中图法分类号:TP37　　**文献标识码**:　　**文章编号**: 1006-8961(XXXX)XX-0001-25

论文引用格式:Wang Sifang, Wu Han, Li Jiawei, Zhang Yuhang. A survey of foundation model-driven 3D spatial intelligence and embodied decision-making[J/OL]. Journal of Image and Graphics, XXXX:1-25. DOI: 10.11834/jig.260374.(王思芳, 吴晗, 李嘉巍, 张宇航. 基础模型驱动的三维空间智能与具身决策综述[J/OL]. 中国图象图形学报,XXXX:1-25. DOI: 10.11834/jig.260374.)[DOI:10.11834/jig.260374]

基础模型驱动的三维空间智能与具身决策综述

王思芳,吴晗,李嘉巍,张宇航

郑州大学计算机与人工智能学院,郑州,450001

摘　要:三维场景理解是具身智能体感知环境、理解指令并支持行动决策的重要基础。随着视觉—语言及多模态基础模型向三维领域拓展,相关研究已从场景感知与空间语义理解延伸至动态预测和具身交互。然而,现有综述多聚焦于单一任务或特定模型,对相关技术的能力边界及相互依赖缺少系统梳理。本文依据模型能力及其信息依赖关系,构建涵盖三维感知、语言驱动理解、世界模型与动态预测、具身智能体决策与交互的四层研究框架。该框架用于刻画不同层次间的能力关联,而非规定固定的任务执行顺序。在此基础上,本文梳理各层次的代表性方法、常用数据集与评价指标,并结合任务设置和评测协议分析不同技术路线的能力及适用范围。分析表明,二维基础模型迁移为开放词汇三维感知提供了重要路径,三维多模态模型促进了场景信息与语言语义的联合建模,部分世界模型进一步探索未来状态预测与动作条件预测,具身智能体则将场景认知应用于导航和机器人操作,并通过环境反馈调整行动。当前,该领域仍面临高质量三维数据不足、跨模态迁移中的几何与语义失配、空间推理稳定性不足及模型部署效率受限等挑战,跨任务评测标准也有待统一。针对这些问题,本文讨论三维原生基础模型、几何约束引导的空间推理、高效模型部署与跨任务统一评测等未来研究方向。本文涉及的代表性算法、数据集和评价指标可通过https://github.com/sunflower0504/Foundation-Models-3D-Spatial-Intelligence-Survey获取。

关键词:三维空间智能;具身决策;基础模型;三维场景理解;世界模型;具身智能

A survey of foundation model-driven 3D spatial intelligence and embodied decision-making

Wang Sifang, Wu Han, Li Jiawei, Zhang Yuhang

School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, Henan, China

Abstract: Three-dimensional (3D) scene understanding is a fundamental capability that enables embodied agents to perceive their surroundings, interpret instructions, reason about spatial relationships, and make decisions. Advances in vision-language and multimodal foundation models have expanded research from conventional geometric reconstruction and category-specific recognition to open-vocabulary 3D perception, language-driven spatial understanding, dynamic prediction, and embodied interaction. However, existing surveys often focus on individual tasks, model families, or application scenarios, leaving the relationships among these directions and the functional boundaries of different approaches insufficiently clarified. A systematic overview is therefore needed to explain how foundation models support 3D spatial intelligence and what challenges remain for reliable embodied deployment. This survey organizes the field according to the func-

收稿日期:2026-06-27;**修回日期**:2026-11-10

*通信作者:吴晗,通信作者,女,讲师,主要研究方向为人工智能,具身智能,计算机视觉。 E-mail: zzuwuhan@zzu.edu.cn

基金项目:河南省科技攻关项目((No. 252102231076))

Supported by:The Science and Technology Project of Henan Province (No. 252102231076)

© 中国图象图形学报版权所有

tional evolution of model capabilities and their information dependencies. It proposes a four-level framework comprising 3D perception, language-driven 3D understanding, world models and dynamic prediction, and embodied agent decision-making and interaction. The framework describes the capability relationships among these levels rather than prescribing a fixed processing sequence. A system may combine selected capabilities and update its scene interpretation through new observations and action feedback. This perspective provides a common basis for comparing methods with different architectures and objectives, while distinguishing the information they estimate, interpret, predict, or use to support action. The first level, 3D perception, establishes geometric and semantic representations of the environment. Representative tasks include semantic and instance segmentation, object recognition, scene representation, and open-vocabulary perception. Conventional methods often rely on predefined categories and task-specific annotations, which can limit their ability to recognize unfamiliar objects. An important research direction transfers knowledge from pretrained two-dimensional (2D) vision or vision-language foundation models to 3D data, linking image features with point clouds, reconstructed scenes, or other three-dimensional representations. However, two-dimensional features do not always preserve 3D geometry and spatial relationships. Viewpoint variation, occlusion, incomplete observations, and ambiguous image-to-scene correspondence can cause geometric-semantic mismatches. Reliable 3D perception therefore requires not only cross-modal transfer but also robust feature association and spatially consistent representations. The second level, language-driven 3D understanding, connects scene representations with natural language descriptions, questions, and instructions. These methods ground linguistic expressions in 3D scenes, interpret object relationships, answer questions about spatial configurations, and identify entities relevant to a user's request. Three-dimensional multimodal large language models support the joint modeling of geometric features, visual observations, and linguistic representations. Representative tasks include referring expression grounding, 3D visual question answering, scene dialogue, and spatial reasoning. Nevertheless, their effectiveness depends on cross-modal alignment and the preservation of spatial information during feature integration. Fluent language output does not necessarily indicate correct grounding or reliable spatial reasoning; evaluation should also examine consistency with scene geometry and object relations. The third level covers world models and dynamic scene prediction. Static scene understanding estimates current geometry, occupancy, or semantic structure, whereas dynamic prediction explicitly models how a scene may evolve over time. These functions should be distinguished: current-state estimation does not itself constitute future-state prediction, and future prediction is not necessarily conditioned on an agent's candidate actions. Dynamic world models may forecast future observations, object motion, scene configurations, or occupancy states based on current or historical information. Action-conditioned models further estimate possible future states under candidate actions, potentially supporting planning and control. Such capabilities are valuable for navigation, trajectory planning, and interactive robotics. Their reliability, however, may decline with incomplete observations, long prediction horizons, distribution shifts, and complex environmental interactions. The extent to which prediction accuracy improves downstream task performance also requires careful assessment. The fourth level concerns embodied agent decision-making and interaction. Embodied agents must translate scene perception and language understanding into executable actions in physical or simulated environments. Relevant research integrates scene representations, language instructions, spatial reasoning, planning, control, and feedback from action execution. Applications include vision-language navigation, object-goal navigation, and robotic manipulation. Scene understanding helps identify target objects and spatial constraints, while predictive models can help estimate the consequences of alternative actions. New observations after execution allow agents to revise their interpretation or plans, forming a perception-reasoning-action-feedback loop. However, errors can propagate across components: inaccurate perception may mislead reasoning, incorrect grounding may distort planning, and unreliable predictions may lead to failed actions. Plausible language responses or action proposals should therefore not be equated with successful task execution. Across the four levels, this survey reviews representative methods, commonly used datasets, and evaluation metrics. It compares approaches in terms of input representations, outputs, supervision requirements, core capabilities, application scopes, and evaluation settings. Metrics vary by task: perception is commonly assessed using detection or segmentation measures; language-driven understanding uses grounding and question-answering metrics; dynamic prediction requires measures suited to future states or trajectories; and embodied tasks may be evaluated through navigation or manipulation success. Results obtained with different datasets, data splits, annotations, and evalua-

© 中国图象图形学报版权所有

tion protocols are not always directly comparable. The survey emphasizes clear reporting of experimental conditions and distinguishes benchmark performance from evidence of generalization, robustness, and end-to-end effectiveness. Several challenges continue to limit progress. High-quality 3D data and detailed annotations are expensive to obtain, restricting training and systematic evaluation. Geometric-semantic mismatch remains a barrier to transferring knowledge from 2D models to 3D scenes, while multimodal grounding and spatial reasoning can be unreliable in complex configurations. Large architectures and processing pipelines also impose substantial computation and memory costs, hindering low-latency deployment on resource-constrained platforms. In addition, inconsistent benchmarks and evaluation protocols complicate fair comparisons and obscure whether improvements in individual components lead to better overall task performance. Future research should develop 3D-native foundation models that better preserve geometry, spatial relations, and scene structure; incorporate geometric constraints into multimodal reasoning to improve grounding consistency; and advance temporal and action-conditioned world models for reasoning about possible outcomes under uncertainty. Model compression, efficient adaptation, and optimized inference may reduce deployment costs. Unified evaluation frameworks are also needed to assess perception, reasoning, prediction, and action under clearly defined protocols, while distinguishing current-state estimation from future prediction, open-loop prediction from action-conditioned prediction, and component-level performance from closed-loop task success. Further study of uncertainty propagation and robustness in unfamiliar and safety-critical environments will be important. In summary, this survey presents a capability-oriented perspective on foundation model-driven 3D spatial intelligence, connecting 3D perception and language-driven understanding with dynamic prediction and embodied decision-making. By synthesizing representative methods, clarifying functional boundaries, reviewing evaluation practices, and identifying persistent challenges, it provides a structured basis for comparing approaches and understanding their complementary roles. This perspective aims to support the development of more geometrically grounded, semantically consistent, predictively reliable, and computationally efficient systems for embodied interaction. Representative methods, datasets, and evaluation metrics covered in this survey have been compiled into a companion repository available at . https://github. com/sunflower0504/Foundation-Models-3D-Spatial-Intelligence-Survey

Key words: 3D spatial intelligence; embodied decision-making; foundation models; 3D scene understanding; world models; embodied intelligence

论文引用格式:【DOI:10. 11834/jig. 260374】

0 引　言

CLIP(Radford 等,2021)、DINO(Caron 等,2021)和SAM(Kirillov 等,2023)等二维视觉—语言及视觉基础模型分别提供跨模态语义、视觉结构与类别无关分割先验,为开放词汇三维感知提供了重要知识来源。随着多模态大语言模型与三维表示结合,相关研究进一步由语义迁移扩展至语言驱动的空间理解、动态场景预测和具身交互。3D场景理解正经历从静态几何重建向动态具身交互的转变(刘烨斌等,2026)。由此,基础模型驱动的三维空间智能主要涉及四个相互关联的方向:开放词汇三维感知、语言驱动理解、世界模型与动态预测,以及具身决策与交互。

尽管相关研究持续发展,三维标注数据不足、二维知识迁移中的几何—语义失配、空间推理幻觉,以及跨场景泛化和部署效率受限等问题,仍制约着模型在复杂具身环境中的可靠应用。相关问题将在第6章进一步讨论。

围绕基础模型支持的三维空间智能与具身智能,已有综述已分别从3D场景问答(Li 等,2026)、视觉—语言—动作模型(Ma 等,2026;刘锐等,2026)、具身智能由感知到行为的演进(Chen 等,2025)、开放词汇检测与分割(Zhu 和Chen,2024)、多模态大语言模型支持的3D任务(Ma 等,2024),以及语言驱动的多模态3D场景理解(Ren 等,2025)等角度进行了系统梳理。这些工作为理解特定任务的模型设计、数据集与评测方法提供了重要基础。不同于单一任务或模型视角,本文以基础模型在不同能力层中的作用变化为主线,分析其对场景表征、语言理解、动态预测和具身行动的支持,并讨论能力边界及可组合关系。部分代表性综述的比较见表1。

本文主要贡献包括:其一、构建覆盖三维感知、语言驱动理解、世界模型与动态预测、具身决策与交

© 中国图象图形学报版权所有

互的四层分析框架，刻画不同任务间的信息依赖与反馈关系。其二、梳理基础模型由跨模态知识源、统一表征接口向空间推理与决策引擎演变的技术路径。其三、比较自动驾驶、室内导航和机器人操作中的空间表示、预测目标与行动空间差异，分析相关方法的可迁移性和适用边界。

表1　代表性相关综述与本文的定位比较

Table 1　Comparison between representative related surveys and this work

综述	主要关注	与本文的重叠部分	本文进一步组织的内容
Embodied Intelligence for 3D Understanding: A Survey on 3D Scene Question Answering（Li等，2026）	具身/情景化3D场景问答、空间问答任务与评测	语言驱动三维理解、具身场景中的语义推理	将问答视为语言理解层的一类任务，并关联其所依赖的场景表征、动态预测和后续行动选择
A Survey on Vision-Language-Action Models for Embodied AI（Ma等，2026）	视觉—语言—动作模型（vision-language-action model，VLA）架构、训练范式与动作生成	多模态输入、具身决策与交互	从三维场景状态表征和世界模型角度分析VLA的空间信息来源及其与动态预测的关系
Embodied AI: A Survey on the Evolution from Perceptive to Behavioral Intelligence（Chen等，2025）	具身智能由感知到行为的整体演进	感知、推理和行为的能力链条	聚焦基础模型驱动的三维空间任务，细化开放词汇感知、语言驱动理解与动态建模的方法脉络
A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future（Zhu和Chen，2024）	开放词汇检测与分割方法	三维感知层中的开放词汇语义表征	将开放词汇感知置于三维空间智能链条中，讨论其如何支撑语言查询、预测和具身任务
When LLMs Step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models（Ma等，2024）	三维多模态大语言模型、任务分类与元分析	三维场景理解、空间推理、规划与导航	进一步系统梳理开放词汇三维感知、世界模型与动态预测，分析基础模型在感知、理解、预测和决策中的角色演变及能力关联
A Survey of Language-Grounded Multi-modal 3D Scene Understanding（Ren等，2025）	语言驱动的三维视觉定位、场景描述与问答	语言驱动的三维理解与空间推理	将语言理解与开放词汇三维感知、动态预测及具身决策置于统一分析框架，讨论不同能力之间的信息依赖与适用边界
本文	基础模型驱动的三维空间智能与具身决策	覆盖上述相关能力	以“外部知识源—多模态统一接口—高层推理与决策引擎”的基础模型角色演变为主线，讨论四层能力的关联、边界与跨场景适用性

1　基础模型与智能体3D场景理解

1.1　基础模型的定义与类型

多模态基础模型的快速迭代，为3D场景理解及具身智能体的发展提供了坚实支撑。基础模型的发展展现了从2D基础模型涌现，到多模态融合与生成，再到3D空间理解扩展的清晰发展历程。

视觉基础模型为三维场景理解提供了可迁移的二维先验：CLIP通过图文对比学习提供开放词汇语义，DINO通过自监督学习提取视觉结构特征，SAM生成类别无关的二维掩膜。随后，BLIP-2（Li等，2023）和LLaVA系列（Liu等，2023；Liu等，2024；Li等，2024）等多模态模型通过视觉编码、轻量跨模态对齐与指令微调建立视觉—语言交互接口，为后续三维多模态模型提供了架构基础。本文不展开这些通用模型的内部训练细节，而聚焦其知识与能力如何迁移至三维空间。

三维适配研究随后形成两条互补路线：OpenShape（Liu等，2023）通过图像、文本与点云的多模态对比学习建立统一三维表征。3D-LLM（Hong等，2023）则将点云特征映射至冻结的2D VLM，并引入3D位置嵌入和离散位置词元，使模型能够理解并输出空间位置。前者侧重三维跨模态表征学习，后者侧重三维信息与语言模型的结合，为后续开放词汇感知和空间推理研究奠定基础。

© 中国图象图形学报版权所有

1.2 三维空间智能与决策的任务体系

结合现有研究(Ren 等,2025;Ma 等,2024;Xie 等,2025),本文将基础模型驱动的三维空间智能与具身决策划分为四层:3D 感知、语言驱动 3D 理解、世界模型与动态预测和具身智能体系统。前三层级分别聚焦场景静态特征解析、跨模态语义理解与时序动态规律推演。第四层级面向具身智能体的任务规划、行动决策与环境交互。四层框架用于组织相关能力及其联系,而非规定所有系统必须遵循固定处理顺序。图 1 给出了本文构建的三维空间智能与具身决策任务体系。

四层框架旨在组织不同任务中的能力关联,而非将各类任务限定为严格串行的处理流水线。三维感知提供场景状态基础,语言驱动理解强化语义查询与解释,世界模型刻画状态演化及其不确定性,具身决策则面向任务目标选择并执行行动。各层之间既存在自下而上的信息支撑,也存在由任务目标和交互反馈驱动的自顶向下调节。表 2 从输入、输出、核心能力、典型任务和层间关系归纳各层功能边界。

1.3 常用数据集与评价指标

三维感知任务常采用 ScanNet(Dai 等,2017)、S3DIS(Armeni 等,2016)、nuScenes(Caesar 等,2020)、ScanNet200(Rozenberszki 等,2022)和 Replica(Straub 等,2019)进行分割与检测评测。语言驱动

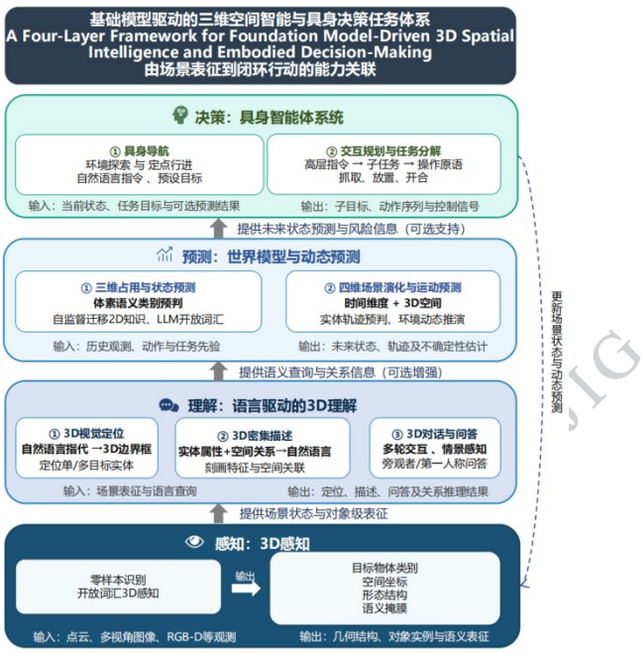


图 1 智能体 3D 场景理解任务体系图

Fig. 1 Hierarchical framework of agent 3D scene understanding tasks

表 2 基础模型驱动的三维空间智能与具身决策四层框架的功能边界与关联关系

Table 2 Functional boundaries and inter-layer relationships of the four-layer framework for foundation model-driven 3D spatial intelligence and embodied decision-making

层级	输入	输出	核心能力	典型任务	层间关系
三维感知与场景表征	点云、多视角图像、彩色与深度图像(red, green, blue and depth,RGB-D)、位姿等	几何结构、语义标签、实例、占据状态及对象级表征	几何重建、检测、分割与语义赋值	三维检测、语义/实例分割、开放词汇三维理解	为后续层提供可计算的场景状态;也可接收语言或任务目标的自顶向下约束
语言驱动的三维理解	场景表征、文本查询或语言指令	对象定位、关系理解、场景描述、问答结果	跨模态对齐、指代消解、空间关系推理	3D Grounding、3D QA、场景描述	在第 1 层表征基础上实现查询、解释与语义增强;其输出可服务于预测和决策,但并非所有任务的必经环节
世界模型与动态预测	当前或历史观测、当前场景状态,以及可选的动作条件与任务信息	未来场景状态、运动轨迹、未来视频及相关不确定性估计	时序状态预测、动作条件推演与预测辅助规划	未来占用预测、轨迹预测、未来场景生成、预测辅助规划	基于前置场景表征进行未来状态推演,为具身决策提供预测性支持
具身决策与交互	当前场景状态、任务目标、约束条件,以及可选的预测结果	子目标、动作序列、控制信号与交互反馈	任务分解、规划、探索、导航、操作与闭环控制	具身导航、VLA、机器人操作、长程任务执行	根据任务需求调用相关感知、理解与预测能力,并通过执行反馈更新场景状态或调整行动策略,形成闭环交互。

© 中国图象图形学报版权所有

的三维理解任务主要使用ScanRefer(Chen等,2020)、SceneVerse(Jia等,2024)、ScanQA(Azuma等,2022)和SQA3D(Ma等,2023)开展视觉定位、场景问答与空间推理,3DSSG(Wald等,2020)则用于三维场景关系建模。世界模型与动态预测相关研究常采用Occ3D-nuScenes(Tian等,2023)和SemanticKITTI(Behley等,2019)等数据集评估场景几何与语义状态表征能力。具身任务常采用ALFRED(Shridhar等,2020)、R2R(Anderson等,2018)、CALVIN(Mees等,2022)和RLBench(James等,2020)评估导航与机器人操作能力。

不同任务采用的评价指标也有所区别:三维分割与检测主要采用mIoU、AP、mAP和OA,nuScenes检测还采用NDS;三维视觉定位常用Acc@0.25/0.5、F1@0.5和mAP;场景问答主要采用CIDEr、BLEU、METEOR、ROUGE和EM;场景图生成采用Recall、mAP和R@K;具身导航与机器人操作则主要采用SR、SPL、Goal Progress、OSR、NE、Success Rate和Completion Rate等指标。总体而言,不同指标分别衡量空间定位精度、语义一致性、关系预测能力和任务执行效果,其评测结果受任务设置与协议影响,不宜直接跨任务比较。各类任务的常用数据集及核心评价指标见表3,数据集的详细信息见表4。

表3 常用数据集与评测基准

Table 3 Commonly used datasets and evaluation benchmarks

任务类型	常用数据集	核心评价指标
3D语义/实例分割	ScanNet, S3DIS,ScanNet200,Replica	mIoU, mAP, OA
3D目标检测	ScanNet, nuScenes	mAP, NDS
3D视觉定位	ScanRefer, ScanNet, SceneVerse	Acc@0.25/0.5, F1@0.5, mAP
3D问答与空间推理	ScanQA, SceneVerse, SQA3D	CIDEr, BLEU-1, METEOR, ROUGE, EM
3D场景图生成	3DSSG	Recall, mAP, R@K
三维占用估计与语义场景补全	Occ3D-nuScenes、SemanticKITTI	IoU、mIoU、RayIoU
具身导航	ALFRED, R2R	SR, SPL, Goal Progress, OSR, NE
交互规划与操作	CALVIN, RLBench	Success Rate, Avg. Reward, Completion Rate

2 开放词汇3D感知

开放词汇3D感知方法广泛采用二维视觉—语言模型与视觉基础模型的知识迁移策略,将研究逐步拓展至三维语义分割、实例分割与目标检测。不同任务在知识迁移机制、空间表示及监督需求方面有所差异。主要技术路线如图2所示。

2.1 3D语义/实例分割

CLIP(Radford等,2021)的图像-文本对比学习,启发了研究者将类似思想应用于3D点云。但直接使用对比损失面临计算开销大、正负样本构造困难等问题。现有方法可按主要知识迁移机制分为两类:一类侧重跨模态特征对齐,包括显式对比学习与预训练语义空间匹配;另一类侧重通过投影、蒸馏或融合将二维基础模型知识迁移至三维表示。两类路线并非完全互斥,部分方法同时采用特征迁移与对比学习机制。

2.1.1 对比学习对齐型方法

对比学习对齐型方法主要利用预训练图像—文本语义空间建立三维特征与语言语义之间的联系。PointCLIP(Zhang等,2022)将点云投影为多视图深度图,并复用CLIP进行零样本分类,但其二维投影难以充分表达三维几何。ULIP(Xue等,2023)进一步引入图像—文本—点云三模态对比预训练,直接学习通用三维表征;ULIP-2(Xue等,2024)利用多模态模型生成更丰富的文本描述以扩展训练数据。与此不同,PointCLIP V2(Xue等,2024)主要改进形状投影和文本提示,在既有CLIP框架下提升分类、分割和检测能力。两条路线分别体现了三维编码器对齐学习与二维语义知识直接迁移的差异。

物体级对齐进一步扩展至场景级语义分割。CLIP2Scene(Chen等,2023)通过语义一致性对比学习约束点云特征与文本类别原型对齐,缓解二维语

© 中国图象图形学报版权所有

表4　代表性三维空间智能与具身任务数据集对比

Table 4　Comparison of representative datasets for 3D spatial intelligence and embodied tasks

数据集	模态	场景	数据规模	主要任务	语言标注	原生语义标注类型	开放词汇评测使用
S3DIS	RGB点云	室内	6个区域、271个房间	语义分割、实例分割	无	固定语义类别	已用于
ScanNet	RGB-D视频、3D点云	室内	1,513个扫描场景、约250万帧RGB-D视图	语义/实例分割、目标检测	无	固定语义类别	已用于
nuScenes	RGB、LiDAR、Radar、IMU/GPS	室外道路	1,000个驾驶场景（每段约20 s）、约140万幅图像、39万帧LiDAR扫描	3D目标检测、跟踪、场景理解	无	固定类别与属性标签	已用于
ALFRED	RGB、动作/状态、文本	室内仿真	120个室内场景、8,055条专家示范、25,743条语言指令	指令跟随、导航、物体操作	有	自然语言任务指令	—
ScanRefer	3D点云、文本	室内	800个场景、11,046个目标物体、51,583条描述	3D视觉定位、指代定位	有	自由文本指代描述	已用于零样本定位
ScanQA	3D点云、文本	室内	800个场景、41,363个问题、58,191个答案	3D场景问答	有	自由文本问答	—
SQA3D	3D点云、文本	室内	650个场景、约6.8k个情境、20.4k条情境描述、33.4k个问题	情境式3D问答、空间推理	有	情境描述与问答文本	—
SceneVerse	3D场景、文本	室内	约68k个3D室内场景、约2.5M视觉-语言对	3D视觉定位、场景描述、问答	有	场景描述与视觉语言标注	用于开放词汇研究
3DSSG	3D点云、场景图	室内	1,482个3D场景图，来自478个室内环境	3D场景图生成、关系推理	无	固定对象及关系标签	—
R2R	360°全景RGB、文本	室内	90个环境、7,189条导航路径、21,567条语言指令	视觉语言导航、指令跟随	有	自然语言导航指令	—
CALVIN	RGB-D、本体感知、文本	室内仿真	4个环境、34项任务、约24 h遥操作交互数据	长时程语言条件机器人操作	有	语言任务指令	—
RLBench	RGB-D、分割掩码、本体感知	室内仿真	100项机器人操作任务，示范可由运动规划器按需生成	机器人操作、模仿学习、多任务/少样本学习	有	可配置任务语言描述	—
ScanNet200	RGB-D、3D点云	室内	基于ScanNet场景划分，包含200个语义类别	语义分割、实例分割、长尾场景理解	无	固定200类语义标签	已用于
Replica	3D网格、纹理、语义标注，可渲染RGB-D	室内数字化场景	18个高保真室内3D场景	语义/实例分割、导航、具身场景理解	无	固定语义类别	已用于

注：RLBench任务可配置语言描述，但其原始核心数据主要由视觉、本体感知和机器人示范构成。“—”表示本表未列出直接对应的开放词汇评测依据，不代表该数据集不能用于相关任务。其中，SceneVerse主要作为开放词汇研究中的视觉语言数据资源，不能直接视为独立的开放词汇评测基准。

义与三维几何联合优化中的特征偏移；MPEC（Wang等，2025）进一步利用跨视图同质实体约束增强实体判别能力，并实现实体特征与文本描述的跨模态对齐。两者体现了从类别级语义匹配向场景实体级语义表征的推进。

© 中国图象图形学报版权所有

2.1.2　像素/特征迁移型方法

像素/特征迁移型方法利用二维基础模型输出的稠密语义特征，通过多视图几何映射、特征蒸馏或场景融合，将像素级语义迁移至三维表示。OpenScene（Peng等，2023）从多视角图像中提取CLIP、LSeg（Li等，2022）等模型的稠密语义特征，并通过几何映射融合至点云；OpenSeg（Ghiasi等，2022）则提供图像级监督下的开放词汇分割能力。进一步地，3D-OVS（Liu K等，2023）将CLIP和DINO特征蒸馏至神经辐射场（neural radiance fields，NeRF，Mildenhall等，2020），通过三维选择性对齐与相关性分布约束抑制CLIP稠密特征歧义，并利用DINO特征补充几何边界线索，以提升分割轮廓完整性与视角一致性。相较于全局向量对齐，该路线能够保留更细粒度的像素级语义及其三维空间对应关系，但其效果受多视图覆盖和二维—三维几何映射精度影响。

与上述两种知识蒸馏思路不同，ConceptFusion（Jatavallabhula等，2023）采用异构特征作为语义来源，CLIP提供开放词汇的语义对齐能力，DINO提供几何与边界先验，AudioCLIP（Guzhov等，2022）引入音频模态的语义关联，并通过视觉SLAM将语义特征绑定到全局地图节点上，形成支持多模态查询的结构化语义地图。RegionPLC（Yang等，2024）则利用多种二维视觉—语言模型生成区域级语言监督，并通过区域感知的点—语言对比学习将其迁移至三维点云。Detic（Zhou等，2022）提供精确的物体类别标签，OFA（Wang等，2022）提供自由短语，GRiT（Wu等，2024）和Kosmos-2（Peng等，2024）提供上下文丰富的属性描述，CLIP则负责跨模态语义对齐。ConceptFusion和RegionPLC分别代表场景地图级多模态融合与区域级语言监督路线。

总体而言，对比学习对齐型方法能够利用预训练语义空间降低稠密三维标注需求，但性能受跨模态对应质量和样本构造影响，语义一致性也不必然带来精细几何边界。稠密特征迁移能够提供更细粒度的点或区域监督，却依赖多视图覆盖、相机位姿和投影精度，二维预测误差可能进一步传递至三维空间。近年来，两类方法呈现融合趋势，核心是在开放词汇泛化与局部几何精度之间取得平衡。表5概括代表方法的任务层级、迁移机制、三维表示和标注需求。

表5　基础模型驱动的3D开放词汇理解方法对比

Table 5　Comparison of foundation model-driven 3D open-vocabulary understanding approaches

方法	任务层级	关键迁移机制	2D基础模型	3D表示	是否需3D标注
(a)对比学习对齐相关方法					
PointCLIP	物体级分类	多视图投影与CLIP特征匹配	CLIP	点云	否
ULIP	通用3D表征	图像—文本—点云对比预训练	CLIP	点云	否
ULIP-2	通用3D表征	自动文本生成与三模态对比预训练	CLIP + LLM	点云	否
PointCLIP V2	物体级分类/分割	形状投影与3D文本提示匹配	CLIP + GPT	点云	否
(b)像素/特征迁移相关方法					
OpenScene	场景级语义分割	多视图特征融合与3D知识蒸馏	CLIP, LSeg, OpenSeg	点云	否
3D-OVS	场景级开放词汇分割	CLIP/DINO特征蒸馏至NeRF	CLIP, DINO	NeRF	否
ConceptFusion	开放集多模态建图	像素级特征融合与SLAM建图	CLIP, DINO, AudioCLIP	点云地图	否
RegionPLC	场景级语义/实例分割	区域语言监督与点—语言对比学习	Detic, OFA, GRiT, Kosmos-2, CLIP	点云	base类需标注

注：表中按主要知识迁移路径分组，部分方法兼具不同路线的技术机制。PointCLIP系列主要利用预训练CLIP进行跨模态特征匹配，不涉及额外的3D对比学习训练；RegionPLC利用二维基础模型构建区域语言监督，并通过对比学习实现点云特征对齐。“是否需3D标注”指相应方法主要训练配置中的3D任务标注需求。

© 中国图象图形学报版权所有

2.1.3　实例分割与全景分割的拓展

在场景级语义分割之外，研究进一步将开放词汇识别扩展至对象级实例分割，以区分同类场景中的不同对象。现有方法的主要差异在于二维基础模型介入三维实例分割的环节。OpenMask3D（Takmaz等，2023）先利用3D实例分割模型生成类别无关的3D实例掩码，再融合多视角CLIP特征，并利用SAM生成的2D掩膜辅助特征提取，实现3D实例与开放词汇文本查询的精准匹配。Open3DIS（Nguyen等，2024）将2D掩膜引导由特征提取阶段扩展至候选实例生成阶段。通过Grounding-DINO（Liu等，2024）提供开放词汇检测，SAM生成精细掩膜，再与3D候选实例融合，以补充小尺度或几何形态模糊的对象。两者均依赖真实多视角图像。与之不同，OpenIns3D（Huang等，2024）在推理阶段仅需三维场景输入，通过虚拟视角生成合成图像，再利用二维开放词汇模型提取语义特征并映射回三维实例。但其候选实例模块仍依赖实例掩膜监督训练，因此三维输入不等同于训练过程无需实例标注。

在此基础上，PanoGS（Zhai等，2025）将3D高斯溅射（3D Gaussian Splatting，3DGS；Kerbl等，2023）引入开放词汇3D全景分割任务。该方法构建与CLIP文本空间对齐的高斯语言特征场，并利用图切割及SAM多视角掩膜约束，将语义相近的高斯点聚合为连贯的全景分割单元。由于全景分割与实例分割的任务目标和评测协议不同，二者结果不宜直接比较。

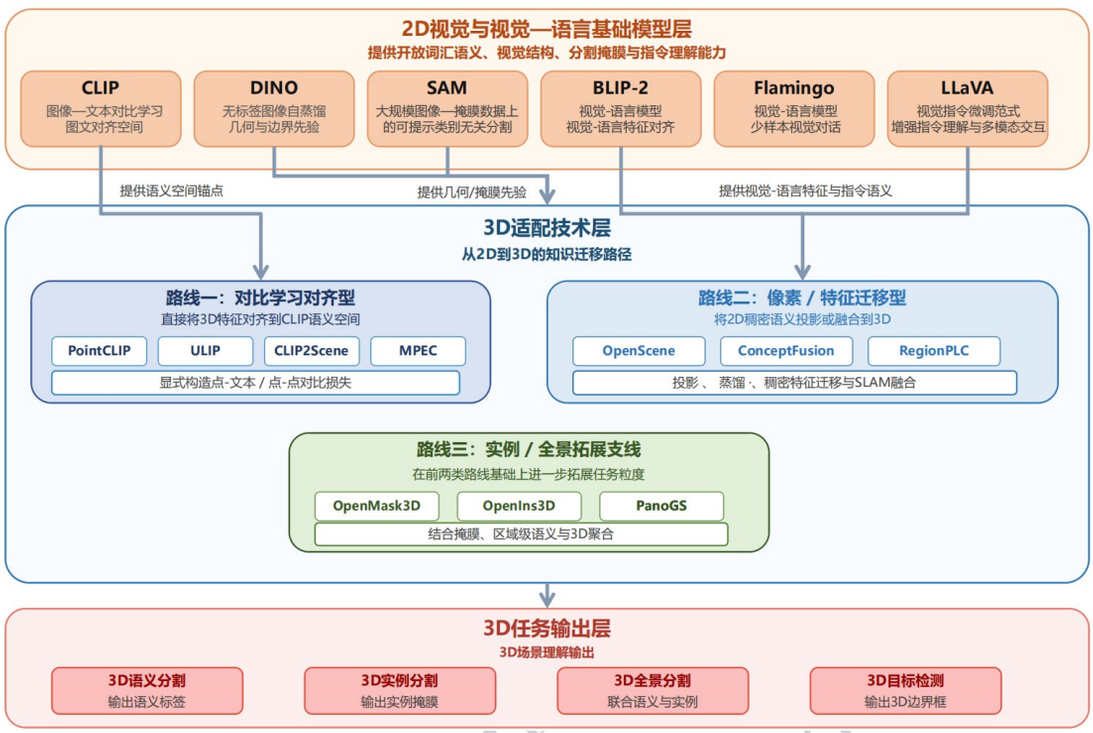


图2　二维基础模型赋能开放词汇三维感知的关系与分类

Fig. 2　Relationship and taxonomy of 2D foundation models for open-vocabulary 3D perception

2.2　3D目标检测

开放词汇3D目标检测沿用跨模态语义迁移思路，但进一步关注新类别候选框发现与三维定位。CLIP2（Zeng等，2023）是开放词汇3D检测的跨模态预训练基础工作，它直接构建文本-图像-点云三元组进行跨模态对比学习，以CLIP的语义空间为对齐基准，通过密集的3D点云-文本对比相似度直接产生物体候选热力图。在此基础上，CoDA（Cao等，2023）将新物体发现与分类统一在同一框架中，通过结合3D几何先验和2D开放词汇语义先验生成新类别伪框，并利用类别无关和类别判别式跨模态对齐，使3D点云特征与图像/文本语义空间逐步对齐。CoDAv2（Cao等，2025）利用3D几何先验与CLIP语义先验发现新物体，并通过在线数据池增强新物体分布。INHA（Jiao等，2024）将对齐扩展到实例、类别和场景三个层级。其中场景级对齐通过PISE模

© 中国图象图形学报版权所有

块将场景内物体特征聚合并与CLIP生成的场景描述文本对比，捕捉物体共现关系。

表6汇总了3D-CLIP、CoDA、INHA和CoDAv2在SUN RGB-D和ScanNetv2上的检测结果。总体来看，开放词汇三维目标检测已由基础跨模态语义对齐，逐步发展为结合新类别发现、几何定位与多粒度特征对齐的技术体系。

表6　开放词汇3D目标检测代表性方法性能比较

Table 6　Quantitative comparison of representative open-vocabulary 3D object detection methods

方法	输入	SUN RGB-D mAP_Novel/%	SUN RGB-D mAP_Base/%	ScanNetv2 mAP_Novel/%	ScanNetv2 mAP_Base/%
3D-CLIP(CoDA baseline)	Image + Point Cloud	3.61	30.56	3.74	14.14
CoDA	Point Cloud	6.71	38.72	6.54	21.57
INHA	Point Cloud	8.91	42.17	7.79	25.10
CoDAv2	Point Cloud	9.17	42.04	9.12	23.35

注：mAP采用IoU=0.25的评价标准。SUN RGB-D划分为10个基础类别和36个新类别，ScanNetv2划分为10个基础类别和50个新类别。结果整理自CoDA原论文表4、INHA原论文表1—2及CoDAv2原论文表IX。3D-CLIP(CoDA baseline)与正文所述CLIP2(Zeng等，2023)指同一方法；表中保留CoDA原文中的基线名称。

3　3D语言驱动理解

3.1　3D视觉定位

随着3D视觉定位领域的兴起，ScanRefer(Chen等，2020)于2020年首次提出了在RGB-D扫描中使用自然语言进行3D物体定位的任务。发展至2023年，M3DRef-CLIP(Zhang等，2023)通过对比学习将3D物体特征与CLIP的文本/图像空间对齐，并利用CLIP的图像编码器提取多视图2D特征辅助定位。3D-LLM(Hong等，2023)进一步将定位由特征匹配转化为语言模型的序列生成：它以冻结2D VLM为视觉骨干，将多视图语义特征聚合为点云特征，并通过3D位置嵌入和离散化AABB坐标词元，使LLM能够自回归输出3D边界框。由此，M3DRef-CLIP侧重跨模态特征对齐，3D-LLM则将空间坐标纳入语言生成过程，代表了两种不同的定位建模方式。

自然语言本身的空间歧义性研究者进一步显式建模对象关系。LEO(Huang等，2024)采用物体级3D表示与显式空间关系编码，将3D视觉定位与导航、操作等任务统一为序列预测。该方法使用PointNet++提取点云特征，再经空间Transformer模块显式建模物体对之间的欧氏距离和角度，并将这些几何偏置注入自注意力机制。同期，Chat-Scene(Huang等，2024)则从语言指代接口入手，通过显式物体标识符消除自然语言的空间歧义。该方法为场景中每个物体分配一个唯一的可学习词元，使LLM能够直接引用和定位物体，并利用Uni3D(Zhou等，2024)和DINOv2(Oquab等，2024)提取点云及图像特征，构建物体级嵌入。两者均试图缓解指代歧义，但LEO侧重以显式几何关系约束推理，Chat-Scene侧重以对象标识符和属性/关系词元增强语言引用能力。

随着定位任务扩展，模型幻觉与过拟合成为新的瓶颈。Robin3D(Kang等，2025)提出RIG数据引擎，通过混合正负样本构造对抗性指令，并利用ChatGPT生成多样化指令，构建包含100万条数据的训练集，以鲁棒指令微调缓解模型幻觉与过拟合。

三为降低三维标注和专用训练需求，研究者进一步探索零样本三维定位。LERF(Kerr等，2023)将CLIP嵌入到NeRF中，CLIP的开放语义对齐能力使LERF无需区域提议或掩膜即可在完成离线场景优化后，可交互式响应文本查询和生成3D相关性热力图。SeeGround(Li等，2025)通过视角适配模块动态选择与文本查询最相关的3D场景渲染视图，并将其与空间文本描述融合后输入VLM，使模型能够在不经过任何3D训练的情况下完成定位。VLM-Grounder(Xu等，2025)通过动态拼接图像序列、迭代式定位-反馈修正、以及多视角预测的3D投影集成，让VLM在2D空间中完成定位后融合出3D边界框。SPAZER(Jin等，2025)采用渐进式推理，将空间候选

© 中国图象图形学报版权所有

筛选与二维视觉语义验证相结合；UniGround（Zhang等，2026）则将几何构建与语义推理解耦，利用无需训练的超点分割和几何聚合构建三维场景，并通过SAM、感知编码器和VLM完成语义理解与定位。总体而言，该方向呈现出基础模型由语义特征提供者向场景理解与定位推理引擎演进的趋势，但不同方法对离线场景优化和上游三维候选生成的依赖仍存在差异。

3D视觉定位还需兼顾多模态融合与计算效率。。M3Fusion（毛潇涵等，2026）以冻结的多模态大语言模型（multimodal large language model，MLLM）为核心推理模块，将多视图图像语义与点云几何特征融合为紧凑的M3词元。MLLM对M3词元进行推理并生成位置词元，再由轻量Transformer解码器回归3D边界框，该方法体现了通过紧凑多模态表示减少中间候选依赖的思路。

总体而言，现有3D视觉定位方法在精度、泛化与效率之间存在权衡。经3D任务训练或适配的方法能够利用点云几何与显式空间关系实现精准定位，但依赖3D标注和跨模态训练，易受数据分布限制。基于预训练2D VLM的零样本方法具有开放词汇与语言推理优势，但依赖多视图或2D预测间接恢复三维信息，易受遮挡和投影误差影响。M3Fusion等方法则通过紧凑多模态表示降低特征构建开销。总体上，该领域仍需兼顾几何精度、泛化能力与部署效率。表7汇总了代表性基础模型驱动的3D视觉定位方法在ScanRefer上的公开结果。本文按照评测设置对结果进行分组展示，相关数值仅用于反映不同技术路线的研究进展，不作跨协议直接排序。

表7　不同评测设置下代表性3D视觉定位方法的公开定量结果

Table 7　Reported results of representative 3D visual grounding methods under different evaluation settings

方法	方法类型	主要基础模型	3D定位任务训练	评测设置	Acc@0.25/%	Acc@0.5/%
ScanRefer验证集常规评测						
M3DRef-CLIP	专用3DVG	CLIP	是	val，Overall	51.9	44.7
3D-LLM	3D-LLM	BLIP-2 FlanT5-XL	是（ScanRefer微调）	val	30.3	—
Chat-Scene	3D-LLM	Uni3D + DINOv2 + Vicuna-7B	是	val，Overall	55.5	50.2
Robin3D	3D-LLM	Vicuna-7B v1.5等	是（指令微调）	val，Overall	60.8	55.1
SeeGround	Zero-shot VLM	Qwen2-VL-72B	否	val，Overall	44.1	39.4
UniGround	Training-free VLM	VLM + 2D基础模型	否	val，Overall	46.1	34.1
250样本子集评测						
VLM-Grounder	Zero-shot VLM	GPT-4o	否	val，250条查询	51.6	32.8
SPAZER	Zero-shot VLM	GPT-4o	否	val，250条查询	57.2	48.8
Top-10评价协议						
M3Fusion	端到端MLLM	LLaVA-1.5-7B	是（三阶段训练）	val，Top-10	74.07	49.84

注：表中结果均来自ScanRefer，Acc@0.25和Acc@0.5分别表示预测框与真实框的IoU达到0.25和0.5阈值的样本比例，数值取自各方法原论文。‘—’表示原论文未报告相应指标。由于训练条件、候选生成方式及评测协议存在差异，表中结果不作直接排序。

3.2　3D问答与空间推理

3D场景语言理解的早期任务包括3D密集描述，如Scan2Cap（Chen等，2021）；随后，研究进一步扩展至自由形式问答ScanQA（Azuma等，2022）和情境化问答SQA3D（Ma等，2023）。

3.2.1　3D大语言模型的探索与多任务统一

2023年，研究者将大语言模型正式融入3D场景理解领域，形成多种问答与空间推理路线。

Chat-3D（Wang等，2023）针对3D-文本数据稀缺的核心问题，借助GPT-4等模型自建了对象中心指

© 中国图象图形学报版权所有

令数据集，将复杂场景问答转化为围绕物体标识符展开的上下文问答。PointLLM（Xu等，2024）专注于打磨大语言模型理解单个彩色物体的能力。它利用GPT-4自动生成73万条点-文本指令数据，使LLM能够对3D物体进行分类和生成物体描述。Chat-Scene（Huang等，2024）针对每个物体学习属性感知令牌和关系感知令牌，使LLM能够在对话中自由引用多个物体并理解它们之间的空间关系，从而突破了Chat-3D一次只能处理一个物体的局限。3D-LLM（Hong等，2023）重点将空间位置纳入语言模型输出：除加入3D位置编码外，还将AABB角点坐标离散为位置词元，使定位任务转化为序列生成问题。

上述模型提升了3D场景问答与理解能力，但训练和推理成本限制了实时部署。3DMIT（Li等，2024）与LL3DA（Chen等，2024）分别从训练效率与计算效率两个维度给出了不同的解答。3DMIT主要降低训练适配成本：将全局场景与物体特征拼接为提示，利用75K指令数据对LLM进行LoRA微调，避免复杂的3D—文本对齐。LL3DA则侧重输入与执行效率，直接处理原始点云，以置换不变的3D编码器提取空间特征，并采用OPT-1.3B作为生成器，减少2D投影的信息损失和计算开销。3DMIT证明了训练阶段的减法同样有效，而LL3DA则展示了直接的感知源头具有潜力。

3.2.2　面向3D场景的空间推理能力强化

首先需要在模型层面注入3D空间知识。SpatialVLM（Chen等，2024）通过构建包含20亿个3D空间推理样本的数据集训练VLM，使其具备距离估计与定性空间推理能力。SpatialReasoner（Ma等，2025）将CLIP与SAM等基础模型作为显式3D表征的构建工具重建出统一3D空间表示，使LLM能在3D感知、计算与推理全流程中以显式方式工作。GR3D（Cheng等，2026）以定位视觉语言模型为基础模型，在单一框架内集成了显式2D定位、隐式2D定位与单目3D定位三种互补能力，并将3D检测器作为定位先验的归纳偏置。

在强化基础能力的同时，低成本的轻量化方案也受到关注。cdViews（Wang等，2025）以预训练2D VLM为基础模型，通过自动筛选关键且多样的2D视角，使VLM在不需要任何3D-语言对齐或微调的情况下零样本解决3D问答任务。CoV（Chain-of-View，Zhao等，2026）进一步引入由粗到细的主动视角探索，让VLM在3D空间中主动调整观察角度。3Dprimitives（Liu等，2026）利用VLM本身的代码生成能力，将可执行代码描述的3D几何图元作为空间推理的中间语言，实现了高效零样本3D问答。

更高级的空间推理能力依赖于与人类认知对齐的训练策略。ShapeLLM（Qi等，2024）以经多视图蒸馏改进的3D编码器ReCon++和LLaMA大语言模型为核心，通过6自由度部件定位将空间推理任务纳入统一序列生成框架。SpatialThinker（Batra等，2025）以预训练多模态大语言模型为基础模型，通过构建场景图与设计密集空间奖励，利用强化学习训练模型进行结构化空间感知与多步推理。Loc3R-VLM（Qu等，2026）以预训练2D VLM为基础模型，通过全局布局重建与显式情境建模两个联合目标，使模型从单目视频中学习构建全局认知地图并锚定自我视角位置。

3.2.3　通用化与高效多模态统一框架

近几年，3D问答与空间推理领域逐渐向通用化、高效化、实时化发展，涌现出多个能够适配多场景、低开销、高响应的统一框架。Scene-LLM（Fu等，2025）复用CLIP作提取语义特征，通过轻量投影和混合点—体素表示压缩视觉词元，侧重降低跨模态对齐及输入表示开销。3UR-LLM（Xiong等，2025）直接以原始3D点云为输入，并通过3DETR提取物体级特征，并压缩为紧凑的查询令牌，减少2D投影和渲染步骤，侧重原生3D输入下的执行效率。与上述两者不同，ORYX（Liu等，2025）提出支持任意分辨率的视觉编码器和动态压缩模块，根据视频、多视图场景或精细图像的需求调整压缩率，重点平衡显存占用与细节保留。。

3D问答与空间推理逐步扩展到任务统一、显式空间推理和高效多模态建模等方向。总体来看，基于3D表征的方法具有较强的空间信息保留能力，但通常依赖跨模态对齐和专用训练。引入场景结构、空间先验或认知地图的方法能够增强几何关系推理，但对前端感知和结构化表示质量较为敏感。复用2D VLM及轻量化建模方案则有利于降低训练与部署成本，但在精细三维关系建模方面仍存在一定局限。因此，现有方法的差异更多体现为任务统一程度、空间信息完整性与计算效率之间的权衡。表8汇总了代表性基础模型驱动3D问答方法在ScanQA和SQA3D上的公开定量结果，表中结果主

© 中国图象图形学报版权所有

要用于反映各方法的任务表现，不作为严格横向排序依据。

表8　基础模型驱动的3D问答方法公开定量结果汇总

Table 8　Reported quantitative results of foundation model-driven 3D question answering methods

方法	方法类型	语言骨干	ScanQA EM@1/%	ScanQA METEOR/%	ScanQA CIDEr(×100)	SQA3D Avg. Acc/%
3D-LLM	3D-LLM	BLIP-2 FlanT5-XL	20.5	14.5	69.4	—
Chat-3D v2	3D-LLM	Vicuna-7B	21.1	16.1	77.1	—
LL3DA	3D-LLM	OPT-1.3B	—	15.9	76.8	—
Scene-LLM	3D-LLM	Llama-2-7B	27.2	16.6	80.0	54.2
3UR-LLM	3D-LLM	Flan-T5-XL	21.3	15.8	76.5	50.0
3UR-LLM+	3D-LLM	Vicuna-7B-v1.5	21.5	18.4	87.7	53.7

注：ScanQA结果对应验证集，采用EM@1、METEOR和CIDEr指标，METEOR以百分数表示，CIDEr为无量纲评分，表中数值按100倍尺度报告；SQA3D结果按原论文报告的评测设置汇总，采用平均问答准确率(Avg. Acc.)。“—”表示未报告相应结果。

3.3　3D场景图生成

3D场景图作为连接几何感知与高层语义的结构化表征，其生成方法的效率近年来因基础模型的引入而获得提升。早期工作如3DSSG(Wald等，2020)奠定了从点云回归图结构的数据基础与基本框架，但受限于封闭集标注，难以泛化至开放世界。后续研究主要从三个方向利用基础模型突破这一局限。

其一，直接注入视觉-语言知识以支持开放词汇生成。CCL-3DSGG(Chen等，2024)利用CLIP设计跨模态对比学习框架。文本3D场景图对比学习和多视图图像对比学习将文本解析与3D场景图结构对齐，实现了对新物体和新关系的零样本识别。Lang3DSG(Koch等，2024)则进一步采用语言对比预训练。通过CLIP文本编码器将文本三元组与预测的图特征对齐，将语言模型的知识蒸馏到图神经网络中。

其二，将基础模型作为核心推理与规划引擎。Concept-Graphs(Gu等，2024)利用SAM和CLIP从RGB-D序列中提取物体实例，并通过多视图融合生成带视觉特征的3D物体节点，借助GPT-4推理空间关系，无需训练即可通过多视图关联构建开集、物体中心的3D场景图。3DGraphLLM(Zemskova等，2025)利用CLIP和Uni3D提取物体的多模态特征，并借助VL-SAT(Wang等，2023)生成物体间的语义关系特征，通过可学习标识符与子图采样将场景图压缩为LLM可处理的词元序列。FOUND-IT(Maggio等，2026)进一步面向在线场景理解，利用几何基础模型从单目视频流中构建层级化场景图，并结合CLIP与SAM 3(Carion等，2026)的开放词汇查询和分割能力，动态调整场景图粒度。

其三，借助基础模型扩展场景图的结构与生成能力。与上述从3D场景中提取场景图的方向相反，GraphDreamer(Gao等，2024)探索了逆向任务。它首先借助ChatGPT将自然语言描述转换为场景图，再以该图为结构化条件，利用预训练的文本-图像扩散模型优化每个物体的神经符号距离函数(signed distance function，SDF)表示。范文刚等人则综合视觉-语言模型与大语言模型，利用VLM增强上下文语义理解并引导LLM推理生成超越空间关系的高层次语义超边，将传统二元关系扩展为语义超图。(范文刚等，2026)。

综上，基础模型已从辅助工具演变为核心引擎，推动3D场景图生成从封闭集、单任务走向开放词汇、实时构建与高阶语义建模的新阶段。

4　世界模型与动态预测

4.1　3D语义占用估计及其扩展

当前3D占用估计主要恢复场景的几何、语义和占用状态，是动态预测的状态表征基础，但本身不等同于未来状态预测。POP-3D(Vobecky等，2023)将2D图像的CLIP像素特征投影到3D空间中的对应体素上，并以此作为训练监督信号，实现了零样本语义

© 中国图象图形学报版权所有

占用估计。GaussTR(Jiang等,2025)采用自监督对齐范式,用Transformer预测稀疏3D高斯基元表示场景。它借助可微高斯泼溅渲染成2D特征图,强制与预训练2D基础模型的特征对齐。LOMA(Cui等,2025)引入显式语言先验,VL感知场景生成器利用VLM为每个3D位置生成富含语义的3D语言特征,并主动注入感知网络,帮助模型区分视觉相似但语义不同的物体。

另一类工作不再局限于知识迁移,而是直接以LLM或MLLM为核心引擎,扩展其3D感知与生成能力。OccLLaMA(Wei等,2024)将连续3D语义占用特征经场景词元器编码,并通过可学习码本进行向量量化,转换为LLM可处理的离散场景词元。DrivePI(Liu等,2025)基于MLLM设计,融合点云、多视角图像和语言指令,同时并行输出3D占用、占用流、轨迹和文本。OccVLA(Liu等,2025)采用逆向监督思路,在VLM中间层插入可学习占用查询,训练时强制模型从2D图像预测3D占用。

总体而言,基础模型通过语义迁移、自监督特征和语言先验增强了三维占用表征能力,但其效果仍受2D—3D映射精度和底层几何建模约束。随着占用表示与时间、动作及规划过程结合,研究逐步由当前状态估计扩展至动态预测与决策支持,但也增加了模型复杂度和训练成本。需要指出,当前占用估计并不等同于未来状态预测,只有进一步利用历史信息预测未来状态,或以候选动作为条件推演其后果,才能体现动态预测与动作条件世界建模能力。

4.2 4D场景演化与运动预测

在基础模型大规模应用于世界模型研究之前,已有大量研究依托自监督学习开展时空预测模型探索。UniWorld(Min等,2023)是几何驱动的代表性方法,但此类模型主要预测场景变化,缺少自然语言理解和交互能力,难以解释预测结果。

随着CLIP等视觉—语言模型和扩散生成模型的发展,研究者开始将跨模态语义表征与生成建模能力引入世界模型架构。GAIA-1(Hu等,2023)使用T5编码文本、DINO蒸馏图像词元,并将图像、文本和动作统一为离散词元序列,通过自回归Transformer预测未来,再由视频扩散解码器生成视频。DriveDreamer(Wang等,2024)将CLIP编码的文本描述和三维框类别信息作为语义条件,通过交叉注意力机制注入扩散模型的U-Net,使生成过程能够响应自然语言指令与场景条件。

前述生成式世界模型虽实现文本可控生成能力,但整体依旧沿用感知-预测-生成的级联式架构,基础模型大多作为辅助特征嵌入模块存在。EMMA(Hwang等,2025)直接将MLLM确立为自动驾驶系统的核心主体,所有传感器输出均统一表示为自然语言文本,通过视觉问答让MLLM直接学习驾驶环境的空间动态与决策逻辑。基础模型驱动的4D场景建模正由跨模态条件生成逐步向统一时空表征与高层决策推理发展。

GAIA-1、DriveDreamer和EMMA等自动驾驶世界模型主要面向道路场景,其多模态表征、时序预测与条件生成机制可为室内具身任务提供方法参考。

综合表9可以看出,相关研究正由几何驱动的自监督时空预测,逐步扩展至基础模型辅助的条件生成以及以MLLM为核心的统一建模。不同路线并非简单替代关系:特征对齐与语言先验方法更侧重提升三维语义表征,生成式世界模型侧重未来场景演化建模,而MLLM统一模型进一步尝试连接感知、预测与决策。与此同时,各路线在几何精度、生成效率、动作闭环能力及跨场景迁移方面仍存在不同约束。对于具身智能任务,已有研究进一步将基于环境动力学模拟与未来状态预演的世界模型视为具身预训练的重要发展方向(穆尧等,2026)。

5 具身智能体决策与交互

5.1 具身导航

具身导航要求智能体在真实3D环境中自主移动至目标位置。LM-Nav(Shah等,2023)将GPT-3、CLIP和ViNG三个预训练基础模型组合成一个导航系统,无需任何微调即可在真实世界中执行长时程户外导航指令,充分展示了预训练基础模型组合使用的潜力。

然而,模块间信息传递依赖人工设计的接口,缺乏统一的特征空间,限制了端到端优化与跨任务泛化。这一局限直接催生了统一模型的探索。NaviLLM(Zheng等,2024)通过基于模式的指令将视觉语言导航、定位、问答等多种任务统一为序列生成问题。同年,LEO(Huang等,2024)将视觉、语言与行动一体化建模,通过LLM接收第一人称2D图像、3D点云和文本指令,并输出离散动作词元。

© 中国图象图形学报版权所有

表9　三维状态表征与世界模型相关技术路线比较

Table 9　Comparison of technical approaches to 3D state representation and world modeling

技术路线	代表方法	基础模型的作用	状态表征或预测能力	主要特点与局限	适用范围
2D知识迁移与当前状态估计	POP-3D、GaussTR	提供预训练视觉语义特征，作为3D表示学习的监督或对齐目标	均以当前多视角观测估计三维占用或语义状态，不使用历史状态序列或动作条件，不预测未来场景	可利用大规模2D预训练知识缓解3D标注不足，但依赖2D－3D对应关系及跨视角对齐精度	静态或弱动态3D场景理解、当前语义占用估计
语言先验增强的占用建模	LOMA	提供语言语义先验，增强类别和场景语义区分	基于当前图像与语言先验估计三维占用，不使用历史观测或动作条件，不预测未来状态	有利于区分视觉相似但语义不同的类别，但精细几何质量仍受底层空间表示约束	当前语义占用估计、开放类别或长尾场景理解
场景词元化与自回归世界建模	OccLLaMA、GAIA-1	作为统一序列建模与生成核心，将场景、语言和动作纳入统一词元空间	OccLLaMA：利用历史占用预测未来场景，联合建模动作，但候选动作条件推演尚未明确验证。GAIA-1：基于历史视频与动作条件生成未来场景	能统一多模态信息并连接预测与任务生成，但离散化可能带来信息损失，且长序列建模成本较高	场景演化预测、生成式世界建模及任务条件预测
条件扩散式世界建模	DriveDreamer	作为文本或视觉条件编码器，为生成过程提供高层语义约束	基于初始场景与道路结构生成未来视频，支持动作条件控制；历史动作主要参与驾驶动作预测	具有较强的高保真生成与条件控制能力，但生成开销较大，且通常不直接形成完整动作闭环	自动驾驶场景生成、未来场景模拟与规划辅助
MLLM统一感知—预测—决策建模	DrivePI、EMMA、OccVLA	由外部知识源进一步转变为多模态理解、预测和任务推理的核心引擎	DrivePI：占用与占用流预测、轨迹生成；EMMA：未来轨迹规划；OccVLA：图像监督占用学习与轨迹规划，均未显式生成未来环境占用状态。	有利于连接语义理解、动态预测与决策过程，但训练与推理成本较高，多任务协同和空间精度仍存在约束	自动驾驶感知与轨迹规划，可为室内具身任务提供方法参考

在统一建模基础上，研究进一步深化导航能力。Open-Nav（Qiao等，2025）利用开源LLM，并通过时空链式推理分解指令理解、进度估计和决策制定，侧重以推理过程增强导航。DualVLN（Wei等，2025）将高层推理与低层动作解耦，采用快慢双系统架构，由VLM预测中程航点实现全局规划，并通过条件扩散策略生成局部连续轨迹，以二维像素航点和隐式目标特征连接高层推理与低层控制。Nav-R1（Liu等，2025）将感知、推理与行动统一于组相对策略优化（group relative policy optimization，GRPO）强化学习框架，通过包含11万条思维链轨迹的数据集和快慢双系统协同架构。PathPainter（Wang等，2026）则利用生成模型构建全局空间先验，根据鸟瞰图预测目标区域和可通行性掩膜，再通过A*搜索生成全局路径，并结合跨视角定位与局部规划实现导航执行。

真实世界中的导航智能体需要面对多样化的硬件平台和任务类型，传统方法难以在不同载体和任务间共享知识。NavFoM（Zhang等，2026）通过时间-视点指示符统一处理多相机输入，使大语言模型理解不同视角与时域，并采用预算感知时间采样动态采样历史帧。SocialNav（Chen等，2026）则关注传统导航较少建模的社会通行规则，构建脑模块和行动模块分层的架构：脑模块基于VLM实现社会可通行区域预测和链式推理，将社会规范编码为语义先验。行动专家采用条件流匹配生成符合社会规范的轨迹。

如图3所示，具身导航基础模型经历了从模块化组合到统一感知与决策，再到推理深化、生成式规划及社会规则理解的发展。模块化方法具有较好的灵活性和可替换性，但不同模块之间依赖人工接口，难以实现整体优化。统一模型能够加强跨任务知识共享，并提升复杂场景下的规划与决策能力，但训练

© 中国图象图形学报版权所有

成本、推理开销以及任务间干扰也随之增加。面向长时程、跨载体和社会导航等更复杂场景，如何兼顾动态适应性、生成可靠性、跨场景泛化与安全约束，仍是重要挑战。

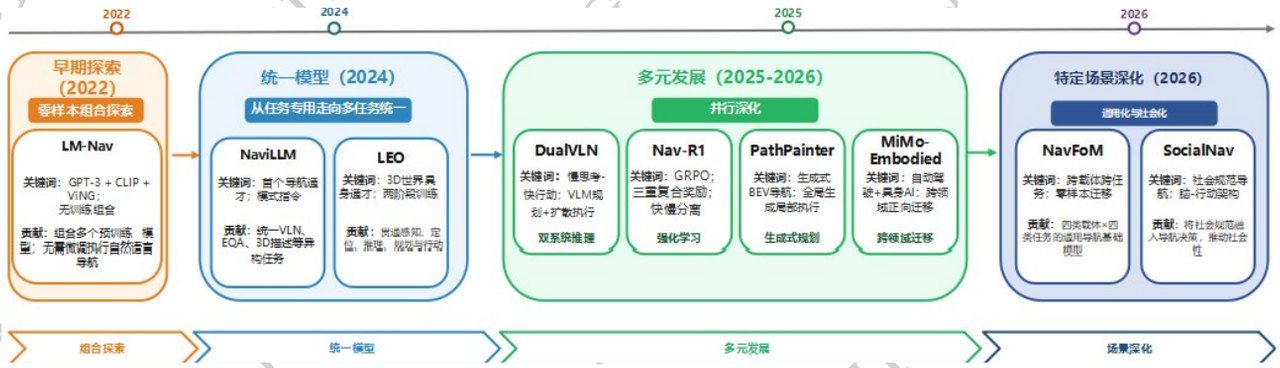


图3　具身导航模型演进图

Fig. 3　Evolution of embodied navigation models.

5.2　交互规划与任务分解

2023年，基础模型驱动的零样本规划取得进展。PaLM-E（Driess等，2023）将图像、机器人状态和3D场景编码后与文本交错输入预训练LLM，实现端到端的高层规划与推理。同期，VoxPoser（Huang等，2023）利用LLM的任务推理与代码生成能力，结合RGB-D观测在三维体素空间中构建可供性和空间约束价值图，并将其转化为运动规划目标与约束，生成六自由度（six degrees of freedom，6-DoF）操作轨迹。

零样本规划已验证LLM在任务分解中的可行性，但模型缺乏对3D场景的通用理解和结构化知识支撑。LEO（Huang等，2024）通过自回归序列预测统一处理2D图像、3D点云和文本指令，完成描述、问答、规划、导航和操作等多种任务。同年，Scene-MMKG（Song等，2024）利用LLM自动设计知识图谱模式，并结合CLIP提取物体属性、空间关系和可供性等多模态知识，构建了面向室内机器人操作与导航的知识图谱。

单一模型可处理多种任务，但跨场景、跨任务的泛化能力仍有瓶颈。2025年，涌现出多个强通用性的统一规划框架。Magma（Yang等，2025）是一个多模态智能体基础模型，通过标记组将可操作区域标记为数字标签，通过轨迹标记预测物体未来运动轨迹，将异构数据转化为统一的预训练语料。RoboBrain2.0（BAAI RoboBrain Team，2025）是面向具身智能的视觉-语言基础模型，统一了感知、推理与规划能力，支持可供性预测、空间指代定位及长时程任务规划。Anticipation-VLA（Zhang等，2026）提出基于预测性子目标的分层决策框架。高层模型根据当前观测生成未来视觉与文本子目标，并通过递归子目标栈组织长时程任务；底层VLA以当前子目标为条件生成连续操作动作，同时依据执行进度动态调整后续目标。不过，该方法的预测主要面向任务子目标，并不等同于显式三维环境状态的动态预测。PRISM（Aissi等，2026）提出感知与决策紧密耦合的框架，通过动态问答机制解决VLM生成任务无关描述的局限性。LLM根据VLM的初始描述提出针对性问题，VLM回答问题后将问答对合成为任务驱动的场景文本，决策模块基于该描述执行动作。

上述方法通过不同形式的空间信息支持行动决策，但其实际利用的空间状态、行动生成方式及反馈机制存在明显差异。表10从空间信息利用、行动输出、闭环反馈和实证依据四个维度，对代表性方法进行归纳。

6　挑战与未来方向

基础模型的出现极大推动了3D空间智能与具身决策领域的发展。但现有研究仍存在多项关键挑战。本节将从五个核心维度系统梳理领域现存问题，分别是2D依赖与3D原生能力、空间推理与幻觉抑制、泛化鲁棒性、效率与实时部署、评测基准与标准化。同时，本节结合现有研究短板，展望各方向的未来研究趋势。

目前，开放词汇3D场景理解大多以2D视觉-语

© 中国图象图形学报版权所有

表10　代表性具身决策方法的空间信息利用、行动输出及实证依据

Table 10　Spatial information utilization, action outputs, and empirical evidence of representative embodied decision-making methods

方法	实际空间状态或预测	行动生成与输出	闭环反馈	可验证影响及局限
LEO	三维任务使用点云与物体级表示；导航主要利用第一人称视觉信息	统一生成规划文本与离散动作词元	依具体任务设置，未统一验证在线反馈机制	在导航与操作基准上验证多任务能力，但显式三维表征的独立贡献尚不明确
DualVLN	当前及历史视觉观测、二维像素航点与隐式目标特征	高层空间目标引导扩散策略生成连续导航轨迹	异步双系统结合高频视觉反馈调整轨迹	R2R成功率64.3%；Social-VLN行人碰撞率35.4%；具备空间目标消融及模块耗时分析
PathPainter	鸟瞰图空间先验、目标区域与可通行性掩膜	A*搜索生成全局路径，局部规划器执行	跨视角定位与局部规划反馈	报告路径规划有效性、计算耗时及真实长距离导航实验；生成阶段仍存在计算开销
SocialNav	视觉场景、社会可通行区域及语义约束	生成符合社会规范的连续导航轨迹	基于环境观测进行闭环导航	闭环导航成功率86.1%，距离合规率82.5%；社会合规表现不等同于全面安全保证
VoxPoser	RGB-D观测构建的三维可供性、价值图及空间约束	三维价值图引导运动规划，生成6-DoF操作轨迹	利用最新观测进行闭环重规划	静态与扰动条件下成功率分别为88%、70%；具有规划方法对照实验，缺少统一安全性评估
Anticipation-VLA	当前视觉状态及预测的视觉、文本子目标	递归子目标引导VLA生成连续操作动作	根据任务进展动态调整子目标	消融验证子目标机制，但未证明完整三维环境预测的独立收益

注：表中空间信息包括显式三维状态、二维空间表征及预测性子目标，具体形式依方法而异。实证结果取自各方法原论文，涉及不同任务与评测条件，不进行跨方法数值排序。

言模型为核心知识来源，主要通过多视图投影算法，将2D语义特征迁移至3D空间。该研究存在两处难以规避的缺陷。一方面，2D特征向3D空间投影的过程，会不可避免地丢失场景几何信息与完整的空间结构特征。另一方面，2D视觉模型本身不具备对3D物理世界的原生认知能力。这一问题导致迁移后的3D语义特征，无法与真实的3D几何结构精准匹配。从数据层面来看，3D数据在规模、场景覆盖和标注质量上远少于2D数据，限制了3D基础模型的独立预训练。部分方法虽尝试直接输入3D数据，但合成视图或语义先验仍可能依赖2D模型。。因此，未来研究需重点突破三大方向，分别是新型3D自监督学习、大规模高质量3D预训练数据集构建以及高效轻量化的3D表征学习算法。

3D大语言模型的落地应用，提升了3D场景理解模型的空间推理能力。但该类模型同时暴露出突出的幻觉问题。模型在推理过程中可能生成场景中不存在的物体，或产生空间位置与布局关系判断错误。除视觉语言模型中常见的事实性偏差外，3D场景理解还突出表现为对象存在性误判、空间关系错误以及几何定位偏差等问题。现有研究已针对该问题开展探索，3D-VCD（Ogunleye等，2026）通过在推理阶段构建失真场景图进行视觉对比验证，缓解幻觉现象。但该方案依赖人工预定义的失真策略，难以全面覆盖属性类幻觉。并且，幻觉检测与幻觉抑制尚未形成统一的几何感知优化框架。未来研究需要聚焦几何感知驱动的幻觉治理机制，构建覆盖多类型幻觉的评测与抑制方法，在推理深度和事实准确性之间取得平衡。

开放词汇3D感知的核心优势是支持任意物体类别的识别任务。但当前主流模型的泛化能力仍存在明显短板，无法适配长尾分布数据、域外陌生场景和复杂开放世界的感知需求。现有公开3D数据集的采集场景较为单一，物体类别覆盖范围有限且数据集内长尾类别的标注样本数量极少。除此之外，2D与3D特征的跨模态对齐过程，极易受到标注噪

© 中国图象图形学报版权所有

声、场景遮挡、分辨率不匹配等因素的干扰。UniGround等无训练方法尝试降低对特定3D训练数据的依赖，但在极端光照、动态干扰和局部遮挡下仍可能退化。未来需建设场景与类别更丰富的3D—文本数据集，改进鲁棒跨模态对齐，并探索在线增量学习和自适应机制。

基础模型驱动的3D感知算法，普遍依赖超大模型参数与稠密算力计算，无法满足实际场景的实时性部署需求。在模型部署效率方面，现有研究逐步探索高效特征构建与实时渲染技术。M3Fusion通过多视角、多模态特征融合，将特征构建时间缩短至0.7 s；FMGS（Zuo等，2025）则利用3D高斯表示实现高效特征渲染，在480×270分辨率下，RGB图像与CLIP、DINO特征图的联合渲染速度达到103.4 FPS，为实时场景语义查询提供支持。但现有高效优化方案大多仅适配离线批次处理场景，无法满足动态流式场景的增量实时推理需求。

当前3D空间智能评测基准大多仅针对单一任务，且不同基准在传感器模态、数据集标注规范与评估指标缺乏统一，制约了跨场景、跨领域的泛化研究。尽管已有部分标准化探索，如UniOcc（Wang等，2025）统一3D占用预测评测标准，Anywhere3D-Bench（Wang等，2025）完善多层级视觉定位评估，IRef-VLA（Zhang等，2025）拓展不完美语言条件下的视觉定位评测，但仍局限于特定任务或交互设置。已有研究从感知、认知、决策到执行系统梳理具身智能评测指标，并提出“静态数据集—仿真平台—真实机器人”三级评测范式（李春一等，2026），但跨任务、跨平台的统一评测协议及结果可比性仍有待完善。未来研究需要搭建多任务、多场景适配的动态评测基准，引入细粒度的3D空间推理能力评估维度，并制定统一规范的大规模3D基础模型评测协议，以提升评估结果的公平性与可比性。

总体而言，3D场景理解正由2D知识迁移走向3D原生智能。未来需在数据层面降低标注成本、扩大场景覆盖，在模型层面兼顾感知精度、推理效率和部署可行性。3D原生基础模型、几何约束推理、高质量预训练数据和轻量化端侧部署的协同发展，将推动相关技术在真实具身场景中应用。

7 结 论

本文系统综述了基础模型驱动的三维空间智能与具身决策领域的研究进展，按照3D感知、语言驱动理解、世界模型与动态预测、具身智能体系统四层框架组织相关工作。

在开放词汇3D感知方面，对比学习对齐和像素/特征迁移构成2D知识迁移至3D空间的两条主要路线，任务已由语义分割扩展至实例、全景分割和目标检测。语言驱动理解由场景语义描述发展至3D定位、问答和空间推理，并向通用、高效的多模态框架演进。世界模型与动态预测由静态占用估计扩展至生成式场景建模和多任务预测；4D场景建模则由几何驱动逐步融合多模态基础模型。具身导航由模块组合走向统一模型，并拓展至跨载体和社会规范感知；交互规划则由零样本任务分解发展至通用规划和层次化决策。

贯穿上述演进的核心驱动力在于基础模型角色的持续演变。从早期作为知识源提供预训练特征，到作为统一接口整合多模态输入与多任务输出，再到作为推理引擎参与高层语义理解与决策生成，基础模型已深度嵌入3D场景理解的各个环节。与此同时，数据稀缺、模态鸿沟、推理幻觉、部署效率与评测体系等核心挑战尚未完全解决，制约着模型的应用。未来研究需在纯3D原生基础模型架构、几何感知的推理机制、大规模3D预训练数据构建、高效轻量化的模型部署方案以及标准化评测体系等方向上持续突破。

参考文献（References）

Aissi M S, Grislain C, Romac C, Soulier L, Chetouani M, Sigaud O, et al. 2026. PRISM: Perception reasoning interleaved for sequential decision making[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2605.05407

Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderhauf N, et al. 2018. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 3674-3683 [DOI: 10.1109/CVPR.2018.00387]

Armeni I, Sener O, Zamir A R, Jiang H, Brilakis I, Fischer M, et al.

© 中国图象图形学报版权所有

2016. 3D Semantic Parsing of Large-Scale Indoor Spaces//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 1534-1543 [DOI: 10.1109/CVPR.2016.170]

Azuma D, Miyanishi T, Kurita S and Kawanabe M. 2022. ScanQA: 3D question answering for spatial scene understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 19107-19117 [DOI: 10.1109/CVPR52688.2022.01854]

BAAI RoboBrain Team, Cao M Y, Tan H J, Ji Y H, Lin M L, Li Z Y, et al. 2025. RoboBrain 2.0 Technical Report[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2507.02029

Batra H, Tu H Q, Chen H, Lin Y Z, Xie C H and Clark R. 2025. SpatialThinker: Reinforcing 3D reasoning in multimodal LLMs via spatial rewards[EB/OL]. [2026-06-10]. https://spatialthinker.github.io

Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, et al. 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 9297-9307

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. 2020. nuScenes: A multimodal dataset for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]

Cao Y, Zeng Y H, Xu H and Xu D. 2023. CoDA: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection//Advances in Neural Information Processing Systems (NeurIPS), 36: 71862-71873 [DOI: 10.52202/075280-3145]

Cao Y, Zeng Y H, Xu H and Xu D. 2025. CoDAv2: Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3D object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47: 10475-10489 [DOI: 10.1109/TPAMI.2025.3593580]

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. 2021. Emerging properties in self-supervised vision transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 9630-9640 [DOI: 10.1109/ICCV48922.2021.00951]

Carion N, Gustafson L, Hu Y T, Debnath S, Hu R, Suris Coll-Vinent D, et al. 2026. SAM 3: Segment anything with concepts//Proceedings of the International Conference on Learning Representations (ICLR).

Chen B, Xu Z, Kirmani S, Ichter B, Sadigh D, Guibas L, et al. 2024. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 14455-14465

Chen D Z, Chang A X and Nießner M. 2020. ScanRefer: 3D object localization in RGB-D scans using natural language//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 202-221 [DOI: 10.1007/978-3-030-58565-5_13]

Chen D Z, Chang A X and Nießner M. 2021. Scan2Cap: Context-aware dense captioning in RGB-D scans//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 3192-3202 [DOI: 10.1109/CVPR46437.2021.00321]

Chen L, Wang X, Lu J, Lin S, Wang C and He G. 2024. CLIP-driven open-vocabulary 3D scene graph generation via cross-modality contrastive learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 27863-27873

Chen R N, Liu Y Q, Kong L D, Zhu X G, Ma Y X, Li Y K, et al. 2023. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 7020-7030 [DOI: 10.1109/CVPR52729.2023.00678]

Chen S J, Chen X, Zhang C, Li M S, Yu G, Fei H, et al. 2024. LL3DA: Visual interactive instruction tuning for omni-3D understanding, reasoning, and planning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 26418-26428 [DOI: 10.1109/CVPR52733.2024.02496]

Chen Y F, Wei M J, Wang X S, Liu Y X, Wang J Z, Song H, et al. 2025. Embodied AI: A survey on the evolution from perceptive to behavioral intelligence. SmartBot, 1(3): e70003 [DOI: 10.1002/smb2.70003]

Chen Z Y, Guo Y N, Chu Z D, Luo M H, Shen Y F, Sun M C, et al. 2026. SocialNav: Training human-inspired foundation model for socially-aware embodied navigation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 28796-28806

Cheng A C, Fu Y, Ji Y T, Zhu L G, Zhan G Q, Zhang Z Y, et al. 2026. Grounded 3D-aware spatial vision-language modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 16688-16700

Cui Y B, Li Z H, Wang J Q and Fang Z. 2025. LOMA: Language-assisted semantic occupancy network via triplane Mamba//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press: 2609-2617 [DOI: 10.1609/aaai.v39i3.32264]

Dai A, Chang A X, Savva M, Halber M, Funkhouser T and Nießner M. 2017. ScanNet: Richly-annotated 3D reconstructions of indoor scenes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 2432-2443

Driess D, Xia F, Sajjadi M S M, Lynch C, Chowdhery A, Ichter B, et

© 中国图象图形学报版权所有

al. 2023. PaLM-E: An embodied multimodal language model//Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR, 202: 8469-8488

Fan W G, Lan Y Q, Yi R J, Zhu C Y and Xu K. 2026. Hierarchical 3D scene graph generation via multimodal foundation models. Journal of Computer-Aided Design & Computer Graphics, 38(1): 90 - 102(范文刚,兰育青,易任娇,朱晨阳,徐凯. 2026. 多模态基础模型驱动的层次化3D场景图生成方法. 计算机辅助设计与图形学学报, 38(1): 90 - 102)[DOI: 10.3724/SP.J.1089.2025-00259]

Fu R, Wen Z H, Liu Z C and Sridhar S. 2025. Scene-LLM: Extending language model for 3D visual reasoning//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE: 2195-2206 [DOI: 10.1109/WACV61041.2025.00220]

Gao G, Liu W Y, Chen A P, Geiger A and Schölkopf B. 2024. GraphDreamer: Compositional 3D scene synthesis from scene graphs//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 21295-21304

Ghiasi G, Gu X, Cui Y and Lin T Y. 2022. Scaling open-vocabulary image segmentation with image-level labels//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 540-557 [DOI: 10.1007/978-3-031-20059-5_31]

Gu Q, Kuwajerwala A, Morin S, Jatavallabhula K M, Sen B, Agarwal A, et al. 2024. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE: 5021-5028 [DOI: 10.1109/ICRA57147.2024.10610243]

Guzhov A, Raue F, Hees J and Dengel A. 2022. AudioCLIP: Extending CLIP to image, text and audio//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE: 976-980 [DOI: 10.1109/ICASSP43922.2022.9747631]

Hong Y, Zhen H Y, Chen P H, Zheng S H, Du Y L, Chen Z F, et al. 2023. 3D-LLM: Injecting the 3D world into large language models//Advances in Neural Information Processing Systems (NeurIPS). Red Hook: Curran Associates: 20482-20494 [DOI: 10.48550/arXiv.2307.12981]

Hu A, Russell L, Yeo H, Murez Z, Fedoseev G, Kendall A, et al. 2023. GAIA-1: A generative world model for autonomous driving [EB/OL]. [2026-06-10]. https://arxiv.org/abs/2309.17080

Huang H F, Chen Y L, Wang Z H, Huang R J, Xu R S, Wang T, et al. 2024. Chat-Scene: Bridging 3D scene and large language models with object identifiers//Advances in Neural Information Processing Systems (NeurIPS). Red Hook: Curran Associates: 113991-114017 [DOI: 10.48550/arXiv.2312.08168]

Huang Z N, Wu X Y, Chen X, Zhao H S, Zhu L and Lasenby J. 2024. OpenIns3D: Snap and lookup for 3D open-vocabulary instance segmentation//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 169-185 [DOI: 10.1007/978-3-031-73033-7_10]

Huang J Y, Yong S L, Ma X J, Linghu X K, Li P H, Wang Y, et al. 2024. An embodied generalist agent in 3D world//Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, 235: 20413-20451

Huang W L, Wang C, Zhang R H, Li Y Z, Wu J J and Fei-Fei L. 2023. VoxPoser: Composable 3D value maps for robotic manipulation with language models//Proceedings of the 7th Conference on Robot Learning (CoRL). PMLR, 229: 540-562

Hwang J J, Xu R S, Lin H, Hung W C, Ji J W, Choi K, et al. 2025. EMMA: End-to-end multimodal model for autonomous driving. Transactions on Machine Learning Research. https://openreview.net/forum?id=kH3t5lmOU8

James S, Ma Z C, Arrojo D R and Davison A J. 2020. RLBench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters, 5(2): 3019-3026 [DOI: 10.1109/LRA.2020.2974707]

Jatavallabhula K M, Kuwajerwala A, Gu Q, Omama M, Iyer G, Saryazdi S, et al. 2023. ConceptFusion: Open-set multimodal 3D mapping//Proceedings of Robotics: Science and Systems XIX. Daegu, Republic of Korea: RSS Foundation [DOI: 10.15607/RSS.2023.XIX.066]

Jia B X, Chen Y X, Yu H Y, Wang Y, Niu X S, Liu T Y, et al. 2024. SceneVerse: Scaling 3D vision-language learning for grounded scene understanding//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 289-310 [DOI: 10.1007/978-3-031-72673-6_16]

Jiang H Y, Liu L, Cheng T H, Wang X J, Lin T W, Su Z Z, et al. 2025. GaussTR: Foundation model-aligned Gaussian transformer for self-supervised 3D spatial understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 11960-11970 [DOI: 10.1109/CVPR52734.2025.01117]

Jiao P K, Zhao N, Chen J J and Jiang Y G. 2024. Unlocking textual and visual wisdom: Open-vocabulary 3D object detection enhanced by comprehensive guidance from text and image//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 376-392 [DOI: 10.1007/978-3-031-73195-2_22]

Jin Z Q, Tu R C, Liao J, Sun W, Luo X, Liu S, et al. 2025. SPAZER: Spatial-semantic progressive reasoning agent for zero-shot 3D visual grounding//Advances in Neural Information Processing Systems (NeurIPS), 38. Red Hook: Curran Associates: 183499-183526 [DOI: 10.52202/085713-5518]

Kang W T, Huang H F, Shang Y Z, Shah M and Yan Y. 2025. Robin3D: Improving 3D large language model via robust instruction tuning//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 3905-3915

© 中国图象图形学报版权所有

［DOI：10.1109/ICCV51701.2025.00372］

Kerbl B，Kopanas G，Leimkühler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics，42(4)：139:1-139:14［DOI：10.1145/3592433］

Kerr J，Kim C M，Goldberg K，Kanazawa A and Tancik M. 2023. LERF：Language embedded radiance fields//Proceedings of the IEEE/CVF International Conference on Computer Vision（ICCV）. Piscataway：IEEE：19672-19682［DOI：10.1109/ICCV51070.2023.01807］

Kirillov A，Mintun E，Ravi N，Mao H，Rolland C，Gustafson L，et al. 2023. Segment anything//Proceedings of the IEEE/CVF International Conference on Computer Vision（ICCV）. Piscataway：IEEE：3992-4003［DOI：10.1109/ICCV51070.2023.00371］

Koch S，Hermosilla P，Vaskevicius N，Colosi M and Ropinski T. 2024. Lang3DSG：Language-based contrastive pre-training for 3D scene graph prediction//Proceedings of the International Conference on 3D Vision（3DV）. Piscataway：IEEE：1037-1047［DOI：10.1109/3DV62453.2024.00076］

Li B，Weinberger K Q，Belongie S，Koltun V and Ranftl R. 2022. Language-driven semantic segmentation//Proceedings of the 10th International Conference on Learning Representations（ICLR）. OpenReview. https://openreview.net/forum? id=RriDjddCLN

Li B，Zhang Y H，Guo D，Zhang R R，Li F，Zhang H，et al. 2024. LLaVA-OneVision：Easy visual task transfer［EB/OL］.［2026-08-29］.
https://arxiv. org/abs/2408.03326［DOI：10.48550/arXiv. 2408.03326］

Li C Y，Zhang J B，Xiao J H，Yan B W，Guo S Y，Ye T R，et al. 2026. Embodied intelligence model evaluation：from perception to execution. Journal of Image and Graphics，31(8)：2705-2730（李春一，张建博，肖嘉豪，闫博闻，郭晟毓，叶桐瑞，等. 2026. 具身智能模型评测：从感知到执行. 中国图象图形学报），31(8)：2705-2730［DOI：10.11834/jig.250550］

Li J，Li D，Savarese S and Hoi S. 2023. BLIP-2：Bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning（ICML）. PMLR，202：19730-19742

Li R，Li S X，Kong L H，Yang X and Liang J. 2025. SeeGround：See and ground for zero-shot open-vocabulary 3D visual grounding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway：IEEE：3707-3717

Li Z C，Yu H S，Ding Y H，Li Y，He Y and Akhtar N. 2026. Embodied intelligence for 3D understanding：A survey on 3D scene question answering. Information Fusion，126：103624［DOI：10.1016/j.inffus.2025.103624］

Li Z J，Zhang C，Wang X Y，Ren R L，Xu Y F，Ma R F，et al. 2024. 3DMIT：3D multi-modal instruction tuning for scene understanding［EB/OL］.［2026-06-10］.
https://arxiv.org/abs/2401.03201

Liu H T，Li C Y，Li Y H and Lee Y J. 2024. Improved baselines with visual instruction tuning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway：IEEE：26286-26296［DOI：10.1109/CVPR52733.2024.02484］

Liu H T，Li C Y，Wu Q Y and Lee Y J. 2023. Visual instruction tuning//Advances in Neural Information Processing Systems（NeurIPS）. Red Hook：Curran Associates：34892-34916［DOI：10.48550/arXiv.2304.08485］

Liu J Z，Qian K，Dubost F，Zhong K，Srinivasan A，Chen N，et al. 2026. 3D primitives are a spatial language for VLMs［EB/OL］.［2026-06-10］.
https://arxiv.org/abs/2605.12586

Liu K，Zhan F N，Zhang J H，Xu M Y，Yu Y C，El Saddik A，et al. 2023. Weakly supervised 3D open-vocabulary segmentation//Advances in Neural Information Processing Systems（NeurIPS），36. Red Hook：Curran Associates：53433-53456［DOI：10.52202/075280-2325］

Liu M H，Shi R X，Kuang K M，Zhu Y H，Li X L，Han S Z，et al. 2023. OpenShape：Scaling up 3D shape representation towards open-world understanding//Advances in Neural Information Processing Systems（NeurIPS）. Red Hook：Curran Associates：44860-44879［DOI：10.52202/075280-1944］

Liu Q X，Huang T，Zhang Z Y and Tang H. 2025. Nav-R1：Reasoning and navigation in embodied scenes［EB/OL］.［2026-06-10］.
https://arxiv.org/abs/2509.10884

Liu R，Wang W G，Wang J and Yang Y. 2026. Survey on embodied intelligence：advances in vision-language-action models. Journal of Image and Graphics，31(8)：2731-2773（刘锐，王文冠，王俊，杨易. 2026. 面向具身智能的视觉—语言—动作模型研究进展. 中国图象图形学报，31(8)：2731-2773）［DOI：10.11834/jig.250544］

Liu R X，Wu J F，Jiang Y J，Li H T，Chen Y Z and Zhang W. 2025. OccVLA：Vision-language-action model with implicit 3D occupancy supervision［EB/OL］.［2026-06-10］.
https://arxiv.org/abs/2509.05578

Liu S，Zeng Z，Ren T，Li F，Zhang H，Yang J，et al. 2024. Grounding DINO：Marrying DINO with grounded pre-training for open-set object detection//Proceedings of the European Conference on Computer Vision（ECCV）. Cham：Springer：38-55［DOI：10.1007/978-3-031-72970-6_3］

Liu Y B，Mu Y，Ye Q，Gao L，Han X G，Chen A P，et al. 2026. Frontier trends and top 10 advances in 3D vision in 2025. Journal of Image and Graphics，31(7)：2279-2316（刘烨斌，穆尧，叶琦，高林，韩晓光，陈安沛，等. 2026. 2025年度三维视觉前沿趋势与十大进展. 中国图象图形学报，31(7)：2279-2316）［DOI：10.11834/jig.260114］

Liu Z，Huang R H，Yang R，Yan S M，Wang Z N，Hou L，et al. 2025. DrivePI：Spatial-aware 4D MLLM for unified autonomous driving understanding，perception，prediction and planning［EB/OL］.

© 中国图象图形学报版权所有

[2026-06-10].
https://arxiv. org/abs/2512.12799 [DOI: 10.48550/arXiv. 2512.12799]

Liu Z Y, Dong Y H, Liu Z W, Hu W, Lu J W and Rao Y M. 2025. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution//Proceedings of the International Conference on Learning Representations (ICLR)[EB/OL]. [2026-10-26]. https://openreview.net/forum?id=ODiY6pbHZQ

Ma W, Chou Y C, Liu Q H, Wang X R, de Melo C, Xie J W, et al. 2025. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning//Advances in Neural Information Processing Systems (NeurIPS). Red Hook: Curran Associates: 156117-156140 [DOI: 10.52202/085713-4703]

Ma X J, Yong S L, Zheng Z L, Li Q, Liang Y T, Zhu S C, et al. 2023. SQA3D: Situated question answering in 3D scenes//Proceedings of the International Conference on Learning Representations (ICLR) [EB/OL]. [2026-10-26]. https://openreview.net/forum?id=IDJx97BC38

Ma X Z, Bhalgat Y, Smart B, Chen S, Li X H, Ding J, et al. 2024. When LLMs step into the 3D world: A survey and meta-analysis of 3D tasks via multi-modal large language models[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2405.10255

Ma Y, Song Z, Zhuang Y, Hao J and King I. 2026. A survey on vision-language-action models for embodied AI. IEEE Transactions on Neural Networks and Learning Systems, 37(7): 3031-3051 [DOI: 10.1109/TNNLS.2025.3650584]

Maggio D, Gorlo N, Hauser K and Carlone L. 2026. FOUND-IT: Foundation-model-first task-driven 3D scene graphs with granularity on demand[EB/OL]. [2026-06-10]. https://arxiv. org/abs/2605.25371v2 [DOI: 10.48550/arXiv. 2605.25371]

Mao X H, Zhu C M, Wang T and Pang J M. 2026. M3Fusion: A unified multi-view multi-modality fusion framework for embodied 3D perception. Scientia Sinica Informationis, 56(2): 312-326 (毛潇涵, 朱宸铭, 王泰, 庞江淼. 2026. M3Fusion: 面向具身3D感知的统一多视角多模态融合框架. 中国科学: 信息科学), 56(2): 312-326 [DOI: 10.1360/SSI-2025-0293]

Mees O, Hermann L, Rosete-Beas E and Burgard W. 2022. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters, 7(3): 7327-7334 [DOI: 10.1109/LRA.2022.3180108]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Min C, Zhao D W, Xiao L, Nie Y M and Dai B. 2023. UniWorld: Autonomous driving pre-training via world models[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2308.07234

Mu Y, Zhao H, Hu R Z, Zhang L, Li H Y, Yang J L, et al. 2026. Frontiers and prospects of embodied AI: evolution of data, models, and systems. Journal of Image and Graphics, 31(6): 2017-2025 (穆尧, 赵昊, 胡瑞珍, 张力, 李弘扬, 杨蛟龙, 等. 2026. 具身智能前沿展望: 数据、模型与系统演进. 中国图象图形学报), 31(6): 2017-2025 [DOI: 10.11834/jig.260059]

Nguyen P, Ngo T D, Kalogerakis E, Gan C, Tran A, Pham C, et al. 2024. Open3DIS: Open-vocabulary 3D instance segmentation with 2D mask guidance//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 4018-4028 [DOI: 10.1109/CVPR52733.2024.00385]

Ogunleye M, Abdelrahman E and Lourentzou I. 2026. 3D-VCD: Hallucination mitigation in 3D-LLM embodied agents through visual contrastive decoding[EB/OL]. [2026-06-10]. https://arxiv. org/abs/2604.08645 [DOI: 10.48550/arXiv. 2604.08645]

Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024: 1-31

Peng S, Genova K, Jiang C M, Tagliasacchi A, Pollefeys M and Funkhouser T. 2023. OpenScene: 3D scene understanding with open vocabularies//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 815-824 [DOI: 10.1109/CVPR52729.2023.00084]

Peng Z, Wang W H, Dong L, Hao Y R, Huang S H, Ma S M, et al. 2024. Grounding multimodal large language models to the world//Proceedings of the 12th International Conference on Learning Representations (ICLR). https://proceedings.iclr.cc/paper_files/paper/2024/hash/e112a4671e8779aa9f640a0e3f81bd26-Abstract-Conference.html

Qi Z K, Dong R P, Zhang S C, Geng H R, Han C R, Ge Z, et al. 2024. ShapeLLM: Universal 3D object understanding for embodied interaction//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 214-238 [DOI: 10.48550/arXiv.2402.17766]

Qiao Y Y, Lyu W Q, Wang H, Wang Z X, Li Z R, Zhang Y, et al. 2025. Open-Nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source LLMs//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE: 6710-6717.

Qu K, Qi H Z, Dusmanu M, Rad M, Wang R and Pollefeys M. 2026. Loc3R-VLM: Language-based localization and 3D reasoning with vision-language models[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2603.18002

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language

© 中国图象图形学报版权所有

supervision//Proceedings of the 38th International Conference on Machine Learning (ICML). PMLR, 139: 8748-8763

Ren R L, Zhao X Y, Xu W C, Cao J, Xu X X and Zhang X. 2025. A survey of language-grounded multimodal 3D scene understanding. Knowledge-Based Systems, 321: 113650 [DOI: 10.1016/j.knosys.2025.113650]

Rozenberszki D, Litany O and Dai A. 2022. Language-grounded indoor 3D semantic segmentation in the wild//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 125-141 [DOI: 10.1007/978-3-031-19827-4_8]

Shah D, Osinski B, Ichter B and Levine S. 2023. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action//Proceedings of the 6th Conference on Robot Learning (CoRL). PMLR, 205: 492-504

Shridhar M, Thomason J, Gordon D, Bisk Y, Han W, Mottaghi R, et al. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 10737-10746 [DOI: 10.1109/CVPR42600.2020.01075]

Song Y X, Sun P L, Liu H Y, Li Z X, Song W, Xiao Y H, et al. 2024. Scene-driven multimodal knowledge graph construction for embodied AI. IEEE Transactions on Knowledge and Data Engineering, 36 (11): 6962-6976 [DOI: 10.1109/TKDE.2024.3399746]

Straub J, Whelan T, Ma L, Chen Y F, Wijmans E, Green S, et al. 2019. The Replica dataset: A digital replica of indoor spaces[EB/OL]. [2026-08-29]. https://arxiv. org/abs/1906.05797 [DOI: 10.48550/arXiv. 1906.05797]

Takmaz A, Fedele E, Sumner R W, Pollefeys M, Tombari F and Engelmann F. 2023. OpenMask3D: Open-vocabulary 3D instance segmentation//Advances in Neural Information Processing Systems (NeurIPS), 36: 68367-68390 [DOI: 10.52202/075280-2989]

Tian X, Jiang T, Yun L, Mao Y, Yang H, Wang Y, et al. 2023. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving//Advances in Neural Information Processing Systems (NeurIPS), 36: 64318-64330

Vobecky A, Siméoni O, Hurych D, Gidaris S, Bursuc A, Pérez P, et al. 2023. POP-3D: Open-vocabulary 3D occupancy prediction from images//Advances in Neural Information Processing Systems (NeurIPS), 36. Red Hook: Curran Associates: 50545-50557 [DOI: 10.52202/075280-2198]

Wald J, Dhamo H, Navab N and Tombari F. 2020. Learning 3D semantic scene graphs from 3D indoor reconstructions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 3960-3969 [DOI: 10.1109/CVPR42600.2020.00402]

Wang F Y, Yu S C, Wu J W, Tang J H, Zhang H W and Sun Q R. 2025. 3D question answering via only 2D vision-language models//Proceedings of the 42nd International Conference on Machine Learning (ICML). PMLR, 267: 65310-65325

Wang P, Yang A, Men R, Lin J, Bai S, Li Z, et al. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR, 162: 23318-23340

Wang T X, Zhang Z F, Zhu Z Y, Fan Y, Xiong J, Li P X, et al. 2025. From objects to anywhere: A holistic benchmark for multi-level visual grounding in 3D scenes//Advances in Neural Information Processing Systems (NeurIPS). Red Hook: Curran Associates: 30457-30498 [DOI: 10.52202/085713-0905]

Wang X F, Zhu Z, Huang G, Chen X Z, Zhu J G and Lu J W. 2024. DriveDreamer: Towards real-world-drive world models for autonomous driving//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 55-72 [DOI: 10.1007/978-3-031-73195-2_4]

Wang Y, Jia B X, Zhu Z Y and Huang S Y. 2025. MPEC: Masked point-entity contrast for open-vocabulary 3D scene understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 14125-14136 [DOI: 10.48550/arXiv.2504.19500]

Wang Y J, Tian Y R, Huang X J, Gai W Q, Zhu M, Zhou X, et al. 2026. PathPainter: Transferring the generalization ability of image generation models to embodied navigation[EB/OL]. [2026-10-28]. https://arxiv.org/abs/2605.07496 [DOI: 10.48550/arXiv.260

Wang Y P, Huang X Y, Sun X K, Yan M X, Xing S, Tu Z Z, et al. 2025. UniOcc: A unified benchmark for occupancy forecasting and prediction in autonomous driving//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 25560-25570 [DOI: 10.48550/arXiv.2503.24381]

Wang Z, Cheng B, Zhao L, Xu D, Tang Y and Sheng L. 2023. VL-SAT: Visual-linguistic semantics assisted training for 3D semantic scene graph prediction in point cloud//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 21560-21569

Wang Z H, Huang H F, Zhao Y, Zhang Z A and Zhao Z. 2023. Chat-3D: Data-efficiently tuning large language model for universal dialogue of 3D scenes[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2308.08769

Wei J L, Yuan S S, Li P F, Hu Q D, Gan Z X and Ding W C. 2024. OccLLaMA: An occupancy-language-action generative world model for autonomous driving[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2409.03272

Wei M, Wan C Y, Peng J Q, Yu X Q, Yang Y Q, Feng D L, et al. 2025. DualVLN: Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2512.08186

Wu J, Wang J, Yang Z, Gan Z, Liu Z, Yuan J, et al. 2024. GRiT: A

© 中国图象图形学报版权所有

generative region-to-text transformer for object understanding//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 207-224

Xie N W, Tian Z Z, Yang L, Zhang X P, Guo M, Li J, et al. 2025. From 2D to 3D cognition: A brief survey of general world models [EB/OL]. [2026-06-10]. https://arxiv.org/abs/2506.20134

Xiong H M, Zhuge Y Z, Zhu J W, Zhang L and Lu H C. 2025. 3UR-LLM: An end-to-end multimodal large language model for 3D scene understanding. IEEE Transactions on Multimedia, 27: 2899-2911 [DOI: 10.1109/TMM.2025.3557620]

Xu R S, Huang Z N, Wang T, Chen Y L, Pang J M and Lin D H. 2025. VLM-Grounder: A VLM agent for zero-shot 3D visual grounding// Proceedings of the 8th Conference on Robot Learning (CoRL). PMLR, 270: 3961-3985

Xu R S, Wang X L, Wang T, Chen Y L, Pang J M and Lin D H. 2024. PointLLM: Empowering large language models to understand point clouds//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 131-147 [DOI: 10.1007/978-3-031-72698-9_8]

Xue L, Gao M F, Xing C, Martín-Martín R, Wu J J, Xiong C M, et al. 2023. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 1179-1189 [DOI: 10.1109/CVPR52729.2023.00120]

Xue L, Yu N, Zhang S, Panagopoulou A, Li J N, Martín-Martín R, et al. 2024. ULIP-2: Towards scalable multimodal pre-training for 3D understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 27081-27091 [DOI: 10.1109/CVPR52733.2024.02558]

Yang J H, Ding R Y, Deng W P, Wang Z and Qi X J. 2024. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 19823 - 19832 [DOI: 10.1109/CVPR52733.2024.01874]

Yang J W, Tan R, Wu Q H, Zheng R J, Peng B L, Liang Y Y, et al. 2025. Magma: A foundation model for multimodal AI agents//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 14203-14214 [DOI: 10.1109/CVPR52734.2025.01325]

Zemskova T and Yudin D. 2025. 3DGraphLLM: Combining semantic graphs and large language models for 3D scene understanding//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 8885-8895

Zeng Y, Jiang C H, Mao J G, Han J H, Ye C Q, Huang Q Q, et al. 2023. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 15244-15253 [DOI: 10.1109/CVPR52729.2023.01463]

Zhai H J, Li H, Li Z Z, Pan X K, He Y J and Zhang G F. 2025. PanoGS: Gaussian-based panoptic segmentation for 3D open vocabulary scene understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 14114-14124 [DOI: 10.1109/CVPR52734.2025.01317]

Zhang H C, Zantout N, Kachana P, Zhang J and Wang W S. 2025. IRef-VLA: A benchmark for interactive referential grounding with imperfect language in 3D scenes//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE: 1677-1683 [DOI: 10.1109/ICRA55743.2025.11127464]

Zhang J X, Wang Y H, Lu W, Wang T W, Xu W S, Zhang S N, et al. 2026. UniGround: Universal 3D visual grounding via training-free scene parsing[EB/OL]. [2026-06-17]. https://arxiv. org/abs/2603.08131 [DOI: 10.48550/arXiv. 2603.08131]

Zhang J Z, Li A Q, Qi Y P, Li M H, Liu J H, Wang S A, et al. 2026. Embodied navigation foundation model//Proceedings of the 14th International Conference on Learning Representations (ICLR). https://proceedings. iclr. cc/paper_files/paper/2026/hash/ceb45c48e29e7f49b0b47edb98e43691-Abstract-Conference.html

Zhang R R, Guo Z Y, Zhang W, Li K C, Miao X P, Cui B, et al. 2022. PointCLIP: Point cloud understanding by CLIP//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 8542-8552 [DOI: 10.1109/CVPR52688.2022.00836]

Zhang Y M, Gong Z M and Chang A X. 2023. Multi3DRefer: Grounding text description to multiple 3D objects//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 15225-15236 [DOI: 10.1109/ICCV51070.2023.01397]

Zhang Z L, Luo W Y, Wang H N, Sheng Y F, Wang Y D, Guo H Y, et al. 2026. Anticipation-VLA: Solving long-horizon embodied tasks via anticipation-based subgoal generation[EB/OL]. [2026-06-10]. https://arxiv.org/abs/2605.01772

Zhao H, Liu A, Zhang Z, Wang W, Chen F, Zhu R, et al. 2026. CoV: Chain-of-View prompting for spatial reasoning//Findings of the Association for Computational Linguistics: ACL 2026. San Diego: Association for Computational Linguistics: 32427-32440 [DOI: 10.18653/v1/2026.findings-acl.1623]

Zheng D, Huang S J, Zhao L, Zhong Y W and Wang L W. 2024. NaviLLM: Towards learning a generalist model for embodied navigation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 13624-13634

Zhou J S, Wang J S, Ma B R, Liu Y S, Huang T J and Wang X L. 2024. Uni3D: Exploring unified 3D representation at scale//Pro-

© 中国图象图形学报版权所有

ceedings of the 12th International Conference on Learning Representations (ICLR). https://proceedings.iclr.cc/paper_files/paper/2024/hash/cc2c03ce028a1d6cbed6f98cede7f5a9-Abstract-Conference.html

Zhou X, Girdhar R, Joulin A, Krähenbühl P and Misra I. 2022. Detecting twenty-thousand classes using image-level supervision//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer: 350-368 [DOI: 10.1007/978-3-031-20077-9_21]

Zhu C W and Chen L. 2024. A survey on open-vocabulary detection and segmentation: Past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 8950-8970 [DOI: 10.1109/TPAMI.2024.3413013]

Zhu X Y, Zhang R R, He B W, Guo Z Y, Zeng Z Y, Qin Z P, et al. 2023. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 2639-2650

Zuo X X, Samangouei P, Zhou Y W, Di Y and Li M Y. 2025. FMGS: Foundation model embedded 3D Gaussian splatting for holistic 3D scene understanding. International Journal of Computer Vision, 133(2): 611-627 [DOI: 10.1007/s11263-024-02183-8]

作者简介

李嘉巍，男，硕士研究生，研究方向为人工智能。E-mail: 15879675062@163.com

张宇航，男，硕士研究生，研究方向为人工智能。E-mail: 1371367598@qq.com

© 中国图象图形学报版权所有