同行专家业内评价意见书编号：20250854469

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名：黄浩凡

学号：22260154

申报工程师职称专业类别（领域）：电子信息

浙江工程师学院（浙江大学工程师学院）制

2025年05月28日

填表说明

一、本报告中相关的技术或数据如涉及知识产权保护、军工项目保密等内容，请作脱密处理。

二、请用宋体小四字号撰写本报告，可另行附页或增加页数，A4纸双面打印。

三、表中所涉及的签名都必须用蓝、黑色墨水笔，亲笔签名或签字章，不可以打印代替。

四、同行专家业内评价意见书编号由工程师学院填写，编号规则为：年份4位＋申报工程师职称专业类别(领域)4位+流水号3位，共11位。

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.对本专业基础理论知识和专业技术知识掌握情况(不少于200字)

在攻读电子信息（新一代电子信息技术）专业学位期间，我系统掌握了本专业领域的核心理论与技术体系。基础理论知识方面，深入学习了扩散模型的数学原理，包括前向噪声扩散与反向去噪的马尔可夫链建模、变分下界优化目标及隐空间特征重构机制，能够熟练推导扩散过程与损失函数。同时，对注意力机制（自注意力、多头注意力）、变分自编码器（VAE）的潜在空间建模方法、CLIP跨模态语义对齐技术等进行了系统性研究，构建了多模态生成的理论框架。此外，结合视频数据的时空特性，掌握了3D卷积网络、光流运动建模、频域分析（如3D-
DCT）等视频处理基础理论，并深入理解视频生成中时序冗余压缩、跨模态融合等核心问题的数学建模方法。

专业技术知识方面，聚焦多模态条件视频生成任务，提出了一系列创新性技术：针对文本生成视频任务，设计了文本-视频共享注意力运动控制网络（TVMC-
Net），通过时空解耦策略实现图像生成模型向视频模型的轻量化扩展；开发了多频谱时空规整网络（MSTAN），利用频域变换提取关键时序模式以增强运动连续性。在图生视频任务中，构建了双流注意力协同网络（DAC-
Net），实现图文特征深度交互，并创新性引入首帧引导自适应层归一化（FFG-
AdaLN）模块，显著提升风格一致性。针对视频空间扩展任务，提出基于通道融合的视频扩展框架（CSVEN），结合条件感知动态选择机制（CADSN）实现多模态引导优化，并通过滑动窗口动量更新策略解决长视频生成难题。在工程实践中，熟练应用PyTorch框架进行大规模视频数据处理，掌握OpenCV光流分析、CLIP语义对齐、RAFT插帧等工具链，具备从算法设计、代码实现到多GPU分布式训练的全流程技术能力。相关成果在DAVIS、YouTube-
VOS等数据集上取得FVD指标6.0的显著提升，验证了理论方法与技术方案的有效性。

2.工程实践的经历(不少于200字)

在杭州拓欣科技有限公司的多模态视频生成项目的工程实践过程中，我全程参与了系统的架构设计、模型研发、数据集构建与实际部署工作。

本项目旨在基于扩散模型技术，构建一套多模态条件驱动的视频生成系统，使其能够根据文本、图像、传感器等多源输入，自动合成具备真实物理特征和时序逻辑的工业视频。项目目标包括：提升工业视频生成的智能化与自动化水平、降低制作成本与周期、增强工业场景下的可视化表达能力，并实现技术在数字孪生、设备预演、应急演练等核心环节的实际落地应用。通过该项目，期望助力企业构建更加高效、安全和可持续的智慧工业体系。在系统搭建阶段，我基于Stable
Diffusion原理对底层模型进行了改造，使其适应多模态输入，并通过引入TVMC-
Net跨模态注意力机制解决了工业语义难以对齐的问题。在工程实现层面，我主导开发了多模态数据预处理与同步模块，使文本、图像、传感器数据能够统一转换为模型可解析的结构化条件输入；同时通过MSTAN模块有效提升了生成视频的时序一致性。在部署实践中，我配合企业IT部门将模型集成至企业数字孪生平台，部署于云服务器上，实现在生产现场的高效调用。针对设备种类多、参数复杂的问题，我还构建了轻量化的DAC-

Adapter机制，提升了模型对不同工业场景的适配性。整个项目涉及的工业视频数据量超2TB，我组织团队制定高效的数据清洗和标注流程，保障训练质量。该系统目前已稳定运行半年，取得良好反馈，并推动企业智能可视化能力迈上新台阶。

3.在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字）

在智慧工业快速发展的背景下，视频生成技术逐渐成为数字孪生、设备监测和培训演练等场景中的关键支撑，企业在设备状态监测、故障预演、安全培训、数字孪生等场景中对高质量工业视频内容的需求日益增长。然而，传统视频生成方法在工业环境中仍面临诸多技术瓶颈。一方面，工业场景中的多模态数据如文本描述、图像资料、传感器信号等难以实现语义对齐，导致生成视频与业务需求脱节；另一方面，工业过程具备强时序性，现有生成模型在动态连贯性建模方面能力不足。此外，工业级长视频生成对实时性和计算资源提出了更高要求，传统方法难以胜任。因此，本文围绕“跨模态对齐”“时序建模”和“计算效率”三大核心挑战，提出一套基于扩散模型的多模态条件视频生成系统，并在智慧工业多个典型场景中实现了落地应用。

本项目设计了一种“文本-图像-视频”渐进式建模框架，通过扩散模型强大的生成能力，引入多项创新性技术构建工业级视频合成路径。首先，提出了“文本-视频共享注意力机制”（TVMC-Net），在传统UNet架构基础上引入跨模态注意力层，实现文本信息对视频生成过程的动态引导。例如在锅炉故障模拟任务中，当输入描述为“锅炉管道发生蒸汽泄漏”时，模型能够准确理解语义并生成对应的蒸汽喷涌动态效果，显著优于仅依赖图像或传感器数据的方案。其次，为增强时序一致性，项目引入“多频谱时序规整网络”（MSTAN），通过3D离散余弦变换提取关键时序特征，提升视频连贯性与物理真实性。在风机振动监测任务中，该模块能够有效重建出风速、振动频率等因素驱动下的自然运动轨迹，模拟精度大幅提升。此外，为解决不同工业输入的差异性，系统集成了动态权重融合（DWF）机制和自适应条件选择策略，能够根据输入数据类型（如文本、图纸、传感器信号）自动调整模态比重，从而生成更具业务针对性的内容。例如在能源管理可视化中，模型能结合厂区布局图与实时电力数据，动态生成能耗热力图，助力调度优化。

针对工业应用中的长视频生成效率问题，项目提出“滑动窗口+动量更新策略”，将长视频序列划分为片段，通过动量平滑算法确保生成片段之间的视觉与时序连续性，有效提升生成速度与稳定性，整体效率提升约40%。同时，引入轻量化适配器（DAC-Adapter）结构，在不显著增加计算负担的前提下，实现对多种场景适配，极大增强了系统的通用性和可部署性。

在实际应用方面，本系统已在多个典型智慧工业场景中完成部署验证。在数字孪生项目中，某热电企业需模拟锅炉泄漏故障以支持应急演练，系统通过输入锅炉设计图、红外热成像图与文本描述“蒸汽从裂缝逸出”，准确生成随压力变化动态增强的蒸汽喷涌效果，且无需人工干预。在风力发电机状态监测任务中，模型结合系统振动频谱与文本参数，如“风速12m/s，叶片振动频率0.5Hz”，生成了真实可视化振动动画，显著提升了运维效率，生成速度为传统仿真软件的五倍。在工业培训场景中，针对无法实地拍摄的危险事件，如反应爆炸，模型基于事故照片与结构图，配合文本描述“安全阀失效导致爆炸”，生成高拟真爆炸与化学品飞溅效果的视频，显著提升培训沉浸感，培训效果提升80%。

从经济与社会价值层面来看，该系统大幅降低了工业视频制作成本，相较传统3D动画方案节

省50%至70%的开支。目前，系统已在项目中投入使用，初步形成一套标准化工业视频生成流程，推动智慧工业从“数据驱动”走向“视频洞察”。该技术成果已发表在2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)国际顶级会议；并且形成了硕士论文，获得了全为优秀的评审意见。

综上所述，本文提出的基于扩散模型的多模态条件视频生成方法，成功解决了工业视频生成中的关键技术难题，并在多个场景中实现了高效应用，充分体现其技术先进性与实践价值。未来，项目计划引入物理仿真引擎，进一步增强生成视频的物理合理性与工程可解释性。同时，开发低代码平台，使非技术用户也能快速定义输入条件生成工业视频，降低应用门槛。此外，结合AR/VR技术探索远程运维、异地培训等新场景，推动智慧工业可视化迈向更高层次的智能化、沉浸化发展。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】

1.
公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
Do Less and Achieve More: Free Condition Video Outpainting with Diffusion Model	会议论文	2024年12月07日	2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	1/6	EI会议收录
基于扩散模型的多模态条件视频生成研究	学位论文送审专家评阅结果全优	2025年04月10日	浙江大学学位论文	1/1	

2. 其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 88 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年（要求1年及以上） 考核成绩： 75 分
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：**	

2226015

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀 ☐良好 ☐合格 ☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：2025年5月28日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过 ☐不通过（具体原因：　　） 工程师学院教学管理办公室审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22260154	姓名：黄浩凡	性别：男	学院：工程师学院	专业：电子信息	学制：2.5年
毕业时最低应获：26.0学分		已获得：28.0学分		入学年月：2022-09	毕业年月：
学位证书号：			毕业证书号：		授予学位：

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2022-2023学年秋季学期	智慧能源工程案例分析		2.0	93	专业学位课	2022-2023学年冬季学期	产业技术发展前沿		1.5	93	专业学位课
2022-2023学年秋季学期	工程技术创新前沿		1.5	91	专业学位课	2022-2023学年春季学期	数学建模		2.0	80	专业选修课
2022-2023学年秋季学期	人工智能算法与系统		2.0	100	专业选修课	2022-2023学年春季学期	自然辩证法概论		1.0	84	公共学位课
2022-2023学年秋冬学期	研究生论文写作指导		1.0	92	专业学位课	2022-2023学年春季学期	研究生英语基础技能		1.0	85	公共学位课
2022-2023学年秋冬学期	数据分析的概率统计基础		3.0	82	专业选修课	2023-2024学年秋季学期	研究生英语应用能力提升		2.0	83	公共学位课
2022-2023学年冬季学期	智慧能源系统工程		2.0	87	专业学位课	2023-2024学年冬季学期	工程伦理		2.0	94	公共学位课
2022-2023学年秋冬学期	高阶工程认知实践		3.0	80	专业学位课		硕士生读书报告		2.0	通过	
2022-2023学年冬季学期	新时代中国特色社会主义理论与实践		2.0	93	公共学位课						

说明：1.研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2.备注中“*”表示重修课程。

学院成绩校核章：

成绩校核人：张梦依

打印日期：2025-06-03

浙江大学研究生院（60）成绩校核章

公开成果代表作 1：以第一作者公开发表论文

Do Less and Achieve More: Free Condition Video Outpainting with Diffusion Model

Publisher: IEEE

Haofun Huang; Yinin Guo; Yening Lv; Sizhe Shan; Yan Zhang; Yushui Wang   All Authors

Abstract:
Video outpainting aims to extend the content of a video beyond its original spatial boundaries. Existing methods tend to condition the generation process on a single frame or caption, failing to address the challenge in long videos with multiple video clips. To address this, we extend the diffusion-based image inpainting to a 3D diffusion model with a motion module for video outpainting. Then, we design four conditioning schemes to seek the appropriate condition strategy: Frame-wise context, Frame-wise Adapter, Text, and Free Condition. Additionally, to enhance the continuity in long video generation, we propose a novel Motion Momentum Update (MMU) method based on a training-free technique. Experimental results demonstrate that our proposed Free Condition strategy, termed Free-Outpaintor, achieves state-of-the-art performance in video outpainting tasks. Ablation studies also show that the condition-free training paradigm is the most effective in avoiding inaccurate external information and fully utilizing inherent video data. Benefiting from the Free Condition strategy, our pipeline reduces the inference time to 5 seconds for 512x512x16 frames with only 10 GB of GPU memory. More results can be found on our demo page https://lilyn3125.github.io/FreeOutpainting/.

Published in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Date of Conference: 06-11 April 2025

DOI: 10.1109/ICASSP49660.2025.10890195

Date Added to IEEE *Xplore*: 07 March 2025

Publisher: IEEE

▸ISBN Information:

Conference Location: Hyderabad, India

⌄ISSN Information:

Do Less and Achieve More: Free Condition Video Outpainting with Diffusion Model

Haofun Huang, Yinin Guo, Yening Lv, Sizhe Shan, Yan Zhang, Yushui Wang
College of Information Science and Electronic Engineering
Zhejiang University
Hangzhou, China

Index Terms—video outpainting, diffusion model, long video generation

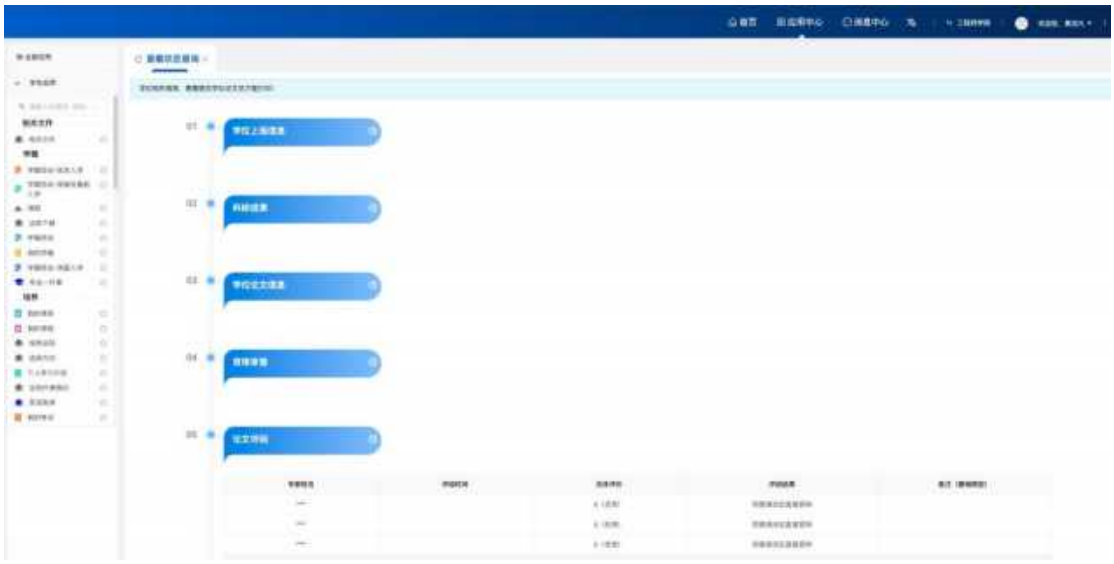
I. Introduction

II. Related Works

Diffusion Model.

Diffusion-based Video outpainting.

公开成果代表作 2：学位论文送审专家评阅结果全部为优秀



分类号：TN912.35　　单位代码：10335

密　级：非涉密　　学　号：

浙 江 大 学

硕士学位论文

（专业学位）

中文论文题目：基于扩散模型的多模态条件视频生成研究

英文论文题目：Research on Diffusion Model-based Multimodal Conditional Video Generation

申请人姓名：

指导教师（组）：

行业导师：

专业学位类别、领域：电子信息、新一代电子信息技术

研究方向：计算机视觉生成

培养类型：全日制非定向

所在学院：工程师学院

论文递交日期　2025 年 4 月 10 日

摘要

近年来，扩散模型在生成式任务中的突破极大推动了人工智能生成技术的发展。作为关键的视觉模态，视频生成已成为当前研究的重点方向。视频生成任务是根据输入条件生成高质量、语义一致的视频内容，在内容创作、数字媒体和智慧工业等领域展现出广阔的应用前景。然而，面临着跨模态语义对齐、时序结构建模、冗余信息压缩以及多模态融合不足等多方面的挑战。本文围绕扩散模型在多模态条件视频生成中的应用开展系统性研究，沿“文本条件—图像条件—视频条件”的渐进式建模路径，逐步解析不同模态条件下视频生成任务的核心问题与应对策略。

在文本生成视频的研究中，本文针对现有方法在运动建模能力弱和时序语义错位等问题，提出一种基于共享注意力的文本驱动生成框架。该方法引入时空卷积层将图像生成模型扩展为可建模时序动态的视频模型。同时，设计文本-视频共享注意力机制，引导不同层次文本语义参与运动生成控制，并构建多阶递进时序建模网络，通过帧级差异提取关键时序模式，从而增强运动连续性与语义一致性。本方法在语义一致性指标达到了79.15，优于多种主流方法，展现出更强的文本对运动控制能力。

在图像生成视频的研究中，本文关注图文融合方式单一、模态引导能力不足的问题，提出图文双流注意力协同机制与动态融合策略，实现参考图像与文本语义之间的深度交互与引导权重的动态调节。同时，设计基于自适应层归一化的图像风格保持模块，增强模型对图像低阶特征和统计信息的建模能力，提升生成视频与首帧图像在风格上的一致性。本方法在CLIPTemp指标上达0.945，优于现有主流方法，验证了该方法的有效性。

在基于视频条件的空间补全的研究中，针对扩展区域内容预测不准确、多模态引导单一以及长视频生成困难的问题，提出基于调度融合的视频扩展框架。该方法通过引入输入视频掩码信息，融合多种条件进行内容引导，提升扩展区域生成的结构一致性与语义合理性。同时，设计条件感知动态选择机制，根据输入模态特征自适应选择最优引导策略，并设计了一种滑动窗口视频扩展方案，实现高效长视频生成。在DAVIS数据集上，本方法在FVD指标上较现有最佳方法下降6.0，在视觉质量和结构保持性方面取得更优结果。

综上，本文针对多模态视频生成任务中存在的建模难题，提出具有针对性的结构设

计，提升了模型的生成能力与可控性，为后续研究提供了方法参考与理论支撑。

关键词： 人工智能生成；视频生成；扩散模型；计算机视觉；多模态融合

目录

6.2 展望

尽管基于扩散模型的多模态生成取得了显著进展，但该领域仍然面临着诸多挑战，同时也蕴含着巨大的发展潜力。

首先是生成内容的可控性和个性化。如何提高生成内容的可控性和个性化程度，以满足用户多样化的需求，是一个重要的研究方向。当前的多模态生成模型，在控制生成内容风格、布局、细节等方面仍然存在一定的局限性。未来研究可以探索更精细的控制机制，例如，引入更多种类的条件输入，如草图、姿势、语义分割图等、可学习的控制模块以及基于用户反馈的迭代生成等技术，提高生成内容的可控性和个性化程度。

然后是计算效率和模型轻量化。扩散模型生成过程的计算成本较高，限制了其在实际应用中的部署和推广。如何降低扩散模型的计算成本，使用轻量化模型，提高生成效率，也是一个重要的研究方向。未来研究可以探索更高效的采样算法、模型压缩和加速技术，如模型剪枝、量化、知识蒸馏以及基于硬件加速的优化方法，提高扩散模型的计算效率和模型轻量化程度。

最后是多模态生成模型的评估。多模态生成模型的评估仍然是一个具有挑战性的问题。现有的评估指标在衡量多模态生成内容的质量和语义一致性方面仍然存在一定的局限性，很多工作依然会使用人类偏好度评分以证明模型的有效性。未来研究可以探索更全面、更客观、更符合人类感知的多模态生成模型评估方法，例如，引入多模态一致性指标、人类偏好评估以及任务驱动的评估指标，更准确地评估多模态生成模型的性能。

未来，基于扩散模型的多模态生成技术将在更广泛的领域得到应用，例如，内容创作、虚拟现实、人机交互、教育娱乐等。随着研究的深入和技术的进步，多模态生成技术将为人类社会带来更多的创新和惊喜。