中图法分类号： 文献标识码： 文章编号：1006-8961(XXXX)XX-0001-36

论文引用格式：. Frontier Trends and Top Ten Advances in 3D Vision in 2025[J/OL]. Journal of Image and Graphics, XXXX:1-36. DOI: 10.11834/jig.260114. (. 2025 年度三维视觉前沿趋势与十大进展[J/OL]. 中国图象图形学报, XXXX:1-36. DOI: 10.11834/jig.260114.)[DOI: 10.11834/jig.260114]

2025 年度三维视觉前沿趋势与十大进展

刘烨斌[1]、穆尧[2]、叶琦[3]、高林[4]、韩晓光[5]、陈安沛[6]、段岳圻[1]、彭思达[3]、邵天甲[3]、张鸿文[7]、张力[8]、廖依伊[3]、许岚[9]、刘希慧[10]、姚遥[11]、胡瑞珍[12]、戈力[1]、郭裕兰[13]、连宙辉[14]、刘子纬[15]、陈宝权[14]

摘　要：三维视觉作为计算机视觉、图形学、人工智能与光学成像的交叉学科，是构建具身通用智能与元宇宙的核心基石。2025 年，以 VGGT（Wang 等，2025d）为代表的前馈三维重建技术的突破，为空间智能提供了坚实的场景三维理解基础，并大幅降低高质量三维内容的制作门槛；三维生成质量逐渐达到工业级扫描水平，技术从单图实例生成向动态复杂场景的多实例前馈重建演进；三维重建与三维生成开始深度融合，逐渐实现复杂场景在稀疏视点输入下的前馈式重建；视频生成技术正融入各式三维表征，推动“感知-生成-交互”一体化的世界模型技术的发展，世界模型已被广泛认为是实现可泛化具身智能与通用人工智能（artificial general intelligence，AGI）的关键路径；蕴含物理常识、因果推理与交互偏好的人类行为与第一人称视频数据开始被广泛使用，成为突破具身智能数据瓶颈、驱动具身智能 Scaling 的核心燃料；具身智能视觉-语言-动作（vision-language-action，VLA）模型正从依赖专家演示的模仿学习，转向融合在线强化学习的复合架构，可在稀疏奖励下显著提升模型的泛化与探索能力。这些技术突破奠定了“多模感知-三维建模-四维生成-实时交互”一体化智能架构的雏形，为空间智能和具身智能的实质性发展提供了关键技术支撑。为促进学术交流，本文分析总结三维视觉领域前沿趋势，并遴选年度十大研究进展，为学术界与产业界提供参考观点。

关键词：三维视觉；具身智能；世界模型；重建与生成；空间智能

Frontier Trends and Top Ten Advances in 3D Vision in 2025

Yebin Liu[1], Yao Mu[2], Qi Ye[3], Lin Gao[4], Xiaoguang Han[5], Anpei Chen[6], Yueqi Duan[1], Sida Peng[3], Tianjia Shao[3], Hongwen Zhang[7], Li Zhang[8], Yiyi Liao[3], Lan Xu[9], Xihui Liu[10], Yao Yao[11], Ruizhen Hu[12], Li Yi[1], Yuan Guo[13], Zhouhui Lian[14], Ziwei Liu[15], Baoquan Chen[14]

Abstract: As an interdisciplinary field spanning computer vision, graphics, artificial intelligence, and optical imaging, 3D vision serves as the core cornerstone for constructing Embodied General Intelligence (EGI) and the Metaverse. As the "Scaling Law" paradigm, upon which AI development relies, faces significantly diminishing marginal returns and encounters bottlenecks, the focus of both academia and industry is pivoting ever more clearly toward foundational subjects closely related to 3D vision, such as "World Models," "Spatial Intelligence," and "Embodied Intelligence," granting 3D vision unprecedented strategic attention and developmental opportunities. In 2025, the primary frontier trends in the field of 3D vision can be summarized as follows: 1) Feed-forward 3D reconstruction that supports spatiotemporal multi-image inputs: with breakthroughs in feed-forward 3D reconstruction technologies such as VGGT, obtaining scene structure and motion information through spatiotemporal multi-image feed-forward methods has become increasingly simple, bringing two pro-

收稿日期：2026-03-02；修回日期：2026-04-08

found impacts: firstly, it provides a solid foundation for 3D scene understanding for spatial intelligence, allowing many traditional 2D vision problems to be solved more fundamentally in 3D space; secondly, combined with efficient rendering technologies such as 3D Gaussian Splatting (3DGS), the threshold for high-quality 3D content production has been significantly lowered, paving the way for large-scale applications such as digital twins and the Metaverse. 2) The gradual fusion of 3D generation and 3D reconstruction: 3D AIGC technologies such as SAM3D support compositional and instance-level object generation under single-image input, with generation quality gradually reaching industrial-grade scanning standards, while simultaneously integrating with feed-forward reconstruction methods to gradually achieve the generation of authentic 3D structures and textures consistent with the input images; this will support feed-forward multi-instance reconstruction of dynamic complex scenes, significantly improving real-time, multimodal perception and understanding capabilities in complex scenarios. 3) The integration from video generation and world models to embodied intelligence: video generation technology is rapidly incorporating explicit or implicit 3D representations and evolving toward multi-view consistency, long sequences, and physical plausibility, directly driving the development of integrated "Perception-Generation-Interaction" world model technologies. These types of world models, combined with feed-forward 3D reconstruction technology, will form a complete "Multimodal Perception-3D Modeling-4D Generation-Real-time Interaction" 4D world model. At the same time, world model methods have begun to serve embodied intelligence, and a unified framework of "understanding-generation-execution" has begun to emerge. World models are widely regarded by the academic community as the key path to achieving generalizable embodied intelligence and ultimately leading to AGI. 4) Human behavior and video data becoming the core fuel driving breakthroughs: human operational spaces and interaction videos constitute a "data goldmine" for training embodied intelligence. The vast amount of human behavior videos on the internet, as well as first-person perspective data collected through simple devices, contain physical common sense, causal reasoning, and interaction preferences that serve as the natural fuel to break through the current data bottlenecks of embodied intelligence. By performing explicit 3D perceptual reconstruction or latent-space action alignment and learning on these data, a "data pyramid" base can be constructed to drive the scaling of embodied intelligence. 5) The evolution of the embodied training paradigm from imitation learning to interaction-driven reinforcement learning: the technical evolution of embodied intelligence VLA models is leaping from a supervised fine-tuning paradigm relying on expert demonstrations to a composite training architecture integrating online reinforcement learning. This shift effectively breaks the dependence on scarce high-quality data, enabling policies driven by sparse rewards to obtain generalization and exploration capabilities surpassing those of imitation learning, solving the challenges of exploration and stable updates in continuous action spaces. Simultaneously, the development of high-performance training systems and action-conditioned world models provides the infrastructure support for large-scale interaction data generation and efficient policy evolution, marking a new "post-training" stage for embodied intelligence centered on "interaction-driven" approaches. The selected top ten research advancements of the year in the field of 3D vision include: 1) Feed-forward 3D reconstruction constructing the foundation models for 3D vision (spatial intelligence); 2) The convergence of reconstruction and generation technical routes (video generation/3D generation), moving from mutual assistance to preliminary integration; 3) 3DGS/4DGS continuously improving representation efficiency, sparking a surge in scene modeling and volumetric video applications; 4) 3D generation: a leap from single-object visual realism to structuralized components/scenes and physical interactivity; 5) From video generation to world models: oriented toward spatiotemporal consistency, physical plausibility, and interactivity; 6) Unified multimodal large models for understanding and generation serving spatial intelligent perception; 7) Frontier shifts in digital humans: from appearance modeling to multimodal interaction; 8) Human data becoming the essential fuel to break through the Scaling Law of embodied intelligence; 9) Embodied intelligence foundation models evolving toward unified models of integrated "understanding-imagination-execution"; 10) The "post-training" moment of embodied intelligence: the paradigm shift of VLA models from imitation learning to online RL. Collectively, these breakthroughs have established the prototype of an integrated intelligent architecture characterized by "Multimodal perception - 3D modeling - 4D Generation - Real-time interaction", providing critical technical support for the substantive advancement of spatial and embodied intelligence. To promote academic discourse, this paper extensively analyzes frontier trends in 3D vision and curates the top ten annual research advances, offering valuable reference perspectives for both academia and industry.

Key words: 3D vision; embodied AI; world model; reconstruction and generation; spatial intelligence

[1] 清华大学; [2] 上海交通大学; [3] 浙江大学; [4] 中国科学院计算技术研究所; [5] 香港中文大学(深圳); [6] 西湖大学; [7] 北京师范大学; [8] 复旦大学; [9] 上海科技大学; [10] 香港大学; [11] 南京大学; [12] 深圳大学; [13] 中山大学; [14] 北京大学; [15] 南洋理工大学

[1]Tsinghua University; [2]Shanghai Jiao Tong University; [3]Zhejiang University; [4]Institute of Computing Technology, Chinese Academy of Sciences; [5]The Chinese University of Hong Kong, Shenzhen; [6]Westlake University; [7]Beijing Normal University; [8]Fudan University; [9]ShanghaiTech University; [10]The University of Hong Kong; [11]Nanjing University; [12]Shenzhen University; [13]Sun Yat-sen University; [14]Peking University; [15]Nanyang Technological University

序言

随着人工智能发展所依赖的‘Scaling Law’范式边际效益显著递减、遭遇瓶颈,学术界与产业界的焦点正日益明确地转向“世界模型”、“空间智能”和“具身智能”等与三维视觉紧密相关的基础课题,三维视觉因此迎来了前所未有的战略关注与发展机遇。多位人工智能(Artificial Intelligence,AI)领域的领军人物也为此提供了理论指引与技术验证:图灵奖获得者杨立昆长期倡导构建能够预测和理解物理规律的世界模型;AI教母李飞飞则系统论述了空间智能是AI不可或缺的下一个前沿,并指出实现它需要具备生成、多模态与交互能力的世界模型;诺贝尔奖获得者谷歌DeepMind的首席执行官德米斯·哈萨比斯认为当前以大型语言模型(large language model,LLM)为主导的AI范式存在根本性局限,未来的突破在于让AI理解并交互于物理世界。

本文承接“2024年度三维视觉前沿趋势与十大进展”报告(Liu等, 2025c),继续整理和总结2025年三维视觉的前沿趋势和十大进展。截至2025年,三维视觉方法已深度拥抱Transformer架构,初步确立了前馈式重建与生成的主流技术范式。这一范式实现了对互联网海量视频数据的端到端时空三维重建与生成,其输出结果不仅包含几何结构,更初步展现了多模态语义信息,从而奠定了“多模感知-三维建模-四维生成-实时交互”一体化智能架构的雏形,为空间智能和具身智能的实质性发展提供了关键技术支撑。

主要的前沿趋势可归纳为以下四点:

前馈三维重建支持时空多图像输入:随着VGGT等前馈三维重建技术的突破,通过时空多图像前馈式获得场景结构和运动信息变得愈发简易,带来两大深远影响:它为空间智能提供了坚实的场景三维理解基础,使得许多传统二维视觉问题得以在三维空间中获得更本质的解决;其次,结合3D高斯泼溅(3D Gaussian Splatting,3DGS)等高效渲染技术,使得高质量三维内容的制作门槛大幅降低,为数字孪生、元宇宙等规模化应用铺平了道路。

三维生成与三维重建逐渐融合:SAM3D等3D人工智能生成内容(artificial intelligence generated content,AIGC)技术支持单图输入下的组件式和实例级物体生成,生成质量逐渐达到工业级扫描质量,同时结合前馈重建方式逐渐做到符合输入图像的真实三维结构和真实纹理的生成,未来将支持动态复杂场景的前馈式多实例重建,显著提高复杂场景的实时、多模态感知和理解能力。

从视频生成、世界模型到具身智能的贯通:视频生成技术正快速融入显式或隐式的三维表征,向多视角一致、长序列、具备物理合理性的方向进化,直接推动了“感知-生成-交互”一体化的世界模型技术的发展。这类世界模型结合前馈三维重建技术,将形成完整的“多模感知-三维建模-四维生成-实时交互”的4D世界模型。同时,世界模型方法也开始服务于具身智能,理解-生成-执行统一框架初现。世界模型被学界广泛认为是实现可泛化具身智能并最终通向AGI的关键路径。

人类行为和视频数据成为驱动突破的核心燃料:人类的操作空间与交互视频构成了训练具身智能的“数据金矿”。互联网上浩瀚的人类行为视频、以及通过简易设备采集的第一人称视角数据,其中蕴含的物理常识、因果推理与交互偏好,是突破当前具身智能数据瓶颈的天然燃料。通过对这些数据进行显式的三维感知重建或隐空间的动作对齐与学习,可以构建起驱动具身智能Scaling的“数据金字塔”基座。

从模仿学习到交互驱动的强化学习具身训练范式演进:具身智能VLA模型的技术演进正从依赖专家演示的监督微调范式,向融合在线强化学习的复合训练架构跃迁。这一转变有效突破了对稀缺高质量数据的依赖,使策略在稀疏奖励驱动下获得超越模仿学习的泛化与探索能力,解决了连续动作空间探索与稳定更新的难题。同时,高性能训练系统与动作条件世界模型的发展为规模化交互数据生成与高效策略进化提供了基础设施支撑,标志着具身智能正进入以“交互驱动”为核心的“后训练”新阶段。

本文将从十个方面细化总结2025年三维视觉领域的十大科研进展。如图1所示,本文给出2025年度三维视觉主要进展的整体框架。

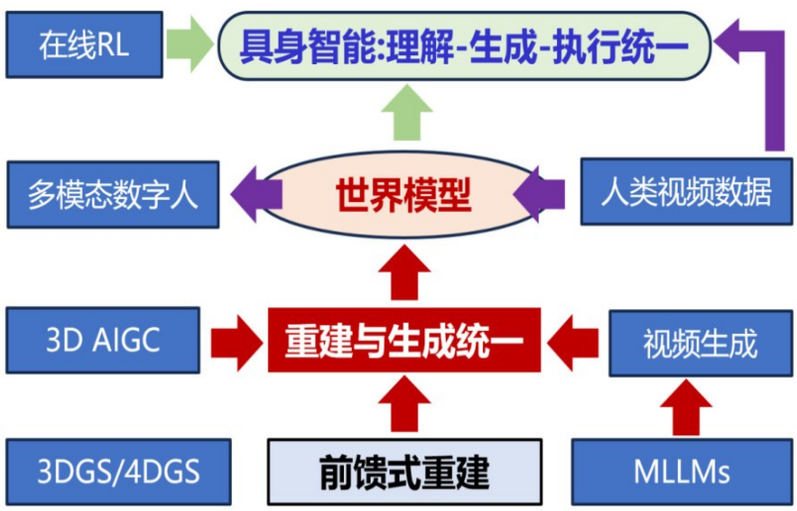

图1　2025年度三维视觉主要进展框架图

Fig. 1　Framework of major advances in 3D vision in 2025

一　前馈式三维重建构筑三维视觉(空间智能)基础模型

近年来,三维重建正从依赖多阶段、模块化迭代优化的传统范式,逐步演进为以大规模数据和模型为核心的前馈式重建范式,并成为空间智能的重要基础能力之一。相较于依赖反复优化的传统结构光束法/多视图立体重建(Structure from Motion/Multi-View Stereo ,SfM/MVS)或NeRF(Mildenhall等,2020)/3DGS(Kerbl等,2023)几何与纹理贯通的端到端可微优化方法,前馈式重建进一步将相机位姿计算、几何重建融为一体,实现了端到端的一体化流程。该方法在计算效率、可扩展性以及系统集成友好性等方面均展现出显著优势。前馈式三维重建从大量数据中学习到的先验知识使其对歧义性和输入变化具有显著更强的鲁棒性,因而正逐步成为空间智能基础模型的重要组成部分。

过去一年最具代表性的前馈式三维重建VGGT(Wang等,2025d)提出了一种统一的前馈Transformer(Vaswani等,2017)架构,能够从单张、多张甚至数百张图像中,一次性直接预测相机参数、深度图、点云图和3D点轨迹等所有关键3D属性。相比于一次仅能处理两张图像的2024年代表工作Dust3R,VGGT通过引入交替注意力机制,在帧内自注意力和全局自注意力之间切换,实现了对任意数量输入图像的灵活、高效处理,推理速度极快。VGGT开创了三维视觉“大一统”模型。它打破了传统上不同3D任务被孤立模型处理的壁垒。其意义在于证明了通过大规模数据驱动,一个极简架构就能实现通用、高效的3D场景理解,为构建实时、鲁棒的3D视觉系统奠定了新的基础,并显著提升了点跟踪、新视图合成等下游任务的性能。

如图2所示,前馈式三维重建的代表性工作见图示。

除了VGGT之外,前馈式三维大模型在多个关键维度取得了系统性进展,主要体现在以下三个方面:

一是重建几何精度的持续提升。在训练数据规模不断扩大、模型结构持续演进以及监督策略逐步完善等多重因素的共同推动下,前馈式三维重建的几何精度在过去一年中实现了显著跃升。自

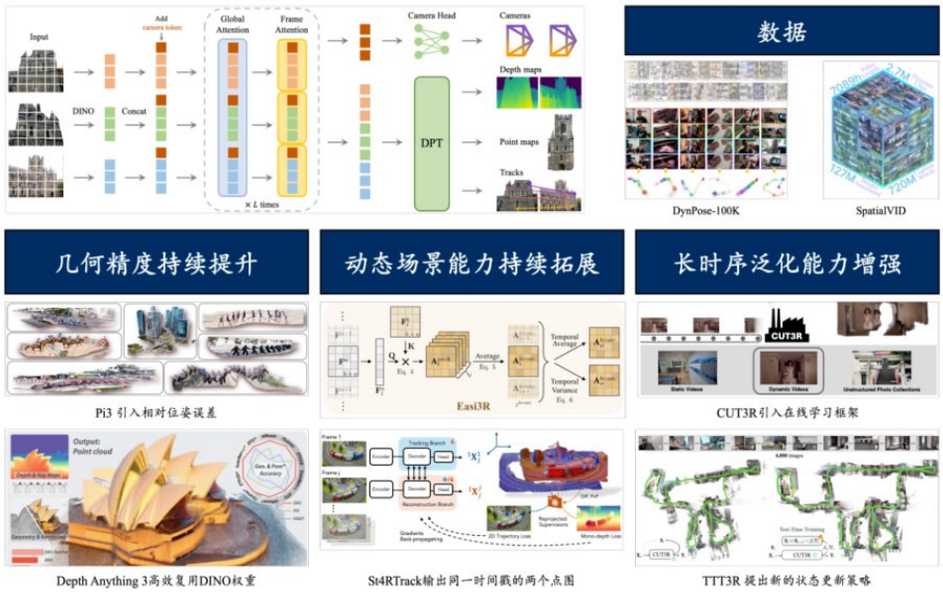

图 2　前馈式三维重建 2025 年度代表性工作

Fig. 2　Representative works of feed-forward 3D reconstruction in 2025

DUST3R（Wang 等，2024）首次在三维视觉任务中验证 Scaling Law 的有效性以来，VGGT（Wang 等，2025d）、CUT3R（Wang 等，2025e）、Pi3（Wang 等，2025i）、Depth Anything3（Lin 等，2025b）等一系列方法不断刷新相机位姿估计与三维几何重建的性能上限。在数据层面，VGGT 与 CUT3R 几乎整合了当前所有公开可用的、带有深度与相机姿态标注的多视角训练数据，大幅提升了模型的泛化能力。在此基础上，Depth Anything3 进一步将关注点从“数据规模”转向“数据质量”，通过引入单目深度模型，对多视角匹配生成的真值深度图中普遍存在的空洞区域进行有效补全，从而显著增强了监督信号的完整性与一致性。与此同时，面向动态场景的数据增量也取得了重要进展。例如，DynPose-100K（Rockwell 等，2025）与 SpatialVID（Wang 等，2025c）通过对 Panda-70M 数据集进行系统清洗，并补充相机位姿与深度标注，构建了大规模动态 视频数据集，进一步丰富了可用于模型训练与评测的数据资源。

二是动态场景建模能力的持续拓展。前馈式重建正逐步从静态三维场景扩展至随时间变化的动态场景估计。针对动态数据稀缺这一核心瓶颈，过去一年的研究主要围绕训练范式迁移与精细化运动建模两个方向展开。

在训练范式方面，研究工作形成了“微调适配”与“免训练迁移”两条演进路径。受限于真实动态数据规模有限，MonST3R（Zhang 等，2024a）与 PAGE-4D（Zhou 等，2025b）采用“静态预训练 + 动态微调”的策略，分别将 DUSt3R 与 VGGT 中学习到的静态几何先验迁移至动态场景。Align3R（Lu 等，2025b）同样通过微调适配，利用 DUSt3R 框架对齐不同帧之间的单目深度，进一步优化全局相机位姿，得到世界坐标系下的动态重建。Stereo4D（Jin 等，2025a）则通过双目视频构建了大规模、带有三维运动标注的数据集，并在此基础上训练 DUSt3R，取得了具有标志性的精度突破。相比之下，Easi3R（Chen 等，2025d）另辟蹊径，通过对 DUSt3R 注意力机制的深入剖析，揭示其交叉注意力在隐式建模刚性视图变换中的关键作用，并据此提出了一种基于注意力调制的免训练动态场景拓展方法。沿着类似思路，VGGT4D（Hu 等，2025b）与 MUT3R（Shen 等，2025）分别在 VGGT 与 CUT3R 框架下，实现了静态模型向动态场景的免训练自适应。

在结构建模层面，研究重心正从“逐帧点云估计”向“全要素点级时序追踪”深化。动态场景重建不仅要求逐像素对齐的几何与相机位姿恢复，还需显式建模时间维度上的几何对应关系。Uni4D（Yao 等，2025b）和 TrackingWorld（Lu 等，2025c）基于 2D tracking 与单目深度估计，进行 3D 的跟踪，而 St4RTrack（Feng 等，2025a）、SpatialTrackerV2（Xiao 等，2025）、Trace Anything（Liu 等，2025b）与

Any4D（Karhade 等，2025）等方法，基于静态前馈模型引入点跟踪分支，实现了对动态物体逐点运动轨迹的精细刻画。最新的 D4RT（Zhang 等，2025a）则进一步革新了解码范式，提出按需查询机制，通过统一的解码接口避免了密集逐帧解码带来的计算冗余，使模型能够灵活查询时空中任意点的三维位置，为高效的时空建模提供了新的技术路径。另外，在场景与人体运动协同重建方面，Human3R（Chen 等，2025f）基于 CUT3R 模型，Josh3R（Liu 等，2025f）基于 DUST3R，UniSH（Li 等，2026）基于 Pi3 进行局部参数适配，实现动态场景与多人体运动的统一前馈重建。

三是长时序泛化能力的显著增强。受限于 GPU 显存容量，早期前馈式重建方法对输入视角数量和序列长度高度敏感。尽管这类模型通常拥有数百至上千兆参数规模，即便在 A100-80G 等高端显卡上，也往往只能在较短序列（如 32 张图像）上进行训练，并在有限长度（如 200 张图像）内测试。如何在“短训长测”条件下实现稳定泛化，一直是前馈式重建面临的核心挑战。过去一年中，研究者通过重构模型架构与优化推理策略，在长时序泛化能力方面取得了显著进展。目前，长序列泛化推理主要围绕两类工作展开：一类是基于流式处理与测试时学习的模型架构，另一类则是将基础重建模型嵌入到现有 SFM/SLAM 框架中。

在模型架构层面，流式处理与测试时学习逐渐成为主流趋势。CUT3R 提出基于循环神经网络的重建范式，将空间记忆编码至压缩状态空间，使模型能够以近似恒定的内存与时间开销处理连续视角输入，从根本上缓解了全局注意力模型随序列长度平方增长的显存瓶颈。在此基础上，TTT3R（Chen 等，2025e）进一步将 CUT3R 的流式模型重构为测试时学习框架，将三维重建表述为空间记忆状态的在线学习过程，并引入置信度引导的自适应状态更新机制，有效缓解了短序列训练、长序列测试所带来的分布外泛化问题。StreamVGGT（Zhuo 等，2025）与 Stream3R（Lan 等，2025b）通过将 VGGT 的全局注意力机制改造为因果注意力机制，实现了对长序列的流式处理。然而，当前模型在处理长序列重建任务时仍面临状态遗忘等挑战，导致重建质量随序列增长而显著下降。为缓解这一问题，此类方法需借助更多训练数据来充分训练模型。

在系统部署层面，分块处理并融入重建框架则成为目前解决长序列漂移问题最有效 的方法。在显式融入重建框架层面，滑动窗口与全局对齐策略为大规模场景重建提供了 可行路径。VGGT-Long（Deng 等，2025b）等方法采用跨窗口分块处理策略，在局部上下 文内维持恒定内存开销，并通过后处理实现不同窗口预测结果的全局对齐，已在公里级 自动驾驶场景的重建上得到成功应用。SLAM3R（Liu 等，2025e）在利用滑动窗口机制的 基础上，引入历史帧检索机制与基于大模型的增量窗口对齐技术，使其在前馈推理中具 备隐式重定位能力，有效抑制长程漂移，并在消费级单卡（4090D）上实现了 20+ FPS 的实时性能。这些工作的共同进展，标志着前馈式重建模型已具备处理超大规模连续数 据的实际能力。

总体来看，过去一年中，前馈式三维重建在几何精度、长时序泛化能力以及动态场景建模等方面均取得了系统性进展。前馈式重建思路正逐步在三维视觉相关任务中得到应用与拓展。同时，三维重建本身也逐步演化为一种预训练任务，其输出的图像与几何特征正成为支撑空间理解、空间生成与具身智能的重要基础能力（见本文第6项进展）。随着更大规模数据、更强模型结构以及重建与生成范式的深度融合，前馈式三维重建有望在未来空间智能体系中扮演核心感知与建模模块的角色。

二、重建技术与生成技术（视频生成/3D 生成）路线相汇，从互助到初步融合

随着生成式人工智能的发展，将欠定信息输入条件下的三维重建理解为一种特殊的三维生成，已成三维视觉领域学者的基本共识。然而，从具体技术实现来看，三维重建方法和三维生成方法仍是独立发展的两条技术路线。三维生成方面，尽管大部分如 Clay（Zhang 等，2024b）或 TRELIS（Xiang 等，2025）的三维生成大模型都能实现单图输入下的三维生成，但输入图像仅作为生成空间的条件，并无法确保获得重建结果与真实三维之间是仿射不变，甚至无法确保重建结果是像素对齐的。因此，大部分单图或稀疏图像输入下的三维生成并不是严格的三维重建。三维重建方面，自 2024 年来，前馈式方法如 Dust3R（Wang 等，2024）、VGGT（Wang 等，

2025d）等得到了广泛的关注，但本质上仍局限在还原输入图像中的每个像素的三维位置以及相机信息，并不涉及对不可见区域的猜想和重建，前馈式三维重建极少引入生成技术。

进入2025年，三维重建与三维生成的技术边界日益模糊，正经历从独立发展迈向深度融合的阶段。面对稀疏视点输入下的欠定重建难题，视点缺失导致对缺失区域的推断的需求，而纯粹的生成模型缺乏对目标还原的忠实度。为此，学术界致力于探索二者的深层互补机制，构建了多条以提升鲁棒性与一致性为核心的技术演进路径。总体而言，这一融合趋势主要体现为生成先验与几何约束的双向耦合，表现为两方面：一方面，“生成辅助重建”：旨在解决稀疏观测下的信息缺失问题。通过引入图像或视频扩散模型的时空一致性和强泛化能力，将传统的重建任务转化为“先验引导下的推断与补全问题”；或者通过引入对象三维生成模型，通过局部对象的理解和生成，完成场景的重建。另一方面，“重建规范生成”：旨在解决重建与生成过程中引入生成机制带来的随机性、不可控性及不忠实度度问题。本报告将该领域的最新进展归纳为以下三个主要方向：生成先验辅助重建、重建信息引导生成和重建生成迭代优化。

如图 3 所示，重建与生成技术的代表性工作如下。

图 3　重建与生成技术 2025 年度代表性工作

Fig. 3　Representative works on reconstruction and generation in 2025

重建信息引导生成：该类型工作可分为针对静态对象采用 3D 原生生成的研究以及针对动态场景采用视频生成的研究。前者使用重建得到的几何线索作为条件，以约束生成网络的输出，使其符合真实几何结构。例如 ReconViaGen（Chang 等，2025）通过将预训练的 VGGT（Wang 等，2025d）网络特征注入到 3D 生成模型中，有效提升了生成细节与整体形态和真实物体的一致性。Hi3DGen（Ye 等，2025a）则利用法向估计来引导生成使网络对法向层面的细节特征更加敏感，并与输入更加一致。此外，在一些特定场景下用户希望通过自己给定的条件控制生成结果。针对这类需求，Hunyuan3D-Omni（Tencent Hunyuan3D Team，2025）集成了包含体素、点云、包围盒、骨架的多模态条件引导，使用用户可以自由调整生成结果的外观、尺寸、形态。类似地，SpaceControl（Fedele 等，2025）则通过在测试时注入额外几何编码实现几何可控，使得用户无需重新训练新的特有模型也能实现可控生成。

对基于视频生成的动态场景研究方面，则侧重于利用重建获得的显式 3D 表示作为一致性约束，以实现高质量的新视角视频生成。例如，ViewCrafter（Yu 等，2025d）、TrajectoryCrafter（Yu 等，2025b）及 Uni3C（Cao 等，2025）等工作首先通过深度估计构建点云，并将其投影至目标视角的相机平

面,利用这种变形的点云视频序列作为条件输入注入到视频生成模型中。DaS(Gu 等, 2025)将三维点追踪数据转化为“追踪视频(Tracking Video)”,并以此作为视频生成模型的控制条件。VerseCrafter(Zheng 等, 2026)提出 4D 几何控制表示,通过在同一世界坐标系下统一描述动态场景几何信息,从而提升相机与多目标运动控制的精度与一致性,实现 4D 几何控制动态视频生成的端到端框架。NeoVerse(Yang 等, 2026a)通过重建 4D 高斯,利用其在新视角下的渲染结果作为条件约束驱动生成过程。值得关注的是,NeoVerse 提出了一种全新的训练范式,支持直接利用单目视频数据进行训练,极大地拓展了数据的来源。这些方法通过结合视频生成基座模型的强大建模能力与显式 3D 表示的几何约束,显著增强了新视角下的生成结果在空间维度上的几何一致性。

生成先验辅助重建:这类工作关注如何借助生成模型的先验来提升欠定信息下的重建完整性和重建质量。根据采用的生成方法,可以再细分为三类。其一是通过图像生成算法提高静态场景的重建质量。Matrix3D(Lu 等, 2025d)通过对深度、相机位姿、RGB 图像共同进行多模态生成式建模,实现了稀疏视点输入下的任意新视点图像与深度同时生成,并用以辅助 3DGS 的重建;UP2You(Cai 等, 2025b)通过图像生成模型构建了一个“数据整流器”,将不规则的图片集合整流为规整视角下的标准图像,间接地实现三维重建。

其二是通过借助原生 3D 生成算法,实现场景类的组件对象三维重建。CAST(Yao 等, 2025c)通过将场景级重建分解为先估计物体位姿再生成 3D 模型两步实现,并通过物理优化实现更真实的到场景的对齐;SAM3D(SAM 3D Team, 2025)则通过 Mixture-of-Transformers(Liang 等, 2025b)架构将位姿估计和三维生成联合建模为共生生成任务,实现单图像场景的示例级对象重建。CUPID(Huang 等, 2025a)通过 3D 形状与 UV 的共生生成间接实现了同时生成 3D 物体与相机位姿估计,此外 CUPID 还利用估计的相机位姿作为更精细化的纹理生成控制。这些通过生成模型的先验进行重建的方式极大地提高了重建的质量和适用范围。Mesh4D(Jiang 等, 2026)进一步将原生 3D 的重建式生成拓展到动态形变物体上,通过变分自编码器(Variational Auto-Encoder, VAE)将动态形变场编码为一个生成式隐空间,并使用扩散模型通过单视频直接生成物体的网格曲面及其动态形变。

最后一类是借助视频生成算法,实现动态场景三维重建。Geo4D(Jiang 等, 2025d)通过对单视频动态场景共生成点云图、深度图和光线图等多种互补的几何视频形态,在推理时对齐并融合这些形态,实现对长视频鲁棒且精确的 4D 重建;4DNex(Chen 等, 2025h)则仅从单图直接共生生成动态场景的 RGB 与点云图像以得到场景动态点云表达;Lyra(Bahmani 等, 2025)则进一步在视频生成模型的基础上进行针对 3D 高斯的自蒸馏,从而在推理时可以直接生成场景级别的 3D 高斯表达。SV4D(Xie 等, 2024)设计了统一的潜在视频扩散模型,通过视角注意力和帧注意力机制的联合推理,从单目视频生成时序一致的多视角输出,进而实现优化式动态 3D 高斯重建。ReconX(Liu 等, 2024)则使用重建的稀疏点云作为引导以生成 3D 一致的新视点视频,并在此基础上通过优化得到 3D 高斯的场景重建。CAT4D(Wu 等, 2025d)、Diffuman4D(Jin 等, 2025b)分别针对一般场景和虚拟数字人提出基于时间与视角采样的多路视频生成模型,实现了将单视频或稀疏视点视频拓展为时间、空间一致的多路视频的功能,并最终将其重建为高精度的动态高斯表达。

重建生成迭代优化:另一个重要的趋势是重建与生成相互迭代优化的框架出现,通过反馈重建和生成的迭代反馈增强整体系统能力。例如 GenFusion(Wu 等, 2025e)、Free360(Bao 等, 2025)、Difix3D+(Wu 等, 2025b)和 GSFixer(Yin 等, 2025)等工作探索了在迭代循环中交替进行重建与生成。其中,Difix3D+ 通过在新采样轨迹上生成含伪影的渲染图像,并利用单步扩散模型对其进行优化,从而提升 3D 重建质量与新视角合成效果。GenFusion 和 Free360 则在新采样轨迹上渲染包含伪影的视频序列,并将其输入视频扩散模型以生成无伪影结果,进而通过迭代反馈逐步优化 3D 表示。这些工作重建提供粗几何结构作为生成的起点,而生成结果又被反馈用于重建优化,使两者协同提升整体输出质量。

总体来看,2025 年以来的这些工作不仅在技术表现上各自推进了重建或生成的边界,更为重要的

是，它们的设计理念都在逐步模糊重建与生成的界限，使得两者在方法架构、条件输入与任务目标上趋于统一。然而，虽然重建与生成的融合已经产生了许多有效的尝试，目前的大多数方法从宏观而言仍然局限于“先重建再生成”、“先生成再重建”、“边生成边重建”，距离一个统一任务的通用综合模型尚有距离。最近 VIST3A（Go 等，2025）、Gen3R（Huang 等，2026）等工作开始探索将 DUSt3R（Wang 等，2024）、VGGT（Wang 等，2025d）等前馈式重建模型与视频生成模型相结合，如 Gen3R 通过结合这二者的隐空间同时实现了场景级生成与重建，为构建通用的重建与生成模型提供了新的思路。展望未来，如何结合重建与生成领域各自前沿发展以设计一个可支持稀疏输入、复杂动态场景的前馈式多模态信息统一重建方法将是下一阶段的主要探索目标。

三、3DGS/4DGS 持续提升表达效率，掀起场景建模和体积视频应用

随着计算机视觉与图形学技术的快速发展，复杂场景的数字化记录已成为连接物理世界与虚拟空间的关键技术。2023 年 3D 高斯泼溅技术凭借显式点云表示和可微分渲染在静态场景重建领域取得突破，其兼顾质量与效率的特性为复杂场景建模提供了新的三维表征、优化和渲染技术。2025 年，3DGS/4DGS 的表达效率持续提升，掀起场景建模和体积视频应用热潮。以下围绕静态场景的 3DGS 技术和动态场景的 4DGS 技术分别总结 2025 年度进展。

进入 2025 年，静态 3D 高斯重建技术在多视图重建质量、训练收敛速度及实时渲染效率等关键维度持续演进，各类新型三维表达范式竞相涌现，显著推动了高斯重建技术从学术研究迈向实际落地应用。在重建质量上，研究者主要聚焦于优化基元核函数与外观表征。Student Splatting and Scooping（Zhu 等，2025）创新性地引入 Student-t 分布替代传统的高斯分布，并允许负不透明度的存在，有效增强了模型对细节的捕捉能力；DBS（Liu 等，2025a）则提出了 Beta 核函数，并将传统的球谐函数（Spherical Harmonics，SH）外观替换为球形 Beta 外观函数，改善了边缘与高光的重建细节。此外，Billboard Splatting（Svitov 等，2024）将高斯基元的单一颜色属性替换为低分辨率纹理图，提升了场景纹理的保真度。

在重建效率优化层面，3DGS^2（Lan 等，2025a）充分利用了高斯优化过程中黑森矩阵的稀疏特性，引入二阶牛顿法进行求解，将高保真场景的重建时间压缩至 3-4 分钟量级。SpeedySplat（Hanson 等，2025）则通过设计高效的剪枝策略，剔除了超过 90% 的冗余高斯基元，在极大幅度降低显存占用的同时，显著加速了训练收敛过程。

在实时渲染效率上，学术界呈现出从体积辐射场向更高效显式表面表达转型的趋势。Gaussian-Enhanced Surfel（Ye 等，2025b）提出了一种混合表达机制，使用一组完全不透明的二维椭圆面片粗略构建场景几何与外观，并辅以高斯基元补充细节。该方法在训练阶段采用渐进式引导策略，实现了半透明面片向完全不透明状态的平滑过渡；在渲染阶段，利用 Z-buffer 技术直接通过传统图形管线渲染面片，而高斯则以排序无关的方式混合叠加，并利用渲染面片得到的深度图进行剔除实现精确遮挡关系。这种无需排序的传统管线机制，在确保高质量渲染的同时实现了三倍于原始高斯重建的加速，并彻底消除了排序抖动引发的闪烁伪影，已成功在 WebGL 端实现轻量化部署。类似地，MeshSplatting（Held 等，2025）直接采用三角形作为基元，结合德布劳内三角化生成网格，通过类似的渐进式不透明度优化策略完成重建，以少量质量损失实现了与现代渲染引擎的完美兼容。

在实时重建方面，GPS-SLAM（Peng 等，2025b）延续 Gaussian-Enhanced Surfel 的混合表达思路，提出了一种名为“Gaussian-Plus-SDF”的双尺度混合场景表达。该方法采用符号距离场（Signed distance filed，SDF）表示粗尺度的几何与颜色信息，并引入稀疏可优化的 3D 高斯修复 SDF 中的色彩失真并补充高频细节。这种表达将传统几何驱动 SLAM 方法的高实时性与辐射场重建方法的高真实感相结合，得益于 SDF 提供的粗尺度几何与颜色信息，所需优化的高斯数量得以大幅减少，从而加速了高斯优化收敛过程。GPS-SLAM 的扫描重建速度相比当前最优方法提升一个数量级以上，实现了超高速（每秒 150 帧以上）且高保真的实时三维场景重建。

如图 4 所示，静态 3D 高斯重建技术代表性工作

见图示。

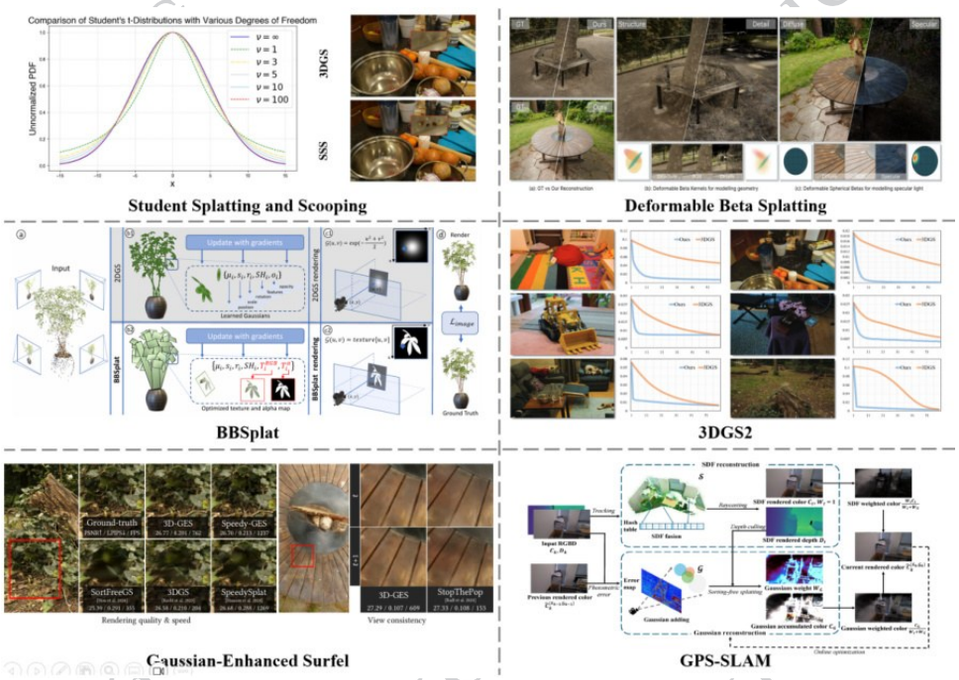

图 4　静态 3D 高斯重建技术 2025 年代表性工作

Fig. 4　Representative works on static 3D Gaussian reconstruction in 2025

尽管静态高斯技术已臻成熟，但在大规模场景重建与数据采集高效性方面仍存瓶颈，特别是如何彻底消除各类伪影以确保任意视角观察的鲁棒性，依然是当前亟待攻克 的难题。展望2026 年，随着大模型技术的深度渗透，利用大规模生成模型提供的强先 验知识来解决过拟合问题，并切实提升重建鲁棒性，将成为核心研究方向（参见本文第 2 项进展）。静态高斯重建技术有望突破“最后一公里”，真正成为高质量的复杂场景建 模工具，落地漫游、VR/AR 等应用服务。

2024 年至 2025 年间，4DGS 技术也是围绕“高效表达”与“高质重建”两大目标快速演进：从原生 4D 高斯的理论探索，到高质量重建、渲染效率提升、存储压缩、长序列处理的工程优化，显著缩小了研究原型与实用系统之间的差距，4DGS 重建基本达到 to B 级别应用。

动态场景建模面临的核心挑战源于时间维度的引入，传统的逐帧重建策略会带来存储需求的线性增长和时间连续性的缺失。4D Gaussian Splatting（Wu 等，2024）和 4D-Rotor Gaussian Splatting（Duan 等，2024）分别独立提出的 4DGS 将动态场景建模为连续的 4D 时空体积，使用原生 4D 高斯直接表示时空结构，自然支持任意时空点的采样和渲染，同时利用 4D 几何的内在属性，用紧凑的参数直接编码运动，避免了冗余的中间表示。通过全局优化和时空密度控制，智能地分配计算资源，用最少的高斯图元覆盖最复杂的动态变化。为后续工作奠定了表示层面的理论基础。

随着研究的深入，工作重点逐渐转向提升方法的实用性。在渲染效率方面，4DGS-1K（Yuan 等，2025）通过时空变异评分剪枝和活跃高斯掩码机制，实现了 1000+ FPS 的渲染速度，同时将存储需求降低了 41 倍；Temporal Gaussian Hierarchy（Xu 等，2024）针对长视频建模，观察到动态场景中不同区域具有不同程度的时序冗余，据此构建多层级 4D 高斯表示，每层只需加载对应片段 4D 高斯进显存，仅用 17. 2GB 显存即可处理 18000 帧的长序列。在功能拓展方面，FreeTimeGS（Wang 等，2025h）允许高斯在任意时间和位置出现的 4D 表示，通过为每个高斯赋予运动函数来模拟物体的物理移动，有助于在时间维度上重用高斯图元，降低表示冗余并提升时序一致性；SharpTimeGS（Liao 等，2026）通过生命周期调制运动与平顶时间可见性函数增强高斯图元表征能力，进一步在时间维度上重用高斯图

元，同时保持动静态建模能力并提升时序一致性；4DSloMo（Chen 等，2025g）引入多相机的时空间插采集，借助 4DGS 建模所具备的时空联合建模能力，实现对非同步相机进行 4D 重建，提升了对高速人体对象运动的高效高质重建；Split4D（Hu 等，2025a）基于流式特征学习策略实现无需视频分割的场景解耦重建，为动态场景的编辑和交互提供了新的可能性。STGS（Li 等，2024）通过将运动建模为非线性复杂运动的同时渲染特征解码图像，实现了高质量重建。4DGV（Dai 等，2025a）通过运动分层表示实现动态分离，提高了具有大幅度运动的物体的渲染质量。总体而言，相比于逐帧的 3DGS 重建，4DGS 联合考虑时空信息，重建结果得以更加时空连续，重建和渲染质量达到可用水平，同时表达效率更高效，初步满足互联网数据传输带宽要求（2MB/s）。

上述4DGS方法能达到满足高质量重建效果以及低功耗终端的实时交互渲染能力。然而，在重建速度和采集便捷性方便仍距离大规模商业落地应用存在差距。目前，重建1分钟的数据仍需要小时级以上时间，并且需要规模化相机阵列。首先，为降低相机数量需求，利用视频生成技术辅助新视点生成，实现动态场景的高斯泼溅重建是当前的主流思路（参见本文第2项进展）。同时，还有一类方法通过引入运动信息进行单视频重建，MoSca（Lei 等，2025）利用从基础视觉模型提取的先验知识得到新的运动支架表示对底层运动和变形进行紧凑平滑的编码，从而实现场景几何形状和外观与变形场解纠缠，完成了对单目视频的动态重建；Shape-of-Motion（Wang 等，2025f）通过具有时变平移和旋转的全局 3D 高斯表示动态元素，使用低维刚性运动基对运动轨迹进行正则化并将噪声观测整合到全局一致的场景估计中，依靠单目视频实现了 2D、3D 追踪和动态场景重建；SplineGS（Park 等，2025）利用创新的运动自适应样条进行动态运动建模，有效地从复杂的单目视频建模动态场景。目前，这类型的重建质量目前仍远低于密集相机阵列的 4DGS 重建方法。

针对快速重建能力研究，前馈重建的兴起将逐场景优化范式转变为数据驱动的直接推理范式，为快速重建和可泛化性提供基础（参见本文第1项进展）。其中，BTimer（Liang 等，2024）采用子弹时间公式化，聚合所有上下文帧的信息预测目标时刻的完整场景，仅需 150 毫秒即可从单目视频完成动态场景的 3D 高斯重建；MoVieS（Lin 等，2025a）使用像素对齐的可变形高斯并显式监督其时变运动，在单一框架内统一了外观、几何与运动的建模，一秒内完成 4D 重建的同时支持场景流估计和运动物体分割等零样本任务；4DGT（Xu 等，2025b）完全使用真实世界单目视频训练，以滑动窗口方式处理 64 帧输入并通过密度控制策略有效扩展至长序列，在跨域泛化能力上取得显著突破。这一范式转变使动态场景重建的时间成本从小时级降至秒级。然而，需看到这类研究目前仅能处理单视点输入，且仅能生成极小视点范围内的自由视点视频内容。

如图 5 所示，动态 3D 高斯重建技术代表性工作如下。

展望 2026 年动态场景高斯建模技术，通过融合生成技术和前馈技术，有望实现少量视点甚至单视点下的高质量内容重建；4DGS 的数据表达效率和压缩效率有望进一步提升，适配当前网络带宽下的实时传输应用；最后，在有先验类别场景，如人体为主的动态场景，实现纯 RGB 相机的实时动态高质量重建能力是应用趋势。随着场景解耦、运动理解和零样本泛化等能力的持续提升，4DGS 有望成为增强现实、影视制作、自动驾驶仿真乃至具身智能等领域的基础设施级技术，推动动态三维内容从“昂贵的专业制作”走向“普惠的即时重建与生成”。

四、三维生成：从单体视觉逼真到部件场景结构化、物理可交互能力跃迁

2025年，三维生成进入从视觉优先向结构与物理优先转变的关键期。技术进展呈现五个互联主线：更精细的几何表征与大规模潜空间推动微观细节恢复；三维原生纹理实现高保真物理基础渲染（Physically Based Rendering，PBR）输出并解决投影伪影；部件级生成在 vecset 与 sparse-voxel 路径上实现可控与高精度拆装；场景生成从语义解耦走向全局协同并借助通用基础模型扩展开放世界能力；与 CAD、装配及绑定流程融合，推动工业与游戏级可装配、物理一致的生成。整体上，表征创新、结构化潜空间与跨模态大模型成为贯穿全链条的核心动力，研究重心由单体视觉逼真向结构一致、物理可

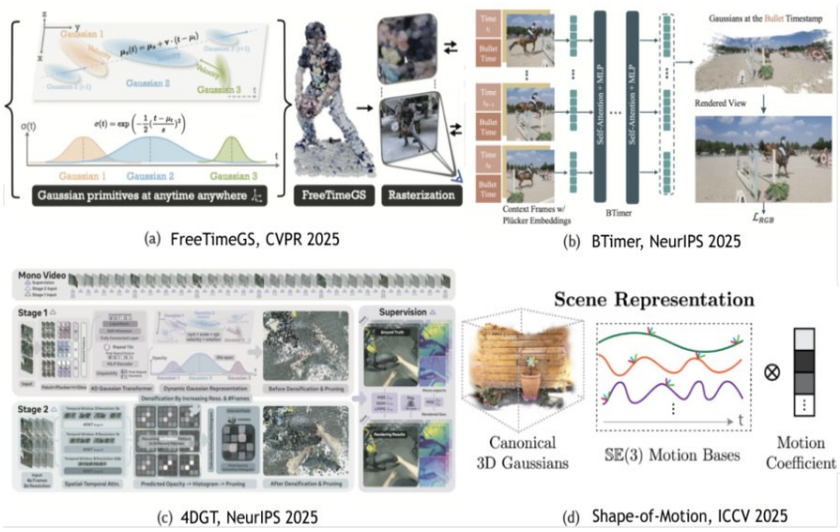

图 5　动态 3D 高斯重建技术 2025 年代表性工作

Fig. 5　Representative works on dynamic 3D Gaussian reconstruction in 2025

交互的系统能力迁移。

如图 6 所示，三维生成代表性工作如下。

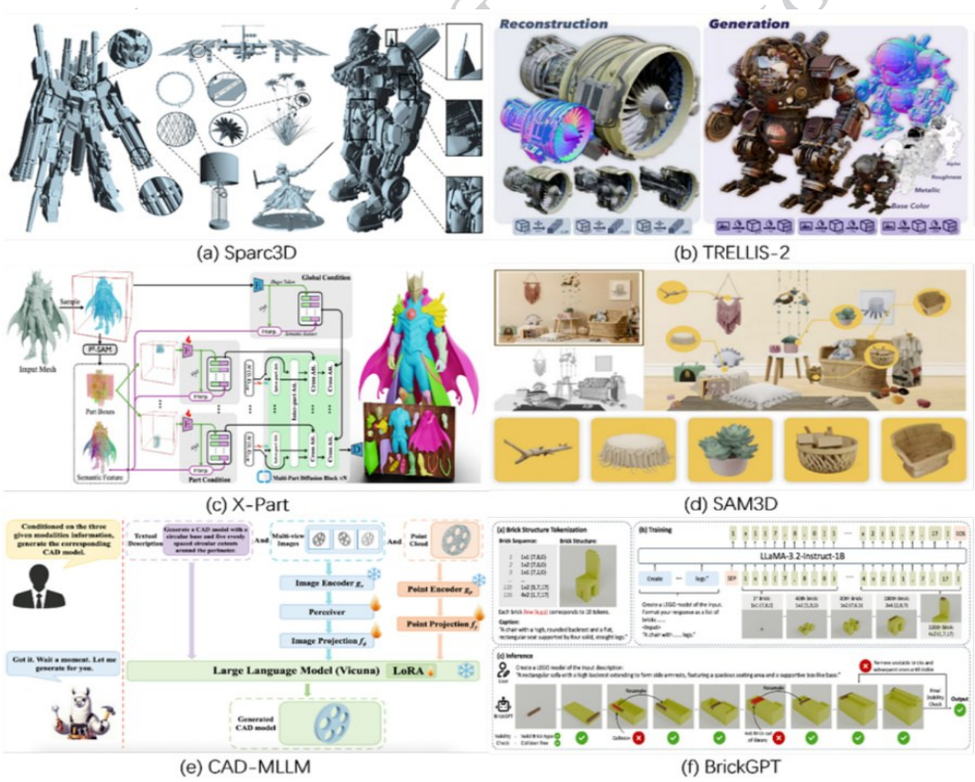

图 6　三维生成 2025 年代表性工作

Fig. 6　Representative works on 3D generation in 2025

● 更精细的三维生成（几何细节、表征与大规模潜空间）：2025 年，三维几何生成的研究重心实现了从描述宏观结构向恢复微观细节的跨越。年初的 Hi3DGen（Ye 等，2025a）率先探索了相关方向，该方法通过利用从图像中得到的 Normal Map 作为中间桥梁实现了精细化的三维生成，为相关研究提供了新的方向。然而，这种依赖中间表示的方法框架并未直接在三维结构上进行优化，针对这一问题，Direct3D-S2（Wu 等，2025）通过空间稀疏注意力（Spatial Sparse Attention，SSA）机制使得在高分辨率

下的生成训练成本大幅降低，Sparc3D（Li 等，2025）从几何表征的角度进行了创新，提出将 Mesh 转为新的几何表征方式，实现了在同一框架下的精细化三维生成，但其较高的训练成本和复杂的数据预处理都限制了该方法的使用。Lattice（Lai 等，2025a）进一步在几何表征上了进行创新，通过设计动态调节的几何表征方式 VoxSet，在精度和效率上实现了兼顾。在此之后，Trellis. 2（Xiang 等，2025）提出了同时编码形状和外观的几何表征结构 O-Voxel，并且减少了对数据预处理的要求，进一步实现了更友好的精细化三维生成。未来研究将更专注于几何表征方面的创新和更简单的方法框架。

● 高保真与三维原生纹理合成：三维纹理生成领域正在经历从“多视图投影拼接”向“3D 原生材质建模”的范式跃迁。TexVerse（Zhang 等，2025b）通过发布超大规模数据集，为高泛化纹理模型的训练奠定了数据基础，促使研究视角从局部纹理扩展至全局特征。为了解决传统投影技术引发的接缝与伪影问题，UniTex（Liang 等，2025）率先提出在 3D 连续表面坐标系中直接生成材质，实现了几何感知的精准对齐。在此基础上，针对表征精度的提升出现了三条互补路径：TEXTRIX（Zeng 等，2025）提出基于潜在属性网格的原生纹理生成方法，NaTex（Lai 等，2025b）利用点集向量（VecSet）建模材质以突破网格分辨率限制，而 LaFiTe（Chen 等，2025）则通过结构化潜空间确保了建模的稳定性。近期，TRELIS. 2（Xiang 等，2025）进一步引入 O-Voxel 潜空间，在原生 3D 框架下实现了高分辨率的 PBR 输出，标志着生成纹理在视觉细节上能已能媲美 2D 扩散模型。当前的研究重心已从单体建模转向场景级材质统合，旨在整合不同表征的泛化力与工业级解耦技术，产出具备物理一致性、可跨场景通用的材质参数集，从而为数字孪生与具身智能提供高保真的视觉与物理交互支持。

● 三维部件生成：部件生成在基于 sparse voxel 以及 vecset 两种原生表征上进行了分别的探索，PartCrafter（Lin 等，2025），PartPacker（Tang 等，2025），BANG（Zhang 等，2025a）基于 vecset 的表征，实现了端到端的自适应的精细部件结构的生成，其中 PartPacker 引入 Dual Volume 的拓扑先验，一定程度上缓解了网络自适应拆件时的几何奇异性。而 X-Part（Yan 等，2025）采用‘先分割后补全’多阶段策略，旨在可控以及高精度的部件生成。Omni-Part（Yang 等，2025）作为 sparse voxel 表征的尝试，引入了分割图条件下的自回归预测部件 bbox，也实现了可控，高效的部件生成。

● 三维场景生成：在场景生成任务中，现有工作均致力于解决多物体间的空间布局与几何一致性难题，但在解决路径上各有侧重。针对复杂场景中物体布局混乱、易穿模的痛点，CAST（Yao 等，2025c）提出了“语义先验解耦”方案，利用大语言模型构建关系图来约束独立生成的物体位姿，通过明确的语义逻辑显著改善了物理合理性。然而，这种分治策略难以捕捉物体间微妙的紧密耦合关系（如相互遮挡与光照影响），MIDI（Huang 等，2025）进而转向“全局协同生成”，通过在扩散模型中引入多实例注意力机制，实现在同一生成流中同步优化多个物体，强化了整体场景的视觉连贯性与几何协调性。为了进一步突破上述方法对特定类别的依赖并解决泛化难题，SAM3D（SAM3D Team，2025）引入了“通用基础模型”驱动的思路，利用大规模分割先验实现了“万物三维化”的几何初始化，弥补了场景生成在开放世界（Open-World）场景下通用性不足的问题，为构建多样化场景提供了高泛化的几何底座。

● 三维生成方向的其他多元化发展（结构化、CAD、绑定与装配）：三维生成技术已不再局限于单纯的形状建模，而是与工业计算机辅助设计（CAD）流程、游戏建模及物理可装配性需求深度融合，呈现多方向并行发展的态势。在工业 CAD 制造领域，多种 CAD 表征方法均取得了阶段性进展：CAD-MLLM（Xu 等，2024）以 CAD 命令序列作为核心表征形式，基于多模态大模型实现 CAD 命令序列的生成，支持多种输入模态到参数化 CAD 模型的统一生成任务；AutoBrep（Xu 等，2025）则采用边界表示（Boundary Representation，BRep）方法，将 BRep 中的面、边及拓扑关系转化为离散标记序列，可支持复杂实体的长序列建模及单阶段训练过程。在游戏中建模的下游应用场景，如布线、多边形网格生成等任务中，RigAnything（Liu 等，2025）模型能够实现无模板约束的关节与骨骼拓扑生成及蒙皮权重分配，有效推动了通用化绑定技术的发展；BPT（Weng 等，2025）通过块级索引与补丁聚合策略对网格序列进行压缩，为高效处理更高面数的网格数据提供

了可行方案；BrickGPT（Pun 等，2025）基于自回归语言模型生成物理稳定的互锁式积木结构，融合物理约束检测与有效性回滚机制，保障了生成结构的可装配性与力学稳定性。

展望未来，三维生成将由模块化研究走向系统化工程。短期内，几何与材质表征的统一、结构化潜空间与多尺度解耦将成为提升精度与泛化的主攻方向；中期将见证跨模态基础模型（融合视觉、语言与物理模拟）带来的普适生成能力，支持开放世界的即插即用三维化；长期看，物理可交互的数字孪生、与CAD/制造流程的无缝对接以及实时可控的高分辨率生成将成为产业化落地的关键。同时，低成本标注、可解释性的评估框架与标准化基准数据集是推动学术向工业转化的必要条件。

五、从视频生成到世界模型：面向时空一致、物理合理与可交互

2025 年，世界模型成为三维视觉和计算机图形学等领域最受关注的焦点之一。随着大语言推理模型、多模态大模型、视频生成和三维重建等技术快速发展，研究者能够以更高层次建构和预测复杂世界状态，使“可理解、可预测、可交互”的数字世界逐渐成为可能。2025 年在世界模型的研究工作方面取得了一系列进展，主要聚焦于物理感知、空间一致、长序记忆等核心问题。该领域正处于蓬勃兴起的快速发展期，新技术迭代迅速，新范式层出不穷，在具身智能、空间智能和智能驾驶等应用场景将产生实用价值。美国工程院院士、斯坦福教授李飞飞指出，世界模型在机器人、教育、医疗保健、制造业、农业、虚拟现实（Visual Reality，VR）和增强现实（Augmented Reality，AR）等多个方面都有广阔应用前景。图灵奖得主 LeCun 也指出，仅仅文本和语言是不够的，AI 系统需要理解现实世界，具备一定的常识，以及推理和规划的能力，拥有持久记忆。

世界模型被定义为编码环境知识并模拟其动态变化的数字引擎，能够在语义、物理、几何与动态等多重复杂世界（无论虚拟或现实）中进行理解、推理、生成和交互的模型。当下，以 Sora 为代表的视频生成模型已经在视觉质量上得到优异的结果，实现“表面忠实”，世界模型技术则进一步推进至“一致性忠实”和“物理内在忠实”，保证时空一致性，支持复杂物理的交互和控制。通过世界模型，可以实现对物理世界未来状态的精确预测和演化模拟，由“视觉合理”迈向“物理真实”。针对使用视频生成模型构建世界模型所面临的挑战，包括物理合理性不足、空间一致性较弱、难以交互式生成、缺乏长序列和记忆四个方面，2025 年出现许多优秀的工作解决上述问题。

如图 7 所示，视频世界模型代表性工作如下。

图 7　视频世界模型 2025 年代表性工作

Fig. 7　Representative works on video world models in 2025

在提升视频生成结果的物理合理性方面,现有方法主要分为两类:隐式先验注入与显式先验注入。隐式先验注入的方法将视频理解模型的物理先验知识蒸馏至视频生成模型,在尽可能保持原始视频模型架构的情况下引入新的优化约束项,隐式提升视频结果物理合理性。DiffPhy(Zhang 等, 2025c)使用大语言模型在物理常识、语义一致性和与预期物理现象等方面添加额外的约束项,并添加新的注意力层关注不合理的物理情境。VideoREPA(Zhang 等, 2025e)将视频理解模型(video foundation models,VFMs)的物理能力迁移到生成模型,提出词元关系蒸馏损失函数,提升物理合理性。对比隐式方法,显式先验注入方法构建场景或物体的几何代理,显式预测物体的运动状态后以引导视频生成。VLIPP(Yang 等, 2025d)基于多模态模型预测二维包围盒的运动轨迹;ReVision(Liu 等, 2026)提取粗糙视频中的三维形状与运动信息,结合参数化物理先验模型(Parameterized Motion Prior model, PMP)优化结果;PhysCtrl(Wang 等, 2025a)则将物体分割并重建为三维点云,再基于扩散模型生成物理轨迹。此外,部分方法在生成视频的时同步生成动作/3D 表征,联合视觉表观先验与物理约束,提高世界模型的几何与物理感知能力。例如,VideoJAM(Chefer 等, 2025)引入光流信息,Tesseract(Zhen 等, 2025)同步生成深度和法向信息,4DNeX(Chen 等, 2025h)则约束生成 4D 点云场景。值得注意的是,商业模型如 Sora 2 和 Veo3 在物理合理的视频生成方面也取得了很好的效果。

除了基于数据驱动方式,直接在重建的三维场景上显式地集成物理仿真引擎,并生成物理真实的结果也出现了许多工作。目前方法主要探索基于三维高斯泼溅的场景加入物理方程约束,例如,RainyGS(Dai 等, 2025b)生成视觉和物理真实的降雨仿真效果,FieryGS(Shen 等, 2026)引入燃烧和炭化方程,模拟符合物理规律的火焰传播和烟雾生成。

多视角一致性是世界模型实现空间连贯感的关键。现有研究也可分为隐式学习与显式学习两类。隐式学习方法采用数据驱动的路径,需要使用渲染引擎或特殊方法采集的多视角数据,并对视频模型结构进行调整,隐式学习多视角特征,实现多视角一致性视频生成。ReCamMaster(Bai 等, 2025a)将相机轨迹输入到 DiT 视频生成模型,实现已有视频的新视角变换。SynCamMaster(Bai 等, 2025b)和 SV4D 2.0(Yao 等, 2025a)同步预测多个视角的视频结果,并扩展引入多视角注意力,提升多视角一致性。与隐式方法不同,显式学习方法则采用"3D 重建-新视角渲染-引导生成"链路,将粗糙投影结果作为条件,引导扩散模型重绘与生成视频细节,实现多视角几何一致且纹理逼真的视频生成。ViewCrafter(Yu 等, 2025d)从稀疏图像出发,实现精确相机控制的场景的新视角合成。TrajectoryCrafter(Yu 等, 2025b)将上述思路扩展至动态场景,由动态点云引导新视角合成。点云以外的三维表示也被用作控制信号,例如 DaS(Gu 等, 2025)采用 3D 轨迹视频,而 GS-DiT(Bian 等, 2025)则采用 4D 高斯表示。

三维高斯的渲染结果真实且本身具备良好的三维一致性,也可用于构建世界模型。例如,GWM(Lu 等,2025a)采用了与潜空间扩散模型(Latent Diffusion Model,LDM)类似的思路,使用 VAE 对高斯场景进行压缩,并在隐态空间中构建扩散模型。另一类方法则更加深入的利用三维高斯的显示特点,对高斯球进行三维变换,生成未来场景。例如,GaussianWorld(Zuo 等,2025b)根据自车移动进行仿射变换,生成新区域并添加物体运动信息,得到未来状态。ManiGaussian++(Yu 等,2025c)针对双手机械臂场景进行设计,构建主导变形模型和跟随变形模型对高斯进行变换,模拟物体抓取等操作。

基于视频生成构建世界模型的关键问题之一是确保长视频生成过程中的时空一致性,需要构建记忆机制。当前的工作主要也可以分为显式记忆和隐式记忆两类。显式记忆方法使用三维重建技术作为载体,直接对 3D 记忆进行渲染作为输入。例如,DeepVerse(Chen 等, 2025b)和 EvoWorld(Wang 等, 2025b)在预测未来视频结果的同时构建点云场景,引导后续生成。Spmem(Wu 等, 2025f)构建了三类记忆模块,包括短期、长期和稀疏情景记忆,实现长时序生成。隐式记忆方法使用 Transformer 的 Context 窗口或键值缓存(Key-Value Cache, KV Cache)存储历史信息,避免重建误差。例如,Context as Memory(Yu 等, 2025a)根据视野范围对历史帧进行检索并提供上下文信息,FramePack(Zhang 等, 2025d)根据重要性对历史帧进行压缩,高效生成时序一致的长视频结果。相关工作也提出混合记忆机制,采用隐式记忆和弱三维重建的路径。例

如，VMem（Li 等，2025b）和 RTFM 基于重建的点云场景或空间位置和朝向信息检索历史帧，引导预测未来状态，期望融合显式和隐式的优势。

视频世界模型交互式生成也是重要研究问题。一类交互式方法关注于从视频数据中学习视角交互信息，以增强模型的空间理解能力。例如，YUME（Mao 等，2025）将相机的相对位置和朝向变化使用文本编码，控制视频生成；NWM（Bar 等，2025）将相机参数作为自适应实例标准化（Adaptive Instance Normalization，AdaIN）参数，并使用交叉注意力融合前序帧特征。另一类交互式方法则以额外的用户动作或机器人动作为输入，实现对视频内容交互生成。例如，Astra（Chen 等，2025c）提出了 ACT-Adapter 的模块，编码动作信息并注入至视频模型；DWS（He 等，2025a）提出了运动强化损失，更加关注由动作引起的视频的区域变化。工业界在交互式视频世界模型领域也快速发展，例如腾讯开源了混元世界模型 1.1；昆仑万维开源 Matrix-Game 2.0。Google 发布了 Genie3，可以实时生成数分钟长的视频结果，并模拟主体和环境的真实的交互效果，也可以确保长时序的一致性。WorldLabs 发布了 Marble，可以通过单张图像或文本创建三维世界，并且支持交互式的编辑、扩展和融合，得到多种表示的场景输出结果。

除了基于视频模型的方法，隐空间世界模型通过在特征空间中建模动态变化，避免直接进行高维像素级预测，从而实现更高的计算效率与建模紧凑性。这类方法通常借助预训练图像编码器或自监督视频表示模型，在隐态特征空间中进行时间序列建模与未来状态预测。典型工作包括 Meta FAIR 团队提出的 DINO-world（Zhou 等，2025a）（在 DINOv2 的特征空间中训练视频世界模型），V-JEPA 2（Assran 等，2025）（采用自监督学习框架，从视频中学习统一的表示空间）。

以上系列研究验证了构建可交互的物理真实的世界模型的可行性，在多个方面已经 展现了惊艳的效果，并在智能驾驶、具身智能和影视游戏等多方面有应用价值。然而，尽管世界模型取得了显著进展，仍有许多值得进一步探讨的问题。首先，虽然已有很多 方法探索生成物理合理的视频合成结果，但是如何在任意情景下确保物理可靠性，并支 持物理参数的精细化控制仍是值得探索的问题。其次，三维一致和交互式的场景生成在 长时序生成方面仍面临瓶颈，如何实现真正无限场景的生成并实现完全准确的历史记忆 仍值得研究。最后，尽管现有方法实现了虚拟场景的构建和模拟，如何对真实的场景进 行虚拟复刻，并且支持一系列探索和交互也是值得探索的方向。

六、理解与生成统一的多模态大模型服务空间智能感知

空间智能（Spatial Intelligence）在 2025 年被学术界公认为视觉语言模型（vision-language model，VLM）从“读图”迈向“世界模型”的关键一步。随着具身智能（Embodied AI）和自动驾驶等领域对物理世界理解需求的爆发，仅具备语义感知的 VLM 已无法满足实际应用中对三维空间推理、导航和操作的要求。这一年，学术界在空间智能的理论体系、数据构建、架构创新及训练范式上均取得了重要进展，标志着视觉语言模型正式从二维图像的被动观察者向三维环境的主动思考者转变。本节将在基础评测基准、训练数据、模型架构、训练方法以及迈向世界模型的超感知探索等方面回顾 2025 年度的代表性研究工作。

如图 8 所示，空间智能感知方向代表性工作如下。

在基础评测基准方面，2025 年的综述论文 Spatial Intelligence in Vision-Language Models 首次为空间智能建立了清晰的认知层级体系，将其划分空间感知（Perception，如 3D 检测）、空间理解（Understanding，如相对位置推理）以及空间外推（Extrapolation，如心理旋转与路径规划）三个递进水平。为了精准评估这些能力，VSI-Bench（Zuo 等，2025a）引入了基于视频的“看、记、忆”（See，Remember，Recall）评估范式，考察模型在观测环境后构建隐式“认知地图”的能力。此外，RealWorldQA（xAI 等，2024）和 Omni-Spatial（Jia 等，2025）等综合基准的推出，覆盖了从基础感知到复杂逻辑推理的全方位能力，揭示了当前模型在跨认知层级整合方面的不足。

在训练数据方面，2025 年见证了空间专用数据集的爆发式增长。据统计，仅在 2023 至 2025 年间就涌现了 21 个主要的空间训练语料库，数据量级呈现出明显的加速趋势。研究重心从通用的图文匹配转

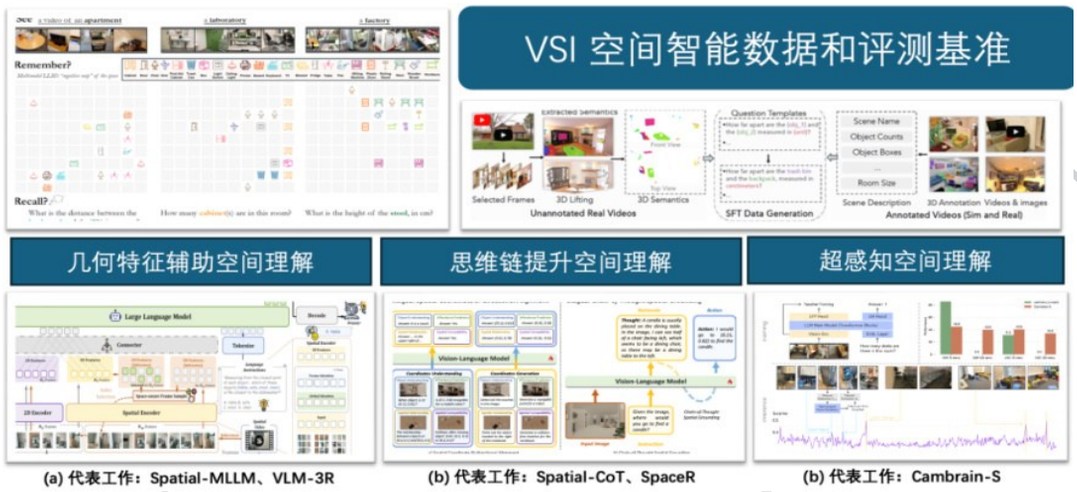

图 8　空间智能感知 2025 年代表性工作

Fig. 8　Representative works on spatial intelligence perception in 2025

向了更具针对性的空间指令微调。例如，为了支持长时程空间记忆的训练，Cambrian-S（Yang 等，2025c）构建了包含 59 万条指令的大规模数据集 VSI-590K，涵盖了空间计数、距离估计等细粒度任务。尽管数据规模在扩大，但综述也指出当前数据仍存在“认知不平衡”和“模态单一”的问题，即侧重于静态 2D 图像的简单关系描述，而缺乏涉及 3D 几何变换和动态预测的高阶数据。

在模型架构方面，核心突破在于打破 2D 视觉与 3D 物理世界之间的壁垒。针对传统 VLM 过分依赖 CLIP 等 2D 语义特征、导致几何感知缺失的问题，Spatial-MLLM（Wu 等，2025a）和 VLM-3R（Yang 等，2025a）等方法引入预训练的视觉几何基础模型（VGGT/CUT3R）作为独立的 3D 空间编码器。这些专用的空间编码器能够从多视角图像（Multi-view Images）或视频流中提取显式的 3D 几何特征，包括深度信息、相机位姿变化以及场景的三维结构流，从而弥补了传统 ViT 在处理空间深度和透视关系上的先天不足。

在训练方法方面，研究发现传统的监督微调难以赋予模型复杂的空间逻辑，因此强化学习（Reinforcement Learning，RL）和思维链（Chain of Thought，CoT）成为了新的破局点。为了增强推理的逻辑性，Spatial-CoT（Liu 等，2025d）训练模型首先生成空间坐标和推理依据，再输出最终答案。更进一步，SpaceR（Ouyang 等，2025）和 ViLaSR（Wu 等，2025c）等工作引入了强化学习框架（如组相对策略优化（Group Relative Policy Optimization，GRPO）算法），通过设计空间特定的奖励函数，鼓励模型在复杂的动态场景中生成符合物理规律的推理路径。这种从“模仿学习”到“强化推理”的转变，显著提升了模型在未见场景中的泛化能力。

作为 2025 年空间智能领域的压轴进展，Cambrian-S（Yang 等，2025c）的提出标志着该领域向“空间超感知”（Spatial Supersensing）与世界模型的演进。针对现有模型在长视频中记忆模糊的痛点，Cambrian-S 提出了“预测即感知”（Predictive Sensing）的理念。该模型不再被动处理每一帧，而是像人类大脑一样，利用下一帧的潜在特征预测误差来动态调节注意力，从而在无限视频流中高效地管理记忆。

同时，多模态模型逐渐从单一的理解或生成功能走向统一理解和生成的多模态输入、多模态输出模型。以 Bagel（Deng 等，2025a）为代表的统一多模态模型（Unified Multimodal Models，UMM），将视觉理解、视觉生成和编辑等任务统一在统一模型中，并进一步展现出了图文交错生成和多模态推理的能力。这些统一多模态模型领域的新进展，给下一代统一感知、理解、重建、预测的空间智能多模态模型奠定了基础。在具身智能领域，WorldVLA（Cen 等，2025）等模型通过统一多模态模型，将视觉和指令理解、图片预测和动作预测统一在 VLA 模型中，提升了 VLA 模型对于视觉环境中的物理理解能力和动作预测能力。

七、数字人前沿转变：从外观建模到多

模态交互

2025年数字人技术方面，基于三维表征的，特别是基于高斯泼溅的头部与全身数字化重建技术继续推进，尤其在基于单张或多张图像输入的高效、高保真重建方面涌现出多种创新方法。在头部重建领域，Flex Avatar（Kirschstein 等，2025a）展现出强大的泛化能力，通过基于 Transformer 的 3D 肖像动画模型，可实现在单目和多视角数据集上进行统一训练，确保泛化性的同时得到 3D 人头的一致性，实现了从单张图像创建高质量且完整 3D 头部虚拟人的方法。RGBAvatar（Li 等，2025a）提出了一种简化的高斯混合蒙皮，可从单目视频输入中快速地重建出人头化身，并且可以支持实时交互。对于多图输入场景，Avat3r（Kirschstein 等，2025b）和 FastGHA（Ji 等，2026）提出了绑定在四张输入图像像素空间的高斯表达可动画三维头部化身；FlexAvator（Peng 等，2025a）与 FastAvator（Liang 等，2025a）；（Wu 等，2025g）均提出了基于前馈网络的快速重建架构，可从任意数量的输入图像中高效复原三维模型；此外，GUAVA（Zhang 等，2025b）和 Bringing Your Portrait to 3D Presence 则提出了一种能够支持头部、半身乃至全身输入的可动画三维化身重建的前馈生成式框架。在全身数字人构建方面，TaoAvator（Chen 等，2025a）和 mmlpHuman（Zhan 等，2025）通过构建轻量化的多层感知机（Multilayer Perceptron，MLP）训练范式，能够从多视点单人采集数据中重建出逼真、可驱动且支持实时渲染的高质量全身高斯数字人；在全身人体 3D 化身建模方面，虽然如 PERSONA、LHM（Qiu 等，2025）、IDOL、PGHM 等融合了二维视频和三维数据进行可泛化重建，但通过单视频或单图重建的数字化化身效果仍不尽人意，难以捕捉服装的高逼真飘动。

如图 9 所示，三维数字化身领域的代表性工作如下。

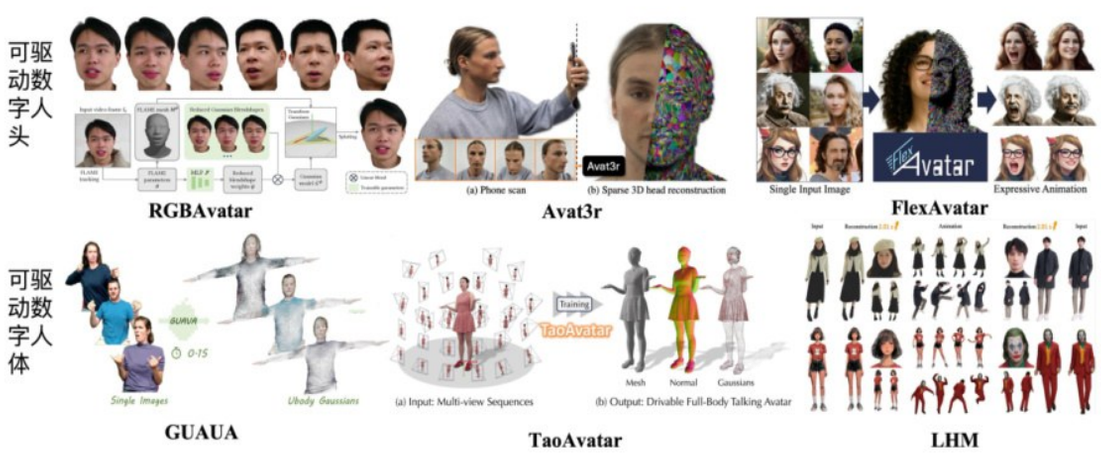

图 9　三维数字化身 2025 年代表性工作

Fig. 9　Representative works on 3D digital avatars in 2025

除了上述 3D 外观建模技术的扎实推进，2025 年数字人生成技术正逐步从纯视觉效果导向的“外观建模”向全方位“基于多模态交互的数字人生成”转型，这一技术变革在产业界表现突出。随着大语言模型、视觉模型与多模态感知体系的协同发展，各大头部厂商已推出具备语音、动作、情感和环境感知能力的多模态数字人系统。以 OpenAI Sora-2、字节跳动 M3-Agent（Long 等，2025）为代表的工作，通过整合语音识别、自然语言理解、情绪识别及动作合成等多模态技术，推动数字人由静态形象迈向具备智能交互与个性化表达的“数字生命体”。

以音视频为核心的二维生成技术是多模态数字人技术的重要分支，主要强调音视频同步、人物和场景声音视频联合建模等高仿真视听效果的呈现。这一技术路线通常包括音频驱动和音视频联合生成两个主要方向。其中，音频驱动数字人视频生成方面，如 InfiniteTalk（Yang 等，2025b），专为无限长序列配音设计，采用流式生成架构，充分利用时间上下文帧实现无缝片段过渡，并通过精细参考帧定位提升音乐、语音与视频片段衔接的流畅性和控制

精度。其凭借无限长度生成能力及唇形、头部、表情与姿态的精准同步，已迅速成为语音驱动虚拟人领域的主流工具之一。AnyTalker（Zhong 等，2025）则将主流的单人说话方法扩展为多人对话视频生成框架，采用灵活的多流结构，仅利用以单人说话为主的数据集即可实现真实一致的多人对话，既能实现 identity 的大规模扩展，又能确保不同 ID 间的无缝互动。

如图 10 所示，以音视频为中心的多模态数字人代表性工作如下。

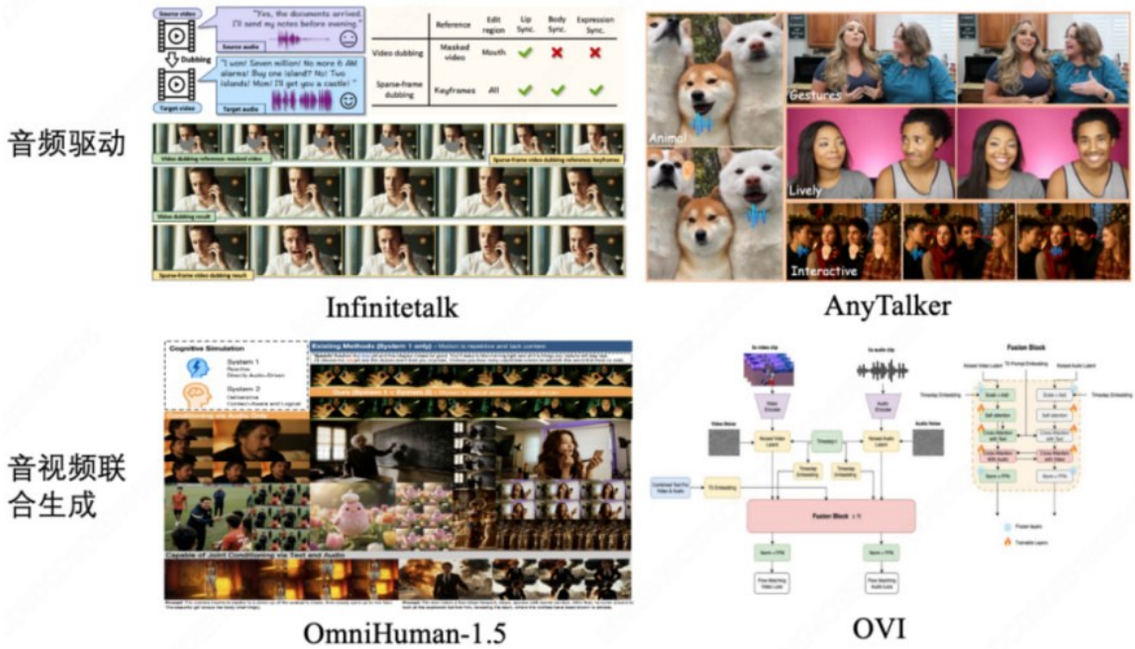

图 10　以音视频为中心的多模态数字人 2025 年代表性工作

Fig. 10　Representative works on audio-visual centered multimodal digital humans in 2025

音视频联合生成方面则以 OmniHuman-1.5（Jiang 等，2025b）和 OVI（Low 等，2025）为代表。OmniHuman-1.5 通过多模态大型语言模型，合成结构化文本表示并提供更强语义引导，实现动作、语音和视频内容的联合建模。该方法不仅能够生成兼具语境理解与情感共鸣的动作，还能够实现在听觉和视觉两个维度上的多感官用户体验，将传统以单模态生成为核心的任务（例如音频驱动的视频生成）扩展至更高维度。随着相关技术的发展，应用场景也得到了广泛拓展，由此前主要依赖音频输入的生成模式，升级为可通过文本输入实现短视频、电影、虚拟数字人等高定制化内容的生成，提升了生成结果的可控性和多样性以及用户体验。

从交互的角度而言，主流方法可划分为多模态交互和物理交互两个研究方向。其中，多模态交互聚焦于对文本、音频及视频等多维信息的深度融合，这些信息既可作为任务的输入，也可作为输出结果。与以内容生成为核心的音视频联合生成任务不同，交互类方法更加强调系统的智能性，要求能够准确理解用户的意图并做出合理响应，从而实现更高水平的人机交互体验。近期主流数字人交互系统，例如 M3-Agent（Long 等，2025）、Qwen2.5-Omni（Xu 等，2025a）、InteractiveOmni（Tong 等，2025）、M.I.O（Cai 等，2025a）、X-Streamer（Xie 等，2025）、ORCA（He 等，2025b）和 FlowAct-R1（Wang 等，2026），普遍采用模块化设计，将整个交互过程分为“思考”与“回复”两个核心环节。其中，“大脑”模块聚焦于理解用户询问及相关语境，“渲染”模块则负责将高维交互信号转化为用户可感知的视频或音频内容，显著优化了响应的自然度和流畅性。在此基础上，M3-Agent 进一步引入记忆模块，赋予系统持久的长时记忆能力，实现上下文连贯且个性化的智能互动。X-Streamer 通过创新性的流式结构，有效支持了无限时长的实时交互需求。M.I.O 则针对人脸与身体动作进行了独立处理，在交互的同时可以实现更加真实拟人的视觉效果。ORCA 通过闭环

的观察-思考-行动-反思循环及分层双系统架构模拟内部世界模型，使虚拟人能够自主推理、验证结果并实时修正错误。FlowAct-R1则通过diffusion forcing训练策略实现了误差更少的长序列交互，并利用蒸馏和系统优化进一步降低了延迟。这些交互方法均以智能为核心，并通过融合长时记忆、实时响应、真实观感等多维功能，进一步增强系统的智能表现。这一技术进步有效推动了数字人从以内容生成为主的模式向以虚拟智能体为核心的模式转变，显著提升了虚拟人系统的交互性与智能水平。

如图11所示，以交互为核心的数字人代表性工作如下。

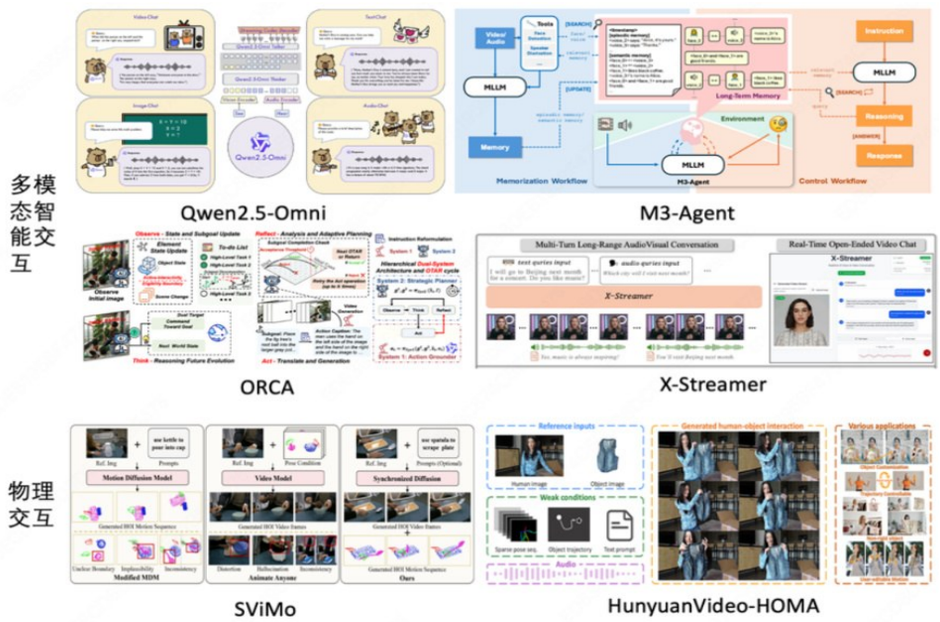

图11　以交互为核心的数字人2025年代表性工作

Fig. 11　Representative works on interaction-centered digital humans in 2025

物理交互则侧重于人与周围物体和环境的实际接触与融合。SViMo（Dang等，2025）将视觉先验与动态约束相结合，实现了在扩散模型框架下同步生成手-物交互的运动信息与视频内容，有效提升了物理合理性与交互真实感。HunyuanVideo-HOMA（Huang等，2025b）提出了基于弱条件约束的多模态驱动方法，通过稀疏且解耦的运动引导策略，达到了对手-物交互过程的灵活可控性。UniMo（Pang等，2025）通过统一的自回归建模框架，将二维人体视频与三维人体运动信息进行整合，探索了二维与三维数据之间的对齐机制。SCAR（Yan等，2025）聚焦于开放环境下手-物交互的泛化能力，通过结构与接触感知的表征增强了生成结果的物理一致性。总体而言，上述方法扩展了传统以运动生成为主的物理交互研究至视频生成领域，进一步强调了多功能融合与生成视频的物理合理性。

展望未来，数字人技术的发展将以多模态融合与智能交互为核心，不断迈向更高的实用性与便捷性。首先，实时性作为交互过程中的关键性能指标，未来的技术演进将力求在保证生成效果的基础上，实现更为高效的实时交互反馈。其次，长序列生成能力也将成为重要的发展方向，有助于提升数字人在复杂场景中的表现力。与此同时，更高精度的控制机制将成为研究重点，为短视频、电影等内容创作带来更加便捷和自动化的技术支持。最后，具备更强认知和推理能力的智能“中枢”将成为数字人技术不可或缺的组成部分，为构建具备自主决策和更高层次交互能力的智能体奠定坚实基础。

八、人类数据成为突破具身智能Scaling Law的重要燃料

高质量机器人本体数据的匮乏已成为制约具身智能Scaling Law生效的最大短板，人类数据不再仅仅是模仿学习的参考，而是跃升为训练具身大脑的重要燃料。人类的操 作空间天然构成了机器人操

作空间的“超集”，人类视频中蕴含的物理常识、因果推理 及交互偏好，正是机器人所缺失的“通用物理直觉”。目前，具身智能领域研究中主要 三大数据组成为人类遥操作数据、手持机械末端（universal manipulation interface，UMI）操作数据和人类视频操作数据。这一年，陆续涌现出基于UMI 和人类操作数据的操作策略学习方法，标志着具身领域逐步从机器人本体数据向人类视频数据中学习转变的趋势。本节将在UMI 数据、人类视频数据、基于人手数据的学习等方面回顾 2025 年度的代表性研究工作。

如图 12 所示，展示了 UMI 与人类操作数据的采集设备。

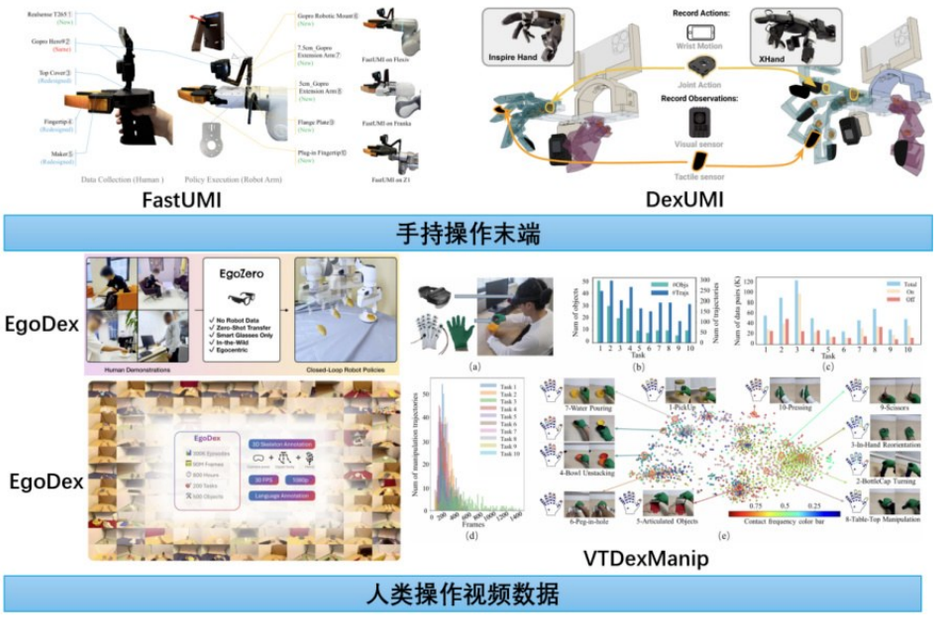

图 12　手持操作末端数据采集设备（UMI）和人类操作数据

Fig. 12　Handheld manipulation end-effector device（UMI）and human operation data

人类遥操作机器人本体执行的操作轨迹是最直接具身学习数据，然而该数据采集方式受本体设备的高昂复杂、采集效率等制约。人手手持末端 UMI 方案在 2025 年受到了更多关注，该方案既保留了人类操作的灵活性，又在物理上天然对齐了机器人本体，极大地降低了全量遥操作的高昂成本与空间消耗。2024 年，UMI（Chi 等，2024）通过手持式夹爪与精心设计的交互接口，实现了便携、低成本、信息丰富的人类操作数据采集。该工作确立了“手持末端”的采集范式。2025 年，FastUMI（Liu 等，2025i；Zhaxizhuoma 等，2025）通过解耦硬件设计并融入广泛机械改进，消除了对专业机器人组件的依赖；DexUMI（Xu 等，2025c）设计了与五指灵巧手同构的手持末端，突破了现有 UMI 的灵巧性瓶颈。这些衍生方案在灵巧度、感知维度与构型适应性上进行了全面升级，而工业界相关工作演示也初步展示了通过大规模 UMI 数据积累进行 scaling 的潜力。Generalist AI 提出的 Gen-0 基础模型则通过大规模扩展这一思路，利用高达 27 万小时的真实世界操作数据进行训练，展现了跨形态、跨场景的强大泛化能力。Sunday Robotics 公司则演示了该类方案对三指夹爪等工业构型的强大适应性。

在人类操作视频方面，采集并解析第一人称视角数据是 2025 年三维视觉在具身领域突破的重要方向。EgoZero（Liu 等，2025k）提出了极具颠覆性的“Smart Glasses Only”方案，证明了仅凭智能眼镜采集的第一人称视频即可训练闭环机器人策略。随着硬件性能提升，EgoDex（Hoque 等，2025）数据集利用 Apple Vision Pro 的空间计算能力，构建了包含全身 3D 姿态及高精度手部追踪的富信息数据。Ego4D（Grauman 等，2025）、EgoLife（Yang 等，2026b）、Epic Kitchen（Damen 等，2018）、100DoH（Shan 等，2020）以及 Motion-X（Lin 等，2024）；（Zhang 等，2025h）等大规模数据集的持续迭代，汇

聚了全球各地的海量第一人称生活视频。针对单纯视频数据缺乏具身操作领域最重要的接触信息这一问题，VTdexManip（Liu 等，2025j）通过自研低成本压阻传感器触觉手套，构建了大规模人类复杂操作的视频和触觉集，并设计了 18 种非预训练和预训练方法进行比较，以研究不同模态和相关策略的有效性。

如图 13 所示，基于人类操作数据的学习范式取得了多项进展。

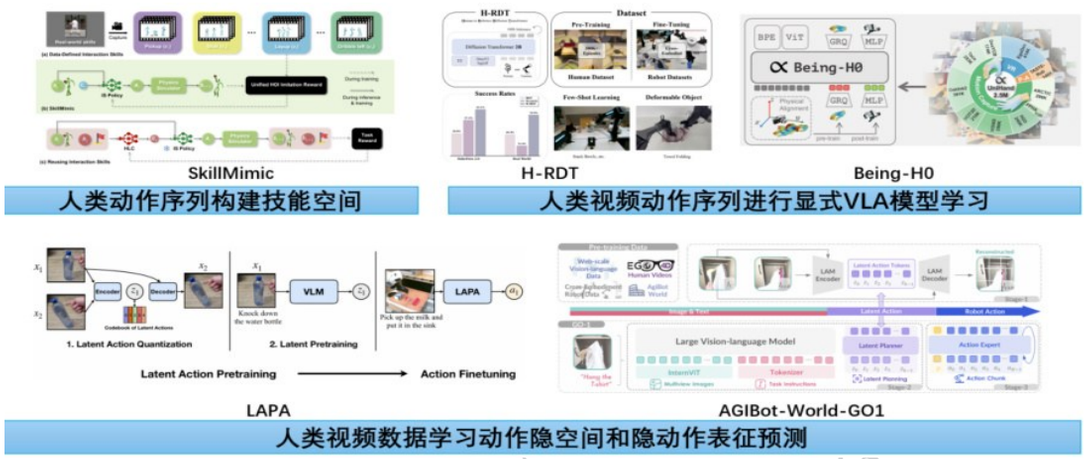

图 13　基于人类操作数据的学习范式进展

Fig. 13　Advances in learning paradigms from human operation data

不同于遥操作和 UMI 数据直接对齐到目标机械手并涵盖机械手本体的物理与动力学信息，人类视频数据缺乏这些对其身本体执行最关键的信息。挖掘人类视频对于具身技能学习的先验是 2025 年的研究热点。

● 人类动作序列构建技能空间：在全身具身人形智能体研究方面，从早期的 Deep-Mimic（Peng 等，2018）到 ASE（Peng 等，2022），研究重心逐渐从单一动作的物理模仿转向构建通用隐式技能空间，使人形智能体具备复用人类基础运动能力的潜能。在此基础上，2025 年的研究进一步聚焦于复杂交互技能迁移，SkillMimic（Wang 等，2025j）与 InterMimic（Xu 等，2026）通过在物理仿真中引入大规模人类演示数据，让智能体习得符合动力学约束的通用人物交互策略，突破了传统模仿学习在物理一致性上的瓶颈。同时，生成式模型正在成为技能数据扩充器，Sitcom-Crafter（Chen 等，2025i）与 Human-X 等工作结合大模型能力，能够生成长序列、情节驱动的复杂交互数据，为具身智能体提供近乎无限的虚拟演练场景。这种高质量合成数据结合 Token-HSI（Pan 等，2025）和 MaskedManipulator（Tessler 等，2025）等多任务联合建模架构，实现了从高层语义理解到全身精细操控的端到端贯通，使得合成人类技能数据同样有潜力作为关键数据支撑，助力智能体在物理世界中实现更通用的灵巧操作和技能迁移。

● 人类视频动作序列进行显式 VLA 模型学习：在人类操作研究方面，从人类视频中直接显式重建人手操作动作并基于该动作序列构建 VLA 模型学习，成为 2025 年从人类视频中学习的主要范式。Being-H0（Luo 等，2025）提出了一种基于显式运动建模的预训练范式，利用部件级动作编码将人类手部作为“基础机械手”，通过物理指令微调实现了从人类视频到机器人灵巧操作的毫米级精度迁移。H-RDT（Bi 等，2025a）验证了大规模人类数据预训练的有效性，证明了在海量人类数据上训练 Diffusion Transformer 骨干网络后再进行跨具身微调，能显著优于仅使用机器人数据的基线。相对于基于 VLA 的预训练后微调方式，构建统一动作空间使得人类视频数据和机器人数据可以进行混合训练。因此，EgoVLA（Yang 等，2025e）提出了“Unified VLA”范式，构建统一动作空间，将人类手部姿态与机器人控制指令映射到同一维度，实现了在人类数据集上的预训练与机器人数据上的微调，大幅提升了跨本体泛化能力。VITRA（Li 等，2025d）通过将灵巧手和人类数据的状态编码到统一空间中进

行本体统一，并在训练过程中将两类数据混合训练。Physical Intelligence 的论文 Emergence of Human to Robot Transfer in VLAs（Kareer 等，2025）也通过数据分析证明了人类数据在充分预训练后能够与具身本体数据做对齐。

● 人类视频数据学习动作隐空间和隐动作表征预测：与上述直接进行显式操作动作提取方式不同，该方向从视频数据学习图像帧间的离散隐动作空间。LAPA（Ye 等，2025d）提出一种从无机器人动作标注的网络规模视频中学习的方法：首先通过基于 VQ-VAE（Van Den Oord 等，2017）目标的动作量化模型学习图像帧间的离散隐动作空间，随后预训练 VLA 模型以根据观测结果和任务描述预测这些隐动作表征，最终在小规模机器人操作数据上对 VLA 模型进行微调，实现从隐动作空间到机器人实际动作的映射。AGIBot-World-GO1（Bu 等，2025）进一步验证了基于隐动作空间预测的方式可以显著提升 VLA 长程任务的能力和泛化性能。

当前，具身智能领域正从依赖有限机器人本体数据加速转向融合多模态人类数据的新范式。UMI 与人类视频数据的规模化应用，以及通过统一动作空间、隐动作表征等方式对数据进行高效对齐与迁移，已初步展现了智能体在复杂长程任务中的泛化与适应能力。展望未来，随着多模态基础模型与物理模拟技术的深度结合，构建从人类经验中提炼“通用物理直觉”的跨本体学习框架将成为关键突破点。这不仅将催生能理解并执行复杂抽象指令的通用机器人，更将推动从“感知-模仿”到类似大语言模型的具身智能涌现本质跨越，最终赋能机器人在开放世界中自主掌握人类级别的操作技能。

九、具身智能基础模型向理解想象执行一体化统一模型演进

围绕“具身智能体如何建立对世界的内在表征”这一核心问题，具身智能基础模型经历了一个从“直接视觉-语言-动作（VLA）模型”向“内在建模”逐步演进的连续过程。传统的视觉-语言-动作（VLA）模型虽然在指令理解上表现出色，但本质上仍属于“基于当前观测的反应式系统”，缺乏对动作后果的物理预测能力，导致在处理长程任务或强动力学交互时表现出短视与脆弱。为了突破这一瓶颈，2025 年的前沿研究呈现出清晰的演进脉络：首先通过“快慢系统”的分层设计，在保留语义泛化的同时引入流匹配（Flow Matching）等生成式策略以提升动作的物理拟合精度；继而引入视频生成模型作为“具身想象器”，利用其内隐的物理规律表征来辅助决策，并解决了从开环想象到闭环控制的实时性难题；最终迈向“理解-想象-执行”的一体化架构，在单一的混合专家（Mixture of Expert，MoT）网络中实现了对过去语义、未来画面与当前动作的联合建模，标志着具身智能体开始具备类似人类的“直觉物理”与“预演规划”能力。

如图 14 所示，具身智能 VLA 与世界模型的融合路径逐步清晰。

● “理解 + 执行”的分层 VLA：针对端到端 VLA 模型在语义广度与动作精度上的权衡难题，新一代模型转向了“系统 1（直觉）+ 系统 2（推理）”的异构分层架构。以 $\pi_{0.5}$（Black 等，2025）为例，其开创性地采用了两阶段训练范式：预训练阶段利用 FAST Tokenizer 将视觉、语言和动作统一为离散 Token，从而能吞吐包括互联网视频和多机器人数据在内的海量异构数据；后训练阶段则切换至流匹配（flow matching）专家模块，专注于将高层语义子任务解码为高频连续动作流，实现了“思维链”式的分层控制。GR00T N1（Bjorck 等，2025）则进一步将这种解耦推向极致，构建了基于 VLM 的语义推理模块与基于扩散 Transformer（diffusion transformer，DiT）的动作生成模块，并通过“数据金字塔”策略，底层利用 VQ-VAE 从人类视频提取潜动作（Latent Actions），顶层使用稀缺真机数据，在保证模型具备通用语义理解的同时，大幅提升了在真实物理环境中的操作鲁棒性。

● 视觉想象引导策略执行：上述“理解 + 执行”的 VLA 模型中的 VLM 基模提供的是静态语义理解能力，而非对物理过程的动态推理能力。模型虽然知道“该抓哪个物体”，却无法预判“抓取后的物理后果”。为了赋予机器人“三思而后行”的前瞻能力，研究者开始利用视频扩散模型作为隐式物理引擎。VPP（Hu 等，2025c）发现视频生成模型的中间层特征隐含了丰富的未来物理动态信息，因此提出直接提取其“预测性视觉表征”来条件化指导动作策略，从而避免了显式生成视频的高昂计算代价。而针对视频预测通常难以用于实时控制的问

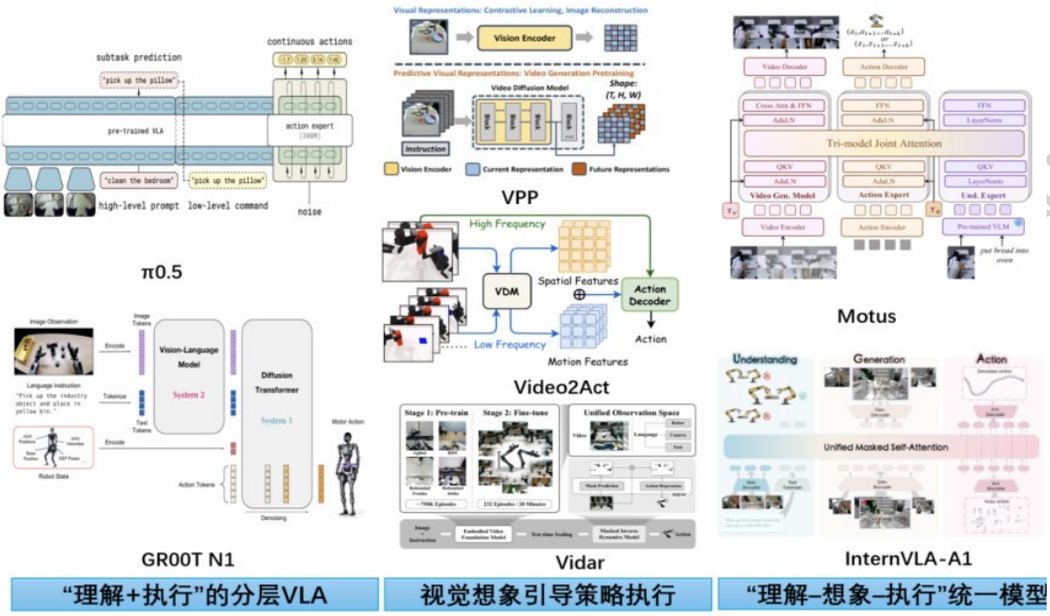

图 14　具身智能 VLA 与世界模型融合演进路线

Fig. 14　Evolution of VLA and world-model integration for embodied intelligence

题，Vidarc（Feng 等，2025b）提出了一种基于 KV Cache 重预填充（Re-prefilling）的自回归推理机制，将实时环境观测反馈注入视频生成过程，实现了从“离线开环想象”到“在线闭环修正”的突破；同时，该模型引入了掩码逆动力学模型（masked inverse dynamics model，Masked IDM）与具身感知损失（embodiment-aware loss），强制模型在生成未来画面时重点关注机械臂与交互对象的物理一致性，显著降低了幻觉对控制精度的影响。

● “理解-想象-执行”统一模型：当前主流视觉-语言-动作模型通常基于多模态大语言模型构建，在语义理解方面展现出卓越能力，但其本质上缺乏对物理世界动态的推理能力。视频预测构建世界模型往往缺乏语义根基，且在处理预测误差时表现出脆弱性。为融合语义理解与动态预测能力，最新演进趋势旨在打破模块间壁垒，构建全链路可微的统一大模型，“理解-想象-执行”高度耦合的统一模型范式在今年逐步涌现。InternVLA-A1（Cai 等，2026）采用了混合专家 Transformer（MoT）架构，通过统一掩码自注意力机制严格定义了从“语义理解专家”到“视觉预见专家”再到“动作执行专家”的信息流向，使得动作生成显式建立在对未来物理后果的预测之上，并在 5.33 亿帧的合成-真实混合数据上验证了其在强动态任务中的优越性。Motus（Bi 等，2025b）则进一步提出了“五位一体”的统一潜动作世界模型，其核心创新在于利用光流提取像素级的“Delta Action”作为通用动作表征，从而能无监督地利用海量互联网视频进行预训练；配合三模态联合注意力机制与 UniDiffuser 调度器，模型不仅能在同一参数空间内灵活切换视频生成、动作预测与语义推理模式，更展现出了对未见物理场景的强大样本泛化能力。

总体而言，VLA 结构的演进呈现出一条清晰的认知层级上移主线：从最初的直接函数逼近（模仿动作），到语义条件控制（理解指令）、显式前瞻预测（预演未来），最终走向统一世界建模。这一过程标志着具身智能正从简单的反应式系统迈向复杂的认知型系统。未来 VLA 范式的核心竞争点，将不再是单一模块的性能堆砌，而是模型内部世界表征的稳定性、可预测性与可控性。当前的统一模型阶段，正是实现这一关键转折的重要里程碑。

十、具身智能的“后训练”时刻：VLA 模型从模仿学习向在线 RL 的范式跃迁

具身智能领域的视觉-语言-动作（VLA）模型正经历着类似于大语言模型后训练阶段的路径演变，即发展重心正从单纯依赖大规模模仿学习，向以在线强化学习为核心的训练范式迁移（Black 等，2025）；（Deng 等，2025c）；（Amin 等，2025）。在早

期阶段，基于多模态预训练与高质量机器人轨迹微调的两阶段监督微调范式占据主导，但随着模型规模扩展，该范式逐渐暴露出数据不可持续与泛化能力受限的内生矛盾。由于高质量人类遥操作数据兼具稀缺性与昂贵性，单纯依赖 SFT 进行规模化面临极高边际成本，且监督学习本质上是对专家演示分布的拟合，导致策略在面对未见场景或长视距任务时往往表现出鲁棒性不足。本节回顾 2025 年在 SFT 基础上通过引入基于环境交互优化机制，提升并突破模仿学习在长时程规划与分布外泛化方面瓶颈的操作策略学习工作。

如图 15 所示，2025 年度在线强化学习的主要方法与挑战概览如下。

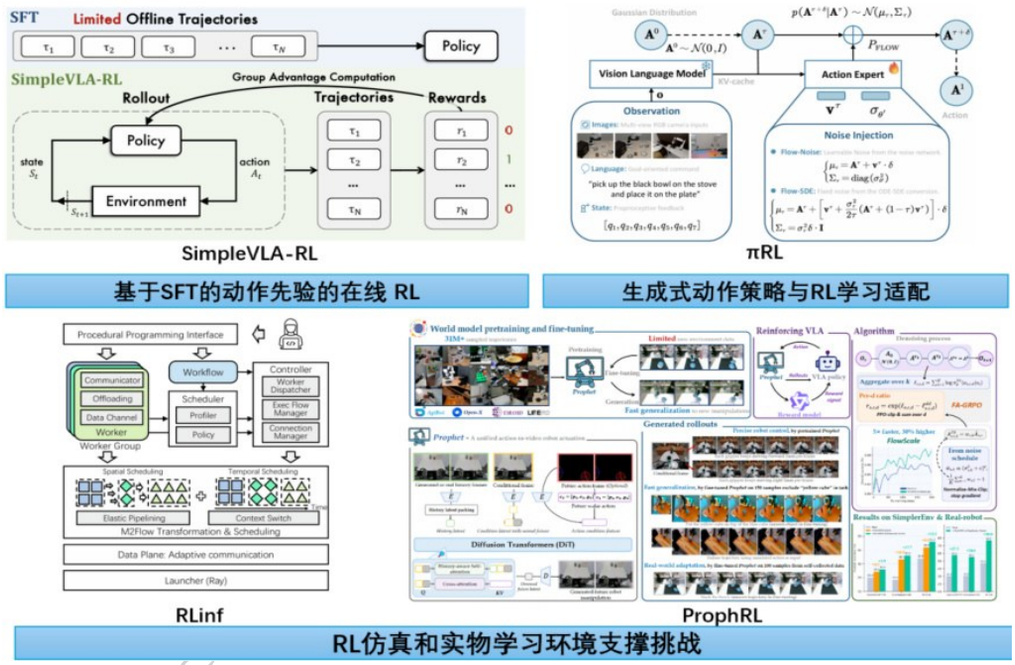

图 15　2025 年度在线强化学习主要方法和挑战

Fig. 15　Major methods and challenges for online reinforcement learning in 2025

● 基于 SFT 动作先验的在线 RL：这一范式核心在于将训练闭环从“拟合专家轨迹”重构为“采样-评估-更新”的强化学习过程。SimpleVLA-RL（Li 等，2025c）中，SFT 角色被重新定义为策略“冷启动”手段，而非能力增长主引擎。实验证明即便在极度稀缺数据设定下（如单条演示轨迹），仅需利用 SFT 提供基础动作先验，后续性能提升即可由在线 RL 完成。通过引入 GRPO 等 On-Policy 算法（Chen 等，2025k）；（Liu 等，2025h）并结合 Outcome-level 稀疏奖励，模型能够在缺乏密集标注情况下通过持续环境交互实现自我进化。这种机制显著缓解了对昂贵专家数据依赖，也促使策略从复刻走向探索，允许模型在奖励信号驱动下搜索更广阔状态-动作空间，并出现如“Pushcut”等涌现行为。通过动态采样（dynamic sampling）策略和提升采样温度，模型得以在更广状态空间探索，显著增强跨任务、跨物体与跨空间泛化能力；相关实证研究也进一步量化了 RL 对 VLA 泛化收益（Liu 等，2025g）。这种从“复刻演示”到“驱动搜索”的转变，确立了在线强化学习作为具身智能能力持续增长核心引擎的地位。

● 生成式动作策略与 RL 学习适配：生成式策略架构演进与 RL 优化接口之间存在显著适配挑战。随着 VLA 动作生成头从自回归架构转向连续空间扩散模型或流匹配，传统 RL 算法在新架构下面临确定性 ODE 采样缺乏探索随机性、以及少量去噪步下难以计算精确对数似然等困境。针对该问题，πRL 等工作（Chen 等，2025j）提出基于随机微分方程（stochastic differential equation，SDE）的改进方案，通过在去噪过程中显式注入噪声并重构概率流，将流匹配过程改造为可计算似然且具备探索能力的马尔可夫决策过程。这种架构层适配使 PPO

等经典 RL 算法能够有效驱动连续动作空间 VLA 模型，在极少监督数据下实现超越全量 SFT 性能。ProphRL（Zhang 等，2025g）针对扩散/流匹配动作头在 RL 场景中的稳定性问题，提出将优势比率与更新目标对齐到“环境动作”粒度，并对去噪内部步梯度贡献进行噪声日程感知重加权，缓解连续生成策略在少步采样下易不稳定、更新效率低等难点。此外，$\pi^{*}0.6$ 的 RECAP 等方法（Zhang 等，2025f）另辟蹊径，利用优势调节（advantage conditioning）规避流匹配模型中复杂策略梯度目标函数，通过在模型输入中注入二值化优势指示符（Indicator），直接在连续动作空间提取更优策略。这些适配方案验证了在线 RL 在不同骨干架构下的普适性，使其在低监督数据条件下依然具备显著优势。

● RL 仿真与实物学习环境支撑：在线强化学习范式高效落地，首要依赖能够处理仿真、采样、奖励计算及策略更新等异构负载的底层系统。以 RLinf 为代表的高性能训练系统（Yu 等，2025e）；（Zang 等，2025），针对 VLA 训练中硬件利用率低和训练缓慢问题，提出了“宏观到微观流转换”（M2Flow）范式。该系统通过将高层逻辑工作流自动分解并重构为优化执行流，实现计算资源在空间（跨加速器分配）和时间（上下文切换）维度的灵活调度。例如在具身任务中，RLinf 可有效应对模拟器（CPU/显存密集）与模型生成（GPU 核心密集）之间资源冲突，通过混合调度将端到端训练吞吐量提升高达 2.43 倍。这种基础设施进步使 VLA“后训练”阶段得以从理论验证走向大规模应用。另一类代表性工作则面向“后训练”交互数据获取与扩展瓶颈：在真实机器人上进行大规模在线探索往往代价高、风险大且难并行，而传统物理仿真又难覆盖复杂视觉与接触动力学。ProphRL（Zhang 等，2025g）通过动作条件世界模型 Prophet 在数据驱动“模型环境”中生成长时序交互推演（rollouts），为 VLA 提供可规模化强化学习训练闭环，从而显著降低后训练对昂贵真机交互与高质量遥操作数据的依赖，并强化长视距任务鲁棒性与泛化能力。

综上所述，VLA 训练范式正在发生更迭，在线强化学习凭借其对数据的低依赖性、对新策略的探索能力以及优越泛化性能，正逐步确立为驱动具身智能能力持续增长的核心范式。与之并行，交互式后训练与推理增强方向也在持续推进（Tan 等，2025）（Ye 等，2025c），共同构成具身智能“后训练时代”的关键技术底座。

结语

本报告系统性梳理与总结了 2025 年度三维视觉领域的关键趋势与十大前沿进展。在核心技术纵深发展与交叉融合的双重驱动下，三维视觉正从一项专项感知和重建技术，演进为构建空间智能与具身智能的核心基础设施。然而，受限于篇幅与报告聚焦点，一些同样至关重要且活跃发展的方向，如三维导航与路径规划、高精度人体运动捕捉与生成、高斯泼溅数据压缩、2D/3D 光影编辑、三维视觉在各领域的应用等，未能在此详尽展开。这些方向作为连接三维感知与最终应用的桥梁，其持续创新对于三维视觉的发展至关重要。由于每个条目的相关工作非常多，许多重要工作也难免遗漏，期待指正。

2025 年的突破性进展，无疑为人工智能突破当前以“Scaling Law”为主导的范式瓶颈注入了崭新而强劲的动力。三维视觉所提供的空间结构与物理理解，正在填补大模型所缺失的“世界常识”。与此同时，伴随“世界模型”、“具身智能”、“空间智能”等前沿概念从学术探讨迅速走向产业风口，三维视觉在人工智能宏大叙事中的战略地位已愈发凸显，不再仅是“视觉”的一个分支，而是通向通用人工智能不可或缺的物理世界编码与交互界面。

展望未来，三维视觉的发展将不再局限于单项技术的精进，而更在于多重技术栈的深度融合与统一范式的确立。这种融合与统一将主要体现在两个层面：

其一，是技术栈的纵向收敛：从分立管线到一体化世界模型。“三维重建（建模）- 三维生成（聚合）- 三维理解（感知）- 视频生成（推演）”的技术链条将加速融合，边界日益模糊。前馈重建将提供实时、鲁棒的环境几何基座；三维生成将聚合场景语义实例对象；视频生成技术将赋予系统对不可见区域的合理想象与对未来状态的物理推演能力。这些技术的深度耦合，正驱动着从静态 3D 模型到动态 4D 场景，再到具备物理感知与实时交互能力的“世界模型”的演进。

其二，是应用层的横向贯通：从孤立场景到泛在智能基座。三维视觉的能力将深度嵌入并赋能千

行百业。在具身智能领域，机器人的“大脑”（VLA模型）将不再仅仅依赖二维图像特征，而是能够直接基于场景的 3D/4D 几何与语义表征进行思考与规划。这些丰富的场景表征将通过统一的特征空间，与语言指令、视觉目标、触觉反馈乃至动作原型进行深度对齐与融合，从而极大提升复杂长程操作任务的泛化性、精确性与物理合理性。

展望 2026，我们正站在一个范式转变的转折点上：三维视觉不再仅仅是“重建”或“生成”一个数字副本，而是致力于“理解”并“构筑”一个可与智能体交互的、符合物理规律的动态数字世界。这正如杨立昆所倡导的“世界模型”、李飞飞所论述的“空间智能”以及德米斯·哈萨比斯所展望的“物理交互 AI”那样，三维视觉将成为实现下一代人工智能——即能真正理解我们所在物理世界的智能——最为关键的基石之一。

参考文献

Assran M, Bardes A, Fan D, Garrido Q, Howes R, Muckley M, et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985.

Bahmani S, Shen T, Ren J, Huang J, Jiang Y, Turki H, et al. 2025. Lyra: Generative 3D Scene Reconstruction via Video Diffusion Model Self-Distillation. arXiv preprint arXiv:2509.19296.

Bai J, Xia M, Fu X, Wang X, Mu L, Cao J, et al. 2025. Recammaster: Camera-controlled generative rendering from a single video// Proceedings of the IEEE/CVF International Conference on Computer Vision: 14834-14844.

Bai J, Xia M, Wang X, Yuan Z, Liu Z, Hu H,
et al. N.d. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints//The Thirteenth International Conference on Learning Representations.

Bao C, Zhang X, Yu Z, Shi J, Zhang G, Peng S, et al. 2025. Free360: Layered Gaussian Splatting for Unbounded 360-Degree View Synthesis from Extremely Sparse and Unposed Views//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 16377-16387.

Bar A, Zhou G, Tran D, Darrell T, and LeCun Y. 2025. Navigation world models//Proceedings of the Computer Vision and Pattern Recognition Conference: 15791-15801.

Bi H, Wu L, Lin T, Tan H, Su Z, Su H, et al. 2025. H-RDT: Human Manipulation Enhanced Bimanual Robotic Manipulation. arXiv preprint arXiv:2507.23523.

Bi H, Tan H, Xie S, Wang Z, Huang S, Liu H, et al. 2025. Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030.

Bian W, Huang Z, Shi X, Li Y, Wang F Y, and Li H. 2025. GS-DiT: Advancing Video Generation with Dynamic 3D Gaussian Fields through Efficient Dense 3D Point Tracking//Proceedings of the Computer Vision and Pattern Recognition Conference: 21717-21727.

Black K, Brown N, Darpinian J, Dhabalia K, Driess D, Esmail A, et al. 2025. π0.5: a Vision-Language-Action Model with Open-World Generalization//9th Annual Conference on Robot Learning.

Bu Q, Cai J, Chen L, et al. 2025. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Cai J, Cai Z, Cao J, Chen Y, He Z, Jiang L, et al. 2026. InternVLA-A1: Unifying Understanding, Generation and Action for Robotic Manipulation. arXiv preprint arXiv:2601.02456.

Cai Y, Chu X, Gao X, Gong S, Huang Y, Kang C, et al. 2025. Towards Interactive Intelligence for Digital Humans. arXiv preprint arXiv:2512.13674.

Cai Z, Li Z, Li X, Li B, Wang Z, Zhang Z, et al. 2025. UP2You: Fast Reconstruction of Yourself from Unconstrained Photo Collections. arXiv preprint arXiv:2509.24817.

Cao C, Zhou J, Li S, Liang J, Yu C, Wang F, et al. 2025. Uni3C: Unifying Precisely 3D-Enhanced Camera and Human Motion Controls for Video Generation//Proceedings of the SIGGRAPH Asia 2025 Conference Papers: 1-12[DOI: 10.1145/3757377.3763842].

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. 2021. Emerging Properties in Self-Supervised Vision Transformers// 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. IEEE: 9630-9640[DOI: 10.1109/ICCV48922.2021.00951].

Cen J, Yu C, Yuan H, Jiang Y, Huang S, Guo J, et al. 2025. World-VLA: Towards Autoregressive Action World Model. arXiv preprint arXiv:2506.21539.

Chang J, Ye C, Wu Y, Chen Y, Zhang Y, Luo Z, et al. 2025. ReconViaGen: Towards Accurate Multi-view 3D Object Reconstruction via Generation. arXiv preprint arXiv:2510.23306.

Chefer H, Singer U, Zohar A, Kirstain Y, Polyak A, Taigman Y,
et al. N. d. VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models//Forty-second International Conference on Machine Learning.

Chen C, Zou Z, Cao Y, Yuan Z, Luo G, Qi X, et al. 2025. LaFiTe: A Generative Latent Field for 3D Native Texturing. arXiv preprint arXiv:2512.04786.

Chen J, Hu P, Chang X, Shi Z, Kampffmeyer M, and Liang X. 2025. Sitcom-Crafter: A Plot-Driven Human Motion Generation System in 3D Scenes. arXiv preprint arXiv:2410.10790.

Chen J, Hu J, Wang G, Jiang Z, Zhou T, Chen Z, et al. 2025. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting//Proceedings of the Computer Vision and Pattern Recognition Conference: 10723-10734.

Chen J, Zhu H, He X, Wang Y, Zhou J, Chang W, et al. 2025. Deepverse: 4d autoregressive video generation as a world model. arXiv preprint arXiv:2506.01103.

Chen K, Liu Z, Zhang T, Guo Z, Xu S, Lin H, et al. 2025. πRL: Online r1 fine-tuning for flow-based vision-language-action models. arXiv preprint arXiv:2510.25889.

Chen S, He P, Hu J, Liu Z, Wang Y, Xu T, et al. 2025. Astra: Toward general-purpose mobile robots via hierarchical multimodal learning. arXiv preprint arXiv:2506.06205.

Chen X, Chen Y, Xiu Y, Geiger A, and Chen A. 2025a. Easi3R: Estimating Disentangled Motion from DUSt3R Without Training [EB/OL]. [DOI: 10.48550/ARXIV.2503.24391].[2026-03-18]. https://arxiv.org/pdf/2503.24391.pdf.

Chen X, Chen Y, Xiu Y, Geiger A, and Chen A. 2025b. TTT3R: 3D Reconstruction as Test-Time Training [EB/OL]. [DOI: 10.48550/ARXIV.2509.26645].[2026-03-18]. https://arxiv.org/pdf/2509.26645.pdf.

Chen Y, Chen X, Xue Y, Chen A, Xiu Y, and Pons-Moll G. 2025. Human3R: Everyone Everywhere All at Once [EB/OL]. [DOI: 10.48550/ARXIV.2510.06219].[2026-03-18]. http://arxiv.org/pdf/2510.06219.pdf.

Chen Y, Guo S, Yang T, Ding L, Yu X, Gu J, et al. 2025. 4dslomo: 4d reconstruction for high speed scene with asynchronous capture// ACM SIGGRAPH Asia. 1-11.

Chen Z, Niu R, Kong H, Wang Q, Xing Q, and Fan Z. 2025. Tgrpo: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. arXiv preprint arXiv:2506.08440.

Chen Z, Liu T, Zhuo L, Ren J, Tao Z, Zhu H, et al. 2025. 4DNeX: Feed-Forward 4D Generative Modeling Made Easy. arXiv preprint arXiv:2508.13154.

Chi C, Xu Z, Pan C, Cousineau E, Burchfiel B, Feng S, et al. 2024. Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots. arXiv preprint arXiv:2402.10329.

Dai P, Zhang P, Dong Z, Xu K, Peng Y, Ding D, et al. 2025. 4d gaussian videos with motion layering. ACM Trans. on Graphics: 1-14.

Dai Q, Ni X, Shen Q, Chen W, Chen B, and Chu M. 2025. Rainygs: Efficient rain synthesis with physically-based gaussian splatting// Proceedings of the Computer Vision and Pattern Recognition Conference: 16153-16162.

Damen D, Doughty H, Farinella G M, Fidler S, Furnari A, Kazakos E, et al. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset//Proceedings of the European Conference on Computer Vision: 720-736.

Dang L, Shao R, Zhang H, Min W, Liu Y, and Wu Q. 2025. Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. arXiv preprint arXiv:2506.02444.

Deng C, Zhu D, Li K, Gou C, Li F, Wang Z, et al. 2025. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv: 2505.14683.

Deng H, Wu Z, Liu H, Guo W, Xue Y, Shan Z, et al. 2025. A Survey on Reinforcement Learning of Vision-Language-Action Models for Robotic Manipulation. Authorea Preprints.

Deng K, Ti Z, Xu J, Yang J, and Xie J. 2025. VGGT-Long: Chunk it, Loop it, Align it - Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences [EB/OL]. [DOI: 10.48550/ARXIV.2507.16443].[2026-03-18]. https://arxiv.org/pdf/2507.16443.pdf.

Duan Y, Wei F, Dai Q, He Y, Chen W, and Chen B. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes//ACM SIGGRAPH 2024: 1-11. FanZ, ZhangJ, LiR, et al. 2025. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction. arXiv preprint arXiv: 2505.20279.

Fedele E, Engelmann F, Huang I, Litany O, Pollefeys M, and Guibas L. 2025. SpaceControl: Introducing Test-Time Spatial Control to 3D Generative Modeling. arXiv preprint arXiv:2512.05343.

Feng H, Zhang J, Wang Q, Ye Y, Yu P, Black M J, et al. 2025. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World [EB/OL]. [DOI: 10.48550/ARXIV.2504.13152].[2026-03-18]. https://arxiv.org/pdf/2504.13152.pdf.

Feng Y, Xiang C, Mao X, Tan H, Zhang Z, Huang S, et al. 2025. Vidarc: Embodied Video Diffusion Model for Closed-loop Control. arXiv preprint arXiv:2512.17661.

Go H, Narnhofer D, Bhat G, Truong P, Tombari F, and Schindler K. 2025. VIST3A: Text-to-3D by Stitching a Multi-view Reconstruction Network to a Video Generator. arXiv preprint arXiv: 2510.13454.

Grauman K, Westbury A, Byrne E, Cartillier V, Chavis Z, Furnari A, et al. 2025. Ego4D: Around the World in 3,600 Hours of Egocentric Video. IEEE Transactions on Pattern Analysis and Machine Intelligence:9468-9509.

Gu Z, Yan R, Lu J, Li P, Dou Z, Si C, et al. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control//Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers: 1-12[DOI: 10.1145/3721238.3730607].

Hanson A, Tu A, Lin G, Singla V, Zwicker M, and Goldstein T. 2025. Speedy-Splat: Fast 3D Gaussian Splatting with Sparse Pixels and Sparse Primitives//IEEE/CVF Conference on Computer Vision and Pattern Recognition: 21537-21546 [DOI: 10.1109/CVPR52734.2025.02006].

He H, Zhang Y, Lin L, Xu Z, and Pan L. 2025. Pre-trained video generative models as world simulators. arXiv preprint arXiv: 2502.07825.

He X, Yang T, Cao K, Wu R, Meng C, Zhang Y, et al. 2025. Active Intelligence in Video Avatars via Closed-loop World Modeling.

arXiv preprint arXiv:2512.20615.

Held J, Son S, Vandeghen R, Rebain D, Gadella M, Zhou Y, et al. 2025. Meshsplatting: Differentiable rendering with opaque meshes. arXiv preprint arXiv:2512.06818.

Hoque R, Huang P, Yoon D J, Sivapurapu M, and Zhang J. 2025. EgoDex: Learning Dexterous Manipulation from Large-Scale Egocentric Video. arXiv preprint arXiv:2505.11709.

Hu Y, Yang Y, Lin H, Wang Y, Dong J, Deng Y, et al. 2025. Split4d: Decomposed 4d scene reconstruction without video segmentation. ACM Trans. on Graphics: 1-15.

Hu Y, Cheng C, Yu S, Guo X, and Wang H. 2025. VGGT4D: Mining Motion Cues in Visual Geometry Transformers for 4D Scene Reconstruction [EB/OL]. [DOI: 10.48550/ARXIV.2511.19971].[2026-03-18].
https://arxiv.org/pdf/2511.19971.pdf.

Hu Y, Guo Y, Wang P, Chen X, Wang Y J, Zhang J, et al. 2024. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. arXiv preprint arXiv:2412.14803.

Huang B, Duan H, Zhao Y, Zhao Z, Ma Y, and Gao S. 2025. CUPID: Generative 3D Reconstruction via Joint Object and Pose Modeling. arXiv preprint arXiv:2510.20776.

Huang J, Yang Y, Yang B, Ma L, Ma Y, and Liao Y. 2026. Gen3R: 3D Scene Generation Meets Feed-Forward Reconstruction. arXiv preprint arXiv:2601.04090.

Huang Z, Guo Y, An X, Yang Y, Li Y, Zou Z, et al. 2025. MIDI: Multi-Instance Diffusion for Single Image to 3D Scene Generation// Proceedings of the Computer Vision and Pattern Recognition Conference: 23646-23657.

Huang Z, Zhou Z, Cao J, Ma Y, Chen Y, Rao Z, et al. 2025. Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. arXiv preprint arXiv:2506.08797.

Intelligence P, Amin A, Aniceto R, Balakrishna A, Black K, Conley K, et al. 2025. $\pi_{0.6}$: a VLA That Learns From Experience. arXiv preprint arXiv:2511.14759.

Ji X, Weiss S, Kansy M, Naruniec J, Cao X, Solenthaler B, et al. 2026. FastGHA: Generalized Few-Shot 3D Gaussian Head Avatars with Real-Time Animation. arXiv preprint arXiv:2601.13837.

Jia M, Qi Z, Zhang S, Zhang W, Yu X, He J, et al. 2026. OmniSpatial: Towards Comprehensive Spatial Reasoning Benchmark for Vision Language Models//The Fourteenth International Conference on Learning Representations.

Jiang H, Tan H, Wang P, Jin H, Zhao Y, Bi S, et al. 2025. RayZer: A Self-supervised Large View Synthesis Model [EB/OL]. [DOI: 10.48550/ARXIV.2505.00702].[2026-03-18].
https://arxiv.org/pdf/2505.00702.pdf.

Jiang J, Zeng W, Zheng Z, Yang J, Liang C, Liao W, et al. 2025. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation. arXiv preprint arXiv:2508.19209.

Jiang L, Mao Y, Xu L, Lu T, Ren K, Jin Y, et al. 2025. AnySplat: Feed-forward 3D Gaussian Splatting from Unconstrained Views. ACM Trans. Graph.: 257:1-257:16[DOI: 10.1145/3763326].

Jiang Z, Zheng C, Laina I, Larlus D, and Vedaldi A. 2025. Geo4D: Leveraging Video Generators for Geometric 4D Scene Reconstruction//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV): 20658-20671.

Jiang Z, Zheng C, Laina I, Larlus D, and Vedaldi A. 2026. Mesh4D: 4D Mesh Reconstruction and Tracking from Monocular Video. arXiv preprint arXiv:2601.05251.

Jin L, Tucker R, Li Z, Fouhey D, Snavely N, and Holynski A. 2025. Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 10497-10509[DOI: 10.1109/CVPR52734.2025.00982].

Jin Y, Peng S, Wang X, Xie T, Xu Z, Yang Y, et al. 2025. Diffuman4D: 4D Consistent Human View Synthesis from Sparse-View Videos with Spatio-Temporal Diffusion Models //International Conference on Computer Vision (ICCV): 11047-11057.

Kareer S, Pertsch K, Darpinian J, Hoffman J, Xu D, Levine S, et al. 2025. Emergence of Human to Robot Transfer in Vision-Language-Action Models. arXiv preprint arXiv:2512.22414.

Karhade J, Keetha N V, Zhang Y, Gupta T, Sharma A, Scherer S A, et al. 2025. Any4D: Unified Feed-Forward Metric 4D Reconstruction [EB/OL]. [DOI: 10.48550/ARXIV. 2512.10935]. [2026-03-18].
https://arxiv.org/pdf/2512.10935.pdf.

Keetha N V, Müller N, Schönberger J, Porzi L, Zhang Y, Fischer T, et al. 2025. MapAnything: Universal Feed-Forward Metric 3D Reconstruction [EB/OL]. [DOI: 10.48550/ARXIV. 2509.13414]. [2026-03-18].
https://arxiv.org/pdf/2509.13414.pdf.

Kerb B, Kopanas G, Leimkühler T, and Drettakis G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph.: 139:1-139:14[DOI: 10.1145/3592433].

Kirschstein T, Giebenhain S, and Nießner M. 2025. FlexAvatar: Learning Complete 3D Head Avatars with Partial Supervision. arXiv preprint arXiv:2512.15599.

Kirschstein T, Romero J, Sevastopolsky A, Nießner M, and Saito S. 2025. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars//Proceedings of the IEEE/CVF International Conference on Computer Vision: 12089-12100.

Lai Z, Zhao Y, Zhao Z, Liu H, Lin Q, Huang J, et al. 2025. Lattice: Democratize High-Fidelity 3D Generation at Scale. arXiv preprint arXiv:2512.03052.

Lai Z, Zhao Y, Zhao Z, Yang X, Huang X, Huang J, et al. 2025. NaTex: Seamless Texture Generation as Latent Color Diffusion. arXiv preprint arXiv:2511.16317.

Lan L, Shao T, Lu Z, Zhang Y, Jiang C, and Yang Y. 2025. 3DG52:

Near Second-order Converging 3D Gaussian Splatting//SIGGRAPH Conference Papers '25: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference. Association for Computing Machinery, [DOI: 10.1145/3721238.3730687].

Lan Y, Luo Y, Hong F, Zhou S, Chen H, Lyu Z, et al. 2025. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer [EB/OL]. [DOI: 10.48550/ARXIV.2508.10893]. [2026-03-18]. https://arxiv.org/pdf/2508.10893.pdf.

Lei J, Weng Y, Harley A W, Guibas L, and Daniilidis K. 2025. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds//IEEE/CVF Conference on Computer Vision and Pattern Recognition: 6165-6177.

Li H, Zuo Y, Yu J, Zhang Y, Yang Z, Zhang K, et al. 2025. Simplevla-rl: Scaling vla training via reinforcement learning. arXiv preprint arXiv:2509.09674.

Li L, Li Y, Weng Y, Zheng Y, and Zhou K. 2025. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars// Proceedings of the Computer Vision and Pattern Recognition Conference: 10747-10757.

Li M, Li P, Zhang Z, Lu J, Zhao C, Xue W, et al. 2026. UniSH: Unifying Scene and Human Reconstruction in a Feed-Forward Pass [EB/OL]. [DOI: 10.48550/ARXIV.2601.01222]. [2026-03-18]. https://arxiv.org/pdf/2601.01222.pdf.

Li Q, Deng Y, Liang Y, Luo L, Zhou L, Yao C, et al. 2025. VITRA: Scalable Vision-Language-Action Model Pretraining for Robotic Manipulation with Real-life Human Activity Videos. arXiv preprint arXiv:2510.21571.

Li R, Torr P, Vedaldi A, and Jakab T. 2025. Vmem: Consistent interactive video scene generation with surfel-indexed view memory// Proceedings of the IEEE/CVF International Conference on Computer Vision: 25690-25699.

Li Z, Chen Z, Li Z, and Xu Y. 2024. Spacetime gaussian feature splatting for real-time dynamic view synthesis//IEEE/CVF Conference on Computer Vision and Pattern Recognition: 8508-8520.

Li Z, Wang Y, Zheng H, Luo Y, and Wen B. 2025. Sparc3D: Sparse Representation and Construction for High-Resolution 3D Shapes Modeling. arXiv preprint arXiv:2505.14521.

Liang H, Ren J, Mirzaei A, Torralba A, Liu Z, Gilitschenski I, et al. 2024. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. arXiv preprint arXiv:2412.03526.

Liang H, Ge Z, Majee S, Tiwari A, Godaliyadda G, Veeraraghavan A, et al. 2025. FastA*vatar*: Instant 3D Gaussian Splatting for Faces from Single Unconstrained Poses. arXiv preprint arXiv:2508.18389.

Liang W, Yu L, Luo L, Iyer S, Dong N, Zhou C, et al. 2025. Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models. Transactions on Machine Learning Research.

Liang Y, Luo K, Chen X, Chen R, Yan H, Li W, et al. 2025. UniTex: Universal High Fidelity Generative Texturing for 3D Shapes. arXiv preprint arXiv:2505.23253.

Liao Z, Zhang J, Tu H, Wang Z, Gao Y, Zhang H, et al. 2026. SharpTimeGS: Sharp and Stable Dynamic Gaussian Splatting via Lifespan Modulation. arXiv preprint arXiv:2602.02989.

Lin C, Lin Y, Pan P, Yu Y, Hu T, Yan H, et al. 2025. Movies: Motion-aware 4d dynamic view synthesis in one second. arXiv preprint arXiv:2507.10065.

Lin H, Chen S, Liew J, Chen D Y, Li Z, Shi G, et al. 2025. Depth Anything 3: Recovering the Visual Space from Any Views [EB/OL]. [DOI: 10.48550/ARXIV.2511.10647]. [2026-03-18]. https://arxiv.org/pdf/2511.10647.pdf.

Lin J, Zeng A, Lu S, Cai Y, Zhang R, Wang H, et al. 2024. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. arXiv preprint arXiv:2307.00818.

Lin Y, Lin C, Pan P, Yan H, Feng Y, Mu Y, et al. 2025. PartCrafter: Structured 3D Mesh Generation via Compositional Latent Diffusion Transformers. arXiv preprint arXiv:2506.05573.

Liu F, Sun W, Wang H, Wang Y, Sun H, Ye J, et al. 2024. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. arXiv preprint arXiv:2408.16767.

Liu I C, Xu Z, Wang Y, Tan H, Xu Z, Wang X, et al. 2025. RigAnything: Template-Free Autoregressive Rigging for Diverse 3D Assets. ACM Transactions on Graphics (TOG): 1-12.

Liu J, Liu G, Liang J, Li Y, Liu J, Wang X, et al. 2025. Flow-GRPO: Training Flow Matching Models via Online Reinforcement Learning. arXiv preprint arXiv:2505.05470.

Liu J, Gao F, Wei B, Chen X, Liao Q, Wu Y, et al. 2025. What can r1 bring to vla generalization? an empirical study. arXiv preprint arXiv:2505.19789.

Liu K, Jia Z, Li Y, Zhaxizhuoma, Chen P, Liu S, et al. 2025. FastUMI-100K: Advancing Data-driven Robotic Manipulation with a Large-scale UMI-style Dataset. arXiv preprint arXiv:2510.08022.

Liu Q, Cui Y, Sun Z, Li G, Chen J, and Ye Q. 2025. VTDexManip: A Dataset and Benchmark for Visual-tactile Pretraining and Dexterous Manipulation with Reinforcement Learning//International Conference on Learning Representations.

Liu Q, He J, Yu Q, Chen L C, and Yuille A. 2025. ReVision: Refining Video Diffusion with Explicit 3D Motion Modeling. arXiv preprint arXiv:2504.21855.

Liu R, Sun D, Chen M, Wang Y, and Feng A. 2025. Deformable Beta Splatting// Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference. New York, NY, USA: Association for Computing Machinery, [DOI: 10.1145/3721238.3730716].

Liu V, Adeniji A, Zhan H, Haldar S, Bhirangi R, Abbeel P, et al. 2025. EgoZero: Robot Learning from Smart Glasses. arXiv preprint arXiv:2505.20290.

Liu X, Xiao Y, Chen D Y, Feng J, Tai Y, Tang C, et al. 2025. Trace Anything: Representing Any Video in 4D via Trajectory Fields [EB/OL]. [DOI: 10.48550/ARXIV.2510.13802]. [2026-03-18]. https://arxiv.org/pdf/2510.13802.pdf.

Liu Y, Su H, Gao L, Li Y, Wang H, Liao Y, et al. 2025. Research Trends and Major Developments in 3D Vision in 2024. Journal of Image and Graphics: 1717-1743[DOI: 10.11834/jig.250057].

Liu Y, Dong D, Wu S, Zhang Z, Hu Y, Zhang L, et al. 2025. Spatial-CoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning. arXiv preprint arXiv:2501.10074.

Liu Y, Dong S, Wang S, Yin Y, Yang Y, Fan Q, et al. 2025. SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 16651-16662 [DOI: 10.1109/CVPR52734.2025.01552].

Liu Z, Lin J, Wu W, and Zhou B. 2025. Joint Optimization for 4D Human-Scene Reconstruction in the Wild [EB/OL]. [DOI: 10.48550/ARXIV.2501.02158]. [2026-03-18]. https://arxiv.org/pdf/2501.02158.pdf.

Long L, He Y, Ye W, Pan Y, Lin Y, Li H, et al. 2025. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. arXiv preprint arXiv:2508.09736.

Low C, Wang W, and Katyal C. 2025. Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284.

Lu G, Jia B, Li P, Chen Y, Wang Z, Tang Y, et al. 2025. Gwm: Towards scalable gaussian world models for robotic manipulation//Proceedings of the IEEE/CVF International Conference on Computer Vision: 9263-9274.

Lu J, Huang T, Li P, Dou Z, Lin C, Cui Z, et al. 2025. Align3R: Aligned Monocular Depth Estimation for Dynamic Videos//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 22820-22830 [DOI: 10.1109/CVPR52734.2025.02125].

Lu J, Xiong W, Deng J, Li P, Huang T, Dou Z, et al. 2025. TrackingWorld: World-centric Monocular 3D Tracking of Almost All Pixels [EB/OL]. [DOI: 10.48550/ARXIV.2512.08358]. [2026-03-18]. https://arxiv.org/pdf/2512.08358.pdf.

Lu Y, Zhang J, Fang T, Nahmias J D, Tsin Y, Quan L, et al. 2025. Matrix3D: Large Photogrammetry Model All-in-One//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 11250-11263.

Luo H, Feng Y, Zhang W, Zheng S, Wang Y, Yuan H, et al. 2025. Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos. arXiv preprint arXiv:2507.15597.

Maggio D, Lim H, and Carlone L. 2025. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold [EB/OL]. [DOI: 10.48550/ARXIV.2505.12549]. [2026-03-18]. https://arxiv.org/pdf/2505.12549.pdf.

Mao X, Lin S, Li Z, Li C, Peng W, He T, et al. 2025. Yume: An interactive world generation model. arXiv preprint arXiv:2507.17744.

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM: 99-106[DOI: 10.1145/3503250].

Murai R, Dexheimer E, and Davison A J. 2025. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 16695-16705 [DOI: 10.1109/CVPR52734.2025.01556]. NvidiaJ B, CastanedaF, CherniadevN, DaX, DingR, FanL, et al. 2025. Gr00t n1: An open foundation model for generalist humanoid robots.

Ouyang K, Liu Y, Wu H, Liu Y, Zhou H, Zhou J, et al. 2025. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning. arXiv preprint arXiv:2504.01805.

Pan L, Yang Z, Dou Z, Wang W, Huang B, Dai B, et al. 2025. TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization. arXiv preprint arXiv:2503.19901.

Pang Y, Zhang Y, Shao R, Deng X, Gao F, Xiaoming X, et al. 2025. UniMo: Unifying 2D Video and 3D Human Motion with an Autoregressive Framework. arXiv preprint arXiv:2512.03918.

Park J, Bui M Q V, Bello J L G, Moon J, Oh J, and Kim M. 2025. Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video//IEEE/CVF Conference on Computer Vision and Pattern Recognition: 26866-26875.

Peng C, Su Z, Wang L, Guo C, Li Z, Long C, et al. 2025. FlexAvATAR: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation. arXiv preprint arXiv: 2512.17717.

Peng X B, Abbeel P, Levine S, and Van de Panne M. 2018. DeepMimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. ACM Transactions on Graphics: 1-14.

Peng X B, Guo Y, Halper L, Levine S, and Fidler S. 2022. ASE: Large-scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. ACM Transactions on Graphics: 1-17.

Peng Z, Zhou K, and Shao T. 2025. Gaussian-plus-SDF SLAM: High-fidelity 3D reconstruction at 150+ fps. Comput. Vis. Media.

Pun A, Deng K, Liu R, Ramanan D, Liu C, and Zhu J. 2025. Generating Physically Stable and Buildable Brick Structures from Text//Proceedings of the IEEE/CVF International Conference on Computer Vision: 14798-14809.

Qiu L, Gu X, Li P, Zuo Q, Shen W, Zhang J, et al. 2025. LHM: Large Animatable Human Reconstruction Model for Single Image to 3D in Seconds//Proceedings of the IEEE/CVF International Conference on Computer Vision: 14184-14194.

Rockwell C, Tung J, Lin T, Liu M, Fouhey D F, and Lin C. 2025. Dynamic Camera Poses and Where to Find Them//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 12444-12455 [DOI: 10.1109/CVPR52734.2025.01161].

Shan D, Geng Z, Rockwell C, and Fouhey D F. 2020. Understanding Human Hands in Contact at Internet Scale//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9736-9745.

Shen G, Deng T, Qin X, Wang N, Wang J, Wang Y, et al. 2025. MUT3R: Motion-aware Updating Transformer for Dynamic 3D Reconstruction [EB/OL]. [DOI: 10.48550/ARXIV.2512.03939]. [2026-03-18]. https://arxiv.org/pdf/2512.03939.pdf.

Shen Q, Tao N, Dai Q, Chen T, Qin M, Zhang Y, et al. N.d. FieryGS: In-the-Wild Fire Synthesis with Physics-Integrated Gaussian Splatting//The Fourteenth International Conference on Learning Representations.

Svitov D, Morerio P, Agapito L, and Del Bue A. 2025. Billboard splatting (bpsplat): Learnable textured primitives for novel view synthesis//IEEE/CVF International Conference on Computer Vision: 25029-25039.

Tan S, Dou K, Zhao Y, and Krähenbühl P. 2025. Interactive post-training for vision-language-action models. arXiv preprint arXiv: 2505.17016.

Tang J, Lu R, Li Z, Hao Z, Li X, Wei F, et al. 2025. Efficient Part-Level 3D Object Generation via Dual Volume Packing. arXiv preprint arXiv:2506.09980.

Team S 3, ChenX, ChuF J, GleizeP, LiangK J, SaxA, et al. 2025. SAM 3D: 3Dfy Anything in Images. arXiv preprint arXiv: 2511.16624.

Tencent Hunyuan3D Team. 2025. Hunyuan3D-Omni: A Unified Framework for Controllable Generation of 3D Assets. arXiv preprint arXiv:2509.21245.

Tessler C, Jiang Y, Coumans E, Luo Z, Chechik G, and Peng X B. 2025. MaskedManipulator: Versatile Whole-Body Manipulation. arXiv preprint arXiv:2505.19086.

Tong W, Guo H, Ran D, Chen J, Lu J, Wang K, et al. 2025. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. arXiv preprint arXiv:2510.13747.

Van Den Oord A and Vinyals O. 2017. Neural Discrete Representation Learning//Advances in Neural Information Processing Systems: vol. 30.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is All you Need//Guyon I, von Luxburg U, Bengio S, Wallach H M, Fergus R, Vishwanathan S V N, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA: 5998-6008.

Wang C, Chen C, Huang Y, Dou Z, Liu Y, Gu J, et al. 2025. Physctrl: Generative physics for controllable and physics-grounded video generation. arXiv preprint arXiv:2509.20358.

Wang J, Ye L, Lu T, Xiao J, Zhang J, Guo Y, et al. 2025. Evoworld: Evolving panoramic world generation with explicit 3d memory. arXiv preprint arXiv:2510.01183.

Wang J, Yuan Y, Zheng R, Lin Y, Gao J, Chen L, et al. 2025. SpatialVID: A Large-Scale Video Dataset with Spatial Annotations [EB/OL]. [DOI: 10.48550/ARXIV.2509.09676].[2026-03-18]. https://arxiv.org/pdf/2509.09676.pdf.

Wang J, Chen M, Karaev N, Vedaldi A, Rupprecht C, and Novotný D. 2025. VGGT: Visual Geometry Grounded Transformer//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 5294-5306[DOI: 10.1109/CVPR52734.2025.00499].

Wang Q, Ye V, Gao H, Zeng W, Austin J, Li Z, et al. 2025. Shape of motion: 4d reconstruction from a single video//IEEE/CVF International Conference on Computer Vision: 9660-9672.

Wang Q, Zhang Y, Holynski A, Efros A A, and Kanazawa A. 2025. Continuous 3D Perception Model with Persistent State//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 10510-10522 [DOI: 10.1109/CVPR52734.2025.00983].

Wang S, Leroy V, Cabon Y, Chidlovskii B, and Revaud J. 2024. DUSt3R: Geometric 3D Vision Made Easy//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. IEEE: 20697-20709 [DOI: 10.1109/CVPR52733.2024.01956].

Wang W, Chen Y, Zhang Z, Liu H, Wang H, Feng Z, et al. 2025. VolSplat: Rethinking Feed-Forward 3D Gaussian Splatting with Voxel-Aligned Prediction [EB/OL]. [DOI: 10.48550/ARXIV. 2509.19297].[2026-03-18]. https://arxiv.org/pdf/2509.19297.pdf.

Wang Y, Zhao Q, Yu R, Tsui H W, Zeng A, Lin J, et al. 2025. SkillMimic: Learning Basketball Interaction Skills from Demonstrations. arXiv preprint arXiv:2408.15270.

Wang Y, Yang P, Xu Z, Sun J, Zhang Z, Chen Y, et al. 2025. Freetimes: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction//IEEE/CVF Conference on Computer Vision and Pattern Recognition: 21750-21760.

Wang Y, Zhou J, Zhu H, Chang W, Zhou Y, Li Z, et al. 2025. pi3: Permutation-Equivariant Visual Geometry Learning. arXiv preprint arXiv:2507.13347.

Wang, Lizhen and Zhu, Yongming and Ge, Zhipeng and Zheng, Youwei and Zhang, Longhao and Hu, Tianshu and Qin, Shiyang and Luo, Mingshuang and Zhang, Jiaxu and Chen, Xin and Wang,

Yulong and Zheng, Zerong and Jiang, Jianwen and Liang, Chao and Chen, Weifeng and Wang, Xing and Zhang, Yuan and Gao, Mingyuan. 2026. FlowAct-R1: Towards Interactive Humanoid Video Generation.

Weng H, Zhao Z, Lei B, Yang X, Liu J, Lai Z, et al. 2025. Scaling Mesh Generation via Compressive Tokenization//Proceedings of the Computer Vision and Pattern Recognition Conference: 11093-11103.

Wu D, Liu F, Hung Y H, et al. 2025. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence//Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems.

Wu G, Yi T, Fang J, Xie L, Zhang X, Wei W, et al. 2024. 4d gaussian splatting for real-time dynamic scene rendering//IEEE/CVF conference on computer vision and pattern recognition: 20310-20320.

Wu J Z, Zhang Y, Turki H, Ren X, Gao J, Shou M Z, et al. 2025. DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 26024-26035.

Wu J, Guan J, Feng K, Liu Q, Wu S, Wang L, et al. 2025. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965.

Wu R, Gao R, Poole B, Trevithick A, Zheng C, Barron J T, et al. 2025. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 26057-26068.

Wu S, Lin Y, Zhang F, Zeng Y, Yang Y, Bao Y, et al. 2025. Direct3D-S2: Gigascale 3D Generation Made Easy with Spatial Sparse Attention. arXiv preprint arXiv:2505.17412.

Wu S, Xu C, Huang B, Geiger A, and Chen A. 2025. GenFusion: Closing the Loop between Reconstruction and Generation via Videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 6078-6088.

Wu T, Yang S, Po R, Xu Y, Liu Z, Lin D, et al. 2025. Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284.

Wu Y, Wu Y, Li W, Lu Y, Feng K, and Chen X. 2025. FastAvATAR: Towards Unified Fast High-Fidelity 3D Avatar Reconstruction with Large Gaussian Reconstruction Transformers. arXiv preprint arXiv:2508.19754.

XAI. 2024. Grok-1.5.

Xiang J, Chen X, Xu S, Wang R, Lv Z, Deng Y, et al. 2025. Native and Compact Structured Latents for 3D Generation. arXiv preprint arXiv:2512.14692.

Xiang J, Lv Z, Xu S, Deng Y, Wang R, Zhang B, et al. 2025. Structured 3D Latents for Scalable and Versatile 3D Generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 21469-21480.

Xiao Y, Wang J, Xue N, Karaev N, Makarov Y, Kang B, et al. 2025. SpatialTrackerV2: 3D Point Tracking Made Easy [EB/OL]. [DOI: 10.48550/ARXIV.2507.12462]. [2026-03-18]. https://arxiv.org/pdf/2507.12462.pdf.

Xie Y, Yao C H, Voleti V, Jiang H, and Jampani V. 2024. SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency. arXiv preprint arXiv:2407.17470.

Xie Y, Gu T, Li Z, Zhang C, Song G, Zhao X, et al. 2025. X-streamer: Unified human world modeling with audiovisual interaction. arXiv preprint arXiv:2509.21574.

Xu J, Wang C, Zhao Z, Liu W, Ma Y, and Gao S. 2024. CAD-MLLM: Unifying Multimodality-Conditioned CAD Generation with MLLM. arXiv preprint arXiv:2411.04954.

Xu J, Guo Z, He J, Hu H, He T, Bai S, et al. 2025. Qwen2.5-omni technical report. arXiv preprint arXiv:2503.20215.

Xu M, Zhang H, Hou Y, Xu Z, Fan L, Veloso M, et al. 2025. DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation. arXiv preprint arXiv:2505.21864.

Xu S, Ling H Y, Wang Y X, and Gui L Y. 2026. InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions. arXiv preprint arXiv:2502.20390.

Xu X, Jayaraman P, Lambourne J, Liu Y, Malpure D, and Meltzer P. 2025. AutoBRep: Autoregressive B-Rep Generation with Unified Topology and Geometry//Proceedings of the SIGGRAPH Asia 2025 Conference Papers: 1-12.

Xu Z, Li Z, Dong Z, Zhou X, Newcombe R, and Lv Z. 2025. 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015.

Xu Z, Xu Y, Yu Z, Peng S, Sun J, Bao H, et al. 2024. Representing long volumetric video with temporal gaussian hierarchy. ACM Trans. on Graphics: 1-18.

Yan H, Yu H, Zhong Z, Yuan W, Gong X, Luo Z, et al. 2025. Open-world Hand-Object Interaction Video Generation Based on Structure and Contact-aware Representation. arXiv preprint arXiv:2512.01677.

Yan X, Xu J, Li Y, Ma C, Yang Y, Wang C, et al. 2025. X-Part: High Fidelity and Structure Coherent Shape Decomposition. arXiv preprint arXiv:2509.08643.

Yang J, Liu S, Guo H, Dong Y, Zhang X, Zhang S, et al. 2026. EgoLife: Towards Egocentric Life Assistant. arXiv preprint arXiv:2503.03803.

Yang J, Yang S, Gupta A W, et al. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces//Proceedings of the Computer Vision and Pattern Recognition Conference: 10632-10643.

Yang R, Yu Q, Wu Y, Yan R, Li B, Cheng A C, et al. 2025. EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos. arXiv preprint arXiv:2507.12440.

Yang S, Kong Z, Gao F, Cheng M, Liu X, Zhang Y, et al. 2025. Infinitetalk: Audio-driven video generation for sparse-frame video dub-

bing. arXiv preprint arXiv:2508.14033.

Yang S, Yang J, Huang P, Brown E, Yang Z, Yu Y, et al. 2025. Cambrian-S: Towards Spatial Supersensing in Video. arXiv preprint arXiv:2511.04670.

Yang X, Li B, Zhang Y, Yin Z, Bai L, Ma L, et al. 2025. Vlipp: Towards physically plausible video generation with vision and language informed physical prior//Proceedings of the IEEE/CVF International Conference on Computer Vision: 12360-12370.

Yang Y, Fan L, Shi Z, Peng J, Wang F, and Zhang Z. 2026. NeoVerse: Enhancing 4D World Model with in-the-wild Monocular Videos. arXiv preprint arXiv:2601.00393.

Yang Y, Zhou Y, Guo Y, Zou Z, Huang Y, Liu Y, et al. 2025. OmniPart: Part-Aware 3D Generation with Semantic Decoupling and Structural Cohesion//Proceedings of the SIGGRAPH Asia 2025 Conference Papers: 1-12.

Yao C H, Xie Y, Voleti V, Jiang H, and Jampani V. 2025. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation //Proceedings of the IEEE/CVF International Conference on Computer Vision: 13248-13258.

Yao D Y, Zhai A J, and Wang S. 2025. Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE: 1116-1126[DOI: 10.1109/CVPR52734.2025.00112].

Yao K, Zhang L, Yan X, Zeng Y, Zhang Q, Xu L, et al. 2025. CAST: Component-Aligned 3D Scene Reconstruction from an RGB Image. ACM Transactions on Graphics: 86: 1-86: 19[DOI: 10.1145/3730841].

Ye A, Zhang Z, Wang B, Wang X, Zhang D, and Zhu Z. 2025. Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623.

Ye C, Wu Y, Lu Z, Chang J, Guo X, Zhou J, et al. 2025. Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging. arXiv preprint arXiv:2503.22236.

Ye K, Shao T, and Zhou K. 2025. When gaussian meets surfel: Ultrafast high-fidelity radiance field rendering. ACM Trans. Graph.: 1-15.

Ye S, Jang J, Jeon B, Joo S J, Yang J, Peng B, et al. 2025. Latent Action Pretraining from Videos//The Thirteenth International Conference on Learning Representations.

Yin X, Zhang Q, Chang J, Feng Y, Fan Q, Yang X, et al. 2025. GSFixer: Improving 3D Gaussian Splatting with Reference-Guided Video Diffusion Priors. arXiv preprint arXiv:2508.09667.

Yu C, Wang Y, Guo Z, Lin H, Xu S, Zang H, et al. 2025. RLInf: Reinforcement Learning Infrastructure for Embodied and Agentic AI. arXiv preprint arXiv:2509.15965.

Yu J, Bai J, Qin Y, Liu Q, Wang X, Wan P, et al. 2025. Context as memory: Scene-consistent interactive long video generation with memory retrieval//Proceedings of the SIGGRAPH Asia 2025 Conference Papers: 1-11.

Yu M, Hu W, Xing J, and Shan Y. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV): 100-111.

Yu T, Lu G, Yang Z, Deng H, Chen S S, Lu J, et al. 2025. ManiGaussian++: General robotic bimanual manipulation with hierarchical Gaussian world model//2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS): 12232-12239.

Yu W, Xing J, Yuan L, Hu W, Li X, Huang Z, et al. 2025. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence: 1-18 [DOI: 10.1109/TPAMI.2025.3613256].

Yuan Y, Shen Q, Yang X, and Wang X. 2025. 1000+ fps 4d gaussian splatting for dynamic scene rendering//Advances in Neural Information Processing Systems.

Zang H, Wei M, Xu S, Wu Y, Guo Z, Wang Y, et al. 2025. RLInf-VLA: A Unified and Efficient Framework for Reinforcement Learning of Vision-Language-Action Models. arXiv preprint arXiv:2510.06710.

Zeng Y, Bao Y, Qian J, Wu S, Lin Y, Zhu H, et al. 2025. TEXTRIX: Latent Attribute Grid for Native Texture Generation and Beyond. arXiv preprint arXiv:2512.02993.

Zhan Y, Shao T, Yang Y, and Zhou K. 2025. Real-time high-fidelity Gaussian human avatars with position-based interpolation of spatially distributed MLPs//Proceedings of the Computer Vision and Pattern Recognition Conference: 26297-26307.

Zhang C, Moing G L, Koppula S, Rocco I, Momenti L, Xie J, et al. 2025. Efficiently Reconstructing Dynamic Scenes One D4RT at a Time [EB/OL]. [DOI: 10.48550/ARXIV.2512.08924].[2026-03-18].
https://arxiv.org/pdf/2512.08924.pdf.

Zhang D, Liu Y, Lin L, Zhu Y, Li Y, Qin M, et al. 2025. Guava: Generalizable upper body 3d gaussian avatar//Proceedings of the IEEE/CVF International Conference on Computer Vision: 14205-14217.

Zhang H, Zhang S, Jin J, Zeng Q, Qiao Y, Lu H, et al. 2025. Balancing signal and variance: Adaptive offline r1 post-training for vla flow models. arXiv preprint arXiv:2509.04063.

Zhang J, Huang Z, Gu C, Ma Z, and Zhang L. 2025. Reinforcing action policies by prophesying. arXiv preprint arXiv:2511.20633.

Zhang J, Herrmann C, Hur J, Jampani V, Darrell T, Cole F, et al. 2025. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion//The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

Zhang K, Xiao C, Xu J, Mei Y, and Patel V M. 2025. Think Before You Diffuse: Infusing Physical Rules into Video Diffusion. arXiv

preprint arXiv:2505.21653.

Zhang L, Zhang Q, Jiang H, Bai Y, Yang W, Xu L, et al. 2025. BANG: Dividing 3D Assets via Generative Exploded Dynamics. ACM Transactions on Graphics (TOG): 1-21.

Zhang L, Wang Z, Zhang Q, Qiu Q, Pang A, Jiang H, et al. 2024. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. ACM Trans. on Graphics: 120:1-120:20 [DOI: 10.1145/3658146].

Zhang L, Cai S, Li M, Wetzstein G, and Agrawala M. 2025. Frame context packing and drift prevention in next-frame-prediction video diffusion models. arXiv preprint arXiv:2504.12626.

Zhang X, Liao J, Zhang S, Meng F, Wan X, Yan J, et al. 2025. Videorepa: Learning physics for video generation through relational alignment with foundation models. arXiv preprint arXiv:2505.23656.

Zhang Y, Lin J, Zeng A, Wu G, Lu S, Fu Y, et al. 2025. Motion-X++: A Large-Scale Multimodal 3D Whole-body Human Motion Dataset. arXiv preprint arXiv:2501.05098.

Zhang Y, Zhang L, Ma R, and Cao N. 2025. TexVerse: A Universe of 3D Objects with High-Resolution Textures. arXiv preprint arXiv: 2508.10868.

Zhao Q, Tan H, Wang Q, Bi S, Zhang K, Sunkavalli K, et al. 2025. E-RayZer: Self-supervised 3D Reconstruction as Spatial Visual Pretraining [EB/OL]. [DOI: 10.48550/ARXIV.2512.10950]. [2026-03-18]. https://arxiv.org/pdf/2512.10950.pdf.

Zhaxizhuoma, Liu K, Guan C, Jia Z, Wu Z, Liu X, et al. 2025. FastUMI: A Scalable and Hardware-Independent Universal Manipulation Interface with Dataset. arXiv preprint arXiv:2409.19499.

Zhen H, Sun Q, Zhang H, Li J, Zhou S, Du Y, et al. 2025. Learning 4D Embodied World Models//Proceedings of the IEEE/CVF International Conference on Computer Vision: 5337-5347.

Zheng S, Yin M, Hu W, Li X, Shan Y, and Fu Y. 2026. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. arXiv preprint arXiv:2601.05138.

Zhong Z, Ji Y, Kong Z, Liu Y, Wang J, Feng J, et al. 2025. Anytalker: Scaling multi-person talking video generation with interactivity refinement. arXiv preprint arXiv:2511.23475.

Zhou G, Pan H, LeCun Y, and Pinto L. 2024. Dino-wm: World models on pre-trained visual features enable zero-shot planning. arXiv preprint arXiv:2411.04983.

Zhou K, Wang Y, Chen G, Chang X, Beaudouin G, Zhan F, et al. 2025. PAGE-4D: Disentangled Pose and Geometry Estimation for 4D Perception [EB/OL]. [DOI: 10.48550/ARXIV.2510.17568]. [2026-03-18]. https://arxiv.org/pdf/2510.17568.pdf.

Zhu J, Yue J, He F, and Wang H. 2025. 3D Student Splatting and Scooping// IEEE/CVF Conference on Computer Vision and Pattern Recognition: 21045-21054 [DOI: 10.1109/CVPR52734.2025.01960].

Zhuo D, Zheng W, Guo J, Wu Y, Zhou J, and Lu J. 2025. Streaming 4D Visual Geometry Transformer [EB/OL]. [DOI: 10.48550/ARXIV.2507.11539]. [2026-03-18]. https://arxiv.org/pdf/2507.11539.pdf.

Zuo S, Zheng W, Huang Y, Zhou J, and Lu J. 2025. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction//Proceedings of the Computer Vision and Pattern Recognition Conference: 6772-6781.

作者简介

刘烨斌,男,教授,主要研究方向为三维视觉和计算成像。E-mail: liuyebin@tsinghua.edu.cn

穆尧,男,助理教授,主要研究方向为具身智能与数字孪生。E-mail: muyao@sjtu.edu.cn

叶琦,女,研究员,主要研究方向为三维视觉和灵巧操作。E-mail: qi.ye@zju.edu.cn

高林,男,研究员,主要研究方向为计算机图形学和三维计算机视觉。E-mail: gaolin@ict.ac.cn

韩晓光,男,助理教授,主要研究方向为三维视觉和图形学。E-mail: hanxiaoguang@cuhk.edu.cn

陈安沛,男,助理教授,主要研究方向为三维视觉和计算图形学。E-mail: chenanpei@westlake.edu.cn

段岳圻,男,副教授,主要研究方向为三维视觉。E-mail: duanyueqi@tsinghua.edu.cn

彭思达,男,研究员,主要研究方向为三维视觉和空间智能。Email: pensida@zju.edu.cn

邵天甲,男,长聘副教授,主要研究方向为计算机图形学和三维视觉。E-mail: tjshao@zju.edu.cn

张鸿文,男,副教授,主要研究方向为三维视觉和图形学。E-mail: zhanghongwen@bnu.edu.cn

张力,男,教授,主要研究方向为计算机视觉、自动驾驶以及具身智能。E-mail: lizhangfd@fudan.edu.cn

廖依伊,女,研究员,主要研究三维视觉。E-mail: yiyi.liao@zju.edu.cn

许岚,男,助理教授,主要研究方向为三维视觉和图形学。E-mail: xulan1@shanghaitech.edu.cn

刘希慧,女,助理教授,主要研究方向为计算机视觉和生成模型。E-mail: xihuiliu@hku.hk

姚遥,男,副教授,主要研究方向为三维计算机视觉。E-mail: yaoyao@nju.edu.cn

胡瑞珍,女,教授,主要研究方向为计算机图形学与具身智能。E-mail: ruizhen.hu@szu.edu.cn

弋力,男,助理教授,主要研究方向为三维视觉和人形机器人学习。E-mail: ericyi0124@gmail.com

郭裕兰,男,教授,主要研究方向为三维视觉。E-mail: guoyulan@sysu.edu.cn

连宙辉，男，长聘副教授，主要研究方向为计算机图形学和三维视觉。 E-mail：lianzhouhui@pku.edu.cn
刘子纬，男，副教授，主要研究方向为三维视觉和多模态学习。 E-mail：ziwei.liu@ntu.edu.sg
陈宝权，男，教授，主要研究方向为计算机图形学和三维视觉。 E-mail：baoquan@pku.edu.cn