditions.

Additive Learning circumvents forgetting by freezing the original model weights and adding a small number of new parameters using learned weight masks [51, 74], pruning [52], or hard attention [80]. Side-Tuning [92] uses a side branch model to learn extra functionality by linearly blending the outputs of a frozen model and an added network, with a predefined blending weight schedule.

Low-Rank Adaptation (LoRA) prevents catastrophic forgetting [31] by learning the offset of parameters with low-rank matrices, based on the observation that many over-

parameterized models reside in a low intrinsic dimension subspace [2, 47].

Zero-Initialized Layers are used by ControlNet for connecting network blocks. Research on neural networks has extensively discussed the initialization and manipulation of network weights [36, 37, 44, 45, 46, 76, 83, 95]. For example, Gaussian initialization of weights can be less risky than initializing with zeros [1]. More recently, Nichol *et al.* [59] discussed how to scale the initial weight of convolution layers in a diffusion model to improve the training, and their implementation of “zero_module” is an extreme case to scale weights to zero. Stability’s model cards [83] also mention the use of zero weights in neural layers. Manipulating the initial convolution weights is also discussed in ProGAN [36], StyleGAN [37], and Noise2Noise [46].

2.2. Image Diffusion

Image Diffusion Models were first introduced by Sohl-Dickstein *et al.* [81] and have been recently applied to image generation [17, 42]. The Latent Diffusion Models (LDM) [72] performs the diffusion steps in the latent image space [19], which reduces the computation cost. Text-to-image diffusion models achieve state-of-the-art image generation results by encoding text inputs into latent vectors via pretrained language models like CLIP [66]. Glide [58] is a text-guided diffusion model supporting image generation and editing. Disco Diffusion [5] processes text prompts with clip guidance. Stable Diffusion [82] is a large-scale implementation of latent diffusion [72]. Imagen [78] directly diffuses pixels using a pyramid structure without using latent images. Commercial products include DALL-E2 [62] and Midjourney [54].

Controlling Image Diffusion Models facilitate personalization, customization, or task-specific image generation. The image diffusion process directly provides some control over color variation [53] and inpainting [67, 7]. Text-guided control methods focus on adjusting prompts, manipulating CLIP features, and modifying cross-attention [7, 10, 20, 27, 40, 41, 58, 64, 67]. MakeAScene [20] encodes segmentation masks into tokens to control image generation. SpaText [6] maps segmentation masks into localized token embeddings. GLIGEN [48] learns new parameters in attention layers of diffusion models for grounded generating. Textual Inversion [21] and DreamBooth [75] can personalize content in the generated image by finetuning the image diffusion model using a small set of user-provided example images. Prompt-based image editing [10, 33, 86] provides practical tools to manipulate images with prompts. Voynov *et al.* [88] propose an optimization method that fits the diffusion process with sketches. Concurrent works [8, 9, 32, 56] examine a wide variety of ways to control diffusion models.

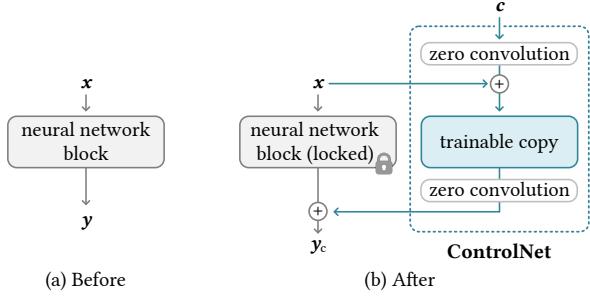


Figure 2: A neural block takes a feature map x as input and outputs another feature map y , as shown in (a). To add a ControlNet to such a block we lock the original block and create a trainable copy and connect them together using zero convolution layers, *i.e.*, 1×1 convolution with both weight and bias initialized to zero. Here c is a conditioning vector that we wish to add to the network, as shown in (b).

2.3. Image-to-Image Translation

Conditional GANs [15, 34, 63, 90, 93, 97, 98, 99] and transformers [13, 19, 68] can learn the mapping between different image domains, *e.g.*, Taming Transformer [19] is a vision transformer approach; Palette [77] is a conditional diffusion model trained from scratch; PITI [89] is a pretraining-based conditional diffusion model for image-to-image translation. Manipulating pretrained GANs can handle specific image-to-image tasks, *e.g.*, StyleGANs can be controlled by extra encoders [71], with more applications studied in [3, 22, 38, 39, 55, 60, 65, 71].

3. Method

ControlNet is a neural network architecture that can enhance large pretrained text-to-image diffusion models with spatially localized, task-specific image conditions. We first introduce the basic structure of a ControlNet in Section 3.1 and then describe how we apply a ControlNet to the image diffusion model Stable Diffusion [72] in Section 3.2. We elaborate on our training in Section 3.3 and detail several extra considerations during inference such as composing multiple ControlNets in Section 3.4.

3.1. ControlNet

ControlNet injects additional conditions into the blocks of a neural network (Figure 2). Herein, we use the term *network block* to refer to a set of neural layers that are commonly put together to form a single unit of a neural network, *e.g.*, resnet block, conv-bn-relu block, multi-head attention block, transformer block, *etc.*. Suppose $\mathcal{F}(\cdot; \Theta)$ is such a trained neural block, with parameters Θ , that transforms an input feature map x , into another feature map y as

$$y = \mathcal{F}(x; \Theta). \quad (1)$$

In our setting, x and y are usually 2D feature maps, i.e., $x \in \mathbb{R}^{h \times w \times c}$ with $\{h, w, c\}$ as the height, width, and number of channels in the map, respectively (Figure 2a).

To add a ControlNet to such a pre-trained neural block, we lock (freeze) the parameters Θ of the original block and simultaneously clone the block to a *trainable copy* with parameters Θ_c (Figure 2b). The trainable copy takes an external conditioning vector c as input. When this structure is applied to large models like Stable Diffusion, the locked parameters preserve the production-ready model trained with billions of images, while the trainable copy reuses such large-scale pretrained model to establish a deep, robust, and strong backbone for handling diverse input conditions.

The trainable copy is connected to the locked model with *zero convolution* layers, denoted $\mathcal{Z}(\cdot; \cdot)$. Specifically, $\mathcal{Z}(\cdot; \cdot)$ is a 1×1 convolution layer with both weight and bias initialized to zeros. To build up a ControlNet, we use two instances of zero convolutions with parameters Θ_{z1} and Θ_{z2} respectively. The complete ControlNet then computes

$$y_c = \mathcal{F}(x; \Theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c; \Theta_{z1}); \Theta_c); \Theta_{z2}), \quad (2)$$

where y_c is the output of the ControlNet block. In the first training step, since both the weight and bias parameters of a zero convolution layer are initialized to zero, both of the $\mathcal{Z}(\cdot; \cdot)$ terms in Equation (2) evaluate to zero, and

$$y_c = y. \quad (3)$$

In this way, harmful noise cannot influence the hidden states of the neural network layers in the trainable copy when the training starts. Moreover, since $\mathcal{Z}(c; \Theta_{z1}) = 0$ and the trainable copy also receives the input image x , the trainable copy is fully functional and retains the capabilities of the large, pretrained model allowing it to serve as a strong backbone for further learning. Zero convolutions protect this backbone by eliminating random noise as gradients in the initial training steps. We detail the gradient calculation for zero convolutions in supplementary materials.

3.2. ControlNet for Text-to-Image Diffusion

We use Stable Diffusion [72] as an example to show how ControlNet can add conditional control to a large pretrained diffusion model. Stable Diffusion is essentially a U-Net [73] with an encoder, a middle block, and a skip-connected decoder. Both the encoder and decoder contain 12 blocks, and the full model contains 25 blocks, including the middle block. Of the 25 blocks, 8 blocks are down-sampling or up-sampling convolution layers, while the other 17 blocks are main blocks that each contain 4 resnet layers and 2 Vision Transformers (ViTs). Each ViT contains several cross-attention and self-attention mechanisms. For example, in Figure 3a, the “SD Encoder Block A” contains 4 resnet layers and 2 ViTs, while the “ $\times 3$ ” indicates that this block is repeated three times. Text prompts are encoded using the

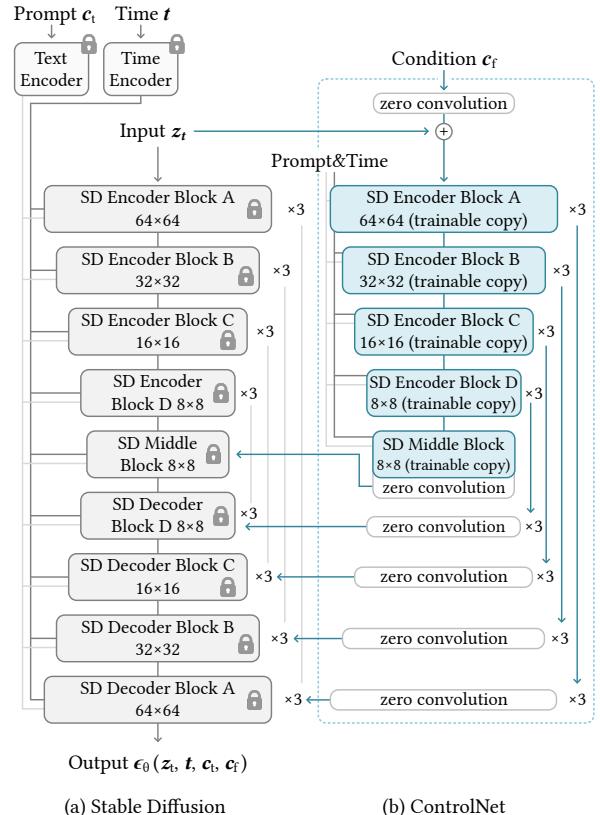


Figure 3: Stable Diffusion’s U-net architecture connected with a ControlNet on the encoder blocks and middle block. The locked, gray blocks show the structure of Stable Diffusion V1.5 (or V2.1, as they use the same U-net architecture). The trainable blue blocks and the white zero convolution layers are added to build a ControlNet.

CLIP text encoder [66], and diffusion timesteps are encoded with a time encoder using positional encoding.

The ControlNet structure is applied to each encoder level of the U-net (Figure 3b). In particular, we use ControlNet to create a trainable copy of the 12 encoding blocks and 1 middle block of Stable Diffusion. The 12 encoding blocks are in 4 resolutions (64×64 , 32×32 , 16×16 , 8×8) with each one replicated 3 times. The outputs are added to the 12 skip-connections and 1 middle block of the U-net. Since Stable Diffusion is a typical U-net structure, this ControlNet architecture is likely to be applicable with other models.

The way we connect the ControlNet is computationally efficient — since the locked copy parameters are frozen, no gradient computation is required in the originally locked encoder for the finetuning. This approach speeds up training and saves GPU memory. As tested on a single NVIDIA A100 PCIE 40GB, optimizing Stable Diffusion with ControlNet requires only about 23% more GPU memory and 34%

more time in each training iteration, compared to optimizing Stable Diffusion without ControlNet.

Image diffusion models learn to progressively denoise images and generate samples from the training domain. The denoising process can occur in pixel space or in a *latent* space encoded from training data. Stable Diffusion uses latent images as the training domain as working in this space has been shown to stabilize the training process [72]. Specifically, Stable Diffusion uses a pre-processing method similar to VQ-GAN [19] to convert 512×512 pixel-space images into smaller 64×64 *latent images*. To add ControlNet to Stable Diffusion, we first convert each input conditioning image (*e.g.*, edge, pose, depth, *etc.*) from an input size of 512×512 into a 64×64 feature space vector that matches the size of Stable Diffusion. In particular, we use a tiny network $\mathcal{E}(\cdot)$ of four convolution layers with 4×4 kernels and 2×2 strides (activated by ReLU, using 16, 32, 64, 128, channels respectively, initialized with Gaussian weights and trained jointly with the full model) to encode an image-space condition c_i into a feature space conditioning vector c_f as,

$$c_f = \mathcal{E}(c_i). \quad (4)$$

The conditioning vector c_f is passed into the ControlNet.

3.3. Training

Given an input image z_0 , image diffusion algorithms progressively add noise to the image and produce a noisy image z_t , where t represents the number of times noise is added. Given a set of conditions including time step t , text prompts c_t , as well as a task-specific condition c_f , image diffusion algorithms learn a network ϵ_θ to predict the noise added to the noisy image z_t with

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0, 1)} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2 \right], \quad (5)$$

where \mathcal{L} is the overall learning objective of the entire diffusion model. This learning objective is directly used in finetuning diffusion models with ControlNet.

In the training process, we randomly replace 50% text prompts c_t with empty strings. This approach increases ControlNet’s ability to directly recognize semantics in the input conditioning images (*e.g.*, edges, poses, depth, *etc.*) as a replacement for the prompt.

During the training process, since zero convolutions do not add noise to the network, the model should always be able to predict high-quality images. We observe that the model does not gradually learn the control conditions but abruptly succeeds in following the input conditioning image; usually in less than 10K optimization steps. As shown in Figure 4, we call this the “sudden convergence phenomenon”.

3.4. Inference

We can further control how the extra conditions of ControlNet affect the denoising diffusion process in several ways.

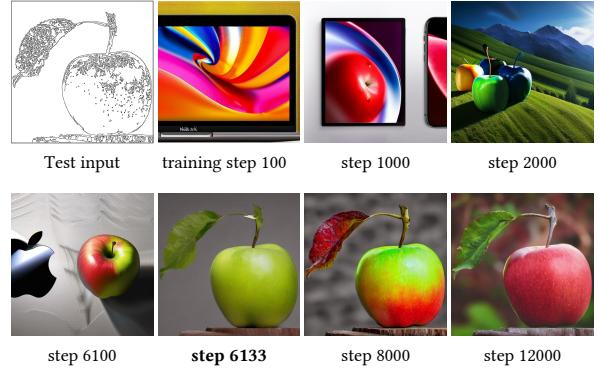


Figure 4: The sudden convergence phenomenon. Due to the zero convolutions, ControlNet always predicts high-quality images during the entire training. At a certain step in the training process (*e.g.*, the 6133 steps marked in bold), the model suddenly learns to follow the input condition.



Figure 5: Effect of Classifier-Free Guidance (CFG) and the proposed CFG Resolution Weighting (CFG-RW).

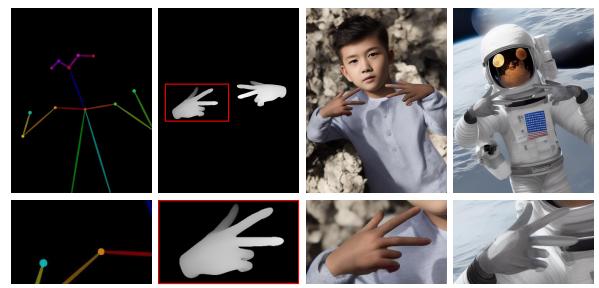


Figure 6: Composition of multiple conditions. We present the application to use depth and pose simultaneously.

Classifier-free guidance resolution weighting. Stable Diffusion depends on a technique called Classifier-Free Guidance (CFG) [29] to generate high-quality images. CFG is formulated as $\epsilon_{\text{prd}} = \epsilon_{\text{uc}} + \beta_{\text{cfg}}(\epsilon_c - \epsilon_{\text{uc}})$ where ϵ_{prd} , ϵ_{uc} , ϵ_c , β_{cfg} are the model’s final output, unconditional output, conditional output, and a user-specified weight respectively. When a conditioning image is added via ControlNet, it can be added to both ϵ_{uc} and ϵ_c , or only to the ϵ_c . In challenging cases, *e.g.*, when no prompts are given, adding it to both ϵ_{uc} and ϵ_c will completely remove CFG guidance (Figure 5b); using only ϵ_c will make the guidance very strong (Figure 5c). Our solution is to first add the conditioning image to ϵ_c and

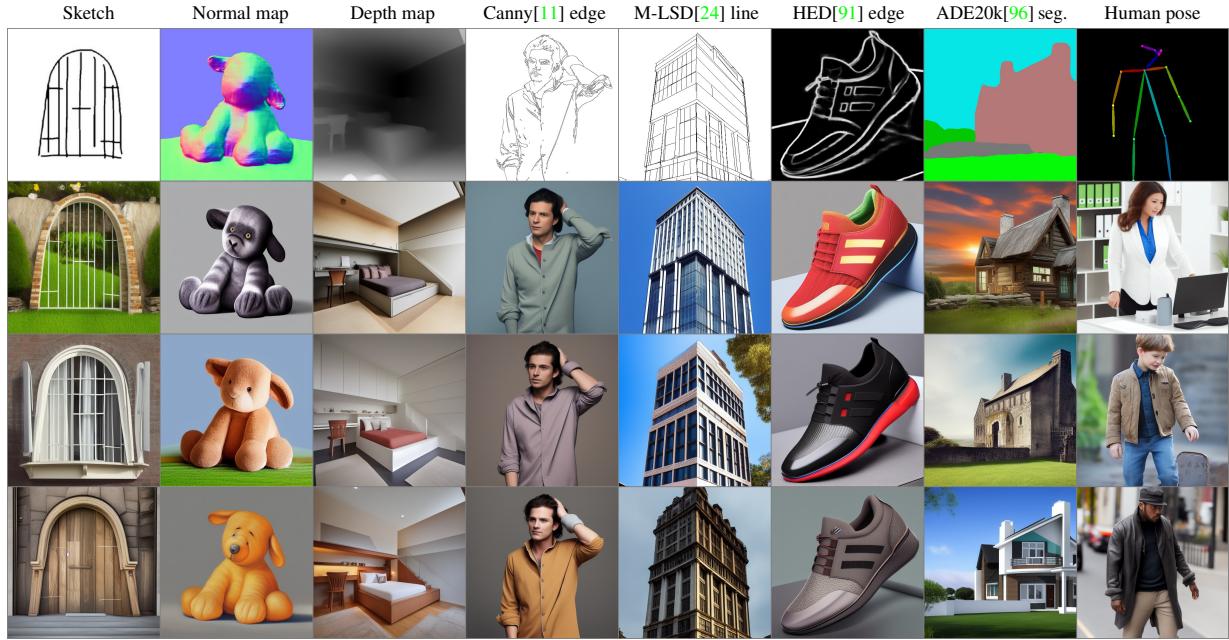


Figure 7: Controlling Stable Diffusion with various conditions **without prompts**. The top row is input conditions, while all other rows are outputs. We use the empty string as input prompts. All models are trained with general-domain data. The model has to recognize semantic contents in the input condition images to generate images.

Method	Result Quality \uparrow	Condition Fidelity \uparrow
PITI [89](sketch)	1.10 ± 0.05	1.02 ± 0.01
Sketch-Guided [88] ($\beta = 1.6$)	3.21 ± 0.62	2.31 ± 0.57
Sketch-Guided [88] ($\beta = 3.2$)	2.52 ± 0.44	3.28 ± 0.72
ControlNet-lite	3.93 ± 0.59	4.09 ± 0.46
ControlNet	4.22 ± 0.43	4.28 ± 0.45

Table 1: Average User Ranking (AUR) of result quality and condition fidelity. We report the user preference ranking (1 to 5 indicates worst to best) of different methods.

then multiply a weight w_i to each connection between Stable Diffusion and ControlNet according to the resolution of each block $w_i = 64/h_i$, where h_i is the size of i^{th} block, *e.g.*, $h_1 = 8, h_2 = 16, \dots, h_{13} = 64$. By reducing the CFG guidance strength , we can achieve the result shown in Figure 5d, and we call this CFG Resolution Weighting.

Composing multiple ControlNets. To apply multiple conditioning images (*e.g.*, Canny edges, and pose) to a single instance of Stable Diffusion, we can directly add the outputs of the corresponding ControlNets to the Stable Diffusion model (Figure 6). No extra weighting or linear interpolation is necessary for such composition.

4. Experiments

We implement ControlNets with Stable Diffusion to test various conditions, including Canny Edge [11], Depth

Map [69], Normal Map [87], M-LSD lines [24], HED soft edge [91], ADE20K segmentation [96], Openpose [12], and user sketches. See also the supplementary material for examples of each conditioning along with detailed training and inference parameters.

4.1. Qualitative Results

Figure 1 shows the generated images in several prompt settings. Figure 7 shows our results with various conditions without prompts, where the ControlNet robustly interprets content semantics in diverse input conditioning images.

4.2. Ablative Study

We study alternative structures of ControlNets by (1) replacing the zero convolutions with standard convolution layers initialized with Gaussian weights, and (2) replacing each block’s trainable copy with one single convolution layer, which we call ControlNet-lite. See also the supplementary material for the full details of these ablative structures.

We present 4 prompt settings to test with possible behaviors of real-world users: (1) no prompt; (2) insufficient prompts that do not fully cover objects in conditioning images, *e.g.*, the default prompt of this paper “a high-quality, detailed, and professional image”; (3) conflicting prompts that change the semantics of conditioning images; (4) perfect prompts that describe necessary content semantics, *e.g.*, “a nice house”. Figure 8a shows that ControlNet succeeds in

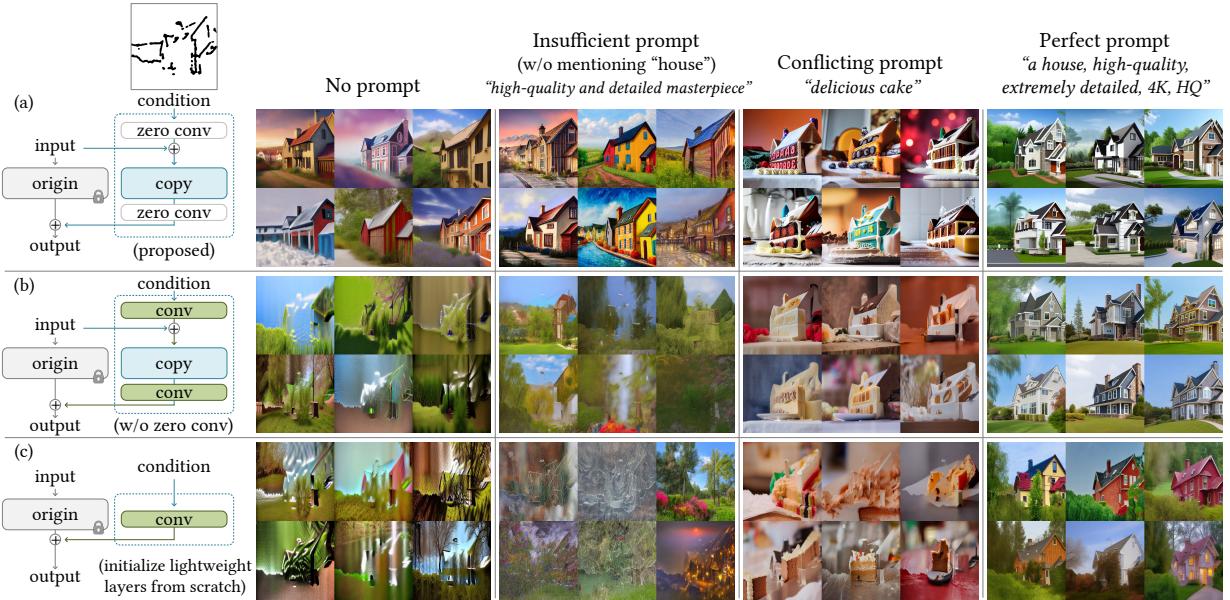


Figure 8: Ablative study of different architectures on a sketch condition and different prompt settings. For each setting, we show a random batch of 6 samples without cherry-picking. Images are at 512×512 and best viewed when zoomed in. The green “conv” blocks on the left are standard convolution layers initialized with Gaussian weights.

ADE20K (GT)	VQGAN [19]	LDM [72]	PITI [89]	ControlNet-lite	ControlNet
0.58 ± 0.10	0.21 ± 0.15	0.31 ± 0.09	0.26 ± 0.16	0.32 ± 0.12	0.35 ± 0.14

Table 2: Evaluation of semantic segmentation label reconstruction (ADE20K) with Intersection over Union (IoU \uparrow).

all 4 settings. The lightweight ControlNet-lite (Figure 8c) is not strong enough to interpret the conditioning images and fails in the insufficient and no prompt conditions. When zero convolutions are replaced, the performance of ControlNet drops to about the same as ControlNet-lite, indicating that the pretrained backbone of the trainable copy is destroyed during finetuning (Figure 8b).

4.3. Quantitative Evaluation

User study. We sample 20 unseen hand-drawn sketches, and then assign each sketch to 5 methods: PITI [89]’s sketch model, Sketch-Guided Diffusion (SGD) [88] with default edge-guidance scale ($\beta = 1.6$), SGD [88] with relatively high edge-guidance scale ($\beta = 3.2$), the aforementioned ControlNet-lite, and ControlNet. We invited 12 users to rank these 20 groups of 5 results individually in terms of “*the quality of displayed images*” and “*the fidelity to the sketch*”. In this way, we obtain 100 rankings for result quality and 100 for condition fidelity. We use the Average Human Ranking (AHR) as a preference metric where users rank each result on a scale of 1 to 5 (lower is worse). The average rankings are shown in Table 1.

Comparison to industrial models. Stable Diffusion V2 Depth-to-Image (SDv2-D2I) [83] is trained with a large-

Method	FID \downarrow	CLIP-score \uparrow	CLIP-aes. \uparrow
Stable Diffusion	6.09	0.26	6.32
VQGAN [19](seg.)*	26.28	0.17	5.14
LDM [72](seg.)*	25.35	0.18	5.15
PITI [89](seg.)	19.74	0.20	5.77
ControlNet-lite	17.92	0.26	6.30
ControlNet	15.27	0.26	6.31

Table 3: Evaluation for image generation conditioned by semantic segmentation. We report FID, CLIP text-image score, and CLIP aesthetic scores for our method and other baselines. We also report the performance of Stable Diffusion without segmentation conditions. Methods marked with “*” are trained from scratch.

scale NVIDIA A100 cluster, thousands of GPU hours, and more than 12M training images. We train a ControlNet for the SD V2 with the same depth conditioning but only use 200k training samples, one single NVIDIA RTX 3090Ti, and 5 days of training. We use 100 images generated by each SDv2-D2I and ControlNet to teach 12 users to distinguish the two methods. Afterwards, we generate 200 images and ask the users to tell which model generated each image. The average precision of the users is 0.52 ± 0.17 , indicating that the two method yields almost indistinguishable results.

Condition reconstruction and FID score. We use the test set of ADE20K [96] to evaluate the conditioning fidelity. The state-of-the-art segmentation method OneFormer [35] achieves an Intersection-over-Union (IoU) with 0.58 on the ground-truth set. We use different methods to generate images with ADE20K segmentations and then apply One-

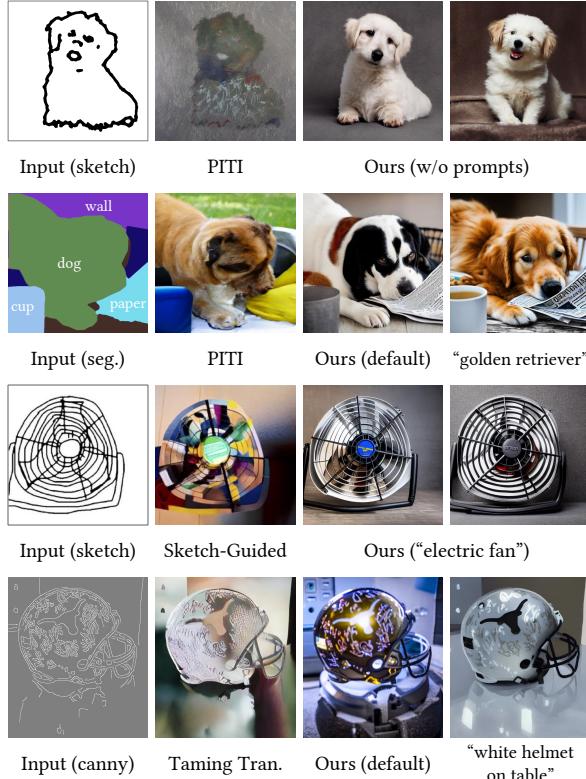


Figure 9: Comparison to previous methods. We present the qualitative comparisons to PITI [89], Sketch-Guided Diffusion [88], and Taming Transformers [19].

Former to detect the segmentations again to compute the reconstructed IoUs (Table 2). Besides, we use Frechet Inception Distance (FID) [28] to measure the distribution distance over randomly generated 512×512 image sets using different segmentation-conditioned methods, as well as text-image CLIP scores [66] and CLIP aesthetic score [79] in Table 3. See also the supplementary material for detailed settings.

4.4. Comparison to Previous Methods

Figure 9 presents a visual comparison of baselines and our method (Stable Diffusion + ControlNet). Specifically, we show the results of PITI [89], Sketch-Guided Diffusion [88], and Taming Transformers [19]. (Note that the backbone of PITI is OpenAI GLIDE [57] that have different visual quality and performance.) We observe that ControlNet can robustly handle diverse conditioning images and achieves sharp and clean results.

4.5. Discussion

Influence of training dataset sizes. We demonstrate the robustness of the ControlNet training in Figure 10. The training does not collapse with limited 1k images, and allows

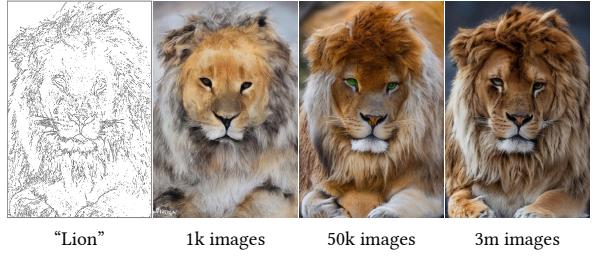


Figure 10: The influence of different training dataset sizes. See also the supplementary material for extended examples.



Figure 11: Interpreting contents. If the input is ambiguous and the user does not mention object contents in prompts, the results look like the model tries to interpret input shapes.



Figure 12: Transfer pretrained ControlNets to community models [16, 61] without training the neural networks again.

the model to generate a recognizable lion. The learning is scalable when more data is provided.

Capability to interpret contents. We showcase ControlNet’s capability to capture the semantics from input conditioning images in Figure 11.

Transferring to community models. Since ControlNets do not change the network topology of pretrained SD models, it can be directly applied to various models in the stable diffusion community, such as Comic Diffusion [61] and Protogen 3.4 [16], in Figure 12.

5. Conclusion

ControlNet is a neural network structure that learns conditional control for large pretrained text-to-image diffusion models. It reuses the large-scale pretrained layers of source models to build a deep and strong encoder to learn specific conditions. The original model and trainable copy are connected via “zero convolution” layers that eliminate harmful noise during training. Extensive experiments verify that ControlNet can effectively control Stable Diffusion with single or multiple conditions, with or without prompts. Results on diverse conditioning datasets show that the ControlNet struc-

ture is likely to be applicable to a wider range of conditions, and facilitate relevant applications.

Acknowledgment

This work was partially supported by the Stanford Institute for Human-Centered AI and the Brown Institute for Media Innovation.

References

- [1] Sadia Afrin. Weight initialization in neural network, inspired by andrew ng, <https://medium.com/@safrin1128/weight-initialization-in-neural-network-inspired-by-andrew-ng-e0066dc4a566>, 2020. ³
- [2] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 7319–7328, Online, Aug. 2021. Association for Computational Linguistics. ³
- [3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40(4), 2021. ³
- [4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. ²
- [5] Alembics. Disco diffusion, <https://github.com/alembics/disco-diffusion>, 2022. ³
- [6] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. *arXiv preprint arXiv:2211.14305*, 2022. ^{2, 3}
- [7] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. ³
- [8] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. ³
- [9] Dina Bashkirova, Jose Lezama, Kihyuk Sohn, Kate Saenko, and Irfan Essa. Masksketch: Unpaired structure-guided masked image generation. *arXiv preprint arXiv:2302.05496*, 2023. ³
- [10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. ^{2, 3}
- [11] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986. ⁶
- [12] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. ⁶
- [13] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. ³
- [14] Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *International Conference on Learning Representations*, 2023. ²
- [15] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. ³
- [16] darkstorm2150. Progen x3.4 (photorealism) official release, <https://civitai.com/models/3666/progen-x34-photorealism-official-release>, 2022. ⁸
- [17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. ³
- [18] Tan M. Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. ²
- [19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. ^{3, 5, 7, 8}
- [20] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022. ^{2, 3}
- [21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. ^{2, 3}
- [22] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. ³
- [23] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. ²
- [24] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. ⁶
- [25] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. ²

- [26] Hethen. Hypernetwork style training, a tiny guide, stable-diffusion-webui, <https://github.com/automatic1111/stable-diffusion-webui/discussions/2670>, 2022. 2
- [27] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3
- [28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 8
- [29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 5
- [30] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799, 2019. 2
- [31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [32] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and controllable image synthesis with composable conditions. 2023. 3
- [33] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023. 3
- [34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 3
- [35] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. 2023. 7
- [36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018. 3
- [37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3
- [38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis*, 2021. 3
- [39] Oren Katzir, Vicky Perepelok, Dani Lischinski, and Daniel Cohen-Or. Multi-level latent space structuring for generative control. *arXiv preprint arXiv:2202.05910*, 2022. 3
- [40] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3
- [41] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3
- [42] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34:21696–21707, 2021. 3
- [43] Kurumuz. Novelai improvements on stable diffusion, <https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac>, 2022. 2
- [44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. 3
- [45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3
- [46] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *Proceedings of the 35th International Conference on Machine Learning*, 2018. 3
- [47] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *International Conference on Learning Representations*, 2018. 3
- [48] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. 2023. 3
- [49] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 2
- [50] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 2
- [51] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *European Conference on Computer Vision (ECCV)*, pages 67–82, 2018. 2
- [52] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018. 2
- [53] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3
- [54] Midjourney. <https://www.midjourney.com/>, 2023. 1, 3
- [55] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3
- [56] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhonggang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning

- adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 3
- [57] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, 2021. 8
- [58] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. 2022. 3
- [59] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3
- [60] Yotam Nitzan, Kfir Aberman, Qirui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022. 3
- [61] ogkalu. Comic-diffusion v2, trained on 6 styles at once, <https://huggingface.co/ogkalu/comic-diffusion>, 2022. 8
- [62] OpenAI. Dall-e-2, <https://openai.com/product/dall-e-2>, 2023. 1, 3
- [63] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3
- [64] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 3
- [65] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 3
- [66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 8
- [67] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3
- [68] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3
- [69] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020. 6
- [70] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018. 2
- [71] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3
- [72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 5, 7
- [73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention MICCAI International Conference*, pages 234–241, 2015. 4
- [74] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):651–663, 2018. 2
- [75] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3
- [76] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986. 3
- [77] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings, SIGGRAPH '22*, New York, NY, USA, 2022. Association for Computing Machinery. 3
- [78] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamvar Seyed Ghasempour, Burcu Karagol Ayan, Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3
- [79] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 8
- [80] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018. 2
- [81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using

- nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3
- [82] Stability. Stable diffusion v1.5 model card, <https://huggingface.co/runwayml/stable-diffusion-v1-5>, 2022. 2, 3
- [83] Stability. Stable diffusion v2 model card, stable-diffusion-2-depth, <https://huggingface.co/stabilityai/stable-diffusion-2-depth>, 2022. 3, 7
- [84] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995, 2019. 2
- [85] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. *arXiv preprint arXiv:2112.06825*, 2021. 2
- [86] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 3
- [87] Igor Vasiljevic, Nick Kolkkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 6
- [88] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. 2022. 3, 6, 7, 8
- [89] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. 2022. 3, 6, 7, 8
- [90] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3
- [91] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1403, 2015. 6
- [92] Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas J. Guibas, and Jitendra Malik. Side-tuning: Network adaptation via additive side networks. In *European Conference on Computer Vision (ECCV)*, pages 698–714. Springer, 2020. 2
- [93] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 3
- [94] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kun-chang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2
- [95] Jiawei Zhao, Florian Schäfer, and Anima Anandkumar. Zero initialization: Initializing residual networks with only zeros and ones. *arXiv*, 2021. 3
- [96] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 6, 7
- [97] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 3
- [98] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 3
- [99] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in Neural Information Processing Systems*, 30, 2017. 3

外文译文

向文本到图像扩散模型添加条件控制

Lvmin Zhang, Anyi Rao, Maneesh Agrawala

斯坦福大学

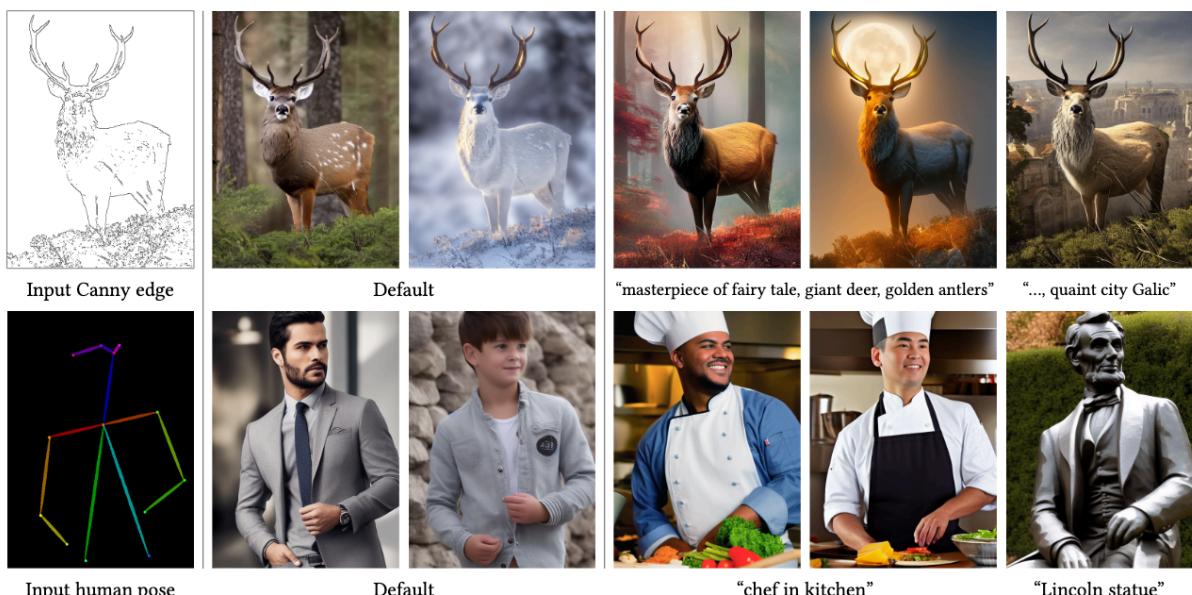


Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), etc., to control the image generation of large pretrained diffusion models. The default results use the prompt “a high-quality, detailed, and professional image”. Users can optionally give prompts like the “chef in kitchen”.

图外 0-1 第一章 引言

我们许多人都有过想要捕捉视觉灵感的时刻，希望将其呈现为独特的图像。随着文本到图像扩散模型的出现，我们现在可以通过输入文本提示创建视觉上令人惊叹的图像。然而，这些文本到图像模型在图像空间组成控制方面存在局限性；仅通过文本提示来精确表达复杂的布局、姿势、形状和形式可能很困难。生成准确匹配我们心中图像的图像通常需要多次的试验和错误，通过编辑提示、检查生成的图像并重新编辑提示来完成。

我们能否通过让用户提供附加图像来直接指定他们期望的图像组成，从而实现更精细的空间控制？在计算机视觉和机器学习中，这些附加图像（例如边缘图、人类姿势骨架、分割图、深度、法线等）通常被视为图像生成过程中的条件。图像到图像翻译模型学习从条件图像到目标图像的映射。研究社区也采取了一些步骤，通过空间掩码、图像编辑指令、个性化微调等控制文本到图像模型。虽然一些问题（例如生成图像变化、修补）可以通过无训练技术解决，如限制去噪扩散过程或编辑注意层激活，但更多问题（如深度到图像、姿势到图像）需要端到端学习和数据驱动的解决方案。

在端到端方式中学习大型文本到图像扩散模型的条件控制是具有挑战性的。特定条件的训练数据量可能显著小于用于一般文本到图像训练的数据。例如，用于各种特定问题（如对象形状/法线、人体姿势提取等）的最大数据集通常约为 10 万，这比用于训练稳定扩散的 LAION-5B 数据集小 5 万倍。直接微调或继续训练大型预训练模型可能导致过拟合和灾难性遗忘。研究人员已经表明，通过限制可训练参数的数量或排名，可以减轻这种遗忘。对于我们的问题，设计更深或更定制化的神经架构可能是处理具有复杂形状和多样化高级语义的自然条件图像的必要条件。

本文提出了 ControlNet，这是一种端到端神经网络架构，可以为大型预训练的文本到图像扩散模型（在我们的实现中为稳定扩散）学习条件控制。ControlNet 通过锁定其参数并同时制作其编码层的可训练副本保留大型模型的质量和能力。这种架构将大型预训练模型视为学习多样化条件控制的强大骨干。可训练副本和原始锁定模型通过零卷积层连接，权重初始化为零，使其在训练过程中逐步增长。这种架构确保在训练开始时不会向大型扩散模型的深层特征添加有害噪声，并保护可训练副本中的大规模预训练骨干不受此类噪声的损害。

我们的实验表明，ControlNet 可以通过各种条件输入（包括 Canny 边缘、Hough 线、用户涂鸦、人类关键点、分割图、形状法线、深度等）控制稳定扩散。我们测试了单个条件图像的使用，有或没有文本提示，并演示了我们的方法如何支持多个条件的组合。此外，我们报告了 ControlNet 在不同大小数据集上的训练是稳健和可扩展的，并且在某些任务（如深度到图像条件）中，在单个 NVIDIA RTX 3090Ti GPU 上的训练结果可以与在大型计算集群上训练的工业模型竞争。最后，我们进行了消融研究，研究了我们模型每个组件的贡献，并通过用户研究将我们的模型与几个强大的条件图像生成基准进行了比较。

总结来说，(1) 我们提出了 ControlNet，这是一种神经网络架构，可以通过高效微调向预训练的文本到图像扩散模型添加空间定位的输入条件；(2) 我们提出了预训练的 ControlNet，用于控制稳定扩散，条件包括 Canny 边缘、Hough 线、用户涂鸦、人类关键点、分割图、形状法线、深度和卡通线条图；(3) 通过消融实验与几种替代架构进行比较，以及针对不同任务的几个先前基准的用户研究验证了该方法。

第二章 相关工作

2.1 微调神经网络

微调神经网络的一种方法是直接使用附加的训练数据继续训练它。但是这种方法可能导致过拟合、模式崩溃和灾难性遗忘。广泛的研究集中在开发避免这些问题的微调策略上。

HyperNetwork 是一种起源于自然语言处理 (NLP) 领域的方法，旨在训练一个小型递归神经网络来影响一个更大网络的权重。它已经应用于生成对抗网络 (GANs) 中的图像生成。Heathen 等人和 Kurumuz 实现了 Stable Diffusion 的 HyperNetworks，用于改变其输出图像的艺术风格。

Adapter 方法广泛应用于 NLP，用于通过嵌入新模块层将预训练的 Transformer 模型自定义为其他任务。在计算机视觉中，适配器用于增量学习和域适应。这种技术通常与 CLIP 一起使用，以将预训练的骨干模型转移到不同任务中。最近，适配器在视觉 Transformer 中取得了成功的结果。与我们同时进行的工作中，T2I-Adapter 将 Stable Diffusion 适应于外部条件。

加法学习通过冻结原始模型权重并使用学习的权重掩码添加少量新参数来避免遗忘。Side-Tuning 使用一个侧枝模型通过线性混合冻结模型和添加网络的输出来学习额外的功能，具有预定义的混合权重计划。

低秩适配通过使用低秩矩阵学习参数的偏移量来防止灾难性遗忘，基于观察到许多过参数化模型位于低内在维度子空间中。

零初始化层被 ControlNet 用于连接网络块。神经网络的研究广泛讨论了网络权重的初始化和操作。例如，高斯初始化权重比零初始化风险更小。最近，Nichol 等人讨论了在扩散模型中缩放卷积层初始权重以改善训练的方式，他们的“零模块”实现是将权重缩放为零的极端情况。Stability 的模型卡片也提到了在神经层中使用零权重。在 ProGAN、StyleGAN 和 Noise2Noise 中也讨论了初始卷积权重的操作。

2.2 图像扩散

图像扩散模型首次由 Sohl-Dickstein 等人引入，最近已应用于图像生成。潜在扩散模型 (LDM) 在潜在图像空间中执行扩散步骤，从而降低计算成本。文本到图像扩散模型通过预训练的语言模型（如 CLIP）将文本输入编码为潜在向量，达到最先进的图像生成结果。Glide 是一个支持图像生成和编辑的文本引导扩散模型。Disco Diffusion 使用 clip 指导处理文本提示。Stable Diffusion 是潜在扩散的大规模实现。Imagen 直接使用金字塔结构扩散像素而不使用潜在图像。商业产品包括 DALL-E2 和 Midjourney。

控制图像扩散模型促进个性化、自定义或任务特定的图像生成。图像扩散过程直接提供对颜色变化和修补的一些控制。文本引导的控制方法集中于调整提示、操作 CLIP 特征和修改交叉注意力。MakeAScene 将分割掩码编码为标记以控制图像生成。SpaText

将分割掩码映射到本地化的标记嵌入。GLIGEN 在扩散模型的注意力层中学习新参数以进行定位生成。Textual Inversion 和 DreamBooth 可以通过微调图像扩散模型并使用一小组用户提供的示例图像来个性化生成的内容。基于提示的图像编辑提供了实际的工具来操作提示生成的图像。Voynov 等人提出了一种优化方法，通过草图拟合扩散过程。同时进行的工作研究了控制扩散模型的各种方法。

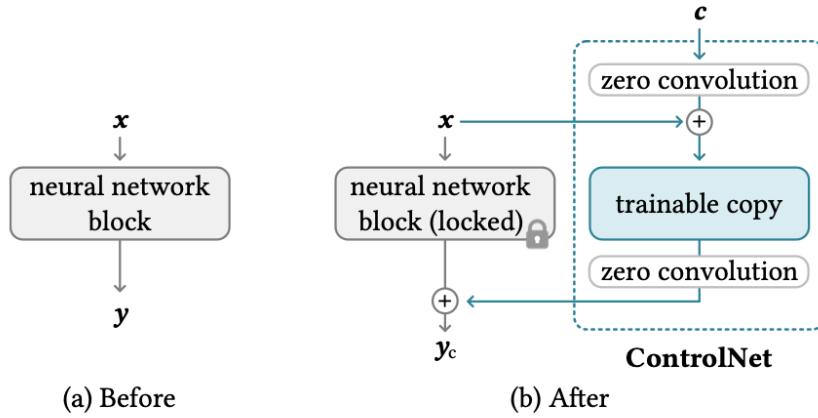


Figure 2: A neural block takes a feature map x as input and outputs another feature map y , as shown in (a). To add a ControlNet to such a block we lock the original block and create a trainable copy and connect them together using zero convolution layers, *i.e.*, 1×1 convolution with both weight and bias initialized to zero. Here c is a conditioning vector that we wish to add to the network, as shown in (b).

图 外 2-1

2.3 图像到图像转换

Conditional GANs 和 Transformers 可以学习不同图像域之间的映射。例如，Taming Transformer 是一种视觉 TRANSFORMER 方法；Palette 是一种从头开始训练的条件扩散模型；PITI 是一种基于预训练的条件扩散模型，用于图像到图像的翻译。操作预训练的 GAN 可以处理特定的图像到图像任务，例如，StyleGANs 可以通过额外的编码器进行控制，还有更多的应用研究在以下文献中：[3, 22, 38, 39, 55, 60, 65, 71]。

第三章 方法

ControlNet 是一种神经网络架构，可以通过空间定位的任务特定图像条件增强大型预训练的文本到图像扩散模型。我们首先在第 3.1 节介绍 ControlNet 的基本结构，然后在第 3.2 节描述我们如何将 ControlNet 应用于图像扩散模型 Stable Diffusion。我们在第 3.3 节详细说明我们的训练，并在第 3.4 节详细介绍推理期间的几个额外考虑，例如组合多个 ControlNet。

3.1 ControlNet

ControlNet 将附加条件注入神经网络的块中（如图外 2-1 所示）。在此，我们使用术语“网络块”指代一组常见组合在一起形成神经网络单元的神经层，例如 resnet 块、conv-bn-relu 块、多头注意力块、变压器块等。假设 $F(\cdot; \Theta)$ 是这样一个训练好的神经块，其参数为 Θ ，它将输入特征图 x 转换为另一个特征图 y ，如下所示：

$$y = F(x; \Theta)$$

在我们的设置中， x 和 y 通常是 2D 特征图，即 $x \in \mathbb{R}^{h \times w \times c}$ ，其中 h, w, c 分别表示图的高度、宽度和通道数（如图外 2-1a 所示）。

为了向这样的预训练神经块添加 ControlNet，我们冻结原始块的参数 Θ ，并同时克隆该块以创建一个具有参数 Θ_c 的可训练副本（如图外 2-1b 所示）。可训练副本接受一个外部条件向量 c 作为输入。当这种结构应用于大型模型（如 Stable Diffusion）时，锁定的参数保留了通过数十亿图像训练的生产就绪模型，而可训练副本重用这样的大规模预训练模型，以建立处理多样化输入条件的深度、稳健且强大的骨干。

可训练副本通过零卷积层与锁定模型连接，记为 $Z(\cdot; \cdot)$ 。具体来说， $Z(\cdot; \cdot)$ 是一个 1×1 卷积层，权重和偏置初始化为零。为了构建 ControlNet，我们使用两个零卷积实例，参数分别为 Θ_{z1} 和 Θ_{z2} 。完整的 ControlNet 计算如下：

$$y_c = F(x; \Theta) + Z(F(x + Z(c; \Theta_{z1}); \Theta_c); \Theta_{z2})$$

其中 y_c 是 ControlNet 块的输出。在第一次训练步骤中，由于零卷积层的权重和偏置参数初始化为零，方程 (2) 中的两个 $Z(\cdot; \cdot)$ 项都等于零，因此：

$$y_c = y$$

这样，在训练开始时，有害噪声无法影响可训练副本中神经网络层的隐藏状态。此外，由于 $Z(c; \Theta_{z1}) = 0$ 且可训练副本也接收输入图像 x ，可训练副本是完全功能性的，并保留了大型预训练模型的能力，使其能够作为进一步学习的强大骨干。零卷积通过在初始训练步骤中消除梯度中的随机噪声来保护这个骨干。我们在补充材料中详细介绍了

零卷积的梯度计算。

3.2 用于文本到图像扩散的 ControlNet

我们使用 Stable Diffusion 作为示例，展示如何将 ControlNet 添加到大型预训练扩散模型中。Stable Diffusion 本质上是一个 U-Net 结构，包含一个编码器、中间块和一个带跳过连接的解码器。编码器和解码器各包含 12 个块，整个模型包含 25 个块，包括中间块。在 25 个块中，有 8 个是下采样或上采样卷积层，其他 17 个是主要块，每个包含 4 个 resnet 层和 2 个视觉 Transformer (ViTs)。每个 ViT 包含几个交叉注意力和自注意力机制。例如，在图外 3-1a 中，“SD 编码器块 A”包含 4 个 resnet 层和 2 个 ViT，而“×3”表示该块重复三次。文本提示使用 CLIP 文本编码器进行编码，扩散时间步使用位置编码的时间编码器进行编码。

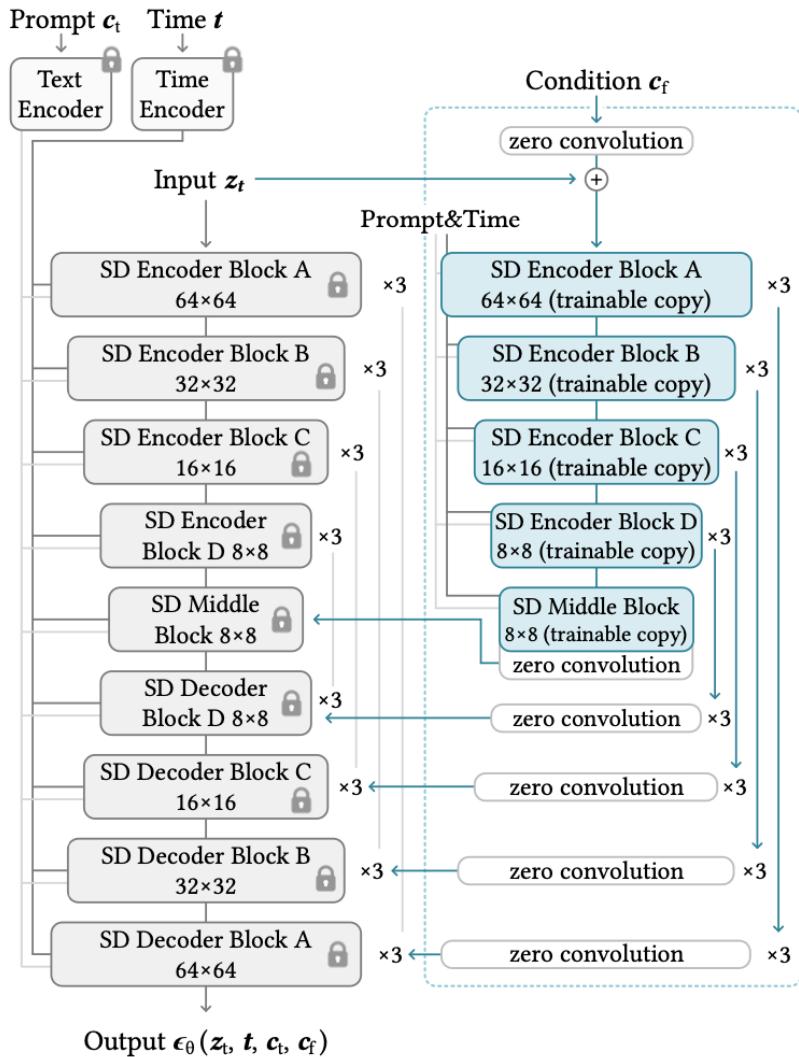
ControlNet 结构应用于 U-net 的每个编码器级别（如图外 3-1b 所示）。特别是，我们使用 ControlNet 创建 Stable Diffusion 的 12 个编码块和 1 个中间块的可训练副本。12 个编码块分为 4 种分辨率 ($64 \times 64, 32 \times 32, 16 \times 16, 8 \times 8$)，每个分辨率重复 3 次。输出被添加到 U-net 的 12 个跳过连接和 1 个中间块中。由于 Stable Diffusion 是典型的 U-net 结构，这种 ControlNet 架构可能适用于其他模型。

我们连接 ControlNet 的方式在计算上是高效的一——由于锁定的副本参数被冻结，微调时无需计算原始锁定编码器的梯度。这种方法加快了训练速度，节省了 GPU 内存。经测试，在单个 NVIDIA A100 PCIE 40GB 上，使用 ControlNet 优化 Stable Diffusion 比不使用 ControlNet 优化 Stable Diffusion 在每次训练迭代中仅多消耗约 23

图像扩散模型学习逐步去噪图像并生成来自训练域的样本。去噪过程可以在像素空间或从训练数据编码的潜在空间中进行。Stable Diffusion 使用潜在图像作为训练域，因为在该空间中工作已被证明可以稳定训练过程。具体来说，Stable Diffusion 使用类似于 VQ-GAN 的预处理方法，将 512×512 像素空间图像转换为较小的 64×64 潜在图像。为了将 ControlNet 添加到 Stable Diffusion，我们首先将每个输入条件图像（例如边缘、姿势、深度等）从输入大小 512×512 转换为与 Stable Diffusion 大小匹配的 64×64 特征空间向量。特别是，我们使用一个小型网络 $E(\cdot)$ ，由四个卷积层组成，卷积核大小为 4×4 ，步幅为 2×2 （使用 ReLU 激活，分别使用 16、32、64、128 通道，高斯权重初始化，并与完整模型一起训练），将图像空间条件 ci 编码为特征空间条件向量 cf ，如下所示：

$$cf = E(ci)$$

然后将条件向量 cf 传递到 ControlNet 中。



(a) Stable Diffusion

(b) ControlNet

Figure 3: Stable Diffusion’s U-net architecture connected with a ControlNet on the encoder blocks and middle block. The locked, gray blocks show the structure of Stable Diffusion V1.5 (or V2.1, as they use the same U-net architecture). The trainable blue blocks and the white zero convolution layers are added to build a ControlNet.

图 外 3-1

3.3 训练

给定输入图像 z_0 , 图像扩散算法逐渐向图像添加噪声并生成一个噪声图像 z_t , 其中 t 表示添加噪声的次数。给定一组条件, 包括时间步 t 、文本提示 c_t 以及任务特定条件 c_f , 图像扩散算法学习一个网络 ϵ_θ 来预测添加到噪声图像 z_t 的噪声, 公式如下:

$$L = E_{z_0, t, c_t, c_f, \epsilon \sim N(0, 1)} [\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2]$$

其中 L 是整个扩散模型的总体学习目标。这个学习目标直接用于使用 ControlNet 进行微调扩散模型。

在训练过程中, 我们随机将 50% 的文本提示 c_t 替换为空字符串。这种方法提高了 ControlNet 直接识别输入条件图像 (例如边缘、姿势、深度等) 中语义的能力, 以替代提示。

在训练过程中, 由于零卷积不会向网络添加噪声, 模型应该始终能够预测高质量的图像。我们观察到, 模型不是逐步学习控制条件, 而是突然成功地遵循输入条件图像; 通常在不到 10,000 次优化步骤内。正如图外 3-2 所示, 我们称之为“突然收敛现象”。

3.4 推理

我们可以通过几种方式进一步控制 ControlNet 的额外条件如何影响去噪扩散过程。

无分类器引导分辨率加权。Stable Diffusion 依赖于一种称为无分类器引导 (Classifier-Free Guidance, CFG) [29] 的技术来生成高质量图像。CFG 的公式为:

$$\epsilon_{prd} = \epsilon_{uc} + \beta_{cfg}(\epsilon_c - \epsilon_{uc})$$

其中 ϵ_{prd} 、 ϵ_{uc} 、 ϵ_c 、 β_{cfg} 分别是模型的最终输出、无条件输出、条件输出和用户指定的权重。当通过 ControlNet 添加条件图像时, 它可以同时添加到 ϵ_{uc} 和 ϵ_c , 也可以只添加到 ϵ_c 。在具有挑战性的情况下, 例如没有提示时, 同时添加到 ϵ_{uc} 和 ϵ_c 会完全移除 CFG 引导 (图外 3-3b); 仅使用 ϵ_c 会使引导非常强 (图外 3-3c)。我们的解决方案是首先将条件图像添加到 ϵ_c , 然后根据每个块的分辨率将一个权重 w_i 乘以 Stable Diffusion 和 ControlNet 之间的每个连接, 其中 $w_i = 64/h_i$, h_i 是第 i 个块的大小, 例如 $h_1 = 8$ 、 $h_2 = 16$ 、...、 $h_{13} = 64$ 。通过降低 CFG 引导强度, 我们可以实现图外 3-3d 所示的结果, 我们称之为 CFG 分辨率加权。

组合多个 ControlNet。为了将多个条件图像 (例如 Canny 边缘和姿势) 应用于一个 Stable Diffusion 实例, 我们可以直接将相应 ControlNet 的输出添加到 Stable Diffusion 模型中 (图外 3-4)。对于这样的组合, 不需要额外的加权或线性插值。

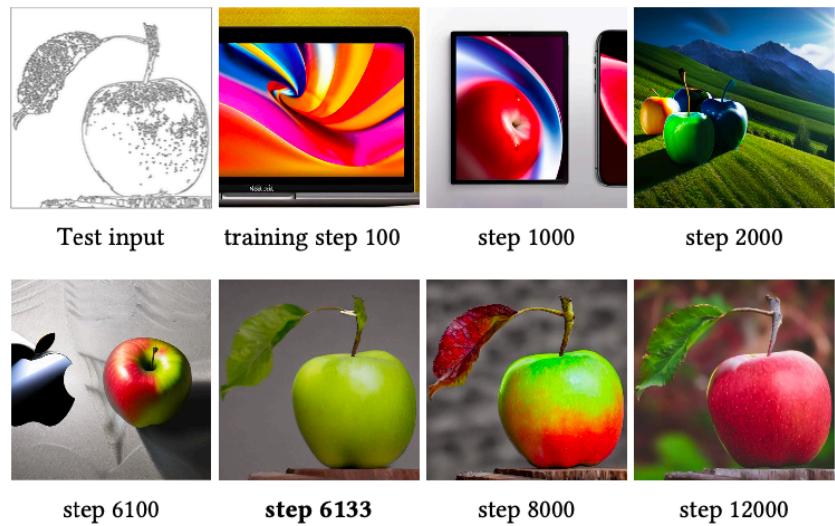


Figure 4: The sudden convergence phenomenon. Due to the zero convolutions, ControlNet always predicts high-quality images during the entire training. At a certain step in the training process (*e.g.*, the 6133 steps marked in bold), the model suddenly learns to follow the input condition.

图 外 3-2



Figure 5: Effect of Classifier-Free Guidance (CFG) and the proposed CFG Resolution Weighting (CFG-RW).

图 外 3-3

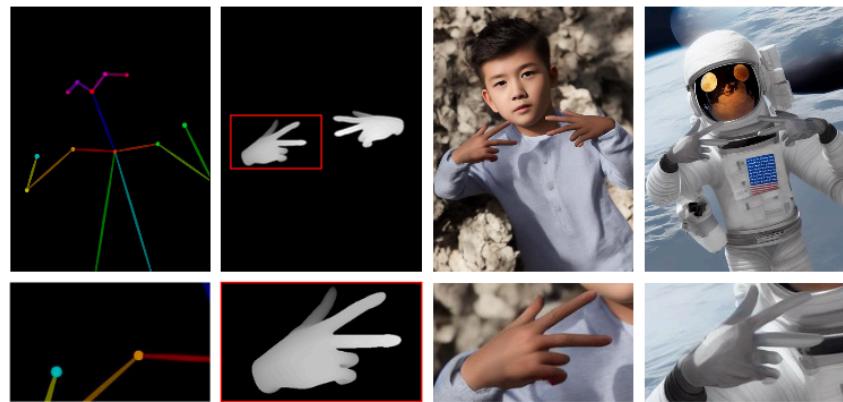


Figure 6: Composition of multiple conditions. We present the application to use depth and pose simultaneously.

图 外 3-4

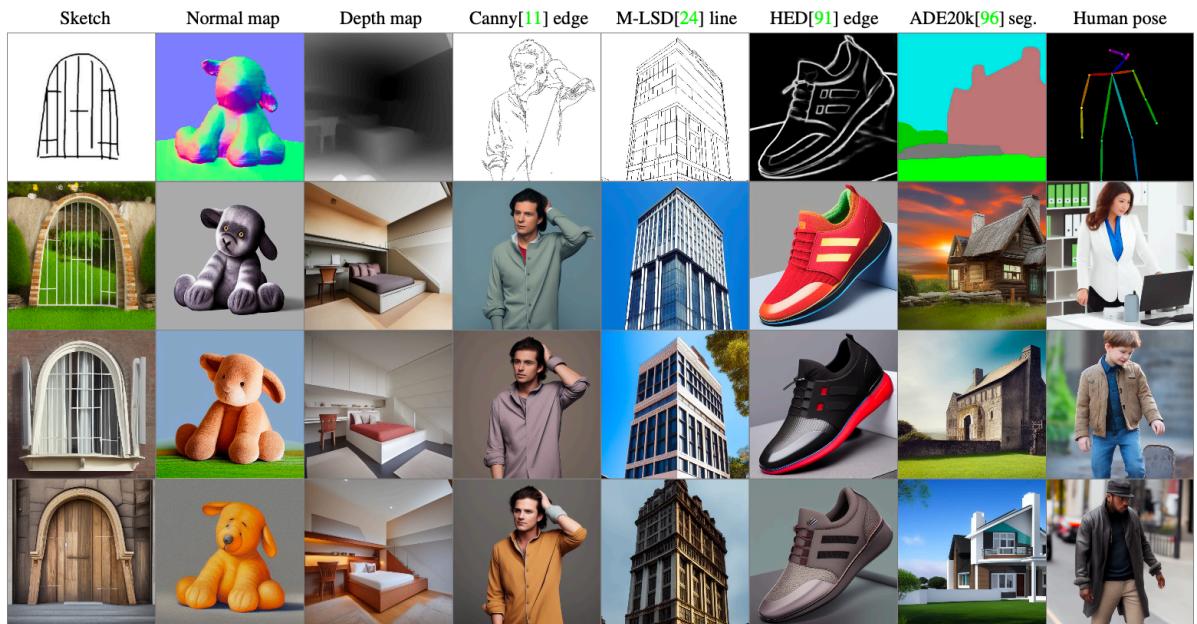


Figure 7: Controlling Stable Diffusion with various conditions **without prompts**. The top row is input conditions, while all other rows are outputs. We use the empty string as input prompts. All models are trained with general-domain data. The model has to recognize semantic contents in the input condition images to generate images.

图 外 3-5

第四章 方法

我们在 Stable Diffusion 中实现了 ControlNet，以测试各种条件，包括 Canny 边缘 [11]、深度图 [69]、法线图 [87]、M-LSD 线 [24]、HED 软边缘 [91]、ADE20K 分割 [96]、Openpose[12] 和用户手绘草图。有关每种条件的示例以及详细的训练和推理参数，请参见补充材料。

4.1 质量结果

图外 0-1 展示了在几种提示设置下生成的图像。图外 3-5 展示了在没有提示的情况下，我们在各种条件下的结果，其中 ControlNet 能够稳健地解释多样化输入条件图像中的内容语义。

4.2 消融研究

我们通过以下两种方法研究了 ControlNet 的替代结构：(1) 用高斯权重初始化的标准卷积层替换零卷积层，(2) 用单个卷积层替换每个块的可训练副本，我们称之为 ControlNet-lite。有关这些消融结构的详细信息，请参见补充材料。

我们提供了 4 种提示设置，以测试现实世界用户的可能行为：(1) 无提示；(2) 不足的提示，未完全覆盖条件图像中的对象，例如本文的默认提示“高质量、详细和专业的图像”；(3) 改变条件图像语义的冲突提示；(4) 描述必要内容语义的完美提示，例如“漂亮的房子”。图外 4-1a 显示，ControlNet 在所有 4 种设置中均成功。轻量级的 ControlNet-lite (图外 4-1c) 不够强大，无法解释条件图像，在不足和无提示条件下失败。当替换零卷积时，ControlNet 的性能下降到与 ControlNet-lite 差不多，表明在微调过程中破坏了可训练副本的预训练骨干 (图外 4-1b)。

4.3 定量评估

用户研究。我们采样了 20 幅未见过的手绘草图，然后将每幅草图分配给 5 种方法：PITI[89] 的草图模型，默认边缘引导比例 ($\beta = 1.6$) 的 Sketch-Guided Diffusion (SGD) [88]，相对较高边缘引导比例 ($\beta = 3.2$) 的 SGD[88]，上述 ControlNet-lite 和 ControlNet。我们邀请 12 名用户根据“显示图像的质量”和“草图的保真度”分别对这 20 组 5 个结果进行排名。通过这种方式，我们获得了 100 个质量排名和 100 个条件保真度排名。我们使用平均人类排名 (AHR) 作为偏好指标，用户对每个结果的排名从 1 到 5 (低分较差)。平均排名如表外 4-1 所示。

与工业模型的比较。Stable Diffusion V2 Depth-to-Image (SDv2-D2I) [83] 使用大型 NVIDIA A100 集群、数千 GPU 小时和超过 1200 万训练图像进行训练。我们为 SD V2 训练了一个 ControlNet，使用相同的深度条件，但只用了 20 万训练样本、单个 NVIDIA

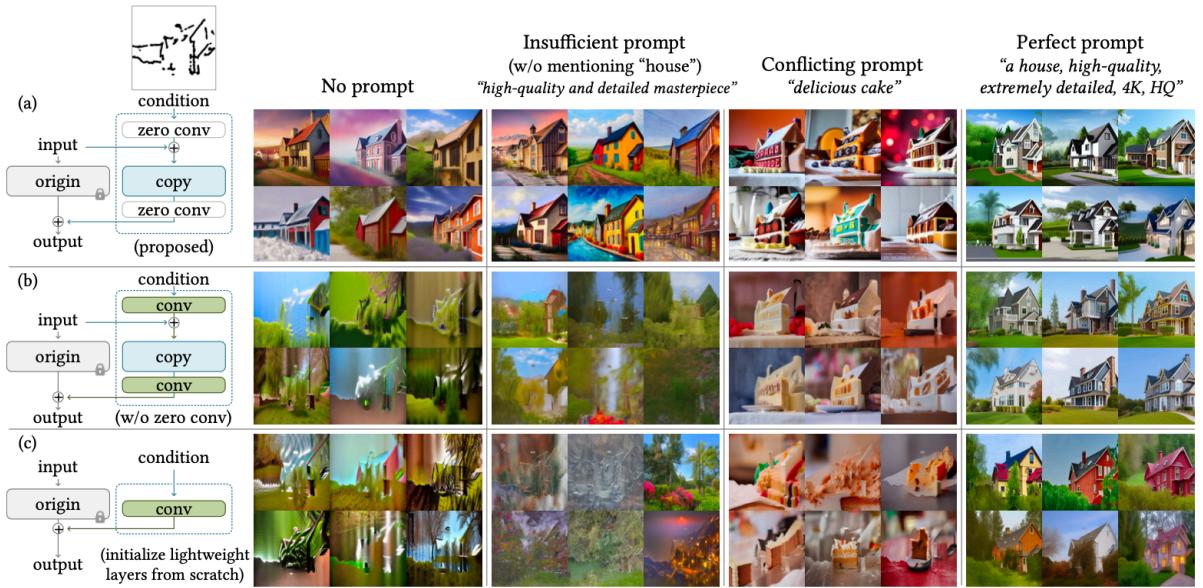


Figure 8: Ablative study of different architectures on a sketch condition and different prompt settings. For each setting, we show a random batch of 6 samples without cherry-picking. Images are at 512×512 and best viewed when zoomed in. The green “conv” blocks on the left are standard convolution layers initialized with Gaussian weights.

图 外 4-1

方法	结果质量↑	条件保真度↑
PITI [89](sketch)	1.10 ± 0.05	1.02 ± 0.01
Sketch-Guided [88] ($\beta = 1.6$)	3.21 ± 0.62	2.31 ± 0.57
Sketch-Guided [88] ($\beta = 3.2$)	2.52 ± 0.44	3.28 ± 0.72
ControlNet-lite	3.93 ± 0.59	4.09 ± 0.46
ControlNet	4.22 ± 0.43	4.28 ± 0.45

表 外 4-1 用户排名平均值 (AUR) 评估结果质量和条件保真度。我们报告了不同方法的用户偏好排名 (1 到 5, 分数越低越差)。

RTX 3090Ti 和 5 天的训练。我们使用每个 SDv2-D2I 和 ControlNet 生成的 100 幅图像来教 12 名用户区分这两种方法。随后，我们生成 200 幅图像，并要求用户辨别每幅图像是由哪种模型生成的。用户的平均准确率为 0.52 ± 0.17 ，表明这两种方法的结果几乎无法区分。

条件重建和 FID 分数。我们使用 ADE20K[96] 的测试集来评估条件保真度。最先进的分割方法 OneFormer[35] 在真实标签集上实现了 0.58 的交并比 (IoU)。我们使用不同的方法生成带有 ADE20K 分割的图像，然后应用 One-Former 再次检测这些分割，以计算重建的 IoU (表外 4-2)。此外，我们使用弗雷歇特嵌入距离 (FID) [28] 来测量使用不同分割条件方法随机生成的 512×512 图像集的分布距离，以及表外 4-3 中的文本图像 CLIP 分数 [66] 和 CLIP 美学分数 [79]。有关详细设置，请参见补充材料。

方法	IoU
ADE20K (GT)	0.58 ± 0.10
VQGAN [19]	0.21 ± 0.15
LDM [72]	0.31 ± 0.09
PITI [89]	0.26 ± 0.16
ControlNet-lite	0.32 ± 0.12
ControlNet	0.35 ± 0.14

表 外 4-2 在 ADE20K 数据集上的语义分割标签重建评估，使用 Intersection over Union (IoU \uparrow)。

方法	FID \downarrow	CLIP-score \uparrow	CLIP-aes \uparrow
Stable Diffusion	6.09	0.26	6.32
VQGAN [19](seg.)*	26.28	0.17	5.14
LDM [72](seg.)*	25.35	0.18	5.15
PITI [89](seg.)*	19.74	0.20	5.77
ControlNet-lite	17.92	0.26	6.30
ControlNet	15.27	0.26	6.31

表 外 4-3 图像生成条件语义分割的评估。我们报告了 FID、CLIP 文本图像分数和 CLIP 美学分数。我们还报告了没有分割条件的 Stable Diffusion 的性能。用“*”标记的方法是从头训练的。

4.4 与以前方法的比较

图外 4-2 展示了基准方法和我们方法 (Stable Diffusion + ControlNet) 的视觉比较。特别是，我们展示了 PITI[89]、Sketch-Guided Diffusion[88] 和 Taming Transformers[19] 的结果。(请注意，PITI 的骨干是 OpenAI GLIDE[57]，其视觉质量和性能不同。) 我们观察到，ControlNet 可以稳健地处理多样化的条件图像，并实现清晰锐利的结果。

4.5 讨论

训练数据集大小的影响。我们在图外 4-3 中展示了 ControlNet 训练的稳健性。即使在有限的 1000 张图像下，训练也不会崩溃，并且允许模型生成可识别的狮子。当提供更多数据时，学习是可扩展的。

解释内容的能力。我们在图外 4-4 中展示了 ControlNet 从输入条件图像中捕捉语义的能力。

向社区模型的转移。由于 ControlNet 不改变预训练 SD 模型的网络拓扑结构，它可以直接应用于 Stable Diffusion 社区中的各种模型，例如 Comic Diffusion[61] 和 Protagen 3.4[16]，如图外 4-5 所示。

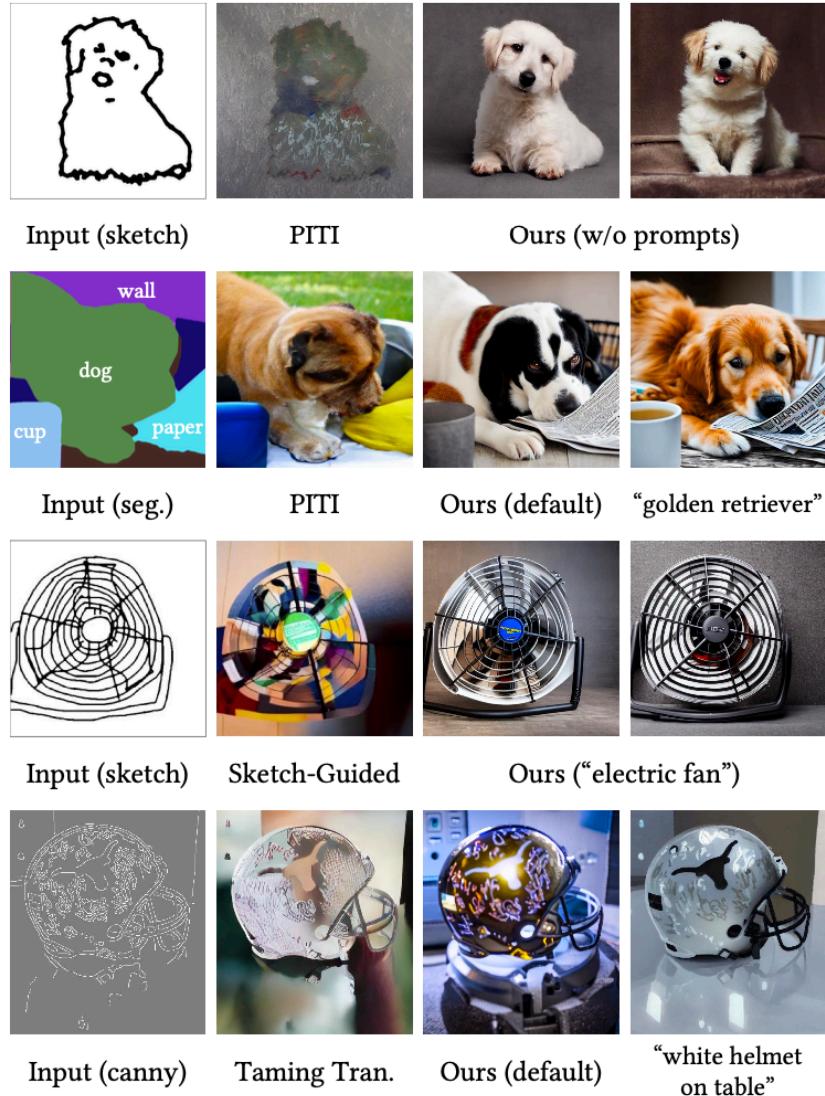


Figure 9: Comparison to previous methods. We present the qualitative comparisons to PITI [89], Sketch-Guided Diffusion [88], and Taming Transformers [19].

图 外 4-2



Figure 10: The influence of different training dataset sizes. See also the supplementary material for extended examples.

图 外 4-3



Input

“a high-quality and extremely detailed image”

Figure 11: Interpreting contents. If the input is ambiguous and the user does not mention object contents in prompts, the results look like the model tries to interpret input shapes.

图 外 4-4



“house”

SD 1.5

Comic Diffusion

Progen 3.4

Figure 12: Transfer pretrained ControlNets to community models [16, 61] without training the neural networks again.

图 外 4-5

第五章 结论

ControlNet 是一种神经网络结构，用于学习大型预训练文本到图像扩散模型的条件控制。它重用了源模型的大规模预训练层，构建了一个深度且强大的编码器，以学习特定条件。原始模型和可训练副本通过“零卷积”层连接，这些层在训练期间消除了有害噪声。广泛的实验证明了 ControlNet 能够在单一或多重条件下，有或没有提示的情况下，有效地控制 Stable Diffusion。在多样化条件数据集上的结果表明，ControlNet 结构可能适用于更广泛的条件，并促进相关应用。

致谢

本工作部分由斯坦福人本 AI 研究院和布朗媒体创新研究所支持。

引用

- [1] Sadia Afrin. Weight initialization in neural network, inspired by andrew ng, <https://medium.com/@safrin1128/weightinitialization-in-neural-network-inspired-by-andrew-nge0066dc4a566, 2020>.
- [2] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7319–7328, Online, Aug. 2021.
- [3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. ACM Transactions on Graphics (TOG), 40(4), 2021.
- [4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18511–18521, 2022.
- [5] Alembics. Discodiffusion, <https://github.com/alembics/disco-diffusion>, 2022.
- [6] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. arXiv preprint arXiv:2211.14305, 2022.
- [7] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18208–18218, 2022.
- [8] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. arXiv preprint arXiv:2302.08113, 2023.
- [9] Dina Bashkirova, Jose Lezama, Kihyuk Sohn, Kate Saenko, and Irfan Essa. Masksketch: Unpaired structure-guided masked image generation. arXiv preprint arXiv:2302.05496, 2023.
- [10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800, 2022.
- [11] John Canny. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, (6):679–698, 1986.
- [12] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
- [13] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12299–12310, 2021.

- [14] Zhe Chen, Yuchen Duan, Wenhui Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. International Conference on Learning Representations, 2023.
- [15] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8789–8797, 2018.
- [16] darkstorm2150. Progen x3.4 (photorealism) official release, <https://civitai.com/models/3666/progen-x34-photorealism-official-release>, 2022.
- [17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780–8794, 2021.
- [18] TanM.Dinh,AnhTuanTran,RangNguyen, andBinh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11389–11398, 2022.
- [19] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12873–12883, 2021.
- [20] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scenebased text-to-image generation with human priors. In European Conference on Computer Vision (ECCV), pages 89–106. Springer, 2022.
- [21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.
- [22] RinonGal,OrPatashnik,HaggaiMaron,AmitHBermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip guided domain adaptation of image generators. ACM Transactions on Graphics (TOG), 41(4):1–13, 2022.
- [23] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544, 2021.
- [24] Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. Towards light-weight and real-time line segment detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022.
- [25] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In International Conference on Learning Representations, 2017.
- [26] Heathen. Hypernetwork style training, a tiny guide, stable-diffusion-webui, <https://github.com/automatic1111/diffusion-webui/discussions/2670>, 2022.
- [27] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-

Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022.

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[30] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, pages 2790–2799, 2019.

[31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[32] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and controllable image synthesis with composable conditions. 2023.

[33] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot textdriven image editing. arXiv preprint arXiv:2302.11797, 2023.

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1125–1134, 2017.

[35] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. 2023.

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. International Conference on Learning Representations, 2018.

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019.

[38] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis, 2021.

[39] Oren Katzir, Vicky Perepelok, Dani Lischinski, and Daniel Cohen-Or. Multi-level latent space structuring for generative control. arXiv preprint arXiv:2202.05910, 2022.

[40] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276, 2022.

[41] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffu-

- sion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2426–2435, 2022.
- [42] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. Advances in Neural Information Processing Systems, 34:21696–21707, 2021.
- [43] Kurumuz. Novelai improvements on stable diffusion, <https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac>, 2022.
- [44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, May 2015.
- [45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
- [46] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. Proceedings of the 35th International Conference on Machine Learning, 2018.
- [47] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. International Conference on Learning Representations, 2018.
- [48] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jian-wei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. 2023.
- [49] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. arXiv preprint arXiv:2203.16527, 2022.
- [50] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. arXiv preprint arXiv:2111.11429, 2021.
- [51] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggy back: Adapting a single network to multiple tasks by learning to mask weights. In European Conference on Computer Vision (ECCV), pages 67–82, 2018.
- [52] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7765–7773, 2018.
- [53] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In International Conference on Learning Representations, 2021.
- [54] Midjourney. <https://www.midjourney.com/>, 2023.
- [55] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In ACM SIGGRAPH 2022 Conference Proceedings, pages 1–9, 2022.
- [56] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xi-

- aohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023.
- [57] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. CoRR, 2021.
- [58] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. 2022.
- [59] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In International Conference on Machine Learning, pages 8162–8171. PMLR, 2021.
- [60] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. arXiv preprint arXiv:2203.17272, 2022.
- [61] ogkalu. Comic-diffusion v2, trained on 6 styles at once, <https://huggingface.co/ogkalu/comic-diffusion>, 2022.
- [62] OpenAI. Dall-e-2, <https://openai.com/product/dall-e-2>, 2023.
- [63] TaesungPark,Ming-YuLiu,Ting-ChunWang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2337–2346, 2019.
- [64] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. arXiv preprint arXiv:2302.03027, 2023.
- [65] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 2085–2094, October 2021.
- [66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.
- [67] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.
- [68] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR, 2021.
- [69] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3):1623–1637, 2020.
- [70] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of

multi-domain deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8119–8127, 2018.

[71] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[72] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention MICCAI International Conference, pages 234–241, 2015.

[74] Amir Rosenfeld and John K Tsotsos. Incremental learning through deep adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(3):651–663, 2018.

[75] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242, 2022.

[76] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323(6088):533–536, Oct. 1986.

[77] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings, SIGGRAPH ’22, New York, NY, USA, 2022.

[78] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022.

[79] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[80] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In International Conference on Machine Learning, pages 4548–4557. PMLR, 2018.

[81] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265. PMLR, 2015.

- [82] Stability. Stable diffusion v1.5 model card, <https://huggingface.co/runwayml/stable-diffusion-v1-5>, 2022.
- [83] Stability. Stable diffusion v2 model card, stable-diffusion-2-depth, <https://huggingface.co/stabilityai/stable-diffusion-2-depth>, 2022.
- [84] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In International Conference on Machine Learning, pages 5986–5995, 2019.
- [85] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vi-adapter: Parameter-efficient transfer learning for vision-and-language tasks. arXiv preprint arXiv:2112.06825, 2021.
- [86] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. arXiv preprint arXiv:2211.12572, 2022.
- [87] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463, 2019.
- [88] Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. Sketch guided text-to-image diffusion models. 2022.
- [89] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. 2022.
- [90] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8798–8807, 2018.
- [91] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1395–1403, 2015.
- [92] Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas J. Guibas, and Jitendra Malik. Side-tuning: Network adaptation via additive side networks. In European Conference on Computer Vision (ECCV), pages 698–714. Springer, 2020.
- [93] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5143–5153, 2020.
- [94] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kun chang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930, 2021.
- [95] Jiawei Zhao, Florian Schäfer, and Anima Anandkumar. Zero initialization: Initializing residual networks with only zeros and ones. arXiv, 2021.
- [96] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition, pages 633–641, 2017.

[97] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11465–11475, 2021.

[98] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

[99] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. Advances in Neural Information Processing Systems, 30, 2017.

北京邮电大学

本科毕业设计（论文）任务书

学院	人工智能学院		专业	智能科学与技术	
学生姓名	罗彬慈	学号	2020212053	班级	2020219107
指导教师姓名	李佩佩	所在单位	人工智能学院	职称	副教授
设计(论文)题目	(中文) 基于 LLM 的交互式多模态图像编辑系统的设计与搭建				
	(英文) Design and Construction of Interactive Multimodal Image Editing System Based on LLM				
题目类型	工程实践类 <input type="checkbox"/> 研究设计类 <input checked="" type="checkbox"/> 理论分析类 <input type="checkbox"/> 文献综述类 <input type="checkbox"/> 其他 <input type="checkbox"/>				
题目来源	题目是否来源于科研项目 是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/>				
	科研项目名称:				
	科研项目负责人:				

主要内容：

一：熟悉图像编辑技术和生成模型的相关知识；

支撑指标点：1. 6 2. 1 2. 3 3. 1 3. 2 4. 1 5. 3 10. 1 11. 2

二：分析传统图像编辑模型的限制和挑战；

支撑指标点：1. 6 2. 1 2. 3 3. 1 3. 2 4. 1 5. 3 10. 1 11. 2

三：设计和构建交互式图像编辑系统；

支撑指标点：1. 6 2. 1 2. 3 3. 1 3. 2 4. 1 5. 3 10. 1 11. 2

四：评估系统性能并进行质量控制；

支撑指标点：1. 6 2. 1 2. 3 3. 1 3. 2 4. 1 5. 3 10. 1 11. 2

主要(技术)要求：

内容一：熟悉图像编辑技术和生成模型的相关知识；

1. 6 掌握图像生成与语言大模型基础知识及原理，能够将其和计算机知识与原理、数学与工程方法以及计算求解能力用于分析和解决复杂工程问题，并能够对解决方案进行比较和综合。

4. 1 能够采用科学方法，通过文献研究和应用案例分析等方法，调研和分析领域图像生成与语言大模复杂工程问题的解决方案。

内容二：分析传统图像编辑模型的限制和挑战；

2.1 针对交互式图像编辑领域的复杂工程问题进行问题识别，分析其功能需求与非功能需求，识别其面临的各种制约条件，对任务目标给出需求描述。

2.3 针对已建立的交互式图像编辑领域的复杂工程问题的抽象模型，论证模型的合理性；并通过文献研究，针对改进的可能性进行分析，确定解决方案，获得有效结论。

10.1 能够以撰写报告、设计文稿、口头陈述等方式，针对交互式图像编辑领域复杂工程问题，与业界同行及社会公众进行有效的沟通和交流。

内容三：设计和构建交互式图像编辑系统；

3.1 了解交互式图像编辑系统开发的一般流程，掌握交互式图像编辑系统开发及工程化的基本方法和技术。

3.2 能够针对特定需求，对交互式图像编辑问题进行分解和细化，具有设计/开发功能模块及智能系统的能力。

11.2 能够在多学科环境下，在设计开发交互式图像编辑系统解决方案的过程中，运用工程项目管理与经济决策方法。

内容四：评估系统性能并进行质量控制；

5.3 能够针对交互式图像编辑系统中的具体问题，开发满足特定需求的现代工具，进行仿真和测试，并能够分析其局限性。

主要参考文献：

[1] ACHIAM, Josh, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] FLORIDI, Luciano; CHIRIATTI, Massimo. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020, 30: 681-694.

[3] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022. p. 10684-10695.

[4] HU, Edward J., et al. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[5] ZHANG, Lvmin; RAO, Anyi; AGRAWALA, Maneesh. Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023. p. 3836-3847.

[6] VAN HUYNH, Nguyen, et al. DeepFake: Deep dueling-based deception strategy to defeat reactive jammers. *IEEE Transactions on Wireless Communications*, 2021, 20.10: 6898-6914.

进度安排：

第1阶段（2023.11.20 – 2024.12.03）：

开始调研基于深度神经网络的图像生成模型，包括对抗生成模型（GANs）、扩散模型等，以了解它们的基本原理和应用领域。

准备开题报告，明确研究目标和方法。

第2阶段（2023.12.04 – 2023.12.17）：

深入研究复现的图像生成模型，理解其优点和限制，探索其在图像编辑任务中的潜在应用。

第3阶段（2023.12.18 – 2023.12.31）：

针对图像编辑任务，提出创新性的问题，明确解决思路。

第4阶段（2024.01.01 – 2024.01.14）：

进行图像编辑模型的实验，收集和分析实验结果，识别性能瓶颈和问题。

研究已有文献和代码，选择合适的图像生成模型进行复现，确保对模型的理解和实现代码能力。

根据图像编辑任务的需求，设计和实施性能提升的方案，可能包括模型改进、数据增强等。

同时开始调研语言大模型，了解其发展现状和应用领域，确保掌握如何使用语言大模型。

第5阶段（2024.02.26 – 2024.03.10）：

结合语言大模型的研究，探讨如何将语言大模型与图像编辑模型相结合，构建交互式图像编辑系统。

第6阶段（2024.03.11 – 2024.03.24）：

进行实验，评估交互式图像编辑系统的性能，进行调整和改进。

第7阶段（2024.03.25 – 2024.04.07）：

文献更新和总结，更新文献调研，将最新研究成果与自己的工作相结合，确保研究与学术前沿保持同步，准备中期报告。

第8阶段（2024.04.08 – 2024.04.19）：

完善研究和论文撰写，完善能量模块的功能，并开始论文撰写，完成研究项目，准备最终论文和答辩

指导教师签字	李佩佩	日期	2023年11月15日
--------	-----	----	-------------

北京邮电大学

本科毕业设计（论文）开题报告

学院	人工智能学院		专业	智能科学与技术	
学生姓名	罗彬慈	学号	2020212053	班级	2020219107
指导教师姓名	李佩佩	所在单位	人工智能学院	职称	副教授
设计（论文）题目	(中文) 基于对话系统的交互式图像编辑的研究与仿真				
	(英文) Research and Simulation of Interactive Image Editing Based on Conversational Systems				

1. 背景和意义

随着图像生成技术的不断发展，图像编辑作为其中的关键技术之一，应用广泛，涵盖了媒体娱乐、数字营销和智能医疗等多个领域。然而，传统的图像编辑模型存在着交互性差和生成图像质量受限的问题，迫使我们探索更先进的方法以提高图像生成的质量和用户交互性。通过深度学习和语言大模型的结合，我们有望构建一个创新的交互式图像编辑系统，为图像编辑领域带来新的可能性。

2. 研究的基本内容和拟解决的主要问题

内容一：熟悉图像编辑技术和生成模型的相关知识

在这一部分，基本内容是深入了解图像生成领域的相关知识，包括图像编辑技术、生成模型的原理，以及语言大模型在图像编辑中的应用。这为后续交互式图像编辑系统的构建奠定了理论基础和技术背景。

内容二：分析传统图像编辑模型的限制和挑战

在这一部分，基本内容是深入研究传统图像编辑模型，识别其存在的问题，特别关注交互性和生成图像质量的限制，为改进图像编辑技术提供方向。

内容三：设计和构建交互式图像编辑系统

在这一部分，基本内容是明确交互式图像编辑系统的设计目标和实现方式。通过考察的不同参数量的多种语言大模型，选择在性能和负载中取得平衡的大语言模型，然后设计系统的核心算法和交互界面。

内容四：评估系统性能并进行部署

在这一部分，基本内容是从多个纬度评估系统性能，包括生成图像的质量、交互系统对用户请

求的处理能力和用户体验。通过不断的改进和优化，确保系统能够满足需求，提高系统的可用性和易用性。选择合适的方式部署系统，是系统能够提供服务。

3. 研究方法及措施

在解决上述问题的过程中，我们将采用以下研究方法和措施：

(1) 深入调研与学习：对深度学习和神经网络的基本原理进行学习，特别关注对抗生成模型和扩散模型等图像生成模型的详细了解。同时，调研语言大模型的最新发展。

(2) 总结与优化：分析当前图像编辑模型的创新之处和局限性，提出改进图像编辑模型的方法，以提高其性能。

(3) 语言大模型应用：学习如何使用语言大模型，并探索其在图像编辑任务中的应用，为系统集成做好准备。

(4) 系统构建与评估：设计并实现一个交互式图像编辑系统，结合深度神经网络的图像生成模型和语言大模型，通过实验评估系统性能。

4. 研究工作的步骤与进度

第 1 阶段（2023.11.20 - 2024.12.03）：背景调研与开题准备

开始调研基于深度神经网络的图像生成模型，了解对抗生成模型（GANs）、扩散模型等的基本原理和应用领域。

准备开题报告，明确研究目标和方法。

第 2 阶段（2023.12.04 - 2023.12.17）：图像生成模型深入研究

深入研究复现的图像生成模型，理解其优点和限制，探索其在图像编辑任务中的潜在应用。

第 3 阶段（2023.12.18 - 2023.12.31）：问题定义与解决思路

针对图像编辑任务，提出创新性的问题，明确解决思路。

第 4 阶段（2024.01.01 - 2024.01.14）：实验与性能评估

进行图像编辑模型的实验，收集和分析实验结果，识别性能瓶颈和问题。

学习语言大模型的使用方法，为系统集成做准备。

第 5 阶段（2024.02.26 - 2024.03.10）：语言大模型与系统集成

结合语言大模型的研究，探讨如何将语言大模型与图像编辑模型相结合，构建交互式图像编辑系统。

第 6 阶段（2024.03.11 - 2024.03.24）：实验与系统调优

进行实验，评估交互式图像编辑系统的性能，进行调整和改进。

第 7 阶段（2024.03.25 - 2024.04.07）：文献更新与中期报告准备

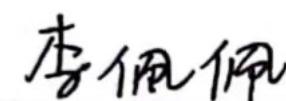
更新文献调研，将最新研究成果与自己的工作相结合，准备中期报告。

第 8 阶段（2024.04.08 - 2024.04.19）：论文撰写与最终总结

完善研究和论文撰写，确保项目完善，准备最终论文和答辩。

主要参考文献：

- [1] ACHIAM, Josh, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [2] FLORIDI, Luciano; CHIRIATTI, Massimo. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 2020, 30: 681-694.
- [3] ROMBACH, Robin, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 10684-10695.
- [4] HU, Edward J., et al. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.
- [5] ZHANG, Lvmin; RAO, Anyi; AGRAWALA, Maneesh. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 3836-3847.
- [6] VAN HUYNH, Nguyen, et al. DeepFake: Deep dueling-based deception strategy to defeat reactive jammers. *IEEE Transactions on Wireless Communications*, 2021, 20.10: 6898-6914.

允许进入毕业设计（论文）下一阶段：是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/>		指导教师	
日期	2023 年 12 月 30 日	签字	

注：可根据开题报告的长度加页

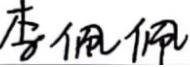
北京邮电大学

本科毕业设计（论文）中期进展情况检查表

学院	人工智能学院		专业	智能科学与技术	
学生姓名	罗彬慈	学号	2020212053	班级	2020219107
指导教师姓名	李佩佩	所在单位	人工智能学院	职称	副教授
设计（论文） 题目	(中文) 基于 LLM 的交互式多模态图像编辑系统的设计与搭建				
	(英文) Design and Construction of Interactive Multimodal Image Editing System Based on LLM				
目前 已 完 成 任 务	<p>目前已完成的工作主要包括背景调研、系统设计、代码实现和系统部署三个方面：</p> <p>一、背景调研</p> <p>图像生成技术和 LLM (Large Language Models) 都是深度学习领域的研究热点。Stable Diffusion 技术的推出和不断的迭代引发了在图像生成领域的浪潮；同时，像 ChatGPT 这样的大语言模型的发布和不断升级也引起了全球范围内的广泛关注。</p> <p>图像编辑是图像生成技术中的关键技术之一，其旨在保留图像的主要信息的同时根据指令对特定内容进行修改。图像编辑技术在媒体娱乐、数字营销、智能医疗等领域具有广泛的应用前景，因此备受关注。然而，传统的图像编辑模型在交互性方面存在一定的局限性。本项目旨在利用大语言模型辅助进行交互式图像编辑，结合先进的图像生成模型以提升生成图像的质量，最终打造一个交互式图像编辑系统。</p> <p>二、系统设计</p> <p>1. 系统框架</p> <p>考虑到本项目需要整合多个模型且要满足用户使用的便利性，本项目使用 Golang 语言搭建了一个后端服务 (middleware) 对相同类型功能的请求进行整合并向 GUI 提供整合后的 API，因此本项目的 GUI 部分仅需安装一个 python 依赖库，降低了使用门槛。图像生成模型的请求通过 API 调用部署在云平台上的 Stable Diffusion 模型和由 OpenAI 提供的 DALL-E 模型；大语言模型的调用目前可选择通过 API 调用 OpenAI 的 GPT3.5turbo/GPT4 模型，后续准备对 ChatGLM 系列模型进行微调后部署在云平台上以通过 API 调用。这样的系统框架增加了用户使用的便捷性、系统迭代的稳定性，可在用户无感的条件下对系统进行升级。</p> <pre style="background-color: #f0f0f0; padding: 10px;"> gradio_web(Image preprocessing, GUI) api ---api---> StableDiffusion middleware <== ---api---> OpenAI (GPT3.5turbo/GPT4, DALLE) ---api---> ChatGLM2-6B </pre>				

图 2-1 系统框架

	<p>2.大语言模型</p> <p>大语言模型主要通过使用给定的 Prompt、图像分割的结果等将用户输入的自然语言转换为 json 格式的指令。系统会在大语言模型返回的文本中自动地抽取指令并结合配置文件提供的规则对指令的合规性进行校验。由于不同的语言大模型的表现并不一致，用户可以在配置文件中添加多个不同的 Prompt 模版以在各个模型上达到更好的效果。</p>			
	<p>3.图像生成模型</p> <p>考虑到 Stable Diffusion 的功能更加丰富以及在开源社区的热度，图像生成主要采用 Stable Diffusion 模型，并在特定的任务中结合不同的模型与插件以满足任务的需求并提高结果的质量。当用户上传图片时，图像分割模型会对图像进行分割，并将分割的结果以对话的方式反馈给用户。在使用图像生成模型对图像进行修改时，首先会结合大语言模型生成的指令和分割的结果对不需要修改的部分进行遮罩，然后在生成的过程中通过 Control Net 对图像的基本框架进行固定以保证生成的质量。由于图像分割模型返回的分割结果不可避免的在一些细节上存在瑕疵（如标签 Hair 在部分区域存在将发丝周围的少量的像素一同分割），在生成遮罩时会对特定的标签进行不同程度的腐蚀--膨胀操作以提高遮罩的质量。同时，可以选择是否使用 ROOP 对图像中的人脸进行替换。当部署在云平台上的 Stable Diffusion 模型不可用时，会使用 DALL-E 模型完成相应功能。</p>			
	<p>三、代码实现</p> <p>目前的总代码量为 3649 行，其中 GUI 代码量为 2323 行， middleware 代码量为 1120 行。</p>			
	<p>1.GUI</p> <p>使用 Python3 实现了 GUI 的构建，主要使用了 gradio 库。在代码实现时将 GUI 分为 webui、controllers、modules、config、test 五个模块，分别实现 UI 布局（各个模块在 UI 界面的分布）、交互逻辑（不同可交互模块的处理逻辑）、请求处理（图像处理、request body 生成等）、系统配置（模版、路由、默认参数等）、代码测试（主要针对 modules 中的函数进行测试）。UI 界面主要由操作面板、图像预览、文本交互界面和图片编辑界面。</p>			
	<p>2.middleware</p> <p>使用 Golang 语言配合 beego 框架实现，将大语言模型调用、图像生成模型调用等数十个 API 整合为 8 个 API 供 GUI 调用，同时在不需要标识请求用户的情况下支持多个 GUI 同时对其进行访问但互不影响（如对话的历史记录等）。</p>			
	<p>3.Stable Diffusion</p> <p>基于开源项目 stable-diffusion-webui 进行修改，在其中添加了 ROOP、Control Net 等插件以满足项目的需求。</p>			
	<p>四、系统部署</p>			
	<p>1.项目运行</p> <p>本项目提供了 Docker、Kubernetes、Terminal 等多种运行方式。</p>			
	<p>2.持续集成、持续部署</p> <p>持续集成（Continuous Integration, CI）是一种软件开发实践，旨在通过频繁地将代码集成到共享存储库中，然后对代码进行自动化构建和测试，以确保代码变化不会破坏整个代码库的稳定性。CI 旨在尽早发现和解决集成问题，从而提高开发效率和软件质量。</p>			
	<p>持续部署（Continuous Deployment, CD）持续部署是自动将经过测试的代码部署到生产环境，没有人工干预。</p>			
	<p>本项目使用 GitHub Action 实现了自动化测试、镜像自动化构建与发布、自动部署至 Azure，提高了代码的可靠性并减少了非开发的工作量。</p>			
	<table border="1"> <tr> <td>是否符合任务书要求进度</td> <td>是 <input checked="" type="checkbox"/></td> <td>否 <input type="checkbox"/></td> </tr> </table>	是否符合任务书要求进度	是 <input checked="" type="checkbox"/>	否 <input type="checkbox"/>
是否符合任务书要求进度	是 <input checked="" type="checkbox"/>	否 <input type="checkbox"/>		

尚需完成的任务	<p>1. 实现建议功能，在对话时提供建议，并且在用户采纳建议后更新指令。由于传统的大语言模型只能接收文本输入，如果使用传统的大语言模型效果不理想，考虑使用 GPT4-V 模型来生成更好的建议。</p> <p>2. 使用 LoRa 微调一个更加适合本任务的大语言模型，在与大语言模型交互时支持切换不同的大语言模型。</p>		
	是否可以按期完成设计（论文） 是 <input checked="" type="checkbox"/> 否 <input type="checkbox"/>		
存在问题和解决办法	存在问题	<p>1. GUI 的交互性仍有待改善。</p> <p>2. 在编辑图像时可选的参数较少。</p> <p>3. 图像生成质量不稳定。</p>	
	拟采取的办法	<p>1. 通过对功能分区分类或简化提升 GUI 的交互性。</p> <p>2. 实现更多的参数可自定义并增加一个设置面板，提高系图像生成的稳定性，将所有设置内容迁移到设置面板中并提供模版，避免在增加参数的过程中增加交互的复杂性。</p>	
指导教师签字		日期	2024 年 3 月 14 日
检查小组评分及意见	<p>评分：26 (总分：30)</p> <p style="text-align: right;">组长签字： 2024 年 3 月 21 日</p>		

注：此表仅供参考，各学院应围绕毕设目标达成度，结合人才培养目标、专业认证要求等进行个性化完善。

北京邮电大学
教师指导本科毕业设计（论文）记录表

学院	人工智能	专业	智能科学与技术	班级	2020219107
学生姓名	罗彬慈	学号	2020212053	班内序号	19
指导教师姓名	李佩佩	职称	副教授		
<p>第 1—2 周记录：</p> <p>说明毕业论文要求，收集选题相关资料，下达任务书。</p>					
指导教师签字	李佩佩		日期	2023 年 10 月 20 日	
<p>第 3—4 周记录：</p> <p>向学生推荐一些选题相关的期刊文献，指导学生完成开题报告。</p>					
指导教师签字	李佩佩		日期	2023 年 11 月 10 日	

第 5—6 周记录：

指导学生修改开题报告，与学生讨论并确定毕业设计实现的总体思路与方法。

指导教师签字	李佩佩	日期	2023 年 12 月 1 日
--------	-----	----	-----------------

第 7—8 周记录：

对学生提出的问题进行解答，确定整体实现思路，辅导学生准备中期检查表。

指导教师签字	李佩佩	日期	2023 年 12 月 22 日
--------	-----	----	------------------

第 9—10 周记录：

基于整体思路对细化各个模块的实现思路，初步验证实现效果并与学生讨论改进的方法，对中期检查表进行审阅并提出修改意见。

指导教师签字	李佩佩	日期	2024 年 3 月 1 日
--------	-----	----	----------------

第 11—12 周记录：

进行中期检查相关工作。

指导教师签字	李佩佩	日期	2024 年 3 月 15 日
--------	-----	----	-----------------

第 13—14 周记录：

进一步验证改进后的实现效果，完成系统各个模块的搭建，收集相关数据。

指导教师签字	李佩佩	日期	2024 年 4 月 5 日
--------	-----	----	----------------

第 15—16 周记录：

与学生讨论论文总体结构，辅导学生撰写论文初稿并提出建议。

指导教师签字	李佩佩	日期	2024 年 4 月 19 日
--------	-----	----	-----------------

第 17—18 周记录：

督促学生对论文进行查重，进一步提出修改意见，与学生讨论答辩相关事宜。

指导教师签字	李佩佩	日期	2024 年 5 月 3 日
--------	-----	----	----------------

第 19—20 周记录：

进行论文答辩相关工作。

指导教师签字	李佩佩	日期	2024 年 5 月 20 日
--------	-----	----	-----------------

注：每 2 周指导内容记录在一个表格中，双面打印。