主办 CCF计算机视觉专业委员会

COMPUTER VISION NEWSLETTER

CCF 计算机视觉专委会简报

03 2025

总第 45 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2025 年第 03 期

总第 45 期

CCF 计算机视觉专委会

主　办　CCF 计算机视觉专业委员会

编委会

CONTENTS
简报目录

CCF 计算机视觉专委会
CCFCV.CCF.ORG.CN
CCFCVN@GMail.com

CCF-CV 走进高校系列报告会

第 145 期 东北大学

2025 年 9 月 16 日，由中国计算机学会主办，中国计算机学会视觉专委会、东北大学联合承办的第 145 期 CCF-CV 走进高校系列报告会在东北大学工业智能与系统优化国家级前沿科学中心 S36 学术报告厅成功举行。本次报告会邀请了中国科学院大学**蒋树强**教授、哈尔滨工业大学**范晓鹏**教授、北京工业大学**马楠**教授、北京航空航天大学**刘军**教授、清华大学**黄高**副教授五位专家学者作特邀报告，东北大学师生齐聚一堂，积极参与这场学术盛宴，探讨具身智能感知交互的最新研究进展和未来发展趋势。出席本次报告会的领导有东北大学党委副书记、副校长**唐立新**院士，本次报告会执行主席是东北大学**杨金柱**教授和**徐特**副教授，主持人是东北大学**杨金柱**教授、东北大学**徐特**副教授、大连大学附属中山医院**王喆**副教授、吉林大学第一医院**李雪峰**教授。

本次报告会聚焦“具身智能感知交互”主题，首先由东北大学党委副书记、副校长**唐立新**院士致辞，唐立新院士代表学校向与会专家和嘉宾表示热烈欢迎，并表示并简要介绍了学校在人工智能、智能传感、交互计算等领域的学科布局与研究成果。并期待通过本次学术交流，进一步凝聚智慧、形成共识，共同推动我国在具身智能感知交互领域的理论突破、技术发展与产业进步，为实现高水平科技自立自强贡献力量。最后预祝本次报告会取得圆满成功！

随后，中国科学院大学**蒋树强**教授，哈尔滨工业大学**范晓鹏**教授，北京工业大学**马楠**教授，北京航空航天大学**刘军**教授，清华大学**黄高**副教授做主题报告，内容涵盖人工智能、智能传感、交互计算等具身智能感知交互技术。最后，五位专家与师生互动，共同探讨、交流，并对师生提出的问题做出详尽的回答，提出了许多有价值的学术见解，论坛现场气氛热烈。

本期 CCF-CV 走进东北大学系列报告会围绕“具身智能感知交互”等前沿技术展开，涵盖了人工智能、智能传感和交互计算等方向的交叉研究。五位专家分别从理论创新、技术突破和系统架构三个维度，全面展现了具身智能感知交互的最新研究成果和实践应用。为与会师生呈现了一场精彩纷呈的学术盛宴。在报告环节中，专家们深入浅出的讲解不仅拓展了与会者的学术视野，更激发了大家对学科交叉研究的浓厚兴趣。随后的互动交流环节气氛热烈，参会师生积极提问，专家们耐心解答，思想的碰撞迸发出智慧的火花。此次报告会极大地鼓舞了师生们的科研动力，为师生搭建了一个难得的学习与交流平台，促进了知识与思想的碰撞与融合。

责任编委 任传贤

CCF-CV 走进企业系列交流会

第 32 期 中国航天科工集团第十研究院

为进一步深化计算机视觉领域的产学研协同创新，推动前沿技术与航天产业深度融合，中国计算机学会计算机视觉专委会（CCF-CV）携手中国航天科工集团第十研究院（以下简称“十院”）举办走进企业活动，该活动于 2025 年 9 月 6 日在贵阳成功举办。

活动伊始，十院科技委秘书长**刘莉**发表开幕致辞，对计算机视觉专委会组织本次活动表示衷心的感谢，并对远道而来的高校专家表示热烈欢迎。并简要介绍了本次活动的举办背景与核心目标，强调了技术交流对航天领域创新发展的重要意义。随后，计算机视觉专委会副秘书长**魏秀参**教授代表计算机视觉专委会致辞，分享了专委会在推动技术成果转化、促进产学研合作方面的努力与规划，为整场活动营造了积极的交流氛围。

西安邮电大学**侯志强**教授、中山大学**郭裕兰**教授、南京邮电大学**周全**教授、南京邮电大学**高广谓**教授、北京航空航天大学**胡峻林**副教授，五位专委会执行委员围绕计算机视觉领域的前沿技术与航天应用方向，先后发表主题报告，分享最新研究成果与实践经验。报告分享结束后，交流研讨环节随即展开。在十院的组织下，参会人员围绕技术落地应用、航天领域特殊需求等话题积极互动，现场讨论氛围热烈，大家各抒己见，为后续技术与产业的结合提出了诸多建设性思路。午餐过后，下午的交流座谈在航天电器举行，十院旗下多家三级单位与到访专家展开面对面沟通，针对航天测试、航天电器等具体领域的合作方向与技术需求深入交流，进一步明确了双方未来合作的重点。

本次 CCF-CV 走进企业之十院活动顺利落下帷幕。此次活动不仅为高校与企业搭建了高效的沟通平台，更推动了计算机视觉前沿技术与航天产业实际需求的精准对接，为双方后续深化合作、共同助力航天事业高质量发展奠定了良好基础。在此，特别感谢十院对本次活动的承办支持，以及各位专家的精彩分享。CCF-CV 走进企业系列活动将持续前行，为推动我国计算机视觉领域技术创新与产业升级贡献力量！

责任编委 潘金山

第 24 期 多模态感知与三维理解：从物理建模到主动认知

CCF-CV 视界无限系列研讨会

2025 年 6 月 15 日，由中国计算机学会（CCF）计算机视觉专委会主办、北京邮电大学人工智能学院承办的 CCF-CV “视界无限”系列研讨会第 24 期——“多模态感知与三维理解：从物理建模到主动认知”在北京成功举办。

本次研讨会汇聚了来自全国多所知名高校与研究机构的专家学者。会议有幸邀请到中国科学院计算技术研究所的**陈熙霖**研究员，中国科学院大学的**蒋树强**研究员，中国科学院自动化研究所的**张兆翔**研究员，大连理工大学的**樊鑫**教授，北京理工大学的**杨健**教授，北京大学的**施柏鑫**长聘副教授，以及西北工业大学的**戴玉超**教授作主题报告，围绕多模态融合、三维建模、主动感知等前沿课题展开深入交流。研讨会由北京邮电大学人工智能学院**郭亨**研究员、**李思**副教授，南开大学计算机学院**杨巨峰**教授共同担任执行主席。

会议伊始，北京大学**查红彬**教授发表开幕致辞。他指出，多模态感知与三维理解已成为智能视觉领域发展的关键方向，亟需在跨模态、跨任务的深度融合中实现创新突破。青年学者要勇于打破学科壁垒，既要注重理论探索，也要面向实际任务，推动研究成果走向应用落地。

蒋树强研究员以《浅谈具身的世界与场景记忆》为题作报告，指出场景记忆是具身智能实现认知与预测的核心，其本质是动态、有参照的空间表示，依赖于持续学习与多模态运动感知。同时，场景建模应具备“多重模型”与“参考系嵌套”能力，从而支持从自我与客体两个视角出发，构建具备推理能力的认知体系。

张兆翔研究员以《世界模型：重建、生成与推演》为题作报告，提出构建具身智能的世界模型，关键在于实现对环境的理解、记忆与预测的有机统一。世界模型可通过三阶段框架建模：真实场景重建、多元环境生成、未来状态推演。此外，世界模型的未来趋势应成为机器人、自动驾驶等 AI 系统感知-认知-行动框架的核心支撑，实现从观测到建模的智能闭环。

樊鑫教授以《海洋环境多模态精准感知的思考与尝试》为题作报告，复杂环境中的水下智能体正逐步从粗粒度感知走向精细状态感知，同时由传统的多模态融合迈向更具时空协同特性的感知机制。为满足水下机器人等系统在复杂任务中的感知与决策需求，构建面向海洋具身智能的多模态感知开放平台显得尤为关键。该平台需具备多模态数据采集、精细地图构建与自主认知能力，为水下智能体提供系统性的感知支撑。

杨健教授以《内镜引导手术导航的研究与应用》为题作报告，指出内镜导航技术对实现微创精准治疗具有重要的现实意义。然而，当前的导航系统仍面临三大技术挑战：一是内镜定位易发生“迷路”问题；二是二维图像缺乏深度信息支撑；三是手术过程中组织形变难以有效量化。突破上述关键问题将显著提升内镜在头颈肿瘤手术中的精准性与安全性，对推动智能医疗设备的发展具有重要意义。

施柏鑫长聘副教授以《神经形态多维度高实时摄像》为题作报告，指出传统成像技术在获取场景几何结构、反射率以及光谱信息等多维数据时，常面临数据冗余、计算耗时等问题。为应对这一挑战，可充分发挥神经形态相机在时间分辨率高、数据稀疏性强等方面的优势，并结合主动光照设计，实现对场景光度变化的高效捕捉，从而推动几何、反射率、光谱等多维信息的高实时成像与应用发展。

戴玉超教授以《事件相机视觉：运动感知与生成》为题作报告，系统介绍了事件相机在高速运动场景中的独特优势及其在多项视觉任务中的应用前景。事件相机具备高时间分辨率和低延迟特性，能够有效应对传统帧式

相机在动态环境下面临的感知瓶颈。深入理解事件视觉的底层工作机制是推动其与传统视觉的深度融合，构建面向具体任务的“感知—生成”闭环系统，实现从信号响应到智能决策转变的重要基础。

在观点分享环节，由中国科学院自动化研究所**王亮**研究员主持，北京理工大学**李长升**教授、浙江大学**崔兆鹏**教授和电子科技大学**刘帅成**教授受邀分享了他们在多模态感知与三维理解方向的思考与观点。自由讨论环节由北京大学**查红彬**教授主持，与会专家学者围绕“多模态感知的融合范式”、“三维重建的几何与语义协同”、“主动认知系统的未来走向”等议题展开深入研讨。

崔兆鹏教授围绕研讨会主题展开分享，指出三维感知正经历从单一图像模态向结构光、黑白相机、红外等多模态融合的演进趋势。多源信息的融合不仅显著提升了建模精度，也为实现更深层次的语义理解与物理属性感知奠定了基础。未来的智能系统不应止步于被动建模，而应具备“主动感知”的能力，即能够基于对环境不确定性的估计，动态决策感知策略，实现探索式建模。例如，在神经辐射场或 3DGS 等系统中引入不确定性估计模块，可有效引导“Next Best View”的视角规划，提升建模效率与完整性。此外，尽管当前如 VGGT 等大模型在多模态输入下展现出较强的数据驱动建模能力，但在建模精度上仍难以达到传统技术的水准，与高精度工业级建模方法在数量级上存在显著差距。因此，在实际应用中需综合考虑建模精度、效率与系统智能性的平衡。

刘帅成教授围绕具身智能时代的技术挑战与产业机遇发言，指出触觉传感器在无视觉或多模态控制等任务场景中，正日益成为关键的感知通道。其团队在 ManiSkill 触觉赛道中，尝试在完全不依赖视觉信息的条件下，通过指尖大小的触觉传感器，仅凭 XY 位置与压力反馈，实现“摸”着用钥匙开锁的复杂操作任务。通过此次实践，团队深刻体会到触觉感知在具身智能系统中的不可替代性。随着具身智能技术的发展，触觉传感器的作用正在从研究走向应用，成为产业界重点探索的方向之一，未来可从材料设计、感知建模、控制策略乃至大模型融合等多个维度推动其技术突破与落地应用。

陈熙霖研究员指出三维重建的价值不应停留于“可视化”的层面，而在于“可应用”——要紧紧围绕任务本质展开建模与理解。物理建模与主动认知并非对立，而是相辅相成：物理建模构筑基础认知结构，主动感知

则引领系统向更智能、更具通用性的方向进阶，两者共同推动对世界的深入理解。建议以“主动认知”为切入点，以“物理建模”服务真实需求，从数据生态的系统化构建到生成模型的引入，每一步都应立足实际，着眼长远。

查红彬教授提出科研应回归本质——让年轻的科研工作者敢于也愿意专注于一件事，沉下心来、步步扎实地深耕细作。在新方向层出不穷、节奏日益加快的当下，科研不应被潮流裹挟，而应坚持聚焦、做透一个问题。因为科研的真正力量，始终藏在“慢一点、深一点”的坚守之中。

在热烈的学术讨论氛围中，本次研讨会圆满落幕。围绕多模态感知、三维理解与物理建模等前沿方向，与会专家学者从理论、方法到实际应用层层深入，激发了广泛而富有价值的思考。通过本次研讨会，与会学生对相关研究方向有了更清晰的认识，进一步了解了当前学术界关注的重点问题和技术进展。

责任编委 杨巨峰

CCF-CV 计算机视觉前沿讲习班

2025 年 8 月 7 日-8 日，第四届 CCF 计算机视觉前沿讲习班在湖北武汉成功举办，吸引了计算机视觉领域的高校教师、研究生、企业技术人员等 200 余人报名参加。本次活动由中国计算机学会（CCF）主办，中国计算机学会计算机视觉专委会（CCF-CV）、武汉大学人工智能学院联合承办，中国科学院计算技术研究所**王瑞平**研究员、武汉大学**夏桂松**教授以及武汉理工大学**朱安娜**副教授担任执行主席。讲习班旨在促进计算机视觉领域的学术交流与高级人才培养，帮助该领域青年从业者提升技术水平，开拓实践视野，掌握最前沿的理论成果和创新应用。本届讲习班共邀请 10 位知名专家报告前沿学术进展，帮助学员全面学习并系统掌握计算机视觉前沿理论、方法与技术。

讲习班开班仪式由执行主席、CCF 计算机视觉专委会副秘书长**朱安娜**主持，她介绍了本次活动的目的、组

织、课程涵盖范围以及报名情况，并对参加讲习班的讲者和学员们表示热烈欢迎。

武汉大学人工智能学院副院长（主持工作）**夏桂松**教授代表本次活动的承办方发表感谢致辞，对参加讲习班的专家和学员表达了热烈欢迎和诚挚谢意，并对 CCF 计算机视觉专委会的信任与支持表示了衷心的感谢，表示会在这份信任和鞭策之下全力做好各项保障工作，为大家提供在交流中碰撞思想、在研讨中凝聚共识的学术交流平台。

随后，讲习班全体成员进行了合影留念。

华中科技大学**王兴刚**教授为学员讲授第一课。王老师以“面向视觉理解生成和规划的高效率表征学习”为主题，从目标检测分割等视觉理解问题、文生图等视觉生成问题、自动驾驶与机器人操作等视觉规划问题，系统讲解对应的视觉表征学习机制，并结合课题组中基于 Transformer 的高效率视觉理解工作 YOLO-World、ViTMatte 和 ViTGaze，高效率跨模态表征理解与推理工作 WeakCLIP 和 WeakSAM，面向高效率视觉生成的表征学习 VA-VAE、DiffusionDrive，视觉表征提取网络线性架构探索工作 Vision GLA 和 Diffusion GLA，面向决策的表征学习和端到端自动驾驶工作 M2Diffuser 和 DiffusionDrive 等展开讲解。

南京大学**王利民**教授以“InternVideo 系列大模型与评测基准”为主题，深入探讨了视频作为人工智能基础数据形态的重要性，并围绕 InternVideo 系列大模型的技术发展、最新进展与评测基准进行介绍与讲解。王老师首先回顾了生物智能的进化历程，指出视频数据因其丰富的时空信息，已成为推动人工智能发展的关键基础。然后，他详细介绍了 InternVideo 系列模型的发展历程及相关技术原理，并展示了最新发布的 InternVideo2.5 在细粒度感知、超长时建模和流式化交

互方面的突出能力。接着，王老师介绍了涵盖短视频、流视频、长视频等场景的通用视频理解全面评测体系 VidBench 的研究动机、整体思想和基准评测结果。最后，他展望了视频大模型与评测基准发展趋势，提出视觉时空智能概念。

香港中文大学教授、上海人工智能实验室领军科学家**欧阳万里**教授以“AI for Science- 机遇与挑战”为题，深入解读了人工智能推动科学研究范式变革的核心思路与当前研究成果。欧阳老师认为，科学领域的重大突破往往源于工具的革新，而非单纯的观点更新，而人工智能正成为助力多个学科实现突破的关键力量。报告中，他介绍了 AI 技术在生命科学、神经科学、物理化学科学、地球科学的进展，比如预测蛋白质、RNG 结构、合成化学分子、气象预报等。随后，他分析了 AI 在工程科学中的应用工作，例如 AI 辅助设计飞行器、预测发电厂发电量等实际应用。最后，他还提出“共建共享共创，打造科学发现的 Scaling Law”，并分享了上海人工智能实验室科学多模态大模型 Inter S1、书生科学发现平台 Intern Discovery 工作。

南京大学、计算机软件新技术全国重点实验室**吴建鑫**教授从工程角度深入探讨了“神经网络量化”这一对工业界尤为重要的课题。吴老师以简明直白的方式解释了神经网络量化的本质，同时指出在当前大模型背景下，算力相比显存、带宽和能耗，并非最紧迫的瓶颈问题，从而引出了“算术密度”(arithmetic density)的概念，明确揭示了量化为何成为技术发展的关键所在。在讲解了经典的神经网络量化方法 Post Training Quantization (PTQ)和 Quantization-Aware Training (QAT)后，吴老师进一步介绍了其团队的研究成果，包括针对 PTQ 的 QwT 系列方法和针对 QAT 的 GPLQ 方法。

北京大学助理教授**袁粒**以“多模态生成、理解及统一架构基础知识与前沿展望”为题，深入探讨了大模型在多模态智能方向的发展路径与底层逻辑。袁老师指出，大模型走向统一架构是技术演进的必然趋势，其中语言与视觉的本质差异决定了它们在建模方式上的不同需求：语言作为人类高度抽象的产物，本身已高度压缩，而视觉信息则需要强力压缩以适应计算框架。接着，他深度解析了自回归和扩散模型两种建模方式的基础知识和各自优势，并针对生成模型为什么重要、如何实现生成和理解统一的原生框架等问题，给出了自己的见解。他认为将自回归模型作为统一架构的主干，用以承担逻辑推理任务，而 Diffusion 模型则作为结构推理的模块嵌入其中，实现对视觉等复杂模态的建模将会是未来多模态模型发展的主流。

清华大学**鲁继文**教授以“视觉感知与自动驾驶”为题，首先回顾了自动驾驶与视觉感知的发展历程，他认为自动驾驶是研究和测试新一代视觉感知系统的绝佳应用。随后，鲁老师从自动驾驶感知任务和自动驾驶生成两部分进行讲解。面向自动驾驶的视觉感知领域，他介绍了多种视觉感知方法在三维目标检测，三维占用预测，三维语义分割等自动驾驶核心感知任务中的优缺点与应用前景。然后指出自动驾驶融入生成方案的重要性并介绍了依靠生成式 AI 产生大量仿真数据的工作。他充分总结了该领域的研究现状，指出了现有工作的缺陷，深入分析自动驾驶的底层逻辑、上限和核心，强调自动驾驶安全，包括感知传感器和场景变换等，是未来自动驾驶领域的重要研究问题。

西北工业大学**戴玉超**教授以“动态场景三维重建与生成”为题，指出动态场景三维重建致力于从连续视频观测恢复所观测场景随时间变化的三维几何结构和外观信息。他向与会学员介绍了动态场景三维重建从显式优化方法到隐式表示方法再到生成方法的发展历程。随后，他围绕动态场景三维重建与生成，在显式优化方法下探讨了单一物体稀疏重建、多物体稠密重建、复杂场景稠密重建，在隐式学习方法下聚焦动态场景新视角合成和三维重建等问题，并对基于生成模型的重建方法、重建与生成的结合和本领域的开放问题与发展趋势进行了展望。

西安交通大学**孟德宇**教授以“机器学习的‘变’与‘不变’”为题，指出在深度学习快速迭代的浪潮中，前沿研究应聚焦于从“变”的角度构建机器学习方法，同时，从机器学习的基础研究视角，也要关注其存在更为本质的“不变性”规律与内涵。他从数据、模型、算法三个层面介绍了课题组针对机器学习中的“不变性”所展开的研究。数据层面，介绍了针对高维标记空间低维不变性隐空间提炼的“标记分布建模”理论与方法；模型层面，介绍了针对网络基础卷积模块旋转 - 尺度 - 仿射等变性结构刻画的“参数化卷积”理论与方法；算法层面，介绍了针对机器学习方法超参设置不变性规律提炼的“模拟学习方法论”理论与方法，凝练机器学习方法如何从“变”中提炼其“不变”内涵的方法论，为机器学习的基础研究与工程应用提供了可参考的视角。

南京大学**俞扬**教授以“大模型背景下的强化学习”为题，介绍了强化学习在大模型时代的发展脉络与前沿进展。俞教授首先简要回顾了强化学习的基本概念，即通过交互学习以最大化长期奖励，并结合 PPO 方法阐述了强化学习在大语言模型优化中的实际应用。他指出，强化学习虽非唯一选择，但在模型调优与行为控制中仍发挥着不可替代的作用。为应对现有方法在奖励设计上的局限，俞教授介绍了 ReMax 等改进型 RL 策略，并指出近年来逐步涌现出专为 LLM 定制的强化学习方法，着力于构建更合理、更通用的奖励机制。俞教授强调，强化学习不仅在提升大模型性能方面日益重要，大模型反过来也为强化学习提供了更强的泛化能力与问题建模能力，预示着两者融合将成为未来智能体发展的重要趋势。

香港大学助理教授**李弘扬**以“Robotic Manipulation Fundamentals and Applications: A Tutorial”为题，为学员们讲授具身智能系统相关问题。他首先介绍了具身智能系统的问题设定、任务分类、基准和评估准则。随后，又全面深入地分析了目前具身智能系统研究在基准评估、数据、算法范式三个层面的挑战，指出基准评估方面需要自动化评估和智能排序；数据方面缺乏高效的采集流程；算法范式方面则缺少更高效的算法范式。针对现存挑战，他向与会学员分享了课题组的代表工作，如大规模标准化的基准数据集 Agibot-World；算法 Go-1-Pro、UniVLA 和 RoboDual，并对未来具身智能系统研究进行了展望，表示该领域有极大的研究空间。

为表达对报告专家的诚挚谢意，CCF-CV 专委会与承办单位精心设计制作了荣誉证书并在报告现场进行颁发。

第四届 CCF 计算机视觉前沿讲习班为计算机视觉领域的广大师生和工程师提供了一个与专家学者近距离交流学习的宝贵机会，大家对专家所讲的课程内容产生了极大兴趣，现场互动非常活跃，学术氛围浓厚。

中国科学院计算技术研究所研究员、讲习班执行主席、CCF 计算机视觉专委会秘书长**王瑞平**进行了总结发言。他首先祝贺本次活动取得了圆满成功，并对各位讲者、学员和承办单位武汉大学人工智能学院表示了衷心感谢。最后，全体学员领取了结业证书并合影留念，记录下了第四届 CCF 计算机视觉前沿讲习班的难忘瞬间。

责任编委 朱安娜

CCF-CV 计算机视觉前沿进展研讨会

2025 年 8 月 9 日，中国计算机学会计算机视觉专委会（CCF-CV）年度学术研讨会 RACV（Recent Advances on Computer Vision）在武汉成功召开。RACV 定位为国内计算机视觉领域的小规模研讨会，通过定向邀请方式汇集领域专家，深度研讨计算机视觉领域中的若干核心问题并形成进展报告。研讨会试图通过务实、开放与平等的对话与讨论，深入发掘相关研究领域潜在的问题，为广大的科研人员提供观察问题的新视角与新观点。

专委会主任、中国科学院计算技术研究所陈熙霖研究员致辞

研讨会开幕式由专委会秘书长、中国科学院计算技术研究所**王瑞平**研究员主持，专委会主任、中国科学院计算技术研究所**陈熙霖**研究员和武汉大学副校长**龚威**教授致开幕辞。

武汉大学副校长龚威教授致辞

专委会秘书长、中国科学院计算技术研究所王瑞平研究员主持研讨会

根据专委常委会前期的讨论票选，本次会议设置了 4 项研讨主题。每项主题由三位专家进行组织，再由引

导发言嘉宾进行报告发言，随后由主题分享嘉宾进行观点分享，之后所有与会人员进行自由讨论。

操晓春教授、左旺孟教授、李弘扬助理教授主持主题一讨论

8 月 9 日上午，首先进行了主题一“大数据驱动 AI 能力的极限与局限：Scaling Laws and Beyond”的研讨。该主题由中山大学**操晓春**教授、哈尔滨工业大学**左旺孟**教授和香港大学**李弘扬**助理教授组织，邀请了中国科学院自动化研究所**张兆翔**研究员、清华大学**代季峰**副教授、智源人工智能研究院**王鑫龙**博士和华中科技大学**白翔**教授 4 位嘉宾进行引导发言。与会嘉宾围绕如何有效延续 Scaling Laws、如何在大数据驱动 AI 研究中发展学习算法、生成理解统一模型领域网络架构有哪些难点、未来大数据驱动 AI 的学习范式是什么、大模型的安全与对齐领域有哪些难点、大数据驱动 AI 有哪些发展和应用领域等问题进行了精彩的讨论和观点分享。

南开大学侯淇彬副教授主持主题二讨论

主题二“多模态大模型的制胜之钥：模型、模态，亦或推理”由百度计算机视觉首席科学家**王井东**博士、中国科学院自动化研究所**赫然**研究员和南开大学**侯淇彬**副教授组织，邀请了 IDEA 研究院讲席科学家**张磊**博士、上海人工智能实验室领军科学家**乔宇**教授和华中科技大学**白翔**教授 3 位嘉宾进行引导发言。嘉宾们围绕多模态大模型的核心未来会如何演化、如何实现更深层次鲁棒的语义对齐、短期内最可能取得突破的方向是什么、如何提升模型生成内容的事实准确性和可验证性、如何提升模型在复杂动态场景中的推理和连贯理解能力、如何高效地从海量异构多模态数据中获取整合和更新知识等议题展开了深入探讨。

南京理工大学舒祥波教授、中山大学郭裕兰教授主持主题三讨论

8 月 9 日下午继续进行了主题三“空间智能：理解、建模与交互新范式”的研讨。该主题由上海科技大学**虞晶怡**教授、南京理工大学**舒祥波**教授、中山大学**郭裕兰**教授组织，邀请了中山大学**李冠彬**教授、浙江大学**崔兆鹏**研究员和香港大学**李弘扬**助理教授 3 位嘉宾进行引导发言。嘉宾们围绕如何构建空间智能理论框架、如何获取高质量数据并高效利用、如何探索高效三维表征形式、如何构建时空统一的 4D 世界模型、如何突破空间记忆建模与任务规划、如何构建符合物理定律的时空因果推理框架等议题展开了深入探讨。

主题四“视觉扩散模型与内容生成”由中山大学**赖剑煌**教授、中山大学**谢晓华**教授、北京大学**袁粒**助理教授组织，邀请了香港大学**赵恒爽**助理教授、香港大学**刘希慧**助理教授和北京大学**袁粒**助理教授 3 位嘉宾进行引

导发言。与会嘉宾们围绕扩散模型和自回归模型的优劣与融合可能、视觉生成扩散模型中的理论缺陷与现实冲突、视频生成的时空建模挑战、视频生成的“抽卡”过程如何优化、视频生成的质量评估体系是否存在“客观真实”与“主观满意”的统一标准、训练数据稀缺瓶颈与出路等议题展开了深入探讨。

中山大学赖剑煌教授、中山大学谢晓华教授、北京大学袁粒助理教授主持主题四讨论

本次研讨会深入探讨了本领域最前沿研究问题，主题分享视角广阔，自由讨论热情激烈，参会嘉宾们纷纷表示本次会议内容丰富，收获良多。按照计划，组委会后续将整理相关主题的发言与讨论文稿，形成观点性文档进行发布，把讨论从线下延伸到线上，欢迎更多专家学者积极参与。

本次研讨会由武汉大学**夏桂松**教授团队承办，赞奇科技和 AutoDL 提供赞助。会议最后安排了特别致谢环节，专委会**陈熙霖**主任、**王亮**副主任、**刘青山**副主任分别为承办单位和赞助单位颁发感谢牌。

责任编委 黄岩

CCF-CV 常务委员会 2025 年度工作会议召开

2025 年 8 月 10 日于武汉召开中国计算机学会计算机视觉专委会 (CCF-CV) 常务委员会 2025 年度工作会议。本次常委会工作会议在武汉东湖宾馆举行，会议邀请到专委会顾问委员**查红彬**教授、**王蕴红**教授参会做工作指导，专委会主任**陈熙霖**研究员主持会议，常务委员会委员参会，秘书处成员列席。

首先，专委会主任**陈熙霖**研究员带领大家学习了 2025 年 4 月 26 日新华社发布的《习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展》全文内容，并进行了深入解读。

随后，**王瑞平**秘书长重点汇报了专委会几项特色品牌活动的进展，包括“走进高校”、“走进企业”、“视界无限”以及 RACV2025 研讨会等。之后，介绍了秘书处纳新和成员职责分工调整情况，PRCV 入选 CCF 高质量国际学术会议培养计划相关情况，专委会关于《CCF 推荐国际学术会议和期刊目录》的调整建议，以及《计算机视觉十讲》出版和《计算机视觉前沿创新战略研究报告》撰写的进展情况。秘书长还介绍了以第四届计算机视觉前沿讲习班为代表的专委会科普教育活动，以及专委会年度奖励、新委员增选等事项的推进情况和后续安排。

接下来，专委会顾问委员和常委们围绕 RACV 议题征集和组织方式、专委会新委员增选、专委会奖励、PRCV 发展规划、专委会委员参与 CCF 事务等议题展开了充分讨论，并就各项活动事项提出了建设性意见和建议。

最后，会议在热烈的交流讨论氛围中结束。

责任编委 马伟

专题综述

CAST：从 RGB 图像重建组件对齐的 3D 场景

姚凯欣[*1,2] 张龙文[*1,2] 严新豪[1,2] 曾焱[1,2] 张启煊[†1,2] 杨卫[3] 许岚[‡1] 顾家远[‡1] 虞晶怡[‡1]
1. 上海科技大学 2. 影眸科技 3. 华中科技大学

本文是上海科技大学、影眸科技和华中科技大学团队合作研究的成果，发表在SIGGRAPH 2025并获得了最佳论文奖。从单个RGB图像中恢复高质量的3D场景是计算机图形学中的一项挑战性任务。现有方法通常受限于特定领域或低质量的物体生成。为了解决这些问题，论文提出了CAST（**C**omponent-**A**ligned 3D **S**cene recons**T**ruction from a Single RGB Image），一种用于3D场景重建的新方法（图1）。CAST首先从输入图像中提取物体级别的2D分割和相对深度信息，然后使用基于GPT的模型分析物体间的空间关系。这使得场景中物体如何相互关联能够被理解，从而确保更整体一致的重建。CAST接着采用一个具备遮挡感知的大规模3D生成模型，独立生成每个物体的完整几何形状，使用Masked Auto Encoder（MAE）和点云条件来减轻遮挡和部分物体信息的影响，确保与源图像的几何形状和纹理准确对齐。为了使每个物体与场景对齐，姿态对齐生成模型计算所需的变换，使得单个生成的物体网格（mesh）能够准确放置并融入到场景的点云中。最后，CAST采用一个基于物理的校正机制，利用细粒度关系图生成约束图，指导物体姿态的优化，确保物理一致性和空间一致性。通过利用有符号距离场（SDF），该方法有效解决了遮挡、物体穿透和浮动物体等问题，确保生成的场景准确反映真实的物理交互。实验结果表明：CAST显著提升了单图像3D场景重建的质量，在场景理解和重建任务中提供了增强的真实感和准确性。CAST具有实际应用价值，在虚拟内容创作中，例如沉浸式游戏环境和电影制作，可以将现实世界无缝集成到虚拟景观中。此外，CAST还可以用于机器人领域，实现高效的真实到模拟工作流，并为机器人系统提供真实、可扩展的模拟环境。

一、导言

人类生活在清晰的关系网络中——家庭、朋友、同事——这些网络指导我们的决策和行为。这些联系塑造了我们的世界并赋予其结构。类似地，空间中的物体也在其自身网络中发挥作用[42]，但较少被注意到。它们并非孤立存在；其放置、设计和材料源于物理限制、功能作用和人类设计意图，并影响我们移动、互动和感知空

图 1　CAST 从单张图像中将多样化的 3D 场景生动呈现，展现出物体间丰富的物理和空间相互作用关系。

间的方式。例如，椅子靠在桌子上以获得支撑，杯子放在碟子上，台灯的光线与周围表面相互作用，投射出塑造整个场景的阴影。识别这些关系对于准确的场景解析、建模以及最近的 3D 生成至关重要，确保虚拟环境感觉真实和与现实世界一致。

在从文本或图像提示生成单个物体方面取得了显著进展。神经渲染方法[62,76]优化了隐式表示，而原生 3D 生成器[80,92]则通过端到端学习直接创建 3D 形状和纹理。虽然这些方法在单个物体方面显示出前景，但将它们应用于通过组合物体来生成整个场景时，面临显著的挑战。一个关键挑战是准确的姿态估计。现有方法通常假设物体是视图对齐的，这在真实世界场景中很少见。物体可能以不同的方向出现，受设计、物理或部分遮挡的限制。然而，大多数现有方法优先考虑几何保真度而非姿态对齐，使这一关键方面未得到充分探索。

一个更根本的问题源于物体间空间关系的缺乏。即使姿态相对准确，生成的场景也常常出现物理上不合理的瑕疵：物体相互穿透、漂浮或未能进行必要的接触。这些错误源于缺乏自然地将物体绑定在一起的空间和物理约束，就像人类关系构建我们的社会一样。虽然最近的一些方法[46,93]隐式编码空间关系，使用编码器-解码器架构，但它们仍局限于室内场景等特定领域。其他场景级生成器[21]将物体放置在全局坐标系中，但忽略了它们的相对姿态和依赖关系，进一步损害了真实性和下游应用的可用性，如编辑、动画和模拟。

为此，我们提出了 CAST，一种用于从单个 RGB 图像生成与图像对齐的组件的 3D 场景重建方法。CAST 为单个物体生成高质量的 3D 网格，并生成它们的相似变换（旋转、平移、缩放），确保与参考图像对齐并强制执行物理上合理的相互依赖关系。CAST 首先通过使用 2D 基础模型（例如 Florence-2[81]、GroundingDINO[49]、SAM[64]、Grounded-SAM[65]）处理非结构化 RGB 图像，以开放词汇方式识别、定位和分割物体。现有的单目深度估计器[75]提供部分 3D 点云和物体间空间关系的初步估计，包括相对变换和尺度。

CAST 的第一个核心组件是 3D 实例生成器，它包含两个模块：一个具备遮挡感知的物体生成模块和一个姿态对齐生成模块。物体生成模块采用基于潜在扩散的生成模型，根据分割出的部分图像（和点云）生成高保真物体网格。该模块包含一个遮挡感知 2D 图像编码器，能够推断被遮挡区域，确保稳健地从图像条件中提取特征。为了提高对真实世界点云条件的鲁棒性，我们在训练过程中模拟了带遮挡区域的部分点云，使模型能够有效处理遮挡。姿态对齐模块采用一个生成模型，生成一个变换后的部分点云，与潜在空间中隐式表示的完整几何形状对齐。相似变换是通过生成的变换点云和从相机估计的部分点云推导出来的。与直接姿态回归方法[35,39]不同，我们的方法通过生成来估计变换，捕捉了姿态对齐的多模态性质。

CAST 的第二个核心组件解决了物体间空间关系问题。尽管像素层面进行了对齐，但如果缺乏对物理约束的显式建模，仍可能出现物理上的不合理，如穿透或漂浮。CAST 引入了一个基于物理的校正过程，以确保空间和物理一致性。GPT-4v[1]被用于识别源图像的常识性物理关系，然后利用这些约束来优化物体姿态。这个过程确保重建的场景表现出真实的物理依赖关系，使其适用于模拟、编辑和渲染等应用。

CAST 可以从各种图像生成真实的 3D 场景，无论这些图像是来源于室内外环境、真实世界拍摄还是 AI 生成。与之前的方法[18,46]不同，CAST 通过精心设计的管线，支持开放词汇重建，甚至支持具有挑战性的室外图像。定量上，CAST 在室内数据集 3D-Front[22]上超越了基线方法，在物体级和场景级几何质量方面表现出色。另外，通过视觉语言模型和用户研究进行验证，它在各种图像上表现出优越的感知和物理真实性。

仅凭一张图像，CAST 就能真实地重建场景，包括细致的几何形状、生动的物体纹理，以及更重要的，物体之间的空间和物理相互依赖关系。这一能力使虚拟创作大众化：一个房间或室外空间的单一快照变成了一个实例化的 3D 环境，包含姿态精确的物体，交互自然，并考虑了遮挡。游戏开发者可以将真实世界的设置集成到沉浸式景观中，电影制作人可以毫不费力地生成复杂的虚拟场景——释放创意潜力。除了娱乐之外，CAST 还为更智能的机器人铺平了道路。通过使机器人研究人

员能够从真实世界演示数据中构建数字副本，它能够促进真实到模拟的流程[44,72]，从而实现更高效、可扩展的物理模拟工作流。

二、相关工作

将真实世界场景转换为数字领域，增强了我们理解、重建和与我们周围 3D 世界互动的能力。这种做法在动画、电影、游戏、建筑和制造等行业中得到广泛应用。它使得沉浸式电影体验、历史文物的数字保存以及交互式游戏环境的开发成为可能。例如，詹姆斯·卡梅隆在《阿凡达》（2009）中采用了开创性的 3D 扫描技术，将潘多拉郁郁葱葱、真实的生境带入生活。同样，在游戏行业中，《巫师 3：狂猎》融合了受波兰真实世界地点启发的逼真地形和建筑，将真实的文化和自然元素与富有想象力的开放世界探索相结合。

摄影测量是一种广泛使用的方法，可以高细节地捕捉物理世界并将其转换为数字形式[3,11,27,36,57,58]，但它需要数十到数百张来自多个视角的图像，这既耗时、资源密集，又难以扩展。相比之下，基于单图像的方法更高效、可扩展，只需要一张图像即可轻松从在线存储库获取，无需昂贵的扫描设备或多视图设置。

2.1 单图像场景重建

从单个图像进行场景级重建面临物体多样性、遮挡以及保持空间关系的挑战。一个例子是单目深度估计，即从单个图像推断深度，进一步得到深度点云[7,61,75,85,87]。虽然这提供了有价值的信息，但它在处理遮挡和场景隐藏部分时会遇到困难。为了解决这个问题，新视图合成方法使用辐射场[71,88]和 3D 高斯[67,68]等表示来学习 3D 数据集中的遮挡先验[10,17,25,66]。尽管取得了进展，单目重建方法仍然难以提供详细而精确的场景表示。

有些方法侧重于直接回归场景中的几何形状及其语义标签[12,15,16,26]。这些方法通常依赖于带有物体真值标注的场景数据集，例如 Matterport3D[22]和 3DFront[22]。这些数据集规模通常较小且仅限于室内房间环境。然而，这些方法的前馈性质导致生成的几何形状通常缺乏足够的细节和质量。

为了更好地将真实世界场景转化为数字，其他方法转向基于检索的方法[18,23,28,38,41]，通过在场景中搜索并替换相似物体来提高场景质量。这些方法结合了 GPT-4[1]、SAM[37,65]和深度先验等先进工具来分解场景。虽然这些方法通过集成真实世界物体提高了场景的真实感，但它们受限于所依赖数据集的丰富性和范围。对于超出数据集领域限制的场景，基于检索的方法要么产生错误结果，要么无法找到合适的替换，显著降低了重建场景的质量。

2.2 重建即生成

随着该领域的不断进步，从各种开放词汇图像或文本提示创建高质量 3D 数字资产的能力显著提高。这一进步促使范式转变，将单视图重建问题演变为生成式 3D 合成框架。这种范式转变允许生成 3D 资产而不受限于固定数据集，从而实现更灵活和可扩展的场景重建。

目前大部分 3D 资产生成研究都集中在从 2D 图像生成模型中提取 3D 几何形状[62,70,76]。最近的发展通过纳入多视图图像进行监督[47,48,51,52,74,78]，通常在大型物体数据集（如 Objaverse[20]）上进行训练以增强生成过程中的视图一致性。一些方法根据输入图像直接回归单个物体的形状和外观[32,69]。虽然这些方法取得了令人满意的视觉结果，但它们经常无法再现精细的几何细节。为了提高 3D 几何质量，越来越多工作已完全脱离 2D 监督，转而直接在 3D 资产上进行训练[19,20]。这些方法通过先进的处理技术[80,92,94]生成高质量的物体级几何形状。然而，这些方法侧重于孤立的物体，未能解决场景级挑战，如建模空间层次结构、物体间关系和环境光照。由于建模物体关系、光照和材料的表示复杂性高，场景生成仍处于不发达状态。尽管取得进展，现有方法仍难以生成完全实现、可编辑的 3D 场景。现有范式要么使用视频扩散模型[8,30,31]生成可导航的 2D 投影[9,89]，要么依赖扩散先验通过 3D 高斯飞溅[24,45,79]进行体素场景近似。虽然这些方法产生吸睛的视觉效果，但它们与传统生产管线不兼容，缺乏可编辑网格、UV 映射和可分解的 PBR 材料。

一种更可行的方法是将场景分解为模块化组件——物体、背景和环境，并将它们生成和重新组合成可编

辑的场景图，以实现更大的灵活性和精度。例如，Gen3DSR[21]使用 DreamGaussian[70]进行开放词汇重建。然而，它在处理遮挡、姿态估计和编辑单个物体方面存在困难，同时依赖 2D 模型导致几何细节差和低保真表示。另一项最近的工作，Midi[33]，学习场景中物体间的空间关系，但需要基于真值 3D 网格和标注的数据集进行训练。这种对特定数据集的依赖限制了其可扩展性和对任意场景的泛化能力。

我们的方法与经典的分析-合成方法[91]共享概念基础，因为两者都旨在通过生成对观察到的图像的解释来推断 3D 结构。然而，分析-合成依赖于迭代渲染和像素级优化，而我们的方法利用预训练的生成模型和学习的先验直接合成可信的 3D 场景，通常绕过显式渲染和优化循环，从而提高了可扩展性、效率和对开放世界场景的适配性。

在此基础上，我们提出了一种新颖的场景重建管线，独立生成每个物体并将其对齐到整体的场景中。与现有方法不同，我们的方法保留了准确的几何形状、纹理和一致的空间关系，从而产生了更真实、可靠和可编辑的重建，提高了质量和灵活性。

2.3 具备物理感知的 3D 建模

生成物理上合理的 3D 资产对于确保动画、游戏和机器人等应用中的真实感和功能性至关重要。虽然最近的 3D 生成模型在创建视觉上真实的物体方面表现出色，但它们通常无法达到物理合理性。为了解决这一限制，物理感知 3D 生成模型被开发出来，将物理原理集成到生成过程中。有些方法使用软体模拟来动画化 3D 高斯[82, 96]，或者通过基于物理的惩罚来指导生成关节物体[50]。而另一些方法则通过刚体模拟[13,55,56]或 FEM[29,83]确保自支撑结构。这些方法利用离线或在线物理模拟来检查生成形状的物理有效性，进而指导生成。然而，这些方法通常局限于单个物体，忽略了场景中多个物体之间的相互影响。

将物理约束纳入场景合成更具挑战性，因为涉及更复杂的物体间接触等关系。Yang 等人[86]将物体碰撞、房间布局和物体可达性等约束集成到其场景级生成管线中。然而，它仅限于室内场景合成，并依赖封闭词汇数据库执行形状检索。Ni 等人[59]解决了多视图神经重建中物理不合理性问题。它利用可微分渲染和物理模拟来学习隐式表示。然而，它需要多视图图像作为输入，专注于单个物体，并且主要只解决稳定性问题。相比之下，我们的方法在开放词汇设置下运行，只需要一张输入图像。此外，它考虑了更复杂的物体间关系，特别是支撑和接触，使其更通用，适用于不同的场景。

三、方法概述

从单个图像进行场景级重建是计算机图形学中的一个基本挑战，在动画、虚拟现实和互动游戏中有广泛应用。与专注于孤立物体的物体级重建不同，场景级重建强调多个实体在真实（或风格化）物理下的排列和关系。通过捕捉每个物体的结构、空间关系和上下文线索，这种整体方法能够实现更沉浸的体验、引人入胜的叙事和高效的工作流程。尽管之前的方法探索了使用固定 3D 模板的前馈管线或基于检索的方法[18, 46]，但这些方法往往难以捕捉细微的场景语义和复杂的物体关系。针对这些局限，我们提出了一种以生成驱动的场景重建方法，强调物体关系，从单个未标注的 RGB 图像构建高保真、上下文一致的 3D 环境，无论是来源于真实世界摄影还是合成数据（见图 2）。

我们方法的关键在于对场景上下文信息进行全面的物体关系分析。首先，我们进行物体分割以识别和定位图像中的组成物体。然后，我们获取初步的几何信息（点云），并探索物体间的语义和空间关系。这些预处理结果，作为上下文，为我们后续的物体级生成管线提供了信息，确保每个重建的物体不仅保持其几何保真度，而且在场景中位置正确。最后，我们合成一个考虑物理合理性的整体 3D 环境，实现结构合理的布局和场景元素间真实的交互。

我们的研究聚焦两个主要目标：探索生成模型如何有效捕捉复杂的物体间关系，以从单个图像生成逼真、场景级重建；以及识别整合几何线索和上下文信息的策略，以最大化 3D 重建的准确性和合理性。通过本文研究，我们证明生成方法提供了比传统前馈和基于检索的技术更灵活和稳健的替代方案。CAST 允许对物体级细节和全局场景构成进行细粒度控制，从而简化了动画、

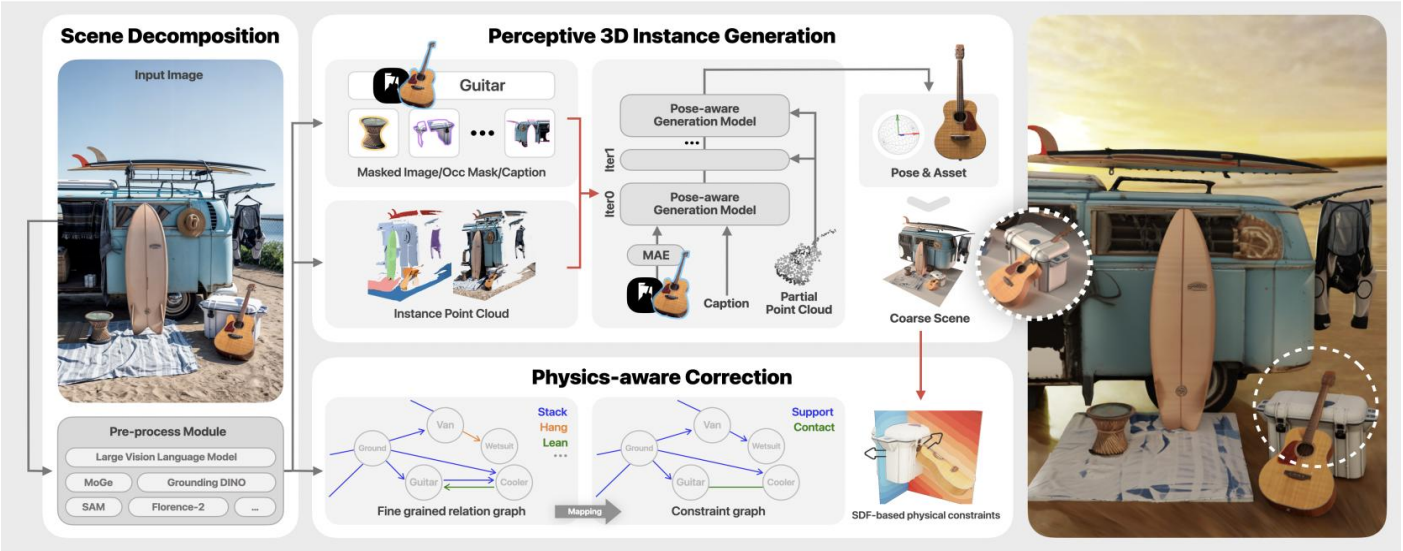

图 2　CAST 管线的概述。首先通过场景分析提取 RGB 输入图像的关键信息，然后通过具备姿态感知的生成过程创建初始 3D 模型。最后基于物理约束的优化促进真实的交互和空间关系，从而生成高质量的网格化 3D 场景。

游戏开发和其他需要 3D 模型的领域的内容创建管线。本文突出了以生成为中心的框架的优势，并为 3D 场景重建的未来发展奠定了基础。它还强调了上下文驱动方法在弥合 2D 图像与沉浸式、交互式虚拟环境之间差距方面日益增长的重要性。

预处理 为了辅助从单个图像进行全面的场景重建，我们首先进行全方位的语义提取，为后续处理提供坚实基础。具体而言，我们使用 Florence-2[81]识别物体，生成它们的描述，并用边界框定位每个物体。然后，我们利用 GPT-4v[1]过滤掉虚假检测，并分离出有意义的组成物体，从而实现不受预定义类别限制的开放词汇物体识别。接下来，我们使用 GroundedSAM-v2[65]为每个识别出的物体$\{o_i\}$生成精细的分割掩模$\{M_i\}$，从而获得精确的物体边界和相应的遮挡掩模，这在物体生成阶段起着关键的辅助作用。除了语义线索，我们还通过提取场景级点云来整合几何信息。我们使用 MoGe[75]生成像素对齐的点云$\{q_i\}$，用于每个物体$\{o_i\}$，　$i \in \{1, \ldots, N\}$，以及场景坐标系中的全局相机参数。这些额外的几何数据随后与每个物体的分割掩模匹配，为最终的 3D 场景重建提供了可靠的结构参考。

四、感知3D实例生成

在从单个 RGB 图像重建高保真 3D 场景的尝试中，一种简单粗暴的方法是使用单图像深度估计或扩散先验等技术直接生成整个场景网格。然而，由于真实世界场景的复杂和交织性质，这种方法在处理遮挡、渲染不可见组件以及准确表示物体关系方面固有地存在困难。因此，如图 3 所示，我们的方法不是直接生成整个场景网格，而是专注于单个物体生成，然后通过精确的对齐来排列物体。这种策略具有以下几个优点：1. 专注于单个物体可确保更高的几何保真度，并允许进行详细建模，从而产生更准确和视觉吸引力的场景组件。2. 在规范化空间中操作可确保生成的资产符合标准化方向和比例，与艺术家定义的坐标系统无缝集成，并促进数字内容创作工具之间的一致性。3. 模块化方法支持各种应用，如编辑、渲染和模拟，从而实现对物体进行独立操作，以获得更大的灵活性和效率。通过将场景重建分解为物体级生成和对齐，我们的方法提高了资产质量和可管理性，同时增强了 3D 场景的整体一致性和功能性。这种方法解决了几何精度和高效后处理等挑战，推动了单图像 3D 场景生成的发展。

场景中的物体级生成同样面临重大挑战，主要原因是场景中物体部分被遮挡以及传感器覆盖范围有限。此外，现有生成方法往往无法协调多个物体，导致场景不一致和不真实。为了克服这些限制，我们提出了一种具备遮挡感知的 3D 物体生成框架，将部分观察结果与全面的场景理解相结合。具体而言，给定图像及其点云，该框架生成一个高质量的 3D 资产，不仅与输入图像相

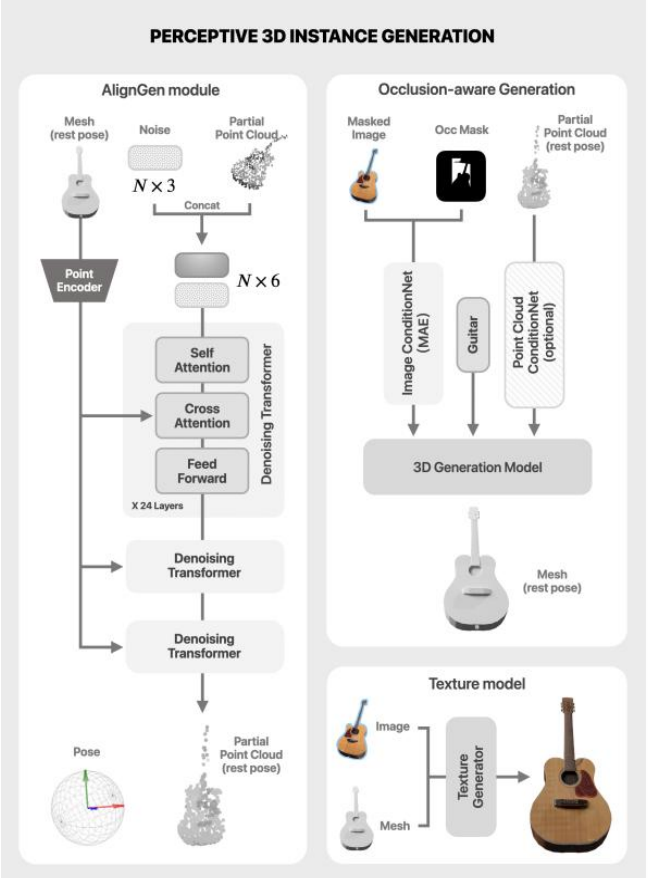

图 3　对齐生成模型（第 4.2 节）、遮挡感知物体生成模型（第 4.1 节）以及纹理生成模型的示意图

似，而且还与规范空间中对应的部分点云准确对齐。此外，我们计算一个变换矩阵，将生成的物体从其规范空间映射回原始场景空间，确保场景内的空间一致性。

我们物体生成过程的一个关键是利用大型生成模型，从部分图像和点云观测中生成整体且高保真度的物体网格。为此，我们首先遵循最先进的原生 3D 生成模型[80,92,94]对大型 3D 生成模型进行预训练。该模型以文本和图像输入为条件。

本文参照现有的生成框架，基于 3DShape2VecSet 表示[92,94]，通过几何变分自编码器（VAE）进行几何生成。这个 VAE 框架将均匀采样的表面点云编码为无序潜在编码，并将这些潜在表示解码为有符号距离场（SDFs）。VAE 编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$ 定义如下：

$$\boldsymbol{Z} = \mathcal{E}(\boldsymbol{X}), \quad \mathcal{D}(\boldsymbol{Z}, \boldsymbol{p}) = \mathrm{SDF}(\boldsymbol{p}), \tag{1}$$

其中 $\boldsymbol{X}$ 表示几何体的表面采样点云，$\boldsymbol{Z}$ 为对应潜在编码，$\mathrm{SDF}(\boldsymbol{p})$表示查询点 $\boldsymbol{p}$ 处的 SDF 值（用于通过 Marching Cubes 方法提取网格）。为了将图像信息有效融入几何生成过程，我们采用 DINOv2[60]作为图像编码器（遵循 Xiang 等人[80]与 Zhang 等人[92,94]的方法论）。几何潜在扩散模型（LDM）形式化表示为：

$$\epsilon_{\mathrm{obj}}(Z_t; t, c) \to Z, \tag{2}$$

其中 ϵ 代表扩散模型（Transformer 架构），Z_t为时间步 t 的带噪几何潜在编码，c 表示 DINOv2 编码的图像特征。我们遵循先前研究[92,94]，在 Objaverse[20]上预训练基础模型。训练后的生成模型ϵ能仅根据图像特征生成精细的三维几何。

4.1 具有遮挡感知的 3D 物体生成

直接使用基于 3D 生成模型面临着相当大的挑战，因为现实世界场景通常存在输入图像中的部分遮挡，这会严重降低生成物体几何形状的质量和准确性。为了解决这个问题，我们利用 DINOv2 的 Masked Auto Encoder（MAE）能力。具体而言，在推理过程中，我们提供一个遮挡掩模$\boldsymbol{M}$ 和输入图像 $\boldsymbol{I}$，使编码器能够通过推断被遮挡区域的潜在特征来处理缺失像素。这形式化为：

$$\boldsymbol{c_m} = \boldsymbol{\mathcal{E}}_{\mathbf{DINOv2}}(\boldsymbol{I} \odot \boldsymbol{M}) \tag{3}$$

其中，$\boldsymbol{M}$ 是一个二进制掩模，指示哪些令牌应该被遮蔽并替换为[mask]令牌。在预训练阶段，**DINOv2** 在随机设置的掩模下进行训练，使其能够根据可见区域稳健地推断缺失部分。因此，在推理过程中，即使物体图像的部分被遮挡，编码器也能有效重建必要的特征，确保生成模型保持高质量和准确性。这种图像条件和遮挡处理的集成对我们的管线至关重要，因为它确保生成的 3D 物体在视觉上与输入图像一致，并在真实地反映集合结构。

规范点云条件 尽管物体生成模型能从输入物体图像产生视觉合理的网格，但由于编码图像条件 $\boldsymbol{c}$ 的高层特性及缺乏像素级监督，生成像素对齐的几何仍具挑战。我们通过额外基于变换到规范坐标系中观测到的部分点云作为生成条件来解决该问题。这种双重条件化确保生成几何不仅与输入图像视觉对齐，更准确反映其内在尺度、形状与深度。在条件化训练期间，我们通过从多

视角渲染每个三维资产来模拟真实局部扫描或估计深度图，从而获取对应 RGB 图像、相机参数与真值深度图。这些 RGB 图像随后通过先进深度估计技术（包括MoGe[75]与 Metric3D[87]）处理，生成估计深度图并投影为部分点云。为保障尺度一致性，我们根据有效深度值的中位数与中位数绝对偏差对 MoGe 与 Metric3D 的估计深度图进行缩放与平移，使其与真值深度图对齐。最终点云被归一化至规范$[-1,1]^3$空间，确保粗略的物体对齐所需的空间表征一致性。

为增强模型鲁棒性及跨现实场景的泛化能力，我们采用数据增强策略：在真值部分点云 $\boldsymbol{p}_{\text{gt}}$（从真值深度图投影以模拟精确深度）与含噪声估计部分点云 $\boldsymbol{p}_{\text{est}}$（从估计深度图投影并对齐以模拟RGB估计噪声深度）间进行插值。数学表示为 $\boldsymbol{p}_{\text{disturb}} = \alpha \cdot \boldsymbol{p}_{\text{gt}} + (1-\alpha) \cdot \boldsymbol{p}_{\text{est}}$，其中$\alpha \in [0,1]$为训练期间均匀采样的权重因子。我们的物体生成器，命名为ObjectGen，在部分点云条件下的形式化表示为：

$$\epsilon(Z_t; t, c, \boldsymbol{p}_{\text{disturb}}) \to Z, \tag{4}$$

其条件化适配方案类似Zhang等人[92,94]的注意力机制。此外，为模拟真实遮挡与数据缺失，我们在不同相机视角的深度图中随机掩码基本图元（如圆形与矩形），从而产生含遮挡与不完整区域的局部点云，进一步提升模型处理不完美输入的能力。本方法的关键设计是保持训练数据集中部分点云与几何的对齐：不同于对增强点云施加随机缩放、平移或旋转的方法，我们对齐的部分点云确保生成模型能更有效契合输入点云。这种对齐约束着模型去紧密遵循物体的实际形状与尺度，从而实现更精确、一致的三维重建。通过对这些良好对齐的部分点云进行条件化，模型在整体尺寸与局部几何细节上均实现卓越对齐。

4.2 生成对齐

每个生成的3D物体都在标准化体积内，并假定一个规范姿态，该姿态可能与图像和场景空间点云不完全对齐。这是因为图像条件使用了例如DINOv2的高级特征，以实现更好的泛化。确保每个物体都被正确变换和缩放以与它在场景中的样子对齐，对于场景组合至关重要。尽管可以采用传统对齐方法，如迭代最近点法（ICP）[2,6]，但它们通常无法考虑语义上下文，导致常出现未对齐和低准确性（见图9）。相反，我们引入了一个对齐生成模型，以场景空间部分点云 $\boldsymbol{q} \in \mathbb{R}^{N\times 3}$ 和规范空间几何潜在编码 $\boldsymbol{Z}$ 为条件。正式地，我们定义我们的对齐生成器AlignGen如下：

$$\epsilon_{\text{align}}(\boldsymbol{p}_t; t, \boldsymbol{q}, Z) \to \boldsymbol{p}, \tag{5}$$

其中 ϵ_{align} 是一个点云扩散变换器，$\boldsymbol{p} \in \mathbb{R}^{N\times 3}$ 是场景空间部分点云转换到规范空间后的版本，与生成的物体网格对齐。$\boldsymbol{Z}$ 是由物体生成模型生成的与 $\boldsymbol{p}$ 对应的物体几何潜在表示。$\boldsymbol{p_t}$ 是时间步 t 处 $\boldsymbol{p}$ 的噪声版本。本质上，生成模型将场景空间部分点云 $\boldsymbol{q}$ 映射到规范的$[-1,1]^3$空间中的 $\boldsymbol{p}$，并与生成的物体网格对齐。我们随后可以使用Umeyama算法[73]从 $\boldsymbol{q}$ 和 $\boldsymbol{p}$ 中恢复相似变换（即缩放、旋转和平移），因为它们是逐点对应的。这一最终步骤在数值上比直接预测变换参数更稳定。

实际上，我们对输入点云 $\boldsymbol{q}$ 和几何潜在 $\boldsymbol{Z}$ 采用不同的条件策略。对于 $\boldsymbol{q}$，我们沿着特征通道维度将输入点云与扩散样本 $\boldsymbol{p_t}$ 拼接起来，使变换器架构能够学习噪声规范帧部分点云与世界空间部分点云之间的显式对应关系。对于几何潜在 $\boldsymbol{Z}$，我们应用交叉注意力机制将其注入点扩散模型（Transformer架构）。这种方法确保模型有效整合空间和几何关系。此外，由于对称性和重复的几何形状，对于给定的 $\boldsymbol{q}$ 和 $\boldsymbol{Z}$ 可能存在多个有效的 $\boldsymbol{p}$。我们的扩散模型通过采样多个噪声实现并聚合结果变换来解决这个问题，以选择置信度最高的表示。

4.3 迭代生成过程

回想一下，在我们的设计中，物体点云最初无法用于物体生成，因为它以场景空间表示，而我们的物体生成模型需要规范空间点云进行条件化。仅仅依赖图像线索进行物体生成通常无法产生像素对齐的几何形状。幸运的是，我们的设计通过联合迭代过程，实现了物体生成和对齐模块的无缝集成。这种集成确保了每个生成的3D物体不仅在视觉上与输入图像一致，而且在场景中准确地定位和缩放。这个迭代的工作流可以概括为以下三个关键步骤（k 表示迭代次数）：

步骤1：物体生成。对于带掩码的物体图像，物体生

成模块（第4.1节）基于DINOv2提取的图像特征 c 与规范坐标系中的对齐点云 $\boldsymbol{p}^{(k)}$，合成几何潜在编码 $\boldsymbol{z}^{(k)}$。我们初始化 $\boldsymbol{p}^{(0)}$为场景空间点云 $\boldsymbol{q}$，并设置点云条件化比例因子 $\beta^{(k)}$ 随迭代进程从0逐步增至1，使部分点云的影响随时间逐步增强。形式化表述为：

$$z^{(k)} = \text{ObjectGen}\left(c, \boldsymbol{p}^{(k)} \otimes \beta^{(k)}\right). \quad (6)$$

步骤2：对齐。随后，生成式对齐模块（第4.2节）接收新生成的几何潜在编码 $\boldsymbol{z}^{(k)}$ 与场景坐标系中的部分点云 $\boldsymbol{q}$，预测变换后的规范空间部分点云 $\boldsymbol{p}^{(k+1)}$：

$$\boldsymbol{p}^{(k+1)} = \text{AlignGen}\left(\boldsymbol{q}, \boldsymbol{z}^{(k)}\right). \quad (7)$$

此变换后的点云 $\boldsymbol{p}^{(k+1)}$作为改进的对齐参考用于下一次迭代。通过生成式变换模型，确保缩放、旋转和平移调整既精确又符合语义理解。

步骤3：细化。利用更新后的部分点云 $\boldsymbol{p}^{(k+1)}$，系统可估算新的相似变换以优化生成几何在场景中的对齐。更新后的点云随后反馈至物体生成模块进行下一次迭代，实现几何精度与空间定位的渐进式提升。

迭代循环在几何生成与变换估计间交替进行，持续直至满足收敛标准。当变换参数变化低于预设阈值或达到最大迭代次数时即实现收敛。最终获得既视觉精确又与输入数据几何对齐的高保真三维物体。通过将物体生成模块与对齐生成模块紧密集成于迭代框架内，我们的方法有效平衡了美学保真度与几何精确度。该联合生成过程综合利用视觉与深度信息，确保每个三维资产均具备高质量特性与精确定位能力。因此，该流程为构建物理正确且视觉一致的三维场景奠定了坚实基础，助力编辑、渲染与动画等下游应用。

确定物体几何后，我们采用最先进的纹理生成模块创建逼真表面细节。遵循成熟纹理合成流程[92,94]，我们分配UV贴图并训练生成网络将细致纹理绘制至三维网格。该模块能稳健处理各种增强条件下的图像，确保最终纹理即使存在遮挡或有限可见性仍与输入外观匹配。

五、基于物理的校正

第4节详述的管线独立生成每个3D物体实例，并根据单个输入图像估计其相似变换（缩放、旋转和平移）。虽然我们提出的模块实现了高精度，但生成的场景有时并不物理合理。例如，如图4所示，一个物体（例如吉他）可能与另一个物体（例如冷藏箱）相交，或者一个物体（例如冲浪板）可能在没有任何支撑（例如来自货车）的情况下不自然地漂浮。

为了解决这些问题，我们引入了一个基于物理的校正过程，该过程优化了物体的旋转和平移，确保场景符合常识性的物理约束。校正过程受物理模拟（第5.1节）启发，并公式化为基于从图像中提取的场景图（第5.3节）的优化问题（第5.2节）。

5.1 刚体模拟简要介绍

我们介绍物理（刚体）模拟的基本原理，这些原理启发了我们的问题建模，并使框架更易于应用于游戏和机器人等下游应用。更详细的介绍请参阅 Bender[5]。

在刚体模拟中，世界被建模为常微分方程（ODE）过程。在每个模拟步骤中，首先应用牛顿-欧拉（微分）方程，它们描述了刚体在无接触情况下的动态运动。进行碰撞检测以找到刚体之间的接触点，这些点是确定接触力的必要条件。对于接触处理和碰撞解决，通常有几个条件：无穿透约束以防止物体重叠，摩擦模型确保接触力在其摩擦锥内，以及互补约束以实现变量之间特定的析取关系。求解器用于解决包含方程和不等式的系统，随后更新每个刚体的速度和位置。

增强物理合理性的直接方法是利用现成的刚体模拟器来处理场景，从之前管线估计的初始状态开始，并在模拟后获得静止状态。然而，这种方法存在几个挑战。

（1）部分场景：由于2D基础模型的限制，某些物体可能缺失，因此无法重建。在完全物理规则下模拟部分场景可能导致次优结果。

（2）不完美几何形状：虽然我们的3D生成模型生成高质量几何形状，但仍可能出现微小缺陷。刚体模拟器通常需要对物体进行凸分解[53,54,77]，这会引入额外的复杂性和超参数。过度细粒度的分解可能导致非平坦、复杂表面，导致物体在模拟过程中掉落或意外移动。相反，粗粒度分解可能由于视觉和碰撞几何体之间的差异而导致视觉上的浮动物体。

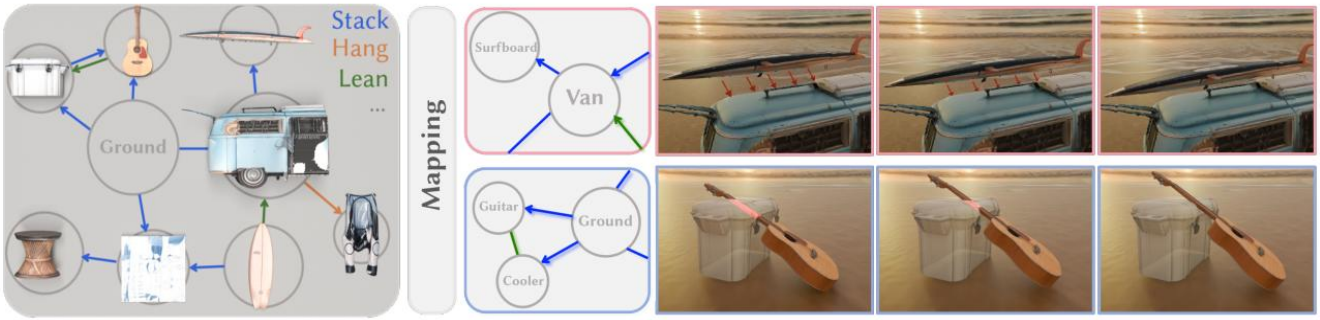

图 4　通过来自细粒度的关系图映射得到的约束图实现基于物理的校正。
右上方：浮动的冲浪板支撑在货车上。右下方：穿透的吉他和冷藏箱分离。

(3) 初始穿透：尽管姿态估计精度高，但在初始状态下可能存在显著的物体间穿透。这些穿透会给标准刚体求解器带来不稳定性，在某些情况下，如果求解器非定制，甚至会导致无解的情况。

因此，我们提出了一种定制和简化的“物理模拟”，以优化物体姿态，确保场景符合从单个图像中推导出的常识性物理原理。请注意，我们的方法不模拟完整的动力学。例如，一个物体可能无法在其当前姿态下长时间保持稳定。然而，我们认为，我们优化后的结果可以作为后续物理模拟的可靠初始化。

5.2 问题公式化和物理约束

我们将物理感知校正过程构建为一个优化问题，旨在最小化针对物体间相互关系约束的总成本：

$$\min_{T=\{T_1,T_2,\ldots,T_N\}} \sum_{i,j} C\left(T_i, T_j; \boldsymbol{o}_i, \boldsymbol{o}_j\right) \tag{8}$$

其中N为物体数量，T_i表示第 i个物体 $\boldsymbol{o}_i$ 的刚体变换（旋转与平移）。C 为成本函数，表征 $\boldsymbol{o}_i$ 与 $\boldsymbol{o}_j$ 间的关系。需注意成本函数随关系类型而变化。

受物理仿真启发，我们将关系分为两类：接触（contact）与支撑（support）。这些关系通过视觉语言模型（VLM）辅助识别，详见第5.3节。

（1） **接触关系**。描述两个物体 $\boldsymbol{o}_i$ 与 $\boldsymbol{o}_j$ 是否处于接触状态。令 $D_i(p)$ 表示物体 $\boldsymbol{o}_i$ 在点p处的符号距离函数（SDF），用于定义约束条件。$D_i(p) = D_j(p) = 0$表明 p 是 $\boldsymbol{o}_i$与 $\boldsymbol{o}_j$ 的接触点。当$D_i(p) = 0$（即 p 为 $\boldsymbol{o}_i$表面点）时，$D_j(p) < 0$表示物体间穿透，而$D_j(p) > 0$则表示物体分离。因此成本函数定义为：

$$C\left(T_i, T_j; o_i \to o_j\right) = -\frac{\sum_{p\in\partial o_j} D_i\left(p(T_j)\right) \mathbb{I}\left(D_i\left(p(T_j)\right) < 0\right)}{\sum_{p\in\partial o_j} \mathbb{I}\left(D_i\left(p(T_j)\right) < 0\right)}$$

$$+ max\left(min_{(p\in\partial o_j)} D_i\left(p(T_j)\right), 0\right)$$

$$C\left(T_i, T_j\right) = C\left(T_i, T_j; o_i \to o_j\right) + C\left(T_i, T_j; o_j \to o_i\right)$$

$$if\ o_i\ and\ o_j\ are\ in\ contact \tag{9}$$

其中 $\partial \boldsymbol{o}_i$表示 $\boldsymbol{o}_i$的表面，$\mathbb{I}$ 为指示函数。该约束确保物体间无穿透且至少存在一个接触点。注意 $p \in \partial o_i$是 T_i 的函数。此处定义的接触约束是双向的，即同时适用于两个物体。

（2） **支撑关系**。作为单向约束，是接触关系的特例。若 $\boldsymbol{o}_i$ 支撑 $\boldsymbol{o}_j$，意味着需要优化 $\boldsymbol{o}_j$ 的位姿 T_j，而假设 $\boldsymbol{o}_i$ 保持静止。此情形通常出现在物体垂直堆叠时。该情况下的成本函数与接触关系类似，但仅涉及单向计算：

$$C\left(T_i, T_j\right) = \left|\min_{p\in\partial o_j} D_i\left(p(T_j)\right)\right|, \quad \text{if } \boldsymbol{o}_i \text{ supports } \boldsymbol{o}_j \tag{10}$$

此外，对于地面或墙壁等平坦支撑表面，我们正则化接触区域附近的SDF值，以确保物体与这些表面紧密接触。这种正则化针对物体部分重建的场景，例如图4中只有两个轮子的货车。

$$C\left(T_i, T_j\right) = \frac{\sum_{p\in\partial o_j} D_i\left(p(T_j)\right) \cdot \mathbb{I}\left(0 < D_i(p) < \sigma\right)}{\sum_{p\in\partial o_j} \mathbb{I}\left(0 < D_i(p) < \sigma\right)} \tag{11}$$

其中 $\mathbb{I}$ 为指示函数，σ 是判定点是否足够接近表面阈值。

5.3 场景关系图

物体间关系的物理线索在图像中直观存在。我们利用视觉语言模型（像 GPT-4v[1]）强大的常识推理能力[14,43,63]来识别第 5.2 节中定义的成对物理约束。给定图

像，我们采用 Set of Mark[84](SoM) 技术，通过视觉提示 GPT-4v 描述物体间关系，并随后从回答中提取场景关系图。为了解决 VLM 固有的采样不确定性，我们采用集成策略，结合多次试验的结果。如果关系在超过一半的样本中出现，则将其定义为正确，以生成相对鲁棒的推断图。更具体地说，我们多次应用随机着色和数值排序的 Set-of-Mark 方法，为进一步基于 GPT 的问答任务提供更可靠和一致的输出。

我们不是直接要求 GPT-4v 识别支撑和接触关系，而是首先提供更细粒度的物理关系，例如堆叠（物体 2 支撑物体 1）、倚靠（物体 1 倚靠物体 2）和悬挂（物体 2 从上方支撑物体 1）。我们指示 GPT-4v 分析 Set-of-Mark 方法中的编号物体，并输出所有基于接触的关系，涵盖六种类型：堆叠、倚靠、悬挂、夹紧、包含和边缘/点接触。提示词指定只有接触物体才具有关系，并且对于模糊情况默认为堆叠。

然后，我们将这些详细关系映射到预定义的支撑和接触类别，以进行进一步优化。具体来说，如果在两个之间存在相互指向的边，则该边被归类为接触；否则，

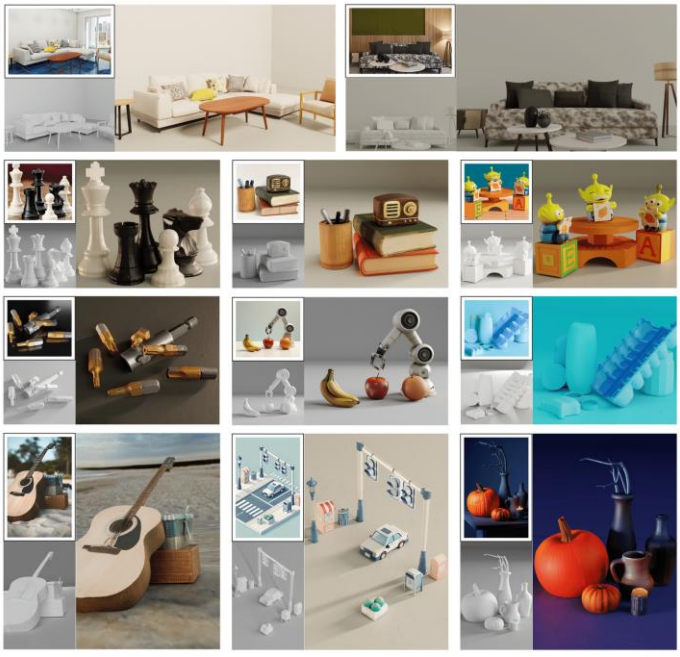

图 5　CAST 将开放词汇场景重新构想为沉浸式数字环境，捕捉每个独特场景的丰富性，将真实世界的生动多样性带入虚拟领域。对于每个场景，图像显示如下：左上角是输入图像，中上角是渲染的几何形状，右侧是带有真实纹理的渲染图像。

归类为支撑。通过向 GPT-4v 提示这些细微的关系，有助于消除二元关系分类中潜在的歧义，并促进 GPT-4v 更准确的推理。由此生成的图示例如图 4 所示。

映射的场景约束图是一个有向图，其中节点表示物体实例，边表示物体间的物理关系。接触关系由双向边表示，而支撑关系由有向边表示。该图作为定义式(8)中使用的成本函数的基础。

5.4 基于物理感知关系图的优化

给定推断关系图定义的物理约束，我们可以实例化式(8)中描述的成本函数。该图允许我们减少需要优化的成对约束的数量，与全局物理模拟不同。

在实现方面，我们从每个物体的静止姿态表面均匀采样固定数量的点。然后根据当前物体的姿态参数对这些点进行变换，并用于查询相对于另一个物体（及其姿态）的 SDF 值。SDF 计算由 Open3D 处理，PyTorch 用于自动微分损失函数。

六、结果

图 5 展示了我们的方法从单视图输入生成的一系列 3D 场景，涵盖了各种开放词汇场景，包括详细的室内环境、物体特写以及 AI 生成的图像。这些示例突出了我们方法的多功能性和鲁棒性，展示了高保真几何、真实纹理和令人信服的场景组合。

6.1 实现细节

ObjectGen（第 4.1 节）模型的预训练遵循 3DShape2VecSet[92]和 CLAY[94]中概述的方法，其中我们利用变分自编码器（VAE）和潜在扩散模型（LDM）来生成 3D 物体几何形状。VAE 和 LDM 模块均采用 24 层 Transformer 实现，总参数量为 1.5 亿。模型在 Objaverse[20]数据集上进行训练，该数据集包含约 50 万个经过过滤的 3D 资产。部分点云条件遵循 CLAY 的适应框架的类似方法。我们将规范空间部分点云编码为位置嵌入，特征维度为 512，并使用交叉注意力机制将其注入 LDM Transformer。对于每个 3D 资产，我们渲染 32 个视图，并使用 MoGe[75]和 Metric3D[87]预计算深度图。这些深度图在训练期间被反投影为点云，并应

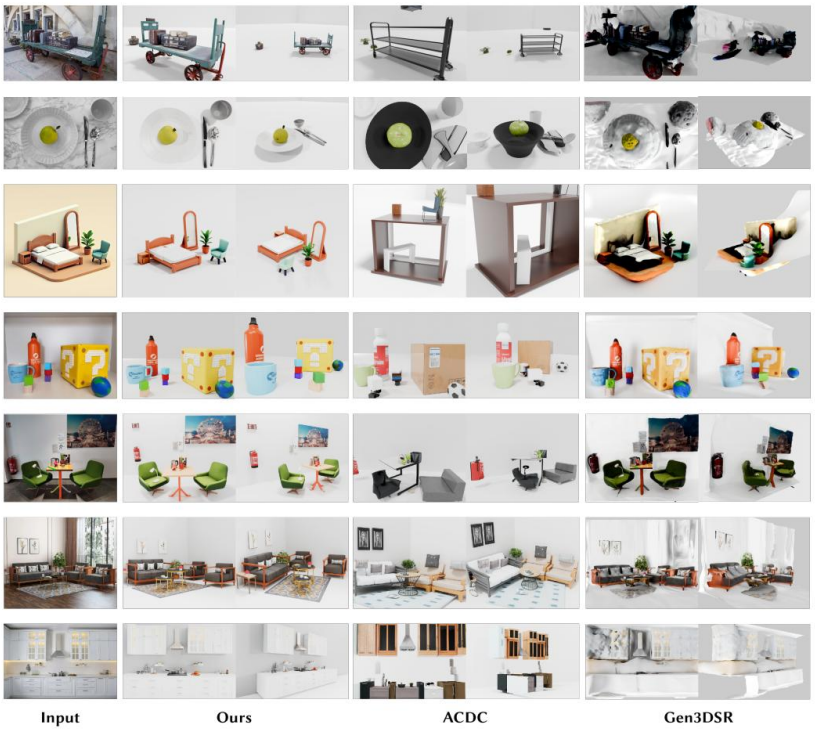

图 6　CAST 与最先进的单图像场景重建方法的定性比较。从左到右：输入图像、CAST、ACDC 和 Gen3DSR。从上到下：随机开放词汇数据集（第 1-3 行）、Gen3DSR 输入（第 4-5 行）、ACDC 输入（第 6-7 行）。

用随机掩模以模拟遮挡。我们使用最远点采样 (FPS) 从点云中采样 2048 个点，作为 LDM 的条件输入。为了增强模型的鲁棒性，我们随机插值真值和预测的部分点云，使得系统能够处理不同质量的数据。条件模块在 20 万个清洗过的 Objaverse 数据上训练 3000 个 epochs，使用 64 块 Nvidia A800 GPU，耗时约一周。采用 AdamW 优化器，学习率为 1e-5。对于单个物体的推理，物体生成约需要 7 秒，纹理生成大约需要 10 秒，均在一块 NVIDIA A6000 GPU 上运行。

AlignGen（第 4.2 节）模块负责生成姿态对齐，使用一个 24 层 Transformer，特征维度为 512，总共有 1.5 亿个参数。在训练期间，我们从事先计算好的深度图反投影得到的点云中随机采样规范空间中的部分点云，并对此点云应用随机变换。变换后的点云以及 ObjectGen 生成的几何潜在编码$\boldsymbol{z}$被用作条件输入。2048 个点通过 FPS 从部分点云中被采样，以确保 Transformer 的固定输入数量。该模块在相同的 20 万个物体的数据集上训练 1500 个 epochs，使用 64 块 Nvidia A800 GPU，大约需要两天时间。采用 AdamW 优化器，学习率为 1e-5。在推理过程中，AlignGen 模块为单个物体生成姿态大约需要 1 秒。

6.2 比较

定性比较　我们首先在开放词汇场景中评估我们的方法 CAST 与最先进的单图像场景重建技术。我们还包括 ACDC 和 Gen3DSR 使用的图像，以进一步展示不同方法的场景重建结果。图 6 展示了三种方法的性能——(1) 基于检索的方法 ACDC[18]，(2) 基于生成的方法 Gen3DSR[21]，以及(3) 我们提出的 CAST——包括参考视图和新视图。我们的结果突出了 CAST 在各种设置中准确重建场景的优越能力，包括室内和室外环境、特写视角以及 AI 生成图像。

如图 6 所示，CAST 通过创新，在 ACDC 和 Gen3DSR 中脱颖而出。ACDC 受限于室内场景，并依赖大型数据集进行物体检索，通常生成与场景中物体相似而非完全相同的物体，而 CAST 支持开放词汇。这使得 CAST 能够准确重建各种复杂环境中的物体。ACDC 使用简单的边界框作为代理，而 CAST 将基于图像的物理先验与网格优化相结合，以有效处理复杂场景。与 Gen3DSR 相比，CAST 通过 Masked Autoencoder 直接进行 3D 生成，消除了容易出错的 2D 修补步骤。这带来了更平滑的网格，显著优于 Gen3DSR 在单物体生

成质量方面的表现，尤其是在具有挑战性的场景中。此外，Gen3DSR 缺乏模拟常常导致物体穿透或浮动等问题，使得场景仅从输入视角看起来一致，并降低了新视图渲染质量。相比之下，CAST 确保了跨视角的场景一致性。CAST 在各种条件下展示了稳健的场景重建，突显了其对广泛的真实世界和生成场景的适用性。

为了评估生成场景的视觉保真度和语义准确性，我们采用了两种互补的评估方法，包括 CLIP 分数 [95] 和 GPT-4 推理。我们计算渲染场景与输入图像之间的 CLIP 分数，以衡量整体重建质量和视觉相似性。为了最小化无关影响，我们在计算分数之前从渲染图像和参考图像中移除了背景。我们还利用 GPT-4 对生成场景进行排名，基于各种语义方面，包括物体排列、物理关系和场景真实感。这种语义反馈有助于识别像素级分数可能不明显的对齐或上下文错误。

除了上述指标，我们还进行了一项用户研究，重点关注视觉质量 (VQ) 和物理合理性 (PP) 两个关键方面。我们随机选择配对的参考、新颖和目标视图，要求参与者选择哪种方法的输出与输入图像在相似度和整体美学方面最匹配。为了减少视觉相似性引入的潜在偏差，参与者在单独的会话中仅查看渲染结果——没有原始输入图像——并根据物理约束和常识（例如，是否有浮动物体或不可能的接触）判断哪个场景更真实。

如表 1 所示，CAST 在所有四个评估指标中均优于 ACDC 和 Gen3DSR，证实了其在生成视觉一致和物理合理场景方面的有效性。

定量比较 尽管 CAST 旨在处理开放词汇场景，但许多此类场景缺乏网格真值，这使得直接定量比较变得困难。为了解决这个问题，我们在 3DFront 数据集[22] 进行了额外评估。该数据集提供了真值网格以及对应的渲染图像，从而能够更精确地评估物体级和场景级重建。我们将 CAST 与 InstPIFu[46]、ACDC[18]和 Gen3DSR[21] 进行比较。我们计算物体级的 Chamfer 距离和 F-Score，以及场景级的 IoU、Chamfer 距离和 F-Score，以评估单个物体几何的保真度及其空间布局的准确性。为了确保公平性，我们用真值掩模替换了其他方法中的分割模块，使得任何差异纯粹源于重建能力而非物体分割。

Method	CLIP↑	GPT-4↓	VQ↑	PP↑
ACDC	69.77	2.7	5.58%	22.86%
Gen3DSR	79.84	2.175	6.35%	5.72%
CAST	**85.77**	**1.125**	**88.07%**	**71.42%**

表 1　将场景重建方法在 CLIP 分数、GPT-4 排名、视觉质量（VQ）和物理合理性（PP）四个指标上的定量比较

如表 2 所示，CAST 不仅实现了更高的物体级生成质量，而且在场景布局精度方面也超越了现有方法。即使在室内数据集的限制下，我们的方法也表现出稳健的性能，优于比较的基线。

6.3 评估

为了阐明 CAST 中关键组件的个体贡献，我们进行了一系列消融研究。这些实验系统地移除或修改了特定组件，以评估它们对整体性能的影响。消融研究侧重于几个关键设计选择：遮挡感知物体生成、点云条件、姿态对齐生成和基于物理的校正过程。

遮挡感知生成消融 遮挡是复杂场景中的一个重大挑战。为了评估 Masked Autoencoder（MAE）在处理遮挡方面的有效性，我们进行了一项消融研究，比较了有无 MAE 组件的生成结果。如图 7 所示，结果突出了遮挡感知模块的重要性。没有 MAE，部分遮挡区域生成的物体表现出显著退化。例如，飞船显示为破碎不完整，而杯子被描绘为破碎带有缺失部分。相比之下，当应用 MAE 条件时，模型成功推断并填充了被遮挡区域，从而产生更准确和视觉一致的生成，与输入图像更好地

Method	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
Vanila	0.079	53.38	0.069	52.83	0.515
+MAE	0.064	53.79	0.066	54.32	0.548
+ PCD	0.056	53.91	0.060	54.60	0.582
+ iter.	**0.052**	**56.18**	**0.057**	**56.50**	**0.603**

表 2　3D-Front 室内数据集上场景重建性能的定量比较。我们根据形状精度计算 Chamfer 距离（CD）、物体级重建质量计算 F-Score（FS）以及场景级重叠计算 Intersection over Union（IoU）评估不同方法。

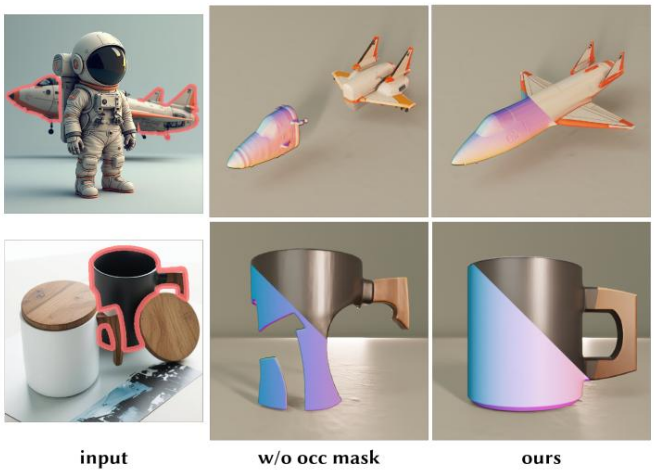

图 7　我们评估了有无遮挡感知生成模块的生成性能。物体渲染图和法线渲染图突显了该模块在确保生成物体的完整性和高质量方面的重要性。

对齐。这表明遮挡感知模块在确保准确重建被遮挡物体、提高最终 3D 场景的完整性和真实感方面具有关键作用。

部分点云条件消融　我们进行了一项消融研究，以研究规范空间部分点云条件在生成过程中的作用。尽管直接从输入图像生成可以产生视觉上合理的结果，但在缺乏像素级对齐的情况下，模型难以保持正确的物体数量和尺度，导致生成不令人满意。为了更有效地展示点

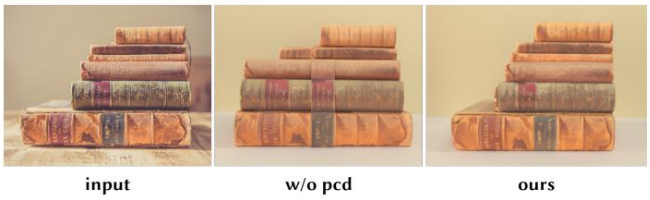

图 8　一叠不同长度和宽度的书籍。没有点云条件，模型直接生成一个复杂的单一物体。这展示了点云条件增强了尺度、维度和局部细节的保留。

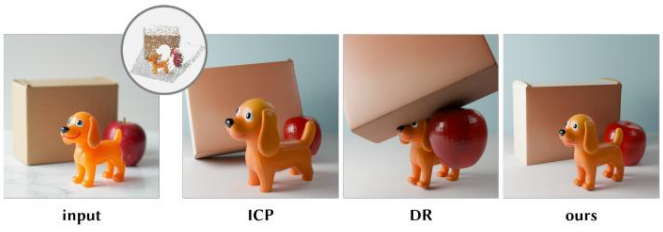

图 9　姿态估计方法的比较。我们的姿态对齐模块相比迭代最近点（ICP）和可微分渲染（DR）显示出优越的对齐精度。

云条件在生成单个实例中的重要性，我们选择直接生成一个更复杂的实例结构：一叠六本长度和宽度各异的书籍。如图 8 所示，当生成过程仅依赖输入图像时，在没有点云条件的情况下，结果频繁出现生成的物体数量和维度不准确。相比之下，点云条件引入了稳健的几何先验，显著提高了生成场景的精度。这种增强确保了具有复杂形状和不同维度的物体能够更准确地重建，与输入图像中描绘的真实世界对应物更加相似。这表明几何先验在通过保留真实尺度和形状来增强 3D 场景生成保真度方面具有关键作用。

姿态对齐生成方法的有效性　为了评估姿态对齐模块的有效性，我们将其与常用的姿态估计算法进行了比较，如迭代最近点（ICP）[2,6]和可微分渲染[40]。生成的网格被提供给不同的姿态估计算法，以使其与参考 RGB 图像及其对应的深度预测对齐。对于 ICP 方法，我们从生成的网格中均匀采样点云，并通过其边界框对采样点云和估计点云进行归一化，避免尺度差异。我们使用了 Open3D[97]中的 ICP 实现来对齐这两个归一化点云。对于可微分渲染，我们优化了旋转和平移参数，使变换后的物体网格的渲染图像与参考 RGB 图像对齐。如图 9 所示，我们的方法在对齐精度方面优于 ICP 和可微分渲染。ICP 在处理点云中的异常值、未知物体尺度以及对称或重复几何形状时，通常难以进行准确的姿态估计，可能导致局部最小值。另一方面，可微分渲染受到 RGB 输入中遮挡的显著影响，干扰了物体姿态的优化，并阻碍了与输入图像的精确对齐。我们的结果表明，我们的姿态对齐模块优于传统的 ICP 和可微分渲染方法，证明了其在准确估计生成网格的物体姿态和改进与输入图像对齐方面的鲁棒性。

物理一致性校正的效果　在 CAST 中，物理约束对于实现真实的物体交互和维持场景内的空间一致性至关重要。虽然我们解决了遮挡和不完整视图等常见挑战，但浮动物体、穿透和未对齐的空间关系等问题仍然存在。如图 10 所示，在没有关系约束的情况下生成的场景可能在物理上不一致；当应用完整的物理模拟时，物体会遵守物理定律，但它们的相对位置和整体排列可能与预期场景显著不同（例如，洋葱可能会从表面掉落，打乱

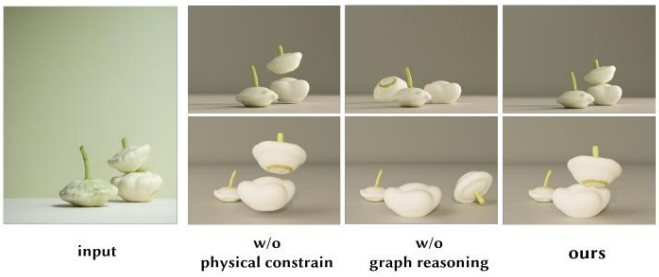

图 10　有无关系图约束的场景重建比较。通过整合关系图约束，我们的方法确保了物理合理性和与预期场景的准确对齐，保持了正确的空间关系。

原始构图）。通过整合关系图约束，我们的方法确保物体不仅符合物理可行性，而且与预期场景布局对齐，同时保留了物理合理性和所需的空间关系。

不同模块的定量消融研究 为了定量评估每个模块的贡献，我们进行了一项全面的消融研究。如表 3 所示，我们评估了移除或修改关键组件对最终场景质量的影响。结果表明，每个组件都对我们方法的整体性能做出了显著贡献。定量分析进一步突出了每个模块在实现高质量、物理一致和逼真的场景重建方面的重要性。

应用 如图 11 所示，CAST 将单个图像转换为一个实例化的 3D 场景，从而支持广泛的应用。这种重建详细环境的能力通过确保真实的物体交互，为基于物理的动画提供了动力。它还支持机器人领域的真实到模拟（real-to-sim）工作流，允许从真实世界数据集进行准确的场景复制。在游戏开发中，CAST 促进了沉浸式环境的创建，使得如实重建的场景可以无缝集成到基于虚幻引擎的交互式世界中。

Method	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
Vanilla	0.079	53.38	0.069	52.83	0.515
+ MAE	0.064	53.79	0.066	54.32	0.548
+ PCD	0.056	53.91	0.060	54.60	0.582
+ iter.	**0.052**	**56.18**	**0.057**	**56.50**	**0.603**

表 3　MAE 模块、点云条件（PCD）和迭代细化策略（iter.）的定量消融研究。为简洁起见，每行仅显示新增关键组件。

七、结论

在本文中，我们介绍了 CAST，一种新颖的单图像 3D 场景重建方法，它结合了几何保真度、像素级对齐和物理约束。通过整合场景分解、感知 3D 实例生成框架和物理校正技术，CAST 解决了姿态未对齐、物体相互依赖和部分遮挡等关键挑战。这种结构化的管线生成了视觉准确且物理一致的 3D 场景，超越了传统以物体为中心方法的局限性。我们通过广泛的实验和用户研究验证了 CAST，证明其在视觉质量和物理合理性方面显著优于现有最先进方法。我们预计 CAST 将为 3D 生成、场景重建和沉浸式内容创建的未来发展奠定坚实基础。

局限性和未来工作 CAST 中场景生成的质量严重依赖于底层的物体生成模型。目前，该模型仍缺乏足够的细节和精度，这一局限性导致生成的物体存在显著不一致，影响它们在场景中的对齐和空间关系。

此外，当前网格表示在处理纺织品、玻璃或织物等材料时仍存在困难，常常显得不自然，并且无法准确描绘透明材料，如图 12 所示。虽然已加入额外模块以增强物体鲁棒性和相似性，但仍需要更先进和鲁棒生成模型。更详细和准确的物体生成器可以显著提高整体场景质量并增强其现实世界适用性。

图 11　CAST 实现了逼真的基于物理的动画、沉浸式游戏环境和高效的真实到模拟过渡，推动各领域创新。

图 12　在某些场景中，透明玻璃、纺织品和织物难以表达，因为网格难以真实地表示它们。

当前方法的一个显著局限是缺乏光照估计和背景建模。没有真实的光照，物体与其周围环境之间的相互作用可能缺乏自然的阴影和光照效果，影响生成的环境的视觉真实感和沉浸感。为了增强视觉真实感，我们采用现成的全景 HDR 生成工具[34]，并结合 Blender 中预设的光照条件进行手动操作。CAST 未来可以受益于整合先进的光照估计和背景建模技术，从而显著丰富场景的上下文深度和视觉保真度。

在更复杂的场景中，当前方法的性能可能会略有下降。复杂的空间布局和密集的物体配置等挑战可能会在一定程度上影响场景重建的准确性。尽管 CAST 目前在重建单个场景方面表现出色，但一个有潜力的方向是利用其输出构建大型数据集，从而促进基于学习的场景生成或视频生成管线。扩展生成场景的多样性和真实感，可以进一步提高 3D 生成模型在电影制作、模拟和沉浸式媒体等领域的鲁棒性和适用性。

致谢

这项工作得到了国家重点研发计划（2022YFF0902301）、国家自然科学基金项目（61976138，61977047）、上海市科学技术委员会（2015F0203-000-06）和上海市教育委员会（2019-01-07-00-01-E00003）的支持。我们还要感谢上海人工智能前沿科学中心（ShangHAI）、教育部智能感知与人机协作重点实验室（上海科技大学）、上海科技大学计算机科学与通信学科平台以及上海科技大学高性能计算平台的支持。

责任编委　**魏秀参**

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).

[2] K Somani Arun, Thomas S Huang, and Steven D Blostein. 1987. Least-squares fitting of two 3-D point sets. IEEE Transactions on pattern analysis and machine intelligence 5 (1987), 698–700.

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo MartinBrualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF international conference on computer vision. 5855–5864.

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5470–5479.

[5] Jan Bender, Kenny Erleben, Jeff Trinkle, and Erwin Coumans. 2012. Interactive Simulation of Rigid Body Dynamics in Computer Graphics. In 33rd Annual Conference of the European Association for Computer Graphics, Eurographics 2012 - State of the Art Reports, Cagliari, Sardinia, Italy, May 13-18, 2012, Marie-Paule Cani and Fabio Ganovelli (Eds.). Eurographics Association, 95–134. https://doi.org/10.2312/CONF/ EG2012/STARS/095-134

[6] Paul J Best. 1992. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Vision 14 (1992), 239–256.

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023).

[8] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023).

[9] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In Forty-first International Conference on Machine Learning.

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015).

[11] Anpei Chen, Minye Wu, Yingliang Zhang, Nianyi Li, Jie Lu, Shenghua Gao, and Jingyi Yu. 2018. Deep surface light fields. Proceedings of the ACM on Computer Graphics and Interactive Techniques 1, 1 (2018), 1–17.

[12] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. 2024a. Single-view 3d scene reconstruction with high-fidelity shape and texture. In 2024 International Conference on 3D Vision (3DV). IEEE, 1456–1467.

[13] Yunuo Chen, Tianyi Xie, Zeshun Zong, Xuan Li, Feng Gao, Yin Yang, Ying Nian Wu, and Chenfanfu Jiang. 2024b. Atlas3D: Physically Constrained Self-Supporting Text-to-3D for Simulation and Fabrication. arXiv preprint arXiv:2405.18515 (2024).

[14] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. 2024. Navila: Legged robot visionlanguage-action model for navigation. arXiv preprint arXiv:2412.04453 (2024).

[15] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. 2023. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4937–4946.

[16] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. 2021. Panoptic 3d scene reconstruction from a single rgb image. Advances in Neural Information Processing Systems 34 (2021), 8282–8293.

[17] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition. 5828–5839.

[18] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. 2024. Automated Creation of Digital Cousins for Robust Policy Learning. arXiv preprint arXiv:2410.07408 (2024).

[19] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems 36 (2024).

[20] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13142–13153.

[21] Andreea Dogaru, Mert Özer, and Bernhard Egger. 2024. Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View. arXiv preprint arXiv:2404.03421 (2024).

[22] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 10933–10942.

[23] Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. 2024b. Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. ACM Transactions on Graphics (TOG) 43, 4 (2024), 1–15.

[24] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo MartinBrualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024a. Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024).

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, 11 (2013), 1231–1237.

[26] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. 2022. Learning 3d object shape and layout without 3d supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1695–1704.

[27] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. 2007. Multi-View Stereo for Community Photo Collections. In 2007 IEEE 11th International Conference on Computer Vision. 1–8. https://doi.org/10.1109/ICCV.2007.4408933

[28] Can Gümeli, Angela Dai, and Matthias Nießner. 2022. Roca: Robust cad model retrieval and alignment from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4022–4031.

[29] Minghao Guo, Bohan Wang, Pingchuan Ma, Tianyuan Zhang, Crystal Elaine Owens, Chuang Gan, Joshua B Tenenbaum, Kaiming He, and Wojciech Matusik. 2024. Physically Compatible 3D Object Modeling from a Single Image. arXiv preprint arXiv:2405.20510 (2024).

[30] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022a. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022).

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022b. Video diffusion models. Advances in Neural Information Processing Systems 35 (2022), 8633–8646.

[32] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023).

[33] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. 2024. MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation. arXiv preprint arXiv:2412.03558 (2024).

[34] Hyper3D. 2025. Omnicraft. https://hyper3d.ai/omnicraft/hdri

[35] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. 2017. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In Proceedings of the IEEE international conference on computer vision. 1521–1529.

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4015–4026.

[38] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. 2021. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 12589– 12599.

[39] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. 2020. Cosypose: Consistent multi-view multi-object 6d pose estimation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer, 574–591.

[40] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics (TOG) 39 (2020), 1 – 14.

[41] Florian Langer, Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2022. SPARC: Sparse render-and-compare for CAD model alignment in a single RGB image. arXiv preprint arXiv:2210.01044 (2022).

[42] Bruno Latour. 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford, UK.

[43] Wenhao Li, Zhiyuan Yu, Qijin She, Zhinan Yu, Yuqing Lan, Chenyang Zhu, Ruizhen Hu, and Kai Xu. 2024b. LLM-enhanced Scene Graph Learning for Household Rearrangement. In SIGGRAPH Asia 2024 Conference Papers. 1–11.

[44] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. 2024a. Evaluating Real-World Robot Manipulation Policies in Simulation. arXiv preprint arXiv:2405.05941 (2024).

[45] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. 2024. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6517–6526.

[46] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. 2022. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In European Conference on Computer Vision. Springer, 429–446.

[47] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems 36 (2024).

[48] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023c. Zero-1-to-3: Zero-shot one image to 3d object. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2025. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision. Springer, 38–55.

[50] Xueyi Liu, Bin Wang, He Wang, and Li Yi. 2023b. Few-Shot Physically-Aware Articulated Mesh Generation via Hierarchical Deformation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 854–864.

[51] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023a. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In arXiv preprint arXiv:2309.03453.

[52] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9970–9980.

[53] Khaled Mamou and Faouzi Ghorbel. 2009. A simple and efficient approach for 3D mesh approximate convex decomposition. In 2009 16th IEEE international conference on image processing (ICIP). IEEE, 3501–3504.

[54] Khaled Mamou, E Lengyel, and A Peters. 2016. Volumetric hierarchical approximate convex decomposition. Game engine gems 3 (2016), 141–158.

[55] Mariem Mezghanni, Théo Bodrito, Malika Boulkenafed, and Maks Ovsjanikov. 2022. Physical simulation layer for accurate 3d modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13514–13523.

[56] Mariem Mezghanni, Malika Boulkenafed, Andre Lieutier, and Maks Ovsjanikov. 2021. Physically-aware generative network for 3d shape modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9330–9341.

[57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision. Springer, 405–421.

[58] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph. 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[59] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. 2024. PhyRecon: Physically Plausible Neural Scene Reconstruction. Advances in Neural Information Processing Systems.

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023).

[61] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. 2024. UniDepth: Universal Monocular Metric Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10106–10116.

[62] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. In arXiv preprint arXiv:2209.14988.

[63] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In 7th Annual Conference on Robot Learning.

[64] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024).

[65] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024).

[66] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. 2018. Pix3d: Dataset and methods for single-image 3d shape modeling. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2974–2983.

[67] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. 2024a. Flash3D: Feed-Forward Generalisable 3D Scene Reconstruction from a Single Image. arXiv preprint arXiv:2406.04343 (2024).

[68] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. 2024b. Splatter image: Ultra-fast single-view 3d reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10208–10217.

[69] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In European Conference on Computer Vision. Springer, 1–18.

[70] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023).

[71] Fengrui Tian, Shaoyi Du, and Yueqi Duan. 2023. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 17903–17913.

[72] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. 2024. Reconciling reality through simulation: A real-to-sim-toreal approach for robust manipulation. arXiv preprint arXiv:2403.03949 (2024).

[73] Shinji Umeyama. 1991. Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence 13, 04 (1991), 376–380.

[74] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2025. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In European Conference on Computer Vision. Springer, 439–457.

[75] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2024b. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. arXiv preprint arXiv:2410.19115 (2024).

[76] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2024a. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems 36 (2024).

[77] Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. 2022. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. ACM Transactions on Graphics (TOG) 41, 4 (2022), 1–18.

[78] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. 2024a. Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image. arXiv preprint arXiv:2405.20343 (2024).

[79] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024b. Reconfusion: 3d reconstruction with diffusion priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 21551–21561.

[80] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2024. Structured 3D Latents for Scalable and Versatile 3D Generation. arXiv preprint arXiv:2412.01506 (2024).

[81] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4818–4829.

[82] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. 2024. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4389–4398.

[83] Qingshan Xu, Jiao Liu, Melvin Wong, Caishun Chen, and Yew-Soon Ong. 2024. PrecisePhysics Driven Text-to-3D Generation. arXiv preprint arXiv:2403.12438 (2024).

[84] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023).

[85] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10371–10381.

[86] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. 2024a. Physcene: Physically interactable 3d scene synthesis for embodied ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16262–16272.

[87] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3d: Towards zero-shot metric 3d prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 9043–9053.

[88] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4578–4587.

[89] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. 2024. Wonderjourney: Going from anywhere to everywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6658–6667.

[90] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems 35 (2022), 25018– 25032.

[91] Alan Yuille and Daniel Kersten. 2006. Vision as Bayesian inference: analysis by synthesis? Trends in cognitive sciences 10, 7 (2006), 301–308.

[92] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 2023. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions on Graphics (TOG) 42, 4 (2023), 1–16.

[93] Jiancheng Zhang, Haijin Zeng, Yongyong Chen, Dengxiu Yu, and Yin-Ping Zhao. 2024b. Improving Spectral Snapshot Reconstruction with Spectral-Spatial Rectification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 25817–25826.

[94] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024a. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. ACM Transactions on Graphics (TOG) 43, 4 (2024), 1–20.

[95] SUN Zhengwentai. 2023. clip-score: CLIP Score for PyTorch. https://github.com/taited/ clip-score. Version 0.1.1.

[96] Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. 2025. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In European Conference on Computer Vision. Springer, 407–423.

[97] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A Modern Library for 3D Data Processing. ArXiv abs/1801.09847 (2018).

姚凯欣

上海科技大学信息科学与技术学院 2023 级博士研究生，导师为虞晶怡教授和许岚教授，主要研究方向为场景生成、3D 生成、3D 高斯。
Email：yaokx2023@shanghaitech.edu.cn

张龙文

上海科技大学信息科学与技术学院 2022 级博士生，导师为虞晶怡教授和许岚教授，主要研究方向为 3D 生成。
Email：zhanglw2@shanghaitech.edu.cn

严新豪

上海科技大学信息学院 2023 级博士研究生，导师为许岚助理教授， 主要研究方向为多模态，3D 生成式模型。
Email：yanxh@shanghaitech.edu.cn

曾 焱

上海科技大学信息科学与技术学院 2024 级博士生，导师为虞晶怡教授，主要研究方向为视频生成。
Email：zengyan2024@shanghaitech.edu.cn

张启煊

上海科技大学研究生，同时担任数字人 AI 公司 DeemosTech 的首席技术官。专攻计算机图形学、计算摄影学和生成式人工智能领域。其研究成果屡获 SIGGRAPH、ICCV、CVPR 等顶尖学术会议录用，相关技术已应用于多部影视作品及游戏项目。
Email：zhangqx1@shanghaitech.edu.cn

杨 卫

华中科技大学副教授、博士生导师、湖北省百人、东湖学者。2017 年毕业于美国特拉华大学，获得计算机科学专业博士学位，师从 Jingyi Yu 教授，毕业后曾在 Google 的先进科技与计划部门和知名初创公司 DGene 工作，从事基于视觉的场景理解与虚拟现实的应用研究。2021 年进入华中科技大学计算机学院智能媒体计算与网络安全实验室从事教学科研工作。主要研究方向包括：三维视觉理解与重建、基于物理特性的视觉和图形学、先进感知与图像传感器等。在 TPAMI、TOG、IJCV、CVPR、 ICCV、ECCV 等期刊会议上发表高水平论文 20 余篇。现任中国图学学会动漫图学工程专业委员会委员，将担任 CVPR 23 的 Area Chair，AAAI 20 & 21，以及 WACV 21，BMVC 18 的程序委员，TPAMI，TIP，TVCJ，CVPR，ICCV，NeurIPS，ECCV 等顶级会议期刊的审稿人。

Email：weiyangcs@hust.edu.cn

许 岚

上海科技大学信息科学与技术学院助理教授、研究员、博士生导师。许岚教授在浙江大学信息与电子工程学系获学士学位；在香港科技大学电子与计算机工程获博士学位。之后加入上海科技大学，任助理教授、研究员。他的研究方向包括计算机视觉、计算机图形学、机器学习、计算摄像学，目前的研究兴趣侧重于动静态三维重建、虚拟现实、数字孪生，终极目标是实现个人数字资产化和沉浸式全息立体通信。他已发表了多篇顶级期刊和会议文章，包括 CVPR、ECCV、ICCP、IROS、IEEE TRO、IEEE TVCG、IEEE TPAMI 等。主要研究内容包括人体动态捕捉、动静态三维重建与理解、数字孪生、虚拟现实、增强现实。

Email：xulan1@shanghaitech.edu.cn

顾家远

上海科技大学信息科学与技术学院助理教授、研究员、博士生导师。顾家远教授 2018 年于北京大学信息科学技术学院智能科学系获得本科学位，并于 2024 年在美国加州大学圣地亚哥分校计算机科学与工程学院获得博士学位。他曾在 Uber ATG、Waymo、Facebook AI、Qualcomm AI、Google DeepMind 等机构担任实习或学生研究员。他的研究方向包括具身智能与三维视觉，在计算机视觉、机器学习、机器人等国际顶会上发表 20 余篇工作。

Email：gujy1@shanghaitech.edu.cn

虞晶怡

上海科技大学副教务长，信息学院院长。在加入上海科技大学前，他任职美国特拉华大学计算机与信息科学系正教授。他于 2000 年获美国加州理工大学应用数学及计算机学士学位， 2003 年获美国麻省理工大学计算机与电子工程硕士学位，2005 年获美国麻省理工大学计算机与电子工程博士学位。他长期从事计算机视觉、计算成像、计算机图形学、生物信息学等领域的研究工作，已发表 120 多篇学术论文，其中超 70 篇发表于国际会议 CVPR/ICCV/ECCV 和期刊 TPAMI。他已获得美国发明专利 20 余项，并于 2009 和 2010 年分别获得美国国家科学基金的杰出青年奖和美国空军研究院的杰出青年奖。他是 IEEE TPAMI、IEEE TIP 和 Elsevier CVIU 的编委，担任 ICPR 2020、CVPR 2021、WACV 2021、ICCV 2027 的程序主席和 ICCV 2025 的大会主席。因为他在计算机视觉和计算成像上的贡献，当选 IEEE Fellow。

Email：yujingyi@shanghaitech.edu.cn

专题综述

多模态生成式 AI 探索：从数据合成到内容创造

同济大学 高俊尧 宋子帆 齐鼎 赵才荣

本文是同济大学VILL实验室在多模态生成领域的一系列工作成果，其中语言/视觉数据生成的相关工作发表在NIPS 2025[1]以及CVPR 2025（做口头汇报）[2]，图像/视频生成的相关工作在github上收获了超过400个stars，并发表在TPAMI 2025[3]和ICLR 2025[4]。随着多模态生成式人工智能的持续演进，从“数据学习”向“内容创造”的范式转变已经成为推动新一轮技术革命的关键驱动力。

在语言数据生成方面，VILL实验室重点关注开源大型语言模型（LLMs）在代码生成方向的精调能力提升，针对当前主流Code LLMs常在单一来源的数据集上进行微调，受限于语言数据质量和风格单一性，无法充分激发预训练模型的泛化能力这一问题，提出了AlchemistCoder框架，从根本上重构了代码微调数据构建的方式。在视觉数据生成方面，VILL实验室重点关注数据集蒸馏（Dataset Distillation, DD）在图像分类之外任务中的泛化与适应能力，创新性地将任务知识挖掘与扩散模型（Diffusion Models）生成过程深度融合以克服现有方法在目标检测与图像分割中的任务局限性，所提出的通用视觉数据蒸馏框架UniDD借助合成高质量的视觉数据，在提升数据利用效率的同时，也为低资源场景下的视觉模型训练提供了新范式。在图像生成方面，VILL实验室聚焦开放域的图像风格迁移，提出StyleShot通过构建一个风格感知编码器和一个大规模的风格数据集，高效地提取丰富的风格表示，并结合内容融合编码器以增强图像驱动的风格迁移能力。在视频生成方面，VILL聚焦现有肖像动画方法在非人类角色（如表情包、玩偶等）时常常发生动画效果失真甚至失败这一问题，提出一个无需训练的肖像动画框架FaceShot，利用扩散模型的强大语义对应关系来生成各种角色类型的动画结果。

一、研究背景

1.1 大语言模型

近年来，大语言模型在自然语言处理领域取得了突破性进展。得益于 Transformer 架构和大规模数据的训练，LLMs 展现出强大的语言理解与生成能力，不仅在对话系统、文本摘要、机器翻译、信息抽取等任务中均表现优异，在代码领域的特定变体（Code LLMs）也取得了显著进展。以往的 Code LLMs 通常在单一来源的数据集上进行微调，这些数据集在质量和多样性上存在局限，这可能无法充分激发预训练 LLMs 的潜力。研究指出，尽管多源数据融合是提升模型能力的关键，但草率地混合不同来源的代码语料库，会因其固有的风格、质量和编程范式冲突，反而导致模型性能下降。因此，如何有效整合多源数据，克服其内在冲突，以充分释放基础模型的代码智能，并进一步提升模型的泛化能力，是当前代码大模型微调领域面临的核心挑战。

1.2 扩散模型

扩散模型近年来在生成建模领域取得了显著进展，逐渐成为继 GAN 和 VAE 之后的主流生成方法。其核心思想源于非平衡热力学过程，通过逐步向数据添加噪声构建前向过程，并在反向过程中学习去噪还原数据分布。与其他生成模型相比，扩散模型在图像、音频、3D 建模等多种模态上表现出更强的生成质量和更高的多样性，

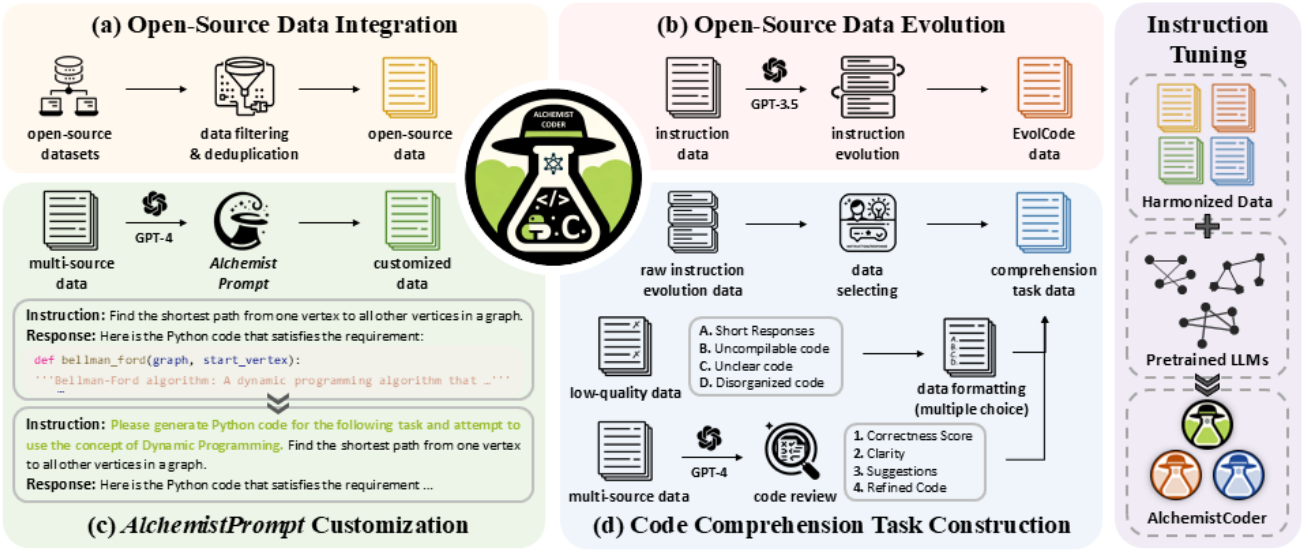

图 1　AlchemistCoder 框架图

尤其在高分辨率图像生成任务中已超越 GAN 的效果。借助扩散模型更清晰的理论基础和更稳定的生成效果，本文挖掘扩散模型在视觉数据生成、图像/视频生成上的进一步潜力，积极探索与任务知识结合的生成范式，推动其在复杂视觉任务中的实用性提升。

二、数据生成

2.1 语言数据生成

首先，为解决多源数据的内在冲突，该研究开创性地引入了 AlchemistPrompts。这是一种基于“事后重标签 (hindsight relabeling)”思想的数据特异性提示（data-specific prompts）。如图 1 所示，具体而言，该研究使用一个强大的模型（如 GPT-4）扮演“炼金术士（Alchemist）”的角色，回顾并分析每个“指令-代码响应”数据对，然后生成一个能更精确、更细致地描述该代码响应特点的新指令。例如，如果原始指令是“编写一个最短路径算法”，而代码实现是 Python 版的贝尔曼-福特算法，AlchemistPrompt 会将其优化为“请使用 Python 语言并结合动态规划思想，为以下任务生成代码”。这种方法不仅通过统一的风格协调了不同数据源之间的差异，还通过增强指令与响应的对齐度，将模型的学习过程从“为相似问题克隆不同答案”转变为“学习遵循多样化的精确指令”。

其次，该研究提出将“数据构建过程”本身也作为训练任务，以提升模型的代码理解（code comprehension）能力。除了传统的代码生成任务，他们设计了三项新的代码理解任务并构建了相应数据：1）指令演进（instruction evolution），让模型学习如何将简单指令优化得更复杂、更明确；2）数据过滤（data filtering），向模型展示低质量代码（如编译错误、不合规等）的反例，训练其辨别和避免生成劣质代码；3）代码审查（code review），要求模型评估代码的正确性和清晰度、提出修改建议并给出优化后的代码。

Model	Params	Base Model	HumanEval (+)	MBPP (+)	Average (+)
Closed-source Models					
GPT-3.5-Turbo [33]	-	-	72.6 (65.9)	81.7 (69.4)	77.2 (67.7)
GPT-4-Turbo [34]	-	-	**85.4 (81.7)**	**83.0 (70.7)**	**84.2 (76.2)**
Open-source Models					
Llama 2-Chat [40]	70B	Llama 2	31.7 (26.2)	52.1 (38.6)	41.9 (32.4)
CodeLlama-Python [35]	70B	Llama 2	57.9 (50.0)	72.4 (52.4)	65.2 (51.2)
CodeLlama-Instruct [35]	70B	CodeLlama	**65.2 (58.5)**	**73.5 (55.1)**	**69.4 (56.8)**
CodeLlama-Python [35]	34B	Llama 2	51.8 (43.9)	67.2 (50.4)	59.5 (47.2)
WizardCoder-CL [30]	34B	CodeLlama-Python	73.2 (56.7)	73.2 (51.9)	73.2 (54.3)
DeepSeek-Coder-Instruct [14]	33B	DeepSeek-Coder-Base	**78.7 (67.7)**	**78.7 (59.7)**	**78.7 (63.7)**
StarCoder [22]	15B	-	34.1 (33.5)	55.1 (43.4)	44.6 (38.5)
CodeLlama-Python [35]	13B	Llama 2	42.7 (36.6)	61.2 (**45.6**)	52.0 (41.1)
WizardCoder-SC [30]	15B	StarCoder	**51.9 (45.7)**	**61.9** (44.9)	**56.9 (45.3)**
Llama 2 [40]	7B	-	14.0 (10.4)	26.1 (17.5)	20.1 (14.0)
StarCoder [22]	7B	-	24.4 (21.3)	33.1 (29.2)	28.8 (25.3)
CodeLlama-Python [35]	7B	Llama 2	37.8 (33.5)	57.6 (42.4)	47.7 (38.0)
WizardCoder-CL [30]	7B	CodeLlama-Python	48.2 (42.1)	56.6 (42.4)	52.4 (42.3)
DeepSeek-Coder-Base [14]	6.7B	-	47.6 (41.5)	70.2 (53.6)	58.9 (47.6)
Magicoder-CL [44]	7B	CodeLlama-Python	60.4 (49.4)	64.2 (46.1)	62.3 (47.8)
MagicoderS-CL [44]	7B	CodeLlama-Python	70.7 (60.4)	68.4 (49.1)	69.6 (54.8)
Magicoder-DS [44]	6.7B	DeepSeek-Coder-Base	66.5 (55.5)	75.4 (55.6)	71.0 (55.6)
DeepSeek-Coder-Instruct [14]	6.7B	DeepSeek-Coder-Base	73.8 (69.5)	72.7 (55.6)	73.3 (62.6)
MagicoderS-DS [44]	6.7B	DeepSeek-Coder-Base	76.8 (65.2)	75.7 (56.1)	76.3 (60.7)
AlchemistCoder-L (ours)	7B	Llama 2	56.7 (52.4)	54.5 (49.6)	55.6 (51.0)
AlchemistCoder-CL (ours)	7B	CodeLlama-Python	74.4 (68.3)	68.5 (55.1)	71.5 (61.7)
AlchemistCoder-DS (ours)	6.7B	DeepSeek-Coder-Base	**79.9 (75.6)**	**77.0 (60.2)**	**78.5 (67.9)**

表 1　HumanEval 和 MBPP 上的评估结果

Model	Python	C++	Go	Java	JS	Avg
Llama 2	14.0	11.0	6.1	11.0	14.0	11.2
CodeLlama	31.7	27.4	12.8	25.6	32.9	26.1
AlchemistCoder-L	**56.7**	**31.1**	**25.6**	**45.1**	**41.5**	**37.1**
CodeLlama-Python	37.8	33.5	30.5	39.6	35.4	35.4
MagicoderS-CL	68.3	47.6	39.6	34.8	**57.9**	49.6
AlchemistCoder-CL	**74.4**	**53.1**	**42.7**	**64.0**	52.4	**57.3**
DeepSeek-Coder-Base	47.6	45.1	38.4	56.1	43.9	46.2
MagicoderS-DS	72.6	**63.4**	51.8	70.7	66.5	65.0
AlchemistCoder-DS	**79.9**	62.2	**59.8**	**72.0**	**68.9**	**68.6**

Model	pd	np	tf	sp	skl	torch	plt	All
Llama 2	2.4	7.3	6.7	6.6	2.6	1.5	7.7	5.0
CodeLlama	12.0	**27.7**	17.8	**13.2**	12.2	**20.6**	15.5	17.0
AlchemistCoder-L	**13.4**	22.7	**31.1**	11.3	**25.2**	8.8	**29.0**	**20.2**
CodeLlama-Python	16.2	16.4	15.6	17.9	20.0	22.1	38.7	21.0
MagicoderS-CL	25.1	40.9	35.6	29.3	36.5	38.2	51.0	36.7
AlchemistCoder-CL	**30.9**	**43.6**	**46.7**	**30.2**	**37.4**	**41.2**	**52.3**	**40.3**
DeepSeek-Coder-Base	21.3	35.0	26.7	23.6	34.8	25.0	34.8	28.7
MagicoderS-DS	30.6	46.8	44.2	30.2	33.0	29.7	45.2	37.1
AlchemistCoder-DS	**32.0**	**51.7**	**44.5**	**33.1**	**38.4**	**33.8**	**49.8**	**40.5**

表 2　HumanEval-X 上的评估结果

最终，通过整合高质量的开源数据、指令演进数据，并策略性地融入由 AlchemistPrompts 协调后的数据以及代码理解任务数据，构建了最终的 AlchemistCoder 微调数据集。

如表 1 所示，在主流的 Python 代码生成基准测试 HumanEval 和 MBPP 上，AlchemistCoder 系列模型在其同等规模 (6.7B/7B) 中取得了全面的领先地位。例如，AlchemistCoder-DS (6.7B) 的性能不仅远超其他同尺寸模型，甚至能够媲美或超越参数量更大的模型（如 15B/33B/70B），显著缩小了与 GPT-3.5-Turbo 等闭源模型的差距。同时，如表 2 所示，模型在多语言代码生成 (HumanEval-X) 和数据科学编程 (DS-1000) 等任务上也表现出强大的泛化能力。

综上所述，AlchemistPrompts 能够有效降低指令与响应之间的语义偏差，提升了数据质量。另外，该方法不仅提升了模型的代码能力，还在通用语言理解（MMLU）、综合推理（BBH）和数学能力（GSM8K）等非代码任务上取得了显著进步。这说明通过协调化的多源数据微调，能够有效缓解领域微调中常见的“灾难性遗忘（catastrophic forgetting）”问题，从而培养出能力更全面的通用模型。

2.2 视觉数据生成

本节简要介绍 UniDD 的通用视觉数据蒸馏生成框架。如图 2 所示，主要包括通用任务知识挖掘和通用任务驱动扩散两个阶段：

通用任务知识挖掘阶段主要负责从大型真实数据集中提取用于指导合成图像生成的重要信息，包括：

(1) 任务特定代理模型训练 (Task-Specific Proxy Training)：训练分类器、检测器和分割器等代理模型 (TSP 模型)。这些模型在真实数据集上进行训练，以捕获和存储任务特定的信息，例如用于分类的类别信息、用于目标检测的边界框位置以及用于图像分割的像素级掩码数据。UniDD 的代理模型选择非常灵活，可以根据目标数据集的性能需求进行调整，例如 ResNet-18 可用于轻量级分类，而 Faster R-CNN 可用于复杂检测任务。

$$\theta = \arg\min_{\theta} \mathcal{L}_{task}(\mathcal{F}(x), y)$$

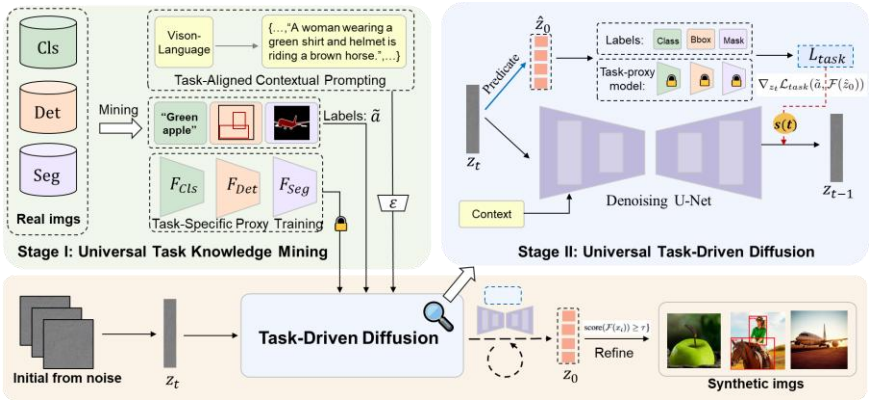

图 2　UniDD 框架图

(2) 任务对齐上下文提示（Task-Aligned Contextual Prompting）：为了解决以往扩散模型在图像生成中忽视自然语言指导作用的问题，UniDD 引入了任务对齐上下文提示。它利用视觉-语言模型来生成与任务相关的自然语言描述。这些提示不仅描述图像中的主要对象，还包含对象之间的关系以及对象与背景的交互，为生成过程提供更丰富的上下文指导。

通用任务驱动扩散利用第一阶段提取的知识借助扩散模型来指导合成视觉图像，具体过程如下：

(1) 任务驱动扩散图像合成：在扩散过程的每个采样步骤中，利用训练好的 TSP 模型对预测的干净图像计算任务损失，并根据这个损失的梯度来调整噪声计算任务损失，并根据这个损失的梯度来调整噪声预测。这种调整机制有效地引导扩散模型生成与给定标签（包括类别标签、边界框和像素级分割掩码）高度一致的数据。与以往需要对扩散模型进行微调的方法不同，UniDD 通过 TSP 模型引导生成过程，从而避免了昂贵的扩散模型微调，显著降低了部署成本。

(2) 代理驱动高真实性细化（Proxy-Driven High-Realism Refinement）：为了确保生成图像的高真实性和准确性，UniDD 采用训练好的 TSP 模型对每张生成的图像进行评分。通过设置明确的阈值，可以过滤掉低质量的图像。剩余的图像会使用相同的 TSP 模型进行重新标注。这种过滤和重新标注的组合过程确保了生成的图像及其标签都具有高质量，并与所需的数据分布紧密匹配。

Dataset	IPC	Method	Accuracy(%)
ImageNet-1K	1000	Full dataset	69.8
	10	TESLA	7.7
		SRe^{2}L	21.3
		D^{4}M	27.9
		RDED	42.0
		MiMxDiff	44.3
		UniDD (Ours)	**50.4 (↑6.1)**
	50	SRe^{2}L	46.8
		D^{4}M	55.2
		RDED	56.5
		MiMxDiff	58.6
		UniDD (Ours)	**62.8 (↑4.2)**

表 3 ImageNet-1K 上的评估结果

Methods	Object Detection			
	Pascal VOC		MS COCO	
	mAP	AP50	mAP	AP50
Ratio	0.5%		0.25%	
Random	0.8±0.2	3.1±0.4	0.5±0.1	1.7±0.3
Uniform	0.9±0.1	3.4±0.3	0.8±0.2	2.4±0.5
K-Center	0.5±0.1	2.1±0.3	0.4±0.1	1.5±0.2
Herding	0.6±0.2	2.4±0.2	0.5±0.1	1.8±0.4
UniDD (Ours)	**8.5±0.4**	**22.3±0.6**	**4.5±0.3**	**10.3±0.4**
Ratio	1%		0.5%	
Random	4.2±0.5	13.7±0.6	3.7±0.2	10.1±0.3
Uniform	5.7±0.2	17.7±0.4	3.4±0.4	9.5±0.6
K-Center	3.6±0.6	12.3±0.3	3.2±0.5	9.3±0.5
Herding	3.5±0.5	11.9±0.5	3.5±0.3	9.7±0.3
UniDD (Ours)	**16.8±0.5**	**38.9±0.7**	**7.1±0.4**	**16.9±0.3**
Ratio	2%		1%	
Random	12.4±0.4	34.3±0.5	7.2±0.8	17.3±0.9
Uniform	13.8±0.3	36.2±0.4	7.4±0.5	17.6±0.5
K-Center	10.9±0.6	29.3±0.6	6.1±0.3	15.4±0.6
Herding	10.4±0.4	28.7±0.7	6.7±0.4	16.3±0.7
UniDD (Ours)	**23.9±0.5**	**48.5±0.6**	**10.8±0.4**	**22.5±0.5**
Full	51.4±0.8	80.3±0.4	32.6±0.7	51.4±0.8

表 4 Pascal VOC 和 MS COCO 上的评估结果

如表 3、表 4 所示，我们在 ImageNet-1K、Pascal VOC 和 MS COCO 等多个基准数据集上进行了广泛实验，结果表明 UniDD 超越了现有最先进的方法。特别是在 ImageNet-1K 数据集上，当每类图像数（IPC）为 10 时，UniDD 相较于之前的基于扩散的方法，性能提升了 6.1%，同时显著降低了部署成本。这一成果为视觉数据生成在更多样化任务中的应用提供了新的思路和方法支持。生成数据的可视化如图 3 所示。

三、图像/视频生成

3.1 风格迁移

图像风格迁移的目标是将一幅参考图像的风格应用到另一幅内容图像上，广泛应用于艺术创作、相机滤镜等场景。然而，现有的风格迁移方法往往依赖于测试时的风格调优，这不仅增加了计算和存储的开销，还可能导致模型过拟合于单一的参考图像。为了解决这些问题，

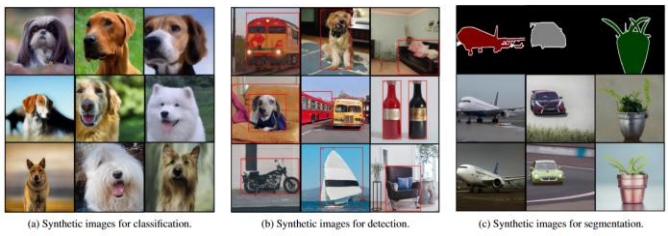

图 3 生成视觉数据的可视化

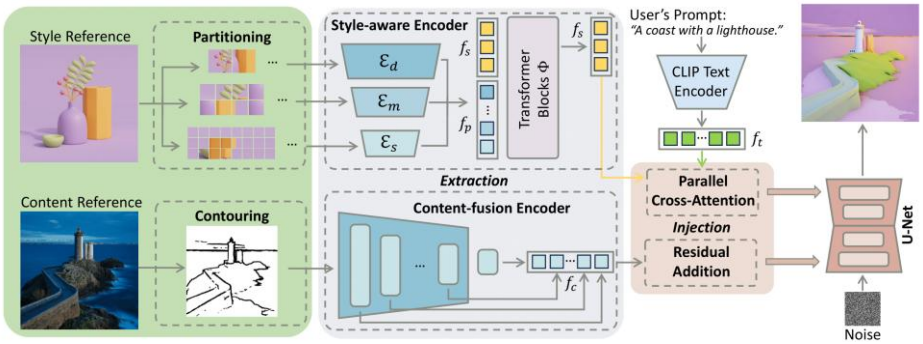

图 4　StyleShot 框架图

本节简要介绍 StyleShot，一种通用风格迁移方法。如图 4 所示，主要包括风格感知编码器、内容融合编码器以及风格数据集三个模块：

一个良好的风格表示对于无测试时调优的情况下的通用风格迁移是非常重要的。由此 StyleShot 通过构建一个风格感知编码器来提取丰富的风格嵌入。与传统的 CLIP 编码器不同，StyleShot 采用了一种多尺度补丁提取策略，可以从不同大小的图像补丁中获取低级和高级风格特征，并且通过 partition 的方式打散内容信息。在获取多尺度补丁后，StyleShot 利用不同深度的网络提取不同尺度的 style 信息，并且使用 transformer 模块进行融合学习得到最终的风格特征。通过这种方式，风格感知编码器能够捕捉到更细腻的风格细节。

为了进一步整合内容信息，实现图像驱动的风格迁移，StyleShot 训练了一个内容融合编码器。首先使用 Contour 的处理方式在保留内容的同时去除内容图像上的风格信息，然后构造一个 ControlNet[5]-like 结构的网络提取空间信息。

最后，StyleShot 构造了一个大规模的风格数据集以提升模型学习富有表现力和广义的风格表征的能力。该数据集一共包含了几万种风格、几百万张风格图片以及对应的福根文本描述。

Human	StyleCrafter	DEADifff	StyleDrop	InST	StyleAligned	StyleShot
text ↑	9.7%	19.3%	6.0%	12.7%	8.0%	**44.3%**
image ↑	14.3%	8.0%	4.0%	6.3%	17.3%	**50.0%**
CLIP	**StyleCrafter**	**DEADifff**	**StyleDrop**	**InST**	**StyleAligned**	**StyleShot**
text ↑	0.202	**0.232**	0.220	0.204	0.213	0.219
image ↑	**0.706**	0.597	0.621	0.623	0.680	0.640

表 5　StyleShot 中的评估结果

如表 5 所示，实验结果表明，StyleShot 能够有效捕捉各种风格特征，从颜色和纹理等基本元素到布局、结构和阴影等复杂元素，最终生成与文本提示一致的风格化图像。这展现了我们风格感知编码器在提取丰富且富有表现力的风格嵌入方面的有效性。

另外，得益于我们的内容融合编码器，StyleShot 还擅长将风格迁移到内容图像上。如图 5、6 所示，我们的 StyleShot 可以将任何风格（甚至包括光影、点画

图 5　文本驱动风格迁移可视化结果

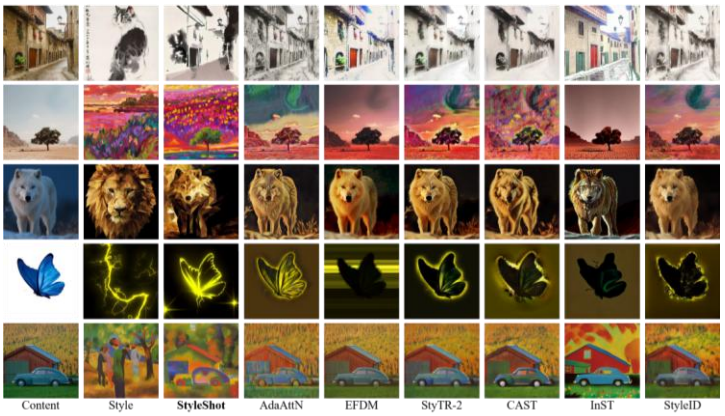

图 6　图像驱动风格迁移可视化结果

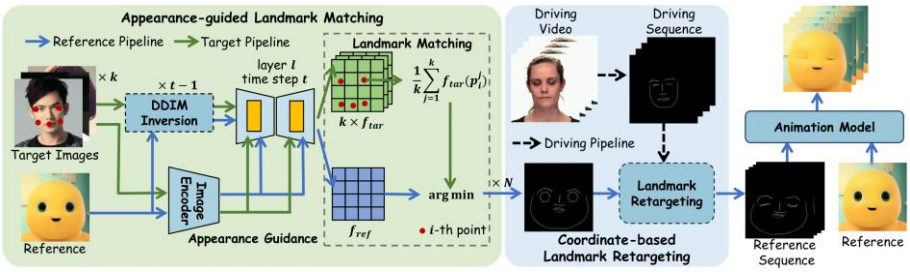

图 7　FaceShot 框架图

法、低多边形和平面等复杂高级风格）迁移到各种内容图像（例如人物、动物和场景）上，而基线方法主要擅长绘画风格，而难以处理这些高级风格。这证明了内容融合编码器在实现卓越风格迁移性能的同时，还能保持内容图像的结构完整性。

3.2 肖像动画

当前的肖像动画方法大致由面部关键点序列控制驱动。这类方法已经表现出了良好的动作可控性与面部保真能力，特别是在人脸驱动任务中能够生成自然、稳定的动画效果，然而由于其核心依赖的人脸关键点检测以及动作序列生成模块大多是在真实高质量面部数据集上进行监督训练，在处理非人类角色（如表情包、玩具、卡通角色）时，往往无法准确识别关键点分布和迁移面部动作，导致动画生成阶段出现结构错位、嘴型崩塌等现象。为了解决这些问题，本节简要介绍 FaceShot，一个无需训练的肖像动画生成框架。如图 7 所示，主要包括语义引导关键点匹配、坐标系建模动作变换以及肖像动画三个模块：

基于扩散模型的强大泛化能力，DIFT[6]发现不同图像的语义相近区域的预训练扩散模型特征是相似的，可以直接进行关键点匹配。然而人脸和非人类角色之间存在较大的域差异，常常会造成错误匹配。为了拉近不同域在特征空间的差异，FaceShot 结合 IP-Adapter[7]，将参考图像作为外观引导注入扩散过程。另外 FaceShot 还构建了包含眼睛、嘴巴、眉毛等五个部分的外观图库，自动选取相近域作为辅助目标，进一步缓解非人类角色与人类语义空间之间的域间差异。实现对非人类角色的面部关键点的精准定位。

为了更精准地捕捉驱动视频中的面部动作，FaceShot 构建了全局与局部坐标系，用于显式建模并迁移整体与局部表情变化。具体而言，FaceShot 利用参考图中面部轮廓两端点定义全局坐标系，通过原点位置与坐标轴角度的变化建模头部的整体位移与旋转；同时，在眼、眉、嘴、鼻等局部区域分别建立子坐标系，通过点在各自坐标系中的相对变化，刻画细节动作的动态变形。

最后将对应肖像的关键点序列输入到任意的关键点驱动的肖像动画模型中，就可以得到最终的动画结果。图 8 和表 6 的实验结果显示，FaceShot 在非人类角色上表现出色。相比现有方法，FaceShot 在身份保持

Methods	Metrics				User Preference		
	ArcFace ↑	HyperIQA ↑	Aesthetic ↑	Point-Tracking ↓	Motion ↑	Identity ↑	Overall ↑
FaceVid2Vid	0.525	33.721	4.267	6.944	3.58	3.83	4.52
FADM	0.633	39.402	4.522	6.993	1.93	2.04	1.96
X-Portrait	0.490	52.357	4.754	7.301	1.47	1.63	1.57
Follow Your Emoji	0.612	52.056	4.906	6.960	6.91	6.67	6.74
AniPortrait*	0.634	55.951	4.928	6.367	5.84	5.64	5.39
MegActor*	0.613	40.191	4.855	7.183	6.53	6.75	6.26
LivePortrait*	0.893	53.587	5.092	7.474	7.33	7.08	7.11
MOFA-Video	0.695	52.272	4.952	14.985	3.27	3.04	3.18
FaceShot	**0.848**	**53.723**	**5.036**	**6.935**	**8.14**	**8.32**	**8.27**

表 6　FaceShot 中的评估结果

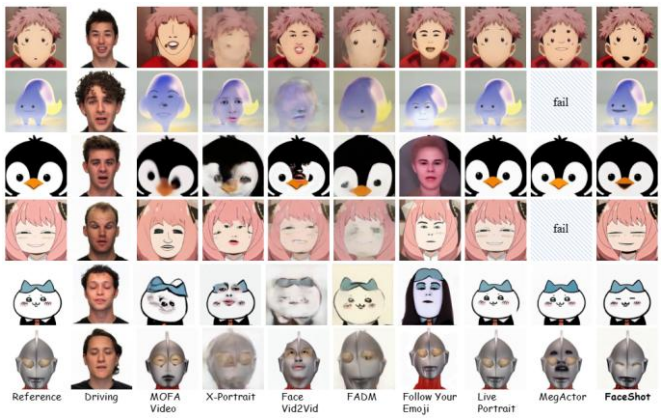

图 8　FaceShot 中的可视化结果

图 9　将 FaceShot 作为插件

图 10　非人类驱动视频可视化结果

（ArcFace）、图像质量（HyperIQA）和动作还原（Point Tracking）等多个指标上均取得领先，尤其在结构不规则、风格差异大的角色（如玩偶、卡通形象、动物）上表现更为稳定。现有方法常常因关键点检测不准或驱动迁移失真而导致动画崩坏、嘴型错位等问题，而 FaceShot 利用语义引导的关键点匹配与坐标建模动作迁移，显著提升了角色动作的还原度和连贯性。

除了作为独立的肖像动画生成框架，FaceShot 还展现出出色的模块化扩展能力。如图 9 所示，在插件化应用方面，FaceShot 可作为关键点序列预测模块集成到现有的关键点驱动方法中（如 MOFA-Video 和 AniPortrait），显著提升其在非人类角色上的动画稳定性与结构一致性。

此外，FaceShot 还支持从非人类驱动视频中提取动作信号，并将其迁移到任意参考角色，实现跨类别、跨风格的开放域角色动画。如图 10 所示，这一能力打破了传统肖像动画对人类驱动数据的依赖，展示了 FaceShot 向通用、开放场景扩展的广阔潜力。

四、总结与未来展望

本文系统梳理了同济大学VILL实验室在多模态生成式人工智能领域的一系列研究成果，涵盖语言数据生成、视觉数据生成、图像风格迁移与肖像动画等多个子方向。实验室针对不同模态下生成技术的关键挑战，提出了具有创新性的技术方案，并在多个顶级会议如NIPS、CVPR、ICLR中发表相关成果，部分工作更是获得了广泛的开源社区认可，展现了团队在该领域的前瞻性研究能力与工程实现水平。

在语言数据生成方面，AlchemistCoder通过构建协调化多源微调框架，有效缓解了Code LLMs在多源数据整合中的冲突问题，并进一步提升了模型的泛化能力和理解能力，不仅在HumanEval和MBPP等标准评测上取得领先，还在跨语言、多任务场景中展现了出色性能。该工作展示了通过精细化提示与任务设计进行数据重构的巨大潜力。

在视觉数据生成方面，UniDD以任务知识挖掘和任务驱动扩散为核心，构建了一种通用的视觉数据蒸馏框架，实现了高质量、高效率的数据合成，显著降低了模型训练对真实数据的依赖，为低资源条件下的视觉模型训练开辟了新路径。尤其是在图像分类、检测和分割等多任务场景中展现了良好适应性，验证了其在实际应用中的可行性和可扩展性。

在图像与视频生成方向，StyleShot和FaceShot分别从风格迁移与动画生成出发，打破了传统方法对测试时调优和人脸驱动数据的依赖。StyleShot通过构建风格感知与内容融合编码器，成功实现了任意风格与任意内容之间的高质量迁移；而FaceShot则通过语义引导的关键点匹配与精细的动作建模，实现了对非人类角色的稳定驱动，并具备良好的插件化集成能力，支持开放域角色间的动作迁移。

展望未来，生成式人工智能将在更多实际应用中发挥更广泛作用。VILL实验室将继续围绕“多模态、高质量、开放域”的核心目标推进研究工作。一方面，在语言生成方面将进一步探索指令构造、模型行为监督与对齐等关键问题，提升模型对人类意图的理解与响应能力；另一方面，在视觉生成方面，将聚焦于大规模多模态数据的结构化建模与跨域泛化能力的提升，拓展生成模型在医疗、工业、娱乐等真实场景中的落地空间。同时，结合扩散模型与结构建模的前沿方法，探索更多有效的通用生成范式，为生成式AI系统的稳定性、可控性与通用性奠定更坚实的基础。

责任编委　王金甲

参考文献

[1] Zifan Song, Yudong Wang, Wenwei Zhang, Kuikun Liu, Chengqi Lyu, Demin Song, Qipeng Guo, Hang Yan, Dahua Lin, Kai Chen, Cairong Zhao. "AlchemistCoder: Harmonizing and Eliciting Code Capability by Hindsight Tuning on Multi-source Data. " Advances in Neural Information Processing Systems(NIPS), 2025, 2185—2214.

[2] Ding Qi, Jian Li, Junyao Gao, Shuguang Dou, Ying Tai, Jianlong Hu, Bo Zhao, Yabiao Wang, Chengjie Wang, Cairong Zhao. "Towards Universal Dataset Distillation via Task-Driven Diffusion." Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 10557-10566

[3] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, Cairong Zhao. "Styleshot: A snapshot on any style." TPAMI:2025.

[4] Junyao Gao, Yanan Sun, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, Cairong Zhao. "FaceShot: Bring Any Character into Life." The Thirteenth International Conference on Learning Representations (ICLR).

[5] Lvmin Zhang, Anyi Rao, Maneesh Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836-3847

[6] Luming Tang and Menglin Jia and Qianqian Wang and Cheng Perng Phoo and Bharath Hariharan. "Emergent Correspondence from Image Diffusion". Thirty-seventh Conference on Neural Information Processing Systems (NIPS). 1363—1389.

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, Wei Yang. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. " arXiv preprint arxiv:2308.06721.

高俊尧

同济大学计算机科学与技术学院 2022 级博士研究生，导师为赵才荣教授。在 ICLR, ICML, NIPS, ICCV, CVPR, AAAI, TPAMI, IJCV 等国际期刊会议发表论文 10 余篇，主要研究方向为图像/视频生成、行人再识别、隐私安全。

Email: junyaogao@tongji.edu.cn

宋子帆

同济大学博士生，师从赵才荣教授，本科毕业于同济大学。在 NeurIPS、ICML、AAAI 等国际期刊会议发表一作论文 4 篇，主要研究方向为多模态学习、Data-centric AI、大模型微调。

Email: 2111139@tongji.edu.cn

齐 鼎

同济大学博士生，师从赵才荣教授。在NeurIPS、CVPR国际会议发表一作论文2篇，主要研究方向为数据集蒸馏、Data-centric AI。
Email: 2011267@tongji.edu.cn

赵才荣

工学博士，同济大学计算机科学与技术学院教授，博士生导师，计算机智能教研室主任。曾任香港理工大学兼职研究员（2016-2017）。目前担任上海市计算机学会计算机视觉专委会主任，中国图象图形学学会青工委秘书长，中国人工智能学会粒计算与知识发现专委常委，中国计算机学会杰出会员，担任IEEE TMM Guest Editor、《中国图象图形学报》、《计算机科学》青年编委。已在TPAMI、IJCV、《中国科学.信息科学》、CVPR、ICML、NIPS等发表学术论文50余篇，研究成果获2022年上海市科技进步一等奖（序4/13）以及2023年上海市自然科学二等奖（序1/4）。
Email: zhaocairong@tongji.edu.cn

顶会观察

CVPR 2025

清华大学 叶栩冰 唐彦嵩

国际计算机视觉与模式识别会议（CVF/IEEE Conference on Computer Vision and Pattern Recognition，CVPR）是计算机视觉和模式识别领域最重要的会议之一。CVPR 于 1983 年在美国华盛顿特区举办，每年举办一次，一般在美国举办。CVPR 2025 于 2025 年 6 月 11 日至 15 日在美国田纳西州纳什维尔的音乐城中心（Music City Center）举办。

一、 会议概况

CVPR 2025 共收到 13,008 篇有效投稿论文，经过严格评审后录用 2,878 篇，录用率 22.1%，两项指标均创历史新高。投稿作者总数超过 42,000 人，本届会议注册参会人数突破 10,000 人，其中线下参会约 9,000 人，来自全球 70 多个国家和地区。美国本土注册人数最多，中国大陆紧随其后，韩国、德国、加拿大和日本分列三至六位。

会议前两天（6 月 10–11 日）为 Tutorial 与 Workshop 时段，共组织了超过 120 场 Workshop 与 20 余场 Tutorial。主会议论文展示继续采用“口头报告+海报”双轨制：口头报告限时 8 分钟，仅少数论文入选；所有录用论文均须以海报形式展示，每轮海报展约 400–500 篇论文，持续 90 分钟。

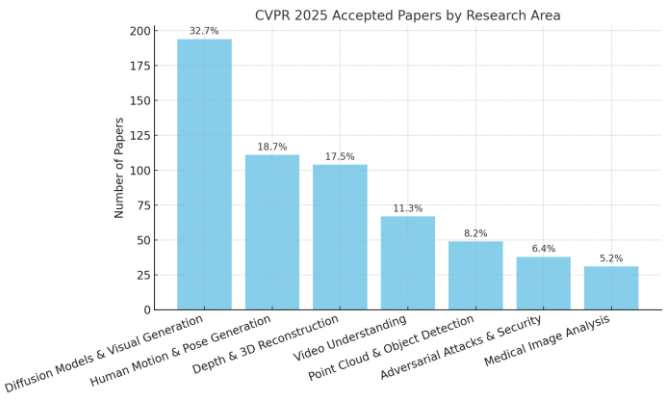

图 1　所有中稿文章的领域统计

图 2　纳什维尔会场大幅海报

CVPR 官方公布了各个细分领域的论文接收情况，如图 1 所示。可以看到，图像与视频生成领域今年度的论文接收数量最多。

根据会方统计，今年大会共收到 4 万多名作者提交的 13008 份论文。相比去年（11532），今年的投稿数量增长了 13%，最终有 2872 篇论文被接收，整体接收率约为 22.1%。在接收论文中，Oral 的数量是 96（3.3%），Highlights 的数量是 387（13.7%）。今年共有 14 篇论文入围最佳论文评选，最终 5 篇论文摘得奖项，包括 1 篇最佳论文、4 篇最佳论文荣誉提名。

此外，大会还颁发了 1 篇最佳学生论文、1 篇最佳学生论文荣誉提名。

计算机视觉技术的火热给大会审稿带来了空前的压力。本届投稿作者数量、论文评审者和领域主席（AC）数量均创下新高。今年前来现场参会的学者也超过

9000 人，他们来自 70 余个国家和地区。

二、 参会感受

飞机落地纳什维尔国际机场，海关大屏就滚动播放“Welcome Nashville”，行李转盘旁竖起了巨型霓虹吉他——会徽与田纳西音乐之城主题第一次同框。接驳大巴一路驶过百老汇大街，沿街酒吧里白天就响起乡村摇滚，司机干脆把车内音响也调到同一频道，途中已有人跟着节奏打拍子，旅途疲劳瞬间减半。

主会场 Hall A 被布置成双层同心圆：内圆 360°环形屏幕滚动播放论文短视频，外圆 380 块海报板呈放射状排布。环形顶部悬挂 36 只音箱，每到整点播放一段 AI 生成的乡村旋律——提醒换场。作者们在“舞台”中央站成一圈，观众像歌迷一样高举手机扫码、递贴纸索要签名。由于场地回音大，讨论基本靠吼，两天下来不少人嗓子沙哑，于是展台免费润喉糖成了最抢手周边。

三、 大会获奖论文

会议共选出了 5 篇论文摘得奖项，包括 1 篇最佳论文、4 篇最佳论文荣誉提名。此外，大会还颁发了 1 篇最佳学生论文、1 篇最佳学生论文荣誉提名。其中的最佳论文、1 篇最佳论文荣誉提名和 1 篇最佳学生论文如下。

最佳论文 1：VGGT：Visual Geometry Grounded Transformer[1]：在计算机视觉领域，3D 场景理解一直是极具挑战性的任务。传统方法，如通过 SfM（结构光运动恢复）、BA（光束平差法）和 MVS（多视图立体视觉）等多阶段流程实现 3D 重构，不仅步骤繁琐，计算成本高，而且耗时较长。而今年 CVPR 最佳论文中的 VGGT（Visual Geometry Grounded Transformer），作为一项开创性成果，为这一领域带来了全新的解决方案。VGGT 旨在实现 “全栈一次性 3D 场景理解”，构建了一个拥有 12 亿参数、24 层交替注意力的纯前馈 Transformer 框架。在这个创新框架下，原本复杂的 3D 重构任务得以大幅简化，仅需一次前向传播，就能在不到 1 秒的时间内，将 1 张到数百张无约束图像，同时解析成相机内外参、稠密深度、视点统一坐标系的点云，以及可跨帧跟踪的 3D 点轨迹 ，极大地提升了处理效率。

图 3　纳什维尔会场的 CVPR 标识

图 4　最佳论文 VGGT 的颁奖现场

在模型构建方面，VGGT 首先利用已冻结的 DINO-V2，将每张输入图像切分成 14×14 的 patch token，以此作为基础视觉单元。为了有效区分参考帧，模型特别引入了可学习的 “相机 token” 与四个注册 token 。其中，“相机 token” 专门负责学习相机相关参数，而注册 token 则用于捕捉全局场景特性。

在主干网络设计上，VGGT 采用了“逐帧自注意力 + 全局自注意力” 交替堆叠 24 次的独特架构。这种设计可谓独具匠心，一方面，逐帧自注意力能够处理每一帧图像内的 patch tokens，确保局部信息的一致性和准确性；另一方面，全局自注意力实现了不同帧间 tokens 的交互，有效整合多视角信息，从而全面理解场景。并且，这种交替设计巧妙地避免了传统交叉注意力可能导致的显存爆炸问题，同时保证了模型对任意帧排列的等变性，极大提升了模型的实用性和稳定性。

为了让模型具备强大的泛化能力，研究团队在 17 个公开的 3D 数据集上，使用 1500 万张图像对 VGGT 进行端到端训练。这些数据集涵盖了丰富多样的场景，包括室内场景（如 ScanNet、Replica）、室外场景（如 MegaDepth、Mapillary）、合成场景（如 Kubric、Objverse），以及手持设备、无人机、车载等多源采集场景 。在训练过程中，采用多任务损失函数来优化模型，它由相机 Huber 损失、深度和点云的不确定性感知 L1 及梯度损失，以及点跟踪 L1 损失共同组成，同时辅以大规模的颜色、模糊、灰度等数据增强策略，进一步提升模型的鲁棒性。研究人员还发现，VGGT 的预训练特征具有出色的通用性，可作为强大的通用 3D 先验。VGGT 为 3D 场景理解开辟了新路径，尽管存在一定局限，但无疑为该领域的后续研究提供了极具价值的参考和方向。

最佳论文荣誉提名 1： MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos [2]：在计算机视觉领域，从动态场景的单目视频里，精准、高效且稳健地估算相机参数与深度图，始终是极具挑战性的研究热点。传统的运动恢复结构（SfM）以及单目 SLAM 技术，大多依赖于静态场景，并且要求输入视频具备大量视差。一旦这些条件无法满足，例如在无约束摄像机运动、未知视野范围，或者存在动态场景干扰时，就极易导致估算结果出现偏差 。近年来，神经网络方法虽尝试打破这些局限，然而在面对复杂动态视频时，却暴露出计算量过大，或者可靠性欠佳的问题。在此背景下，研究团队创新性地提出了 MegaSaM 这一视觉 SLAM 框架，致力于攻克野外动态场景下单目视频的相机跟踪与深度估计难题 。

MegaSaM 框架的核心亮点，在于对现有深度视觉 SLAM 架构进行了深度扩展与优化 。一方面，它巧妙借鉴了 DROID-SLAM 等系统中可微分捆集调整（BA）层的优势。通过迭代更新场景几何与相机姿态变量，同时借助相机和光流监督，从海量数据中学习中间预测结果，为在复杂动态场景中实现精准的相机姿态估计筑牢根基 。另一方面，MegaSaM 开创性地将单目深度先验和运动概率图融入可微分 SLAM 范式。这种融合策略，极大地增强了模型对动态场景的适应能力，使其能够更好地应对复杂多变的实际情况 。

MegaSaM 深入剖析了视频中结构和相机参数的可观测性，并据此引入了不确定性感知的全局 BA 方案。当相机参数受输入视频的约束较弱时，该方案能够显著提升系统的稳健性，同时还达成了在测试时无需对网络进行微调，就能高效获取一致视频深度的目标 。

MegaSaM 在合成数据集与真实世界数据集上开展了大量实验。结果显示，MegaSaM 在相机姿态与深度估计的精度方面，远超先前以及同期的方法。并且，在运行时间上，MegaSaM 也表现出色，要么比其他方法更快，要么与之相当 。这充分验证了 MegaSaM 在处理无约束相机路径、复杂动态场景以及低视差视频等具有挑战性的场景时的有效性，为动态场景下的单目视觉定位与建图提供了创新的解决方案，有望推动相关领域迈向新的发展阶段 。下的单目视觉定位与建图提供了革新性的解决方案。

最佳学生论文 1：Neural Inverse Rendering from Propagating Light [3]：此论文首次提出了基于物理的多视角动态光传播神经逆渲染系统。其核心创新点在于对神经辐射缓存技术进行了时间分辨维度的拓展 。神经辐射缓存作为一种加速逆向渲染的技术，通过存储从任意方向抵达任意点的无限反射辐射来实现加速效果。研究团队创新性地将时间因素融入其中，构建出能够精确计算直接和间接光传输效应的模型 。在实际应用场景中，当面对闪光激光雷达系统捕获的测量结果时，该模型优势尽显，即便是在强间接光存在的复杂场景下，也能够实现当前最先进水平的三维重建 。

从具体实现过程来看，在模型搭建初期，系统借助先进的算法，对多视角视频中的光线传播数据进行细致分析与处理，构建起初始的光线传播模型。随后，基于拓展后的神经辐射缓存技术，模型能够持续跟踪光线在不同时间、空间维度下的传播路径与反射情况。在处理间接光时，模型通过对多次反射光线的精准捕捉与分析，有效克服了传统方法中易出现的光线信息丢失或错误计算的问题 。多视图时间分辨重新照明这一创新功能，更是允许用户在不同时间维度下，对捕获场景进行重新

照明模拟，进一步挖掘场景中的光线细节与潜在信息。

四、 总结展望

CVPR 2025 再次刷新规模与质量纪录，投稿量、录用论文、参会人数均创新高。研究主题呈现“三维化、多模态、大模型、可解释”四大趋势：NeRF 与 3D GS 继续深化，扩散模型走向高效与可控，多模态大模型参数突破百亿，同时可解释性与评测方法成为焦点。数据、算力与标注成本持续攀升，催生“开放权重+开放数据”的新研究范式；工业界与学术界合作更加紧密，现场招聘与技术 Demo 成为会议标配。如何在数据墙与资源墙日益逼近的背景下保持创新，将是 CV 社区在 2026 及以后必须回答的问题。

责任编委 张 青

参考文献

[1] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, David Novotny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 5294-5306

[2] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10486-10496

[3] Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, David B. Lindell; Neural Inverse Rendering from Propagating Light. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10534-10544

叶栩冰

清华大学深圳国际研究生院硕士生，研究方向多模态大模型。CVPR 2025 录用论文为 VoCo-LLaMA: Towards Vision Compression with Large Language Models 和 ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models，两篇论文对视觉语言大模型进行视觉 token 剪枝和压缩，取得了显著效果。

Email: yxb_tongji@163.com

唐彦嵩

清华大学深圳国际研究生院副教授、博士生导师、科研处副处长。分别在清华大学自动化系获得工学学士和博士学位，并于英国牛津大学从事博士后工作。主要从事具身智能、计算机视觉、模式识别等领域的相关工作，以第一/通讯作者发表 TPAMI 等 IEEE 汇刊和 CVPR 等 CCF-A 类会议论文 40 余篇，主持广东省杰青、国家重点研发计划课题、中国科协青年托举工程等项目，获 2024 年公安部科学技术奖一等奖、2024 年广东省科学技术奖（科技进步）二等奖和国际顶会竞赛冠军 3 项，担任 CVPR、FG 等国际会议领域主席、国际期刊 JVCI 编委以及中国人工智能学会模式识别专业委员会（CAAI-PR）常务委员兼副秘书长等学术职务。

Email：tang.yansong@sz.tsinghua.edu.cn

南通大学李洪均教授访谈

2025 年 8 月 20 日，《CCF-CV 专委简报》在线采访了南通大学博士生导师李洪均教授。下面是采访实录。

李老师，您好！首先，请您分享一下您的个人学习和研究经历。

我在南京工业大学完成本科和硕士阶段的学习，随后进入南京航空航天大学攻读博士学位，研究方向聚焦人工智能与模式识别。博士毕业后，我入职南通大学信息科学技术学院，2013 年赴加拿大 Concordia 大学孙靖夷（Ching Y.Suen）教授团队访问交流一年。期间围绕手写汉字识别和人脸识别开展了相关交流合作。我的研究长期围绕视频行为分析与理解展开，具体涉及：

（1）复杂场景下的目标检测与跟踪；（2）跨模态视频语义理解，融合视觉、文本与音频信息，实现事件级解析；（3）小样本与自监督学习，解决行业数据稀缺导致的模型泛化瓶颈；（4）面向公共安全的实时视频分析系统研制与产业转化。

多学科交叉学习与工作的经历，使我形成“机理-数据-场景”三维并重的研究范式：既重视数学建模与算法可解释性，又擅长利用大数据与深度学习技术提升系统性能，更能从终端应用需求出发反向优化模型设计。未来，我将继续围绕“可信、可控、可用”的新一代视觉智能开展研究，推动科研成果在智慧城市与公共安全领域的落地。

您主要从事动作识别、视频异常检测等领域的研究，能否对这些领域的研究现状及其发展潜力进行一些分析？是什么契机使您选择了这些研究领域呢？

动作识别与视频异常检测领域正处在一个从“技术驱动”迈向“场景驱动”与“价值驱动”的关键转折点，从技术指标的“刷榜”，转向解决实际问题的“落地”。

前几年的研究焦点是如何设计更复杂的模型，在几个标准数据集上把准确率提高零点几个百分点。但现在大家意识到，这种“内卷”的边际效益正在降低。当前的主流框架，比如 3D 卷积神经网络和 Video Transformer，已经比较成熟，研究正转向更精细的改进，比如如何让模型更高效、如何利用无标签数据。

未来的巨大潜力在于与千行百业的深度融合。比如，在工业质检中实时发现产品缺陷，在智慧养老中智能监测，这些才是技术真正创造价值的地方。同时，与多模态大模型融合也是必然趋势，让系统不仅能“看”，还能“听”和“读”，能用自然语言和我们交互，成为未来通用人工智能的“视觉核心”。

至于我个人的研究契机，源于两次触动。一是十多年前第一次看到深度学习在识别领域的发展，让我确信“机器看懂世界”将成为现实。二是了解到安防监控员需要日夜紧盯屏幕后，产生的责任思考——我想开发能充当“永不疲倦的 AI 助手”的技术，让他们从繁重的监视中解脱出来，去做更重要的决策，这种“科技向善”

的可能性，让我看到了科研的社会价值。因此，我的研究始终围绕着两个轴心：一是追求技术前沿，二是探索如何让这些先进技术解决真实世界中有挑战、有价值的问题。这种结合让我始终保持着高度的研究热情。

您主持了多项国家级、省级自然科学基金项目，发表了高水平学术论文 80 余篇，授权专利 20 余件，请问在这些科研成果中，您觉得最值得骄傲的成果是哪一项？能否介绍一下？

我们近期取得了一些与应用场景结合比较紧密的研究成果，将研究用于实际工程中，服务地方经济发展。包括视频异常行为识别、图像异常检测等，也没有什么值得骄傲的成果，只能说每个成果或多或少能够解决企业的一些实际问题。

除了自然科学基金项目外，您还主持了多项产学研合作项目，能否介绍一下您在科研成果转化方面的经验，分享一下您的经历？

我最近做了一个农业口的项目，将图像处理和模式识别技术应用于水产养殖业。利用水面无人船自动监测水质环境信息并及时预警，利用无人机之间的信息交互，进行精准数字化饲养投饵增氧。

在项目推进过程中，我们遇到了不少挑战。比如水面无人船在复杂水域环境下的航行稳定性问题，水流的冲击、水下障碍物等都会影响其正常行驶和数据采集的准确性。为此，我们通过优化无人船的动力系统和导航算法，增强了它应对复杂环境的能力。对于无人机之间的信息交互，信号干扰和数据传输延迟等主要难题，我们采用了更先进的通信技术和信号增强设备，确保无人机之间能够高效、稳定地进行信息共享。

经过一段时间的努力，项目取得了显著成效。水质环境监测的及时性和准确性大幅提高，能够提前发现水质恶化等问题，为水产养殖的健康发展提供了有力保障。精准数字化饲养投饵增氧，不仅提高了饵料和氧气的利用效率，降低了养殖成本，还提高了养殖产量和质量，实现了绿色养殖。最让养殖户高兴的是不再需要 24 小时无间断巡塘，可以在手机上直接查看养殖情况，睡上安稳觉！

您在教学方面主持江苏省本科高校课程思政典型案例 1 项，主持教育部重点领域首批人工智能教学应用示范项目等，获“中国移动”教学名师等荣誉称号，能否请您结合这些成果，分享一下在教学改革与实践中的主要经验与心得？

教学改革和实践经验谈不上，主要是平时做了一些工作，得到了同行们的认可。作为一名老师，教学是首要任务，如何上好一门课，说实话很难，很费时间。如何让学生喜欢一门课或者认真听讲，是难上加难。网络的发展，课堂教学的吸引能力骤降，如何让学生愿意去听，主动去听，自觉地思考，这些都是教学改革的难点。我们教学团队通过多年的教学积累，不断紧跟技术发展，做了一些工作，获得一些奖项。获奖固然是好，但教学的本质还是希望学生能够学到真知识，把知识用于实践中，学行交替。

您在科研、教学之余，还承担了行政工作及一些学术服务，每一项工作都必然占据您大量时间，能否跟大家分享一下您是如何协调这些工作的？

很多时候教学、科研和行政等工作很难协调，只能说哪个优先级高就先处理哪个。时间是固定的，很多时候占用大量的休息时间。有时候自己也想有一个合理的时间管理，但很多时候事与愿违，如何分配，各有所需吧。

当前，计算机视觉领域普遍被认为存在较高的竞争强度与内卷现象，您如何评价这一现状？在您看来，该领域仍存在哪些关键性科学问题或技术瓶颈，值得研究者进一步深入探索与突破？

高竞争强度说明该领域充满活力，但“内卷”确实存在，主要体现在大家扎堆在热门任务上，靠微调模型来换取微弱的性能提升。要破局，关键在于走出舒适区。

与其在旧赛道上挤破头，不如去定义新任务、挑战真实世界的新场景。评价标准也不能只看准确率，更要看效率、鲁棒性和可解释性。一个又快又稳又容易理解的模型，远比一个笨重但精度高 0.1%的模型更有价值。

除了眼前的“内卷”，我认为还有几个更本质的“硬骨头”值得去啃：小数据与零样本学习、因果推理、极致效率与边缘部署、安全与可信赖性。这些问题的任何一个突破，其意义都远大于在现有数据集上刷高分数，也正是避免内卷、开创下一轮浪潮的关键。

您是南通大学智能信息处理团队负责人，能否介绍一下您的团队，以及您是如何管理团队的？

团队除了我之外，还有 7 名老师，目前在校研究生 20 余人。团队目前主要在视频行为分析方面开展研究。我本人非常重视学术交流，有相关的学术会议就会组织青年教师和学生参加。团队组会每月不定期开展，每天只要有时间，我都要和学生进行简短的交流。青年教师是团队的核心力量，我们团队通过青蓝工程带好每一位年轻教师。欢迎具有共同学术理想和共同发展目标的青年教师加入我们团队。

如果吐露研究工作者的心声，您最想说的是什么？

砥砺前行勿停步，攀登绝顶待时来。

责任编委 余烨 赵振兵

李洪均

李洪均，博士，教授，南通大学信息科学技术学院副院长，电子信息工程专业负责人，智能信息处理团队负责人。2011 年博士毕业于南京航空航天大学，2013 年赴加拿大 Concordia 大学师从 Ching Y.Suen 教授。主要从事人工智能、模式识别、机器学习、图像处理和视频理解等方面的研究，在 IEEE TMM、PR 等期刊发表 SCI/EI 收录论文 60 余篇，授权专利 20 余件；主持或参与多项国家级、省部级自然科学基金项目；曾担任计算机国际会议 ICBDA 会议主席，IAPR TC3 Workshop on ANNPR 宣传主席和 ICIAI 技术委员；担任 IEEE ACCESS 的专刊编辑，担任 IEEE TPAM、IEEE TIP、IEEE TMM、IEEE TCSVT 等期刊审稿人；获江苏省教育科学研究成果奖三等奖 1 项，自动化科技进步奖二等 1 项。

教学方面：国家一流本科专业电子信息工程专业负责人，主持江苏省一流本科课程 1 门，主编教材《数字信号处理》获通信学会信息通信科普教育精品教材，主持江苏省本科高校课程思政典型案例 1 项，主持教育部重点领域首批人工智能教学应用示范项目等，获“中国移动”教学名师等荣誉称号。

指导学生方面：指导研究生发表 SCI/EI、中文核心等论文 50 余篇；申请、授权专利 30 余项；指导学生获中国研究生数学建模竞赛二等奖 5 项、三等奖 7 项；获电子设计大赛一等奖 1 项、三等奖 6 项，优秀研究生论文 1 项等；指导本科生获江苏省本科优秀毕业设计三等奖 2 项，获全国人工智能大赛二等奖 1 项等。

社会任职方面：现任江苏省高教学会电子信息类专业教学研究会常务理事，任中国自动化学会、中国人工智能学会、中国计算机学会、中国图象图形学学会旗下人工智能和计算机视觉领域专业委员。

委员好消息

✪ 2025 年 7 月 24 日，2024 年度陕西省科学技术奖励结果揭晓，重庆邮电大学**肖斌**等完成的“智能模型协同安全构建理论与关键技术”获自然科学二等奖。

✪ 2025 年 8 月 19 日，CCF-CV 专委会执行委员、哈尔滨工程大学**刘海波**获第五届全国高校教师教学创新大赛课程思政赛道一等奖。

✪ 2025 年 8 月 21 日，教育部公示了第三批国家级一流本科课程认定结果，CCF-CV 专委会 9 位执行委员的课程入选。浙江大学**赵洲**的《机器学习：模型与算法》入选线上一流课程，北京航空航天大学**刘祥龙**的《数据结构与程序设计（信息类）》、中国科学院计算技术研究所/中国科学院大学**王瑞平**、**陈熙霖**的《概率论与数理统计》、苏州科技大学**胡伏原**《人工智能基础》和中南大学**赵于前**《数字图像处理》入选线下一流课程，中国石油大学**李宗民**《大学计算机》和西安电子科技大学**董伟生**的《人工智能系统实验》入选线上线下混合式一流课程。

✪ 2025 年 8 月 22 日，CCF-CV 专委会执行委员、北京大学**彭宇新**获 2025 年青年科学基金项目 A 类（原国家杰青项目）延续资助。

✪ 2025 年 8 月 26 日，2024 年度上海市科学技术奖获奖名单公布，上海海事大学**周日贵**等完成的“特殊生物资源跨境检测与溯源关键技术及仪器装备研发”获技术发明二等奖。2024 年度共 206 项（人）获奖。

责任编委 刘海波

流视频长上下文理解数据集及模型开源代码

中国科学院 钱胜胜 杨振宇

传统的视频理解针对输入的视频，只能回答若干时序无关的独立问题，缺乏对事件间时序关系的理解，因此从这个角度而言，传统视频理解更像“看完再理解”：对一段录播视频只回答一些彼此独立的问题，缺少对事件先后、因果和人物状态变化的把握。因此在快速变化的场景里，它往往只能给出片段式信息，辅助决策的作用有限。而流式视频理解则更符合人类看视频的习惯，是一种“边看边理解”的方式。模型与画面同步，一边识别刚发生的事件，一边沿同一时间线连续追问一系列问题，把零散画面串成连贯叙事，更适合开放世界中的不确定与突发。这种能力天然适配实时场景：比如说直播解说与视频描述生成、可穿戴设备的第一视角辅助、安防监控的及时预警、机器人与自动驾驶的在线感知决策，以及需要快速分析的智能军用场景。

1、流视频长上下文理解数据集 SVBench

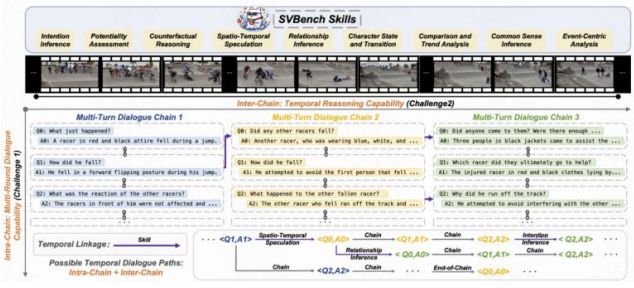

图 1 流视频时序理解与持续推理能力评测图示

尽管大型视觉语言模型在传统视频理解基准上表现突出，但其在长上下文流式视频理解中的能力仍缺乏系统评估。现有基准多基于单轮、孤立的问答任务，难以衡量模型在连续时序推理中的表现。为此，本文推出SVBench——一个专为流式视频理解设计的新型基准，如图 1 所示，通过时序多轮问答链全面评估 LVLMs 的时序理解与持续推理能力。

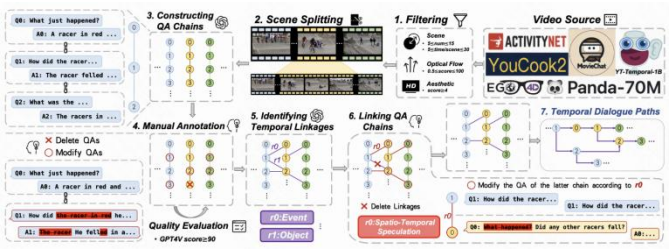

图 2 SVBench 数据集构建流程图

如图 2 所示，本文从 YTTemporal-1B、YouCook2 等 12,989 个公开视频数据源中筛选高质量视频，依据场景复杂度、光学流畅度等标准，通过 PySceneDetect 工具进行场景分割，将视频切分为 5-15 个场景的片段，并调整片段边界以保持连续性。为评估模型对视频时序对话的理解能力，本文设计半自动化标注流程：首先利用 GPT-4 等大型视觉语言模型为每个视频片段生成 5-6 轮初始 QA 对，形成“QA 链”。随后，人工标注者对生成的 QA 进行修正，确保问题连贯、指代明确（如统一使用第三人称），并与视频内容严格对齐。整个过程耗时 3 个月，涉及 30 余名专业标注者。

通过 GPT-4 对标注结果进行 7 维度量化评估（准确性、逻辑一致性、时间关联性等），并设置总分≥90 的严格阈值。未达标者需重新修订，保证数据的高可靠性。这一机制确保 QA 链具备深度推理价值。为支持跨片段的时间推理，本文通过大型语言模型分析相邻 QA 链的潜在关联（如相同实体、事件发展），构建关系五元组（包含问题、答案及关系类型）。人工进一步调整后续 QA 对，形成逻辑闭环：既保持单链内问答的连贯性，又建立跨链的逻辑联系。例如，将后续问题修改为

基于前链信息的深化提问，推动多轮对话的纵向推理。

如图 3 所示，数据集包含 12 个主类别与 36 个子类别，覆盖广泛场景；以及 9 类专项评估问题，系统测评多模态大模型的核心能力。最终形成训练集 1,353 个多样化流视频，标注 49,979 对 QA，平均每视频 36.94 对和评估集（200 个视频，7,374 QA 对），远超现有数据集。

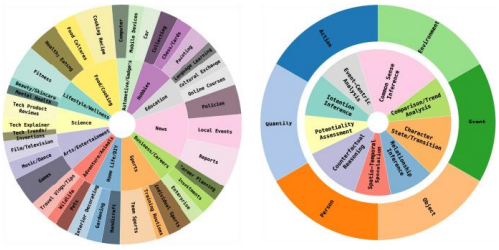

图 3 视频类别和问答对类型图

2、流视频理解模型 StreamingChat

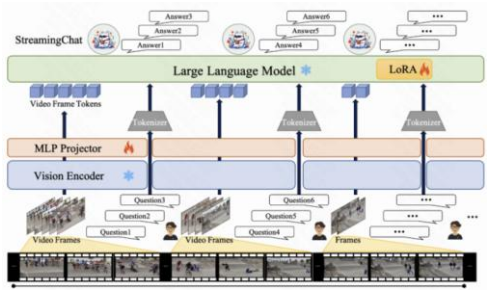

图 4 Streaming Chat 模型架构图

我们推出了基于 InternVL2 框架的 Streaming-Chat 模型，如图 4 所示，模型采用三阶段架构以高效处理长视频流对话任务。模型首先通过专门预训练的 InternViT 视觉编码器以每秒 1 帧提取视频特征，支持长达数分钟的视频流并适配 32k 上下文长度；特征转换模块借助 MLP 投影将视觉 token 转化为语言模型可识别序列；语言核心则基于 InternLM2 大模型，创新地在所有线性层嵌入 LoRA 适配器进行高效微调，并通过视觉-语言 token 交错拼接实现跨模态融合。训练数据采用自建的时序对话数据集，以“视频片段+多轮问答”链式结构组织，模拟真实边播边问场景，并引入动态分段与滑动窗口机制突破长视频上下文限制。

为全面评估模型在流媒体视频理解中的表现，我们设计了两种实验模式：对话评估要求模型基于历史对话连续回答，检验长程上下文跟踪与多轮连贯能力；流媒体评估则引入时间跳转机制，使模型在 80%情况下直接跳至后续相关问题，考验其跨片段事件推理与动态整合能力。评估不仅涵盖 METEOR 和 GPT4-Score 等基础指标，还包括多维度对话框架：语义准确性（SA）、上下文连贯性（CC）、逻辑一致性（LC）、时间理解（TU）、信息完整性（IC）与综合得分（OS），从事实匹配、逻辑一致、时序推理等多个角度系统衡量模型性能。

如图 5 所示，在开源模型中，StreamingChat 表现尤为突出，其 OS 得分相比原版模型大幅提升 28.79%，显著优于其他同类开源模型（如 MiniCPM-V 2.6，提升 26.20%），充分验证了 StreamingChat 在流媒体视频理解任务中的有效性和优越性。值得注意的是，所有模型在流媒体评估中的表现普遍低于对话评估，这主要是由于该任务对动态时序信息的理解与跨片段推理能力提出了更高要求，而 StreamingChat 在这一更具挑战性的设定中仍展现出强大的性能优势。

Table 2: Evaluation results of various models on SVBench in dialogue and streaming evaluation.

Model	Dialogue Evaluation						Streaming Evaluation					
	SA	CC	LC	TU	IC	OS	SA	CC	LC	TU	IC	OS
Open-source LVLMs												
MovieChat	20.46	20.05	27.76	21.81	22.21	21.89	17.99	16.42	20.37	15.77	19.08	17.43
Video-ChatGPT	31.86	32.58	40.28	35.32	36.26	33.80	27.98	29.54	33.81	27.95	31.00	28.88
Video-LLaVA	35.62	36.52	42.93	38.63	38.84	37.34	32.22	32.83	36.35	32.46	34.54	32.79
ShareGPT4Video	39.01	40.42	47.89	41.42	43.18	40.70	34.65	36.70	41.07	35.76	37.22	35.79
VideoLLaMA2	39.13	40.33	47.60	42.36	41.80	40.60	35.68	36.40	42.23	34.65	36.70	35.84
TimeChat	36.19	37.06	44.72	40.42	37.12	37.22	35.72	37.88	42.65	36.23	36.34	36.32
InternVL2	45.91	46.30	52.67	49.81	46.25	46.13	43.55	44.10	48.91	40.95	44.17	42.71
VILA	46.83	48.41	54.92	48.30	50.12	48.51	46.19	47.95	51.60	44.84	48.56	46.26
InternLM-XComposer2.5	51.57	53.93	59.69	51.57	56.28	52.31	52.22	53.39	58.14	48.05	54.79	51.46
MiniCPM-V 2.6	53.50	55.42	60.88	55.03	55.78	54.30	53.33	54.30	58.97	49.64	54.71	52.19
StreamingChat	**59.48**	**61.31**	**66.05**	**58.61**	**61.09**	**59.41**	**55.10**	**56.66**	**60.72**	**51.78**	**55.87**	**53.90**
Closed-source LVLMs												
Gemini 1.5 Pro	54.89	56.05	61.45	53.08	56.06	54.29	49.06	50.05	54.62	45.73	49.84	48.02
GPT-4V	65.56	68.02	71.78	63.80	68.01	65.19	58.82	59.55	64.29	54.08	60.61	57.35
GPT-4o	**65.73**	**68.10**	**71.95**	**66.54**	**68.40**	**66.29**	**59.52**	**60.42**	**65.45**	**55.10**	**61.36**	**58.17**

图 5 模型在 SVBench 上评估图

项目主页：
https://yzy-bupt.github.io/SVBench/
论文链接：
https://arxiv.org/abs/2502.10810
代码链接：
https://github.com/yzy-bupt/SVBench
模型链接：
https://huggingface.co/yzy666/StreamingChat_8B
数据集链接：
https://huggingface.co/datasets/yzy666/SVBench
Leaderboard 链接：
https://huggingface.co/spaces/yzy666/SVBench
Leaderboard 提交链接：
https://forms.gle/tmY8PmM5KWSvTGcn7

3、流视频思维链数据集 StreamingCoT

介绍：随着 5G 和边缘计算的发展，流媒体视频日益成为信息传播的主要形式，对视频的时空推理能力提出了更高要求。为克服现有视频问答数据集中标注静态、缺乏显式推理过程等局限，本文提出了 StreamingCoT——一个面向流媒体视频问答与多模态思维链任务的新数据集。该数据集通过分层时序标注流程构建，包括视频筛选、逐秒描述生成与合并、动态问答对生成与人工校验，以及多模态思维链构建，确保推理步骤具备视觉依据。StreamingCoT 包含六类问题类型，能够很好地支持复杂的时序推理任务。融合了动态流式问答、精确时间戳和多模态思维链，能够支持答案随时间演变的推理，并提供可解释的推理依据，弥补了现有数据集在时序推理和可解释性方面的不足。

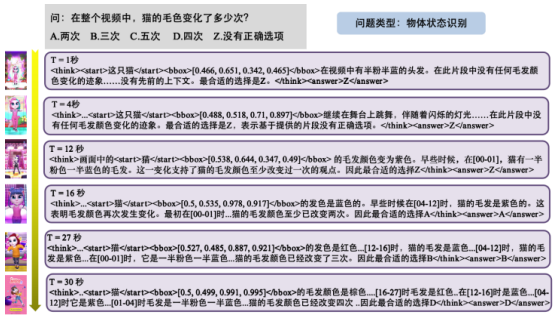

图 6 StreamingCoT 数据集示例

代码链接：
https://github.com/Fleeting-hyh/StreamingCoT

4、在线视频理解模型 VideoChat-Online

介绍：尽管 MLLMs 在离线视频理解领域取得了显著进展，但在需要实时处理连续在线视频流的现实场景中受到限制，原因在于在线视频流的实时性以及对无限长视觉信息的处理。本文提出一个专为在线视频流设计的问答基准 OVBench，涉及过去、现在和未来三个时间上下文，以全面测试模型的时空感知、记忆和推理能力。并且本文提出了新型模型架构 Pyramid Memory Bank（PMB），有效地保留视频流中的关键时空信息，本文开发了 VideoChat-Online 模型，结合 PMB 实现对视频流中关键时空信息的动态保留，并通过交错式对话数据训练增强其在线推理能力。实验表明，VideoChat-Online 在 OVBench 以及多个主流离线视频理解基准上达到最先进性能，同时保持了低计算开销和高部署灵活性，验证该方法在实时视频理解任务中的有效性与实用性。

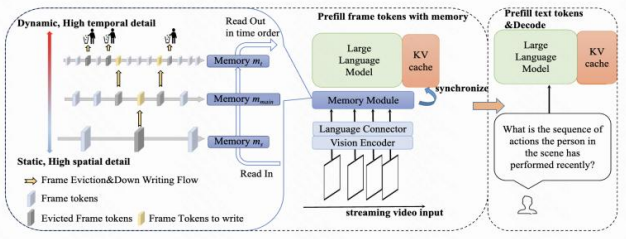

图 7 VideoChat-Online 架构图

论文链接：
https://arxiv.org/abs/2501.00584

代码链接：
https://github.com/MCG-NJU/VideoChat-Online

责任编委 王田 李策

钱胜胜

中国科学院自动化研究所副研究员，主要研究方向为多媒体内容理解、跨模态检索、多模态大模型。

电子邮箱: shengsheng.qian@nlpr.ia.ac.cn

杨振宇

中国科学院自动化研究所博士研究生，研究方向为视频理解大模型、跨模态检索。

电子邮箱: yangzhenyu2022@ia.ac.cn

开放世界目标检测数据集

香港理工大学 付陈平 大连理工大学 樊鑫

在追求让机器感知和解读世界的过程中，计算机视觉已取得长足进步。该领域最初通过将这一复杂挑战分解为更小、更易处理的问题来应对。早期研究聚焦于边缘检测、图像分类和物体识别等专项任务，由此诞生了一系列专用算法。虽然这些努力取得了成功，但也凸显出孤立处理视觉问题的局限性。随着机器学习技术的日益精进与计算能力的指数级增长，计算机视觉领域正朝着更全局化的方法转向。近年来，在海量数据集整合与大规模多模态基础模型发展的推动下，能够同时执行多种视觉任务的模型不断涌现。其中开放世界目标检测得到了研究人员的热切关注。

开放世界目标检测(Open World Object Detection, OWOD) 由约瑟夫等人于 2021 年首次提出，该概念将传统目标检测扩展至能感知给定场景中所有可见物体。其突破性在于：无论目标类别是否在训练阶段被明确呈现给模型，系统都能实现与类别无关的开放世界检测。这一范式转变得益于多个关键视觉细分领域的数据集突破，本文将重点阐释这些支撑性数据集，包括以下 3 个代表性开放世界目标检测数据集，分别是：ODinW（2022），Objects365 （2019），和 LVIS （2019）数据集。

1、ODinW 数据集

介绍： ODinW（Open world Object Detection in the Wild）基准数据集是开放世界目标检测领域的一个标志性评测基准，由 Chunyuan Li 等人于 2022 年在其开创性论文中首次提出并构建。该数据集并非一个全新的独立数据集，而是巧妙地整合了多个现有且广泛认可的经典计算机视觉数据集，例如 CIFAR100、PASCAL VOC、MountainDewCommercial 、以及更复杂场景的 OxfordPets（breed）和 OpenPoetryVision 等，共计包含了 35 个不同的子数据集和 314 个目标类别。此外，该数据集有 132K 张训练图片，20K 张测试图片。图 1 展示了该数据集中图片例子。

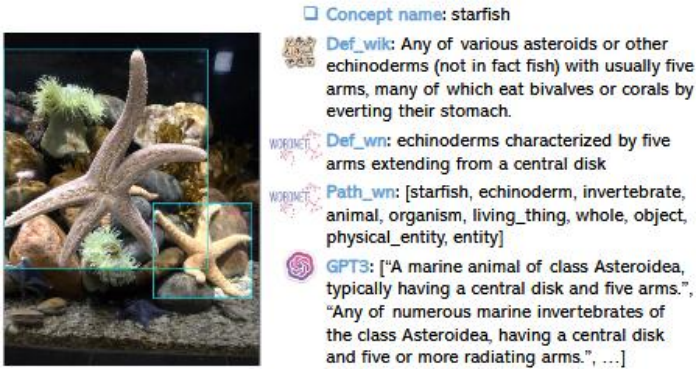

图 1 ODinW 数据集图片示例

ODinW 的核心设计理念在于系统性地评估模型在“开放世界”环境下的性能。与传统检测数据集要求模型只在预定义的封闭类别中进行识别不同，ODinW 旨在测试模型面对“未知”类别对象的能力。其评测流程模拟了智能体的真实学习过程：模型首先在一组已知类别上训练，然后在测试时会遇到包含已知类和未知类的图像。模型不仅要准确地检测出所有已知对象，还必须能够发现并定位那些在训练中从未见过的未知对象，同时避免将其错误地归类为已知类别。这一设定带来了前所未有的挑战，它迫使模型必须超越简单的模式匹配，发展出真正的视觉理解与泛化能力。ODinW 的建立极大地推动了开放世界检测研究的发展，为衡量模型在复杂、动态的真实世界中的感知能力提供了一个

rigorous（严谨）且统一的评测标准，促使研究者们开发出更具适应性、更接近人类感知智慧的检测系统。

数据集地址
https://computer-vision-in-the-wild.github.io/ELEVATER/

2、Objects365 数据集

介绍: Objects365 是一个专为物体检测任务设计的大规模、高质量数据集，由旷视科技研究团队于 2019 年提出。该数据集旨在克服当时主流数据集(如 MS COCO)在规模和质量上的局限，为训练更鲁棒、更通用的目标检测模型奠定新的数据基础。

图 2 Objects365 数据集图片示例

Objects365 核心价值体现在“大规模”与“高质量”两个维度。首先，在规模上，Objects365 V1 版本包含了超过 60 万张图像、1000 万个高质量的边界框标注，覆盖了 365 个日常生活中常见的物体类别。这一庞大的标注数量远超同期数据集，为训练深度模型提供了丰富的多样性，能有效减少模型过拟合，提升其泛化能力。此外，该数据集遵循“高质量”的标注标准。简单而言，Objects365 数据集的所有图像都通过了一个精心设计的、两阶段的众包标注流程。首先，专业标注员需要对图像中的物体进行详尽标注（exhaustive annotation），即尽可能标注出所有可见的指定类别物体，而非只标注主要物体，这极大地减少了标注遗漏。随后，还有专门的验证团队对标注结果进行严格质检，确保了边界框位置和类别标签的高度准确性。因此，Objects365 不仅是一个庞大的资源库，更是一个高质量的基准。本文在图 2 中展示了 Objects365 数据集中的图片例子。

在 Objects365 上预训练的深度学习模型展现出了卓越的迁移学习能力，在包括 MS COCO、Pascal VOC 在内的其他下游检测任务上实现了显著的性能提升，迅速成为许多顶尖检测模型和预训练模型的首选训练集，深刻影响了目标检测领域的研究与发展。

3、LVIS 数据集

数据集地址 http://www.objects365.org/

介绍: LVIS（Large Vocabulary Instance Segmentation）数据集是实例分割领域一个里程碑式的大规模基准数据集。其名称中的“Large Vocabulary”直接指明了其核心特征：极其丰富的类别数量。LVIS V1 版本包含了超过 1200 个物体类别，远超同时代主流数据集 MS COCO（仅 80 类），旨在推动模型向“大词汇量”视觉感知发展。本文在图 3 中展示了 LVIS 数据集中的图片例子。

图 3 LVIS 数据集图片示例

该数据集的构建并非为了追求绝对的数据量，而是为了解决计算机视觉中一个至关重要的挑战——长尾分布（Long-Tail Distribution）问题。在现实世界的视觉场景中，物体类别的出现频率极不均衡；常见物体（如“人”、“汽车”）的图片很多，而大量罕见物体（如

“绞盘”、“官帽”）的图片则非常少。LVIS 通过从多个现有数据源（如 MS COCO）中系统性地收集和整合图像，并采用一种新颖的数据驱动的标注流程，真实地复现了这种长尾分布。其标注过程通过迭代挖掘图像中的罕见对象，确保了类别的充分覆盖。

此外，该数据集为每个实例都提供了高质量的像素级掩码（mask）标注，精度极高。更重要的是，LVIS 根据每个类别的图像出现频率，明确地将所有类别分为三类：常见(Frequent)、罕见(Rare)和一般(Common)。这种细粒度的划分使得研究者能够精准地评估模型在不同分布类别上的性能，尤其关注模型在识别和分割长尾末端罕见类别方面的能力，这比仅仅一个整体平均精度（AP）指标更具洞察力。

因此，LVIS 的提出极大地挑战并推动了相关领域的研究。它迫使模型必须超越对常见模式的简单记忆，学会更好地泛化到稀有类别上，从而催生了许多针对长尾分布、零样本/少样本学习的新型算法。它不仅是一个检测与分割基准，更成为了研究和解决视觉模型泛化性与公平性问题的重要实验平台，对整个领域的发展方向产生了深远影响。

数据集地址

https://link.zhihu.com/?target=https%3A//www.lvisdataset.org/

责任编委 贾同 王田

付陈平

博士后，香港理工大学航天航空工程学院，研究方向为计算机视觉，目标检测，图像增强，水下成像。

樊 鑫

博士生导师，大连理工大学国际信息与软件学院从事教学与科研工作，担任中日国际信息与软件学院院长。研究方向为计算机视觉与图像处理、医学影像分析。

个人主页：http://faculty.dlut.edu.cn/Xin_Fan/zh_CN/index.htm

好 文 推 荐

北京大学、南洋理工大学和上海人工智能实验室“DST-Det: Open-Vocabulary Object Detection via Dynamic Self-Training”最新成果发表在 IEEE Transactions on Circuits and Systems for Video Technology 2024。

论文：Xu S, Li X, Wu S, et al. DST-Det: Open-Vocabulary Object Detection via Dynamic Self-Training[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 35:5037-5050.

开放词汇目标检测（OVOD）的目的是突破训练阶段有限基类的约束，准确检测和识别未用于训练的新类。现有的基于伪标签的 OVOD 方法需要先利用额外的外部数据生成伪标签，比如使用视觉语言模型（VLM）从无标注数据或者图像文本对数据集获得伪标签，然后再使用基类的真值和新类的伪标签训练开放词汇目标检测器。这种方式通常包含很多手工提取的步骤。

本文基于两阶段检测器 Mask R-CNN 构建了端到端的架构，利用预训练的 VLM，比如 CLIP，通过零样本分类识别潜在的新类。具体而言，它使用冻结的 CLIP 视觉编码器作为骨干网络，并在训练过程中引入伪标签生成模块 PLM 来动态生成伪标签。该模块同时嵌入于 RPN 与 RoI Head 中，用于筛选负样本候选框。当负样本候选框与基类真值框的重叠度较低，但其区域嵌入与新类文本嵌入具有较高相似度时，PLM 就将其标注为新类别的伪标签。因此，在训练阶段同时需要基类和新类的类别名称。DST-Det 的整体框架结构及其中 PLM 模块的设计如图 1 所示。

文章在 LVIS、V3Det 和 COCO 三个数据集上进行了系统性的实验评估，实验结果表明 DST-Det 在无需增加额外参数或推理阶段计算开销的情况下，相较基线模型实现了显著性能提升。此外，所提出的方法能够灵活应用于多种基线模型，并取得不同程度的性能提升，进一步验证了所提方法的有效性。

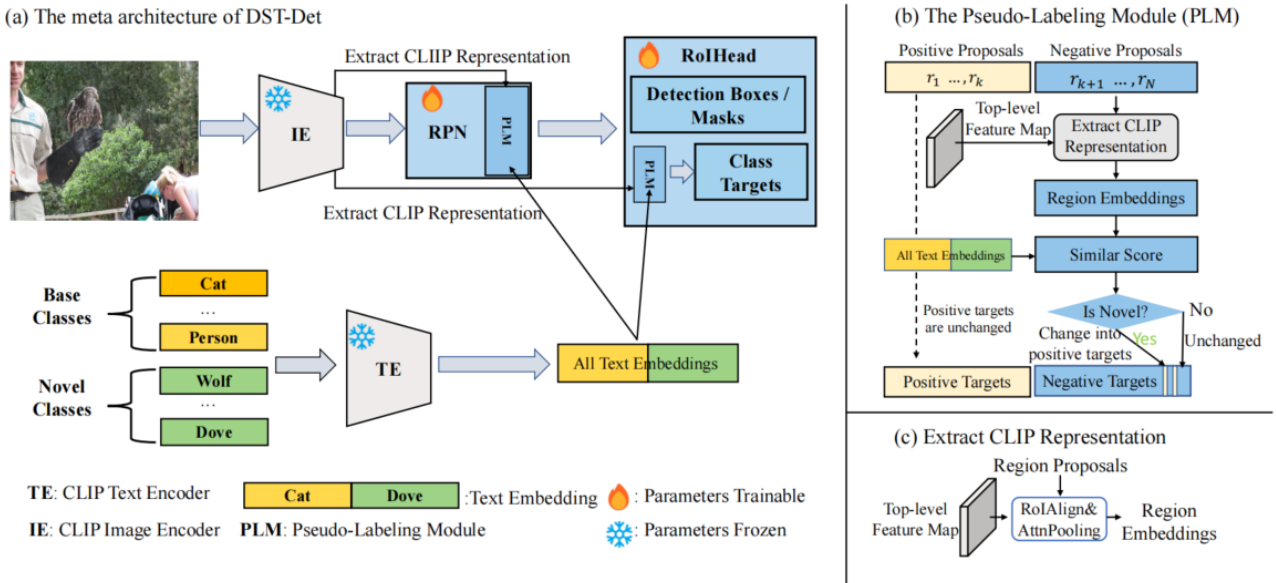

图 1 提出的 DST-Det 网络结构图

责任编委 樊鑫 贾同

好 文 推 荐

厦门大学、伦敦大学学院和之江实验室“Dual-Mode Learning for Multi-Dataset X-Ray Security Image Detection”最新成果发表 IEEE Transactions on Information Forensics and Security 2024。

论文：Yang F, Jiang R, Yan Y, et al. Dual-Mode Learning for Multi-Dataset X-Ray Security Image Detection[J]. IEEE Transactions on Information Forensics and Security, 2024, 19:3510-3524.

随着深度学习的快速发展，大量用于 X 光安检图像的违禁品检测方法被提出。通常，这些方法仅在单一的 X 光图像数据集上进行训练，而单一的数据集往往只包含有限类别的违禁品。为了检测更多类别的违禁品，理想的做法是利用由多个数据集组合而成的多数据集训练模型。然而，由于不同数据集之间存在较大的域差异，且违禁品图像中普遍存在遮挡问题，将现有方法直接应用于多数据集时，往往难以取得理想效果。

为解决上述问题，文章提出了一个双模态学习网络 DML-Net 来高效检测多数据集上的所有违禁品类别。网络总体结构如图 1 所示。具体而言，DML-Net 网络架构采用增强的 RetinaNet，并在其中引入晶格外观增强的子网络 LAE 来提升特征表征能力，从而缓解遮挡难题。在此基础上，DML-Net 的学习过程包括公共模态学习（检测各数据集共有的违禁品类别）和特有模态学习（检测各数据集特有的违禁品类别）两个部分。公共模态学习通过对抗原型对齐模块，在域不变的特征空间中对齐来自不同数据集的特征原型。特有模态学习采用特征蒸馏，强制学生模型去模仿多个预训练教师模型所提取的特征。通过紧密结合这两种模态，各数据集间的域差异得以有效消除。

文章使用 OPIXray、SIXray 和 HiXray 数据集进行实验。在多个组合 X 光图像数据集上的大量实验结果表明，本文方法在性能上明显优于多种最新的方法，充分验证了所提出的 DML-Net 的有效性。

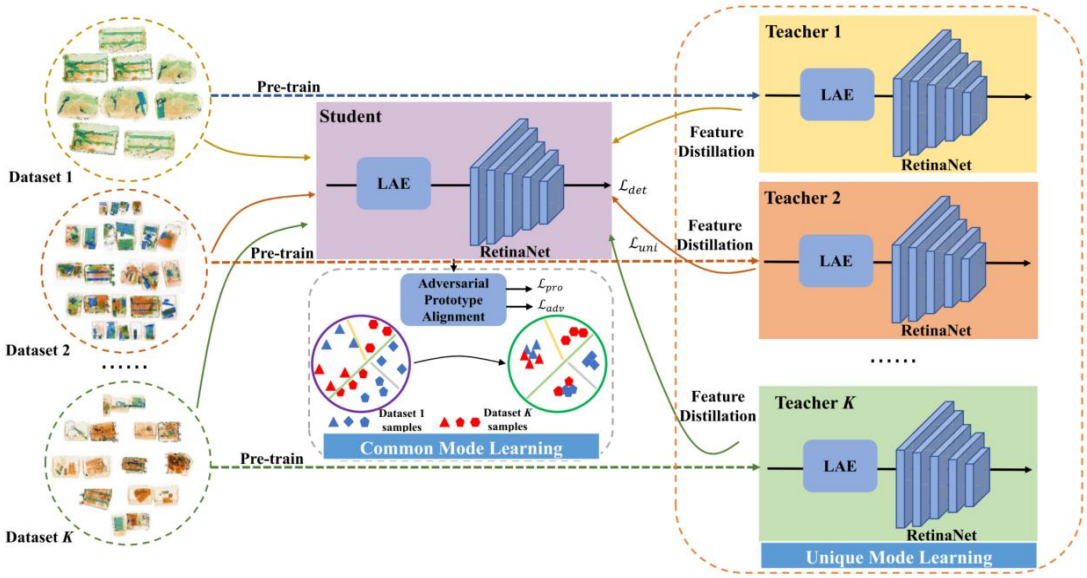

图 1 提出的 DML-Net 网络结构图

责任编委 李策 贾同

好 文 推 荐

哈尔滨工业大学提出的”Optimus-2: Multimodal Minecraft Agent With Goal-Observation-Action Conditioned Policy”最新成果发表在 CVPR 2025。

论文：Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, Liqiang Nie. Optimus-2: Multimodal Minecraft Agent With Goal-Observation-Action Conditioned Policy, Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 9039-9049.

使智能体能够学习人类行为模式以完成复杂任务，一直是人工智能领域的长期目标。现有的 Minecraft 智能体通常采用基于目标条件的策略来完成任务。然而，在复杂的开放世界场景中，单一的目标条件策略改进未能充分考虑观察和动作之间的因果关系，导致智能体在执行复杂任务时容易发生错误。因此，为了解决这一问题，本文提出了 Optimus-2 智能体。该智能体通过引入目标-观察-动作条件策（Goal-Observation-Action Conditioned Policy，GOAP）以及大规模语言模型（Multimodal Large Language Model, MLLM）相结合，来建模观察、动作和子目标之间的因果关系。这使得智能体能够更准确地理解并执行任务。GOAP 方法通过引入行为引导编码器来动态建模历史观察-动作序列并整合成固定长度的行为标记，从而有效克服长时间跨度任务中的信息丢失问题。与先前的工作相比，Optimus-2 在执行复杂任务时，通过对历史序列的高效处理和语言理解能力的提升，展现了显著的性能优势，特别是在长时间跨度和开放式指令任务中表现优异。

图 1 中展示了 Optimus-2 的工作流程。首先，智能体从任务指令开始，利用 MLLM 作为规划器生成一系列子目标。接着，这些子目标与当前的图像和动作信息一起被输入到 GOAP 中，GOAP 通过引入行为引导编码器来处理这些信息。在 GOAP 模块中，行为引导编码器首先通过交叉注意力机制将由视觉编码器提取的图像特征与动作嵌入进行交互，增强观察与动作之间的因果关系。最后，这些经过编码的行为标记会与生成的子目标信息一起输入到 MLLM 中，MLLM 会根据这些输入信息生成自动回归的下一步动作预测。具体来说，MLLM 结合当前的子目标和历史行为标记，生成下一个动作的预测。通过这种方式，Optimus-2 能够在复杂环境中执行任务，理解并适应开放式的自然语言指令。实验结果表明，Optimus-2 在 Minecraft 环境中，尤其在长时间跨度和开放式指令任务中具有突出优势。

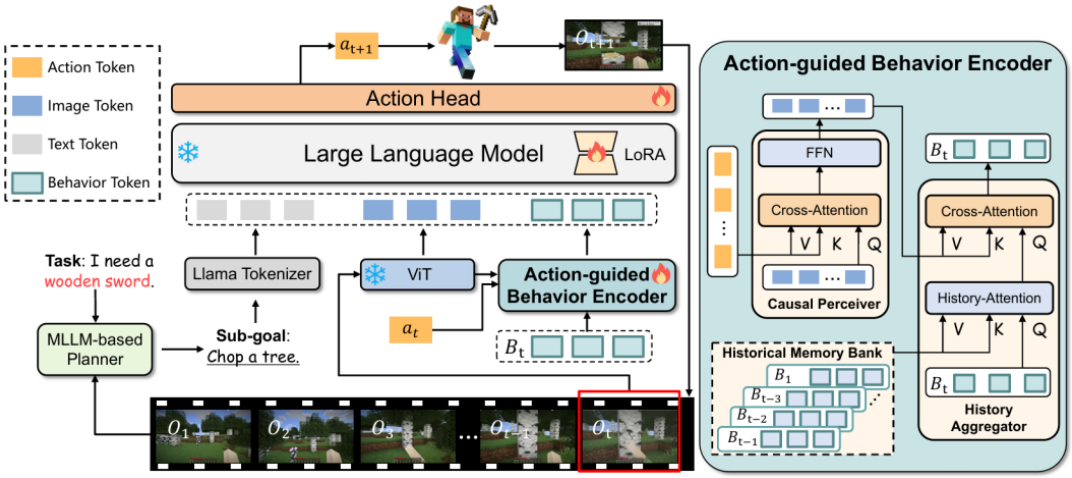

图 1 目标-观察-动作条件策略流程图

责任编委 李策 王田

征文通知

1 会议征文

计算机视觉领域相关国内外会议的征文通知如表 1 所示。同时，可继续关注每个会议举办的 workshop 或 special session。

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 Image and Vision Computing，Journal of Visual Communication and Image Representation，Electronics 和 IET Image Processing。

3 会议简介

中国模式识别与计算机视觉学术会议 PRCV (Chinese Conference on Pattern Recognition and Computer Vision)，由中国计算机学会（CCF）、中国自动化学会（CAA）、中国图象图形学学会（CSIG）和中国人工智能学会（CAAI）联合主办，定位国内顶级的模式识别和计算机视觉领域学术盛会。

第八届 PRCV 将于 2025 年 10 月 16 日至 10 月 19 日在上海举办，由上海交通大学承办。本届会议将秉持团结模式识别与计算机视觉领域科技工作者的宗旨，进一步推动开放合作，广泛吸引学术界和工业界的人才，提升会议的国际化水平，力求打造一个高品质的学术交流平台。大会的举办将为学术界与工业界提供更多产学研合作机会，推动模式识别与计算机视觉领域的协同创新和可持续发展。

责任编辑：刘帅奇

表 1 计算机视觉领域相关国内外会议

会议名称	会议时间	会议地点	截稿日期	会议网站
ICICML 2025	2025.11.21-23	Chongqing, China	2025.10.26	https://icicml.org/
CVPR 2026	2026.06.02-06	Denver, USA	2025.11.15	https://cvpr.thecvf.com/Conferences/2026
ICPR 2026	2026.08.17-21	Lyon, France	2025.12.20	https://icpr2026.org/index.html

表 2 计算机视觉领域相关国内外期刊专刊

期刊名称	专刊题目	投稿网址	截稿日期
IMAVIS	Advancing Visual Data Analytics for Disaster Management	https://www.sciencedirect.com/special-issue/322678/advancing-visual-data-analytics-for-disaster-management	2025.10.31
JVCI	Multimodal Learning for Visual Intelligence: From Emerging Techniques to Real-World Applications	https://www.sciencedirect.com/special-issue/322635/multimodal-learning-for-visual-intelligence-from-emerging-techniques-to-real-world-applications	2025.11.30
Electronics	Image Segmentation, 2nd Edition	https://www.mdpi.com/journal/electronics/special_issues/09RKSO8442	2025.10.15
IET Image Processing	Computer Vision for Earth Observation and Environmental Monitoring	https://ietresearch.onlinelibrary.wiley.com/hub/journal/17519667/homepage/call-for-papers/si-2025-000252	2025.12.01

COMPUTER VISION NEWSLETTER

03 2025
总第 45 期

计算机视觉专委会简报

CCF 计算机视觉专委会