**Computer Science and Application** 计算机科学与应用, 2025, 15(12), 118-124
Published Online December 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1512327

Hans汉斯

# 多模态大模型驱动的空间智能：技术进展、评估体系与未来挑战

**王承伟[1]，赵虹阳[1]，刘小华[1,2*]**

[1]新疆理工职业大学人工智能学院，新疆 图木舒克
[2]深圳职业技术大学人工智能学院，广东 深圳

收稿日期：2025年10月30日；录用日期：2025年11月28日；发布日期：2025年12月5日

## 摘　要

**近年来，随着多模态大语言模型(Multimodal Large Language Models, MLLMs)的迅猛发展，空间智能(Spatial Intelligence)作为连接感知、推理与行动的核心能力，正成为人工智能迈向物理世界的关键突破口。本文系统梳理了多模态大模型在三维视觉理解、空间感知与推理、具身交互等方面的技术演进路径，重点分析了以视频、深度图、点云等多源异构数据为基础的空间表征方法，并归纳了当前主流的评估基准与典型应用。同时，本文指出模型在跨视角一致性、组合推理、动态场景建模等方面仍面临显著挑战，并对未来研究方向提出展望，旨在为空间智能系统的构建提供理论支撑与技术路线参考。**

## 关键词

**多模态大语言模型，空间智能，具身智能，评估基准**

# Spatial Intelligence Powered by Multimodal Large Language Models: Technological Advances, Evaluation Frameworks, and Future Challenges

**Chengwei Wang[1], Hongyang Zhao[1], Xiaohua Liu[1,2*]**

[1]College of Artificial Intelligence, Xinjiang Vocational University of Technology, Tumushuke Xinjiang
[2]School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen Guangdong

Received: October 30, 2025; accepted: November 28, 2025; published: December 5, 2025

*通讯作者。

文章引用: 王承伟, 赵虹阳, 刘小华. 多模态大模型驱动的空间智能: 技术进展、评估体系与未来挑战[J]. 计算机科学与应用, 2025, 15(12): 118-124. DOI: 10.12677/csa.2025.1512327

## Abstract

**The rapid development of Multimodal Large Language Models (MLLMs) has established spatial intelligence as a key enabler for AI to interact with the physical world, connecting perception with reasoning and action. This paper systematically reviews the technical progress of multimodal models in 3D understanding, spatial reasoning, and embodied interaction. We analyze representation learning methods based on diverse data like video, depth maps, and point clouds, and summarize key benchmarks and applications. Critical challenges in cross-view consistency, compositional reasoning, and dynamic scene understanding are discussed, followed by an outlook on future research to provide a foundation and reference for the development of spatial intelligence systems.**

## Keywords

**Multimodal Large Language Models, Spatial Intelligence, Embodied AI, Evaluation Benchmarks**



## 1. 引言

人类对三维世界的理解不仅依赖于视觉输入，更涉及对空间关系、物体布局及自身与环境相对位置的综合认知。这种能力由心理学家霍华德·加德纳提出，称为“空间智能”，指个体通过视觉进行空间判断、场景心理表征与导航规划的心智技能。在人工智能领域，空间智能被进一步拓展为智能体在三维环境中感知、推理并自主行动的能力，涵盖场景感知、任务规划与执行三大核心环节。

传统视觉语言模型(Vision-Language Models, VLMs)虽在图像描述、目标定位等 2D 任务上表现卓越，但在处理深度、遮挡、视角转换等 3D 空间信息时存在明显局限[1]。其根本原因在于：主流 MLLMs 多基于 CLIP 范式在图像–文本对上预训练，缺乏对几何结构与空间拓扑的显式建模[2]。这一瓶颈严重制约了其在机器人操作、自动驾驶、增强现实等具身智能场景中的落地应用。值得注意的是，这一挑战并非新问题，它根植于计算机视觉的经典难题，即如何从二维投影恢复三维结构，早期工作如 Marr 的 2.5D 草图理论[3]和 SLAM (Simultaneous Localization and Mapping)技术[4]便致力于此。MLLMs 的机遇在于，能否利用其强大的语义理解和生成能力，将此类几何先验与大规模语言知识相融合。

为此，学界近年来围绕“如何赋予 MLLMs 真正的空间智能”展开系统性探索，形成了三大研究主线：(1) 空间感知，即通过引入深度图、点云或相机参数等 2.5D/3D 先验提升模型的空间意识；(2) 多帧/多视角融合，利用视频或多图像序列构建全局空间表征；(3) 评估与诊断体系构建，设计专门基准揭示模型在空间推理中的缺陷模式。本文将围绕上述方向，结合最新研究成果进行综述。

## 2. 评估体系揭示的核心挑战：从表面统计到几何推理的鸿沟

可靠的评估不仅是技术进步的标尺，更是揭示模型认知盲区的探针。近年来，空间智能的评估体系经历了从理想化合成任务向现实复杂性、从单图静态理解向跨图动态推理、从语义正确性向几何一致性的三重跃迁，系统性地暴露了当前 MLLMs 在空间认知上的根本局限。

### 2.1. 对统计先验的过度依赖：现实干扰下的推理崩溃

早期空间推理基准如 SpatialBench [5]聚焦于干净合成图像中的基本空间谓词(如“left of”、“on top

of”)，缺乏真实世界的视觉噪声与语义模糊性。尽管这类任务便于控制变量，却严重高估了模型在真实环境中的泛化能力，因为模型往往仅依赖训练数据中的表面统计关联(例如“桌子上有杯子”的共现模型)作答，而非真正的几何推理。MIRAGE [6]首次系统引入现实干扰因子，构建更具挑战性的评估场景，涵盖以下维度：(1) 遮挡密度：物体被部分或完全遮挡(如“被书挡住的杯子数量”)；(2) 指代歧义：多个外观相似目标共存(如“拿那个红色杯子”，但有三个)；(3) 组合约束：需同时满足颜色、材质、空间位置与动作意图(如“拿起最靠近我的非金属绿色物体”)。实验揭示了一个关键问题：当前最强开源模型 Qwen2.5VL-72B 在基础关系任务中准确率达 56.62%，但在组合任务中骤降至 36.94%。错误分析进一步表明，模型在面对反常识布局(如悬空杯子、倒置椅子)时，模型的推理能力急剧退化，这说明其空间判断高度依赖训练数据中的语义先验，而未能建立对物理空间的鲁棒表征。这一发现呼应了计算机视觉早期对“感知 vs 理解”的讨论，也表明：仅靠大规模数据无法自动催生对几何结构的理解能力；必须通过精心设计的评估机制，引导模型学习底层的空间与物理规律。

## 2.2. 跨视角整合与长期记忆的缺失：从单图到多图的挑战

静态单图提供的空间线索有限，而人类的空间认知天然依赖多视角观察与时间积累。然而，绝大多数 VLMs 仍以单图作为默认输入单元，导致其在需要跨图像推理的任务中能力显著不足。MMSI-Bench [7]是首个专门评估多图像空间智能(Multi-Image Spatial Intelligence)的基准。该基准包含 1000 道人工设计的多选题，要求模型在 1990 张图像构成的集合中进行多跳推理，例如：“图 3 中的椅子扶手是否与图 7 中的桌子属于同一套家具？请参考图 1 的品牌标识和图 5 的材质纹理。”此类任务不仅考验细粒度识别能力，更要求模型建立跨图像实体对齐(cross-image entity alignment)与情境一致性判断。实验结果表明：人类在此基准上准确率达 97%，而当前最强开源模型仅 30%。进一步的人工推理链标注揭示四大典型错误，如图 1 所示：(1) 定位失败：无法在复杂背景中检测小目标或局部部件；(2) 匹配失败：不能将不同视角/光照下的同一物体关联起来；(3) 情境混淆：将厨房物品误用于浴室等不兼容场景；(4) 空间逻辑错误：违反基本物理常识(如认为透明物体不可见即不存在)。这些缺陷直指当前 MLLMs 的根本短板：缺乏显式的跨图记忆机制与情境建模模块。相比之下，人类会主动构建“心理地图”并持续更新，而模型仍停留在“逐图独立处理”阶段。

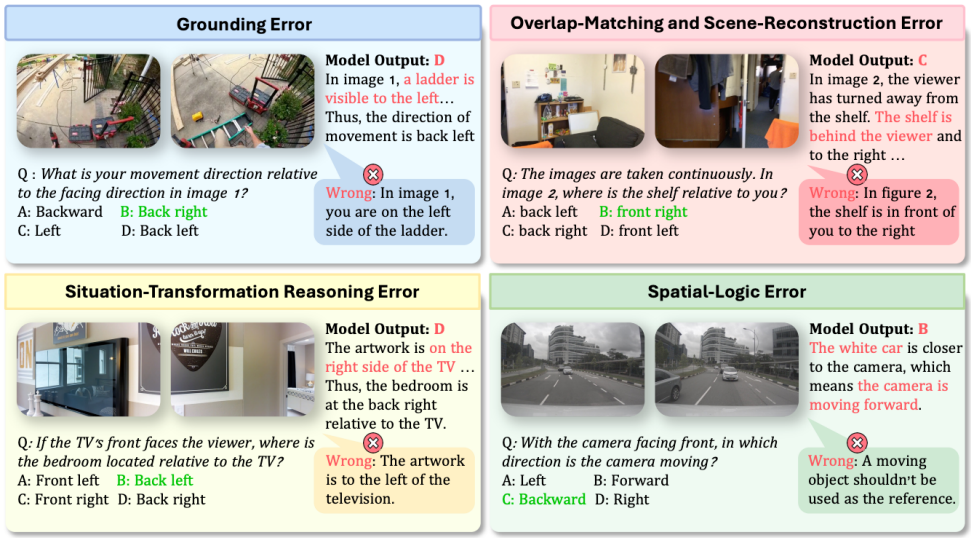


**Figure 1.** Four error types in MLLM spatial reasoning identified in the MMSI-Bench dataset
**图 1.** 在 MMSI-Bench 数据集中发现的多模态大语言模型空间推理的四种错误类型示意图

## 3. 技术路径综述：赋能 MLLMs 空间智能的探索

面对第二章评估所揭示的挑战，研究者们从不同技术路径出发，试图为 MLLMs 注入空间感知与推理能力。这些工作大致可分为两类：一是增强对静态场景的几何理解，二是利用动态序列信息构建空间认知。

### 3.1. 静态场景理解：融合几何先验与相机感知

#### 3.1.1. 深度信息的融合与编码

视觉语言模型(VLMs) [8]-[10]严重依赖 RGB 图像作为唯一输入模态，导致其在处理涉及绝对/相对距离估计、遮挡推理、三维布局理解等任务时表现受限。这一瓶颈源于二维投影丢失深度信息的本质缺陷。为突破此限制，近期研究从显式深度输入与隐式几何重建两条路径推进空间感知能力。SpatialBot 首次将深度图作为独立输入通道引入多模态大模型架构。为解决室内外场景深度值动态范围差异显著的问题，作者提出一种三通道 uint8 编码策略。基于此，团队构建 SpatialQA 数据集，包含三级任务，如图 2 所示：

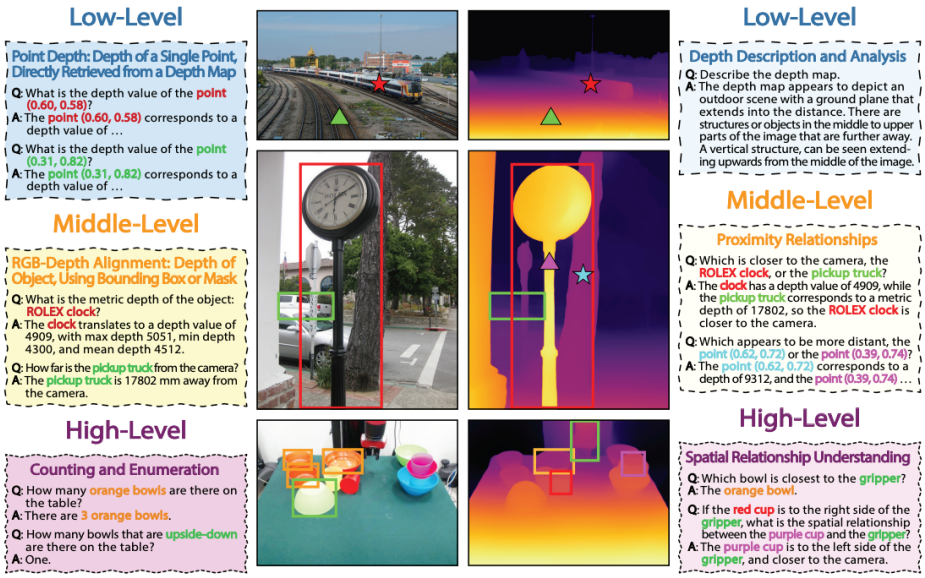


**Figure 2.** SpatialQA dataset task level
**图 2.** SpatialQA 数据集任务级别

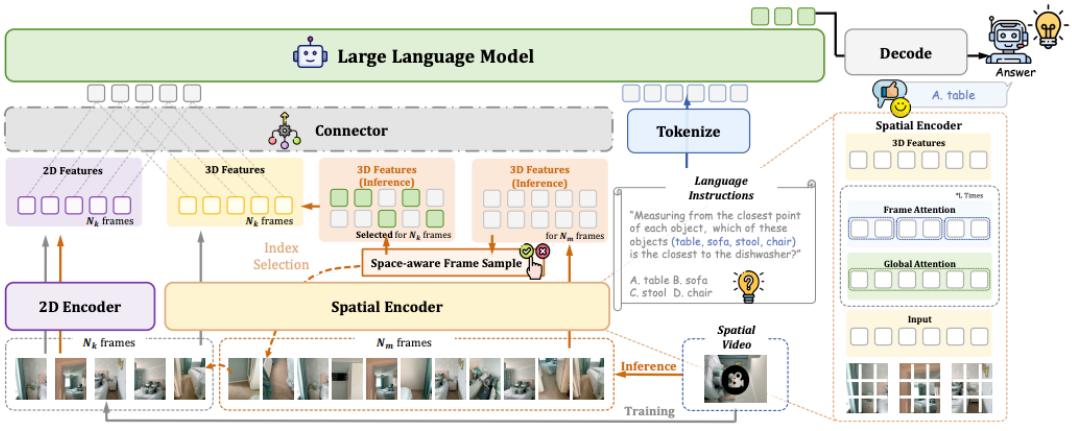


**Figure 3.** Spatial-MLLM model architecture diagram
**图 3.** Spatial-MLLM 模型架构图

实验表明，在真实机器人抓取任务中，引入深度通道使成功率提升 18.2%。值得注意的是，该工作呼应了早期计算机视觉中“RGB-D 融合”的思想如 KinectFusion [11]，但首次将其规模化集成至语言引导的通用推理框架中。类似地，Spatial-MLLM [12]采取更轻量级方案：不依赖外部深度传感器，而是利用预训练的视觉几何基础模型从单目视频帧中恢复隐式 3D 结构。具体而言，该编码器输出 per-pixel 的法向量、表面曲率与相对深度，这些几何特征随后与 CLIP-style 语义特征在 token 级融合，这种“几何先验注入”显著提升了模型在遮挡补全、视角外推等任务上的鲁棒性，其模型架构如图 3 所示。

#### 3.1.2. 相机参数的语言化表达与可控生成

在像素级几何表征的基础上，Puffin [13]提出了更高层次的抽象：将相机成像过程本身纳入语言建模范式。其核心思想是“用相机思考”，即把相机内参和外参编码为特殊语言 token，插入到 prompt 或生成序列中。借此，模型可在统一 Transformer 架构下联合实现相机理解与可控生成两类能力。为支撑该范式，团队构建 Puffin-4M 数据集，包含 400 万组(图像，文本描述，相机参数)三元组，其概览如图 4 所示。实验结果表明，Puffin 在视角一致性(view consistency)与几何合理性上显著优于 ControlNet [14]、Stable Diffusion [15]等专用生成模型。该工作首次实现理解与生成的几何统一建模，为 AR/VR 内容创作、自动驾驶仿真等需精确视角控制的应用开辟新路径。


**Figure 4.** Overview of Pufffn-4M dataset
**图 4.** Puffin-4M 数据集概述

### 3.2. 动态过程建模：利用多帧与多视角序列

静态图像提供的空间线索有限，而时间连续性(运动视差)与视角多样性(多视图几何)是构建全局、一致空间认知的关键。近期研究系统探索如何从视频或多图像序列中提取并整合跨帧空间信息。Yang 等人[16]首次系统评估 MLLMs 在视频级空间推理上的能力。其构建的 VSI-Bench 涵盖自我中心到环境中心坐标系转换、长期物体追踪与位置记忆、基于运动的深度推断等任务。研究发现，现有 MLLMs [17]虽能有效记忆局部物体位置(短时记忆)，但在跨视角整合与长期空间建模上表现薄弱。更关键的是，传统的思维

链(CoT) [18]提示对空间任务几乎无效，而要求模型显式生成认知地图可使距离估计误差降低 31%。这一结果表明，有效的空间推理依赖结构化中间表示，而非纯语言链式推导。

为系统提升多帧理解能力，Meta 团队提出 Multi-SpatialMLLM [19]。其核心是构建 MultiSPA 数据集(2700 万样本)，涵盖深度感知、视觉对应、动态感知三大能力维度。模型采用时空联合注意力机制，在 token 级别对齐不同帧的语义与几何特征。据初步研究显示，其在 12 项空间任务上平均提升 36%，且展现出明显的能力涌现现象。尤为突出的是，该模型可作为机器人学习的多帧奖励标注器，体现强具身智能潜力。另一创新方向来自 Video-3DLLM [20]，其提出“视频即 3D”(Video-as-3D)的表示哲学：将 3D 场景视为视频动态流，并在视频 token 中嵌入三维坐标的位置编码，构建位置感知的视频表征。该思路受 NeRF [21]和 DynamicNeRF [22]等神经渲染技术的启发，但首次将其与语言模型深度融合。该方法在 ScanQA、SQA3D 等基准上达到 SOTA，验证了无需显式点云/网格输入也能实现高阶 3D 推理的可行性。

## 4. 综合讨论与未来展望

通过对多模态大模型在空间智能领域的技术路径与评估体系的系统梳理，本研究揭示了该领域正从单一的静态图像理解迈向融合几何、时序与交互的复杂空间认知。尽管已有研究在深度信息融合、多视角理解等方面取得了显著进展，但迈向通用、鲁棒的空间智能仍面临一系列系统性挑战。这些挑战并非孤立存在，而是相互关联，共同指向了下一代空间智能系统所需突破的关键方向。

### 4.1. 共性挑战

首要的共性挑战源于数据、模型架构与评估范式之间的相互制约。在数据层面，当前高质量、大规模、多模态的空间标注数据依然稀缺，现有数据集往往偏向理想化的合成场景或存在特定的场景偏差(如室内与室外环境分布不均)，这严重限制了模型在真实复杂环境中的泛化能力。因此，如何高效地构建和利用数据，例如通过高保真仿真环境(如 Habitat [23]，AI2-THOR [24])生成海量、多样化的训练数据，成为一个关键问题。在模型架构层面，主流的 Transformer 架构在处理长序列视频数据和进行复杂的组合空间推理时，仍面临计算效率与建模效果的双重挑战。未来的研究需要探索能显式建模空间关系(如图神经网络)、并有效支持长期记忆与跨模态对齐的新颖模型架构。相应地，评估范式也需同步演进，从封闭的问答任务走向开放、动态的具身交互环境。未来的评估体系应进一步融合物理引擎、可微分渲染技术(如 PyTorch3D)与先进的具身仿真平台，构建能够同时评估语义正确性、几何一致性和物理合理性的“闭环式空间智能评测平台”。

### 4.2. 未来研究方向展望

基于对上述挑战的深刻理解，未来的研究可重点关注以下几个有前景的方向。其一是探索统一的空间表征学习框架，旨在无缝融合 2D 视觉特征、3D 几何先验、语言语义乃至动作指令，为各种下游任务构建强大而通用的基础模型。其二是发展因果与反事实推理能力，引导模型超越对表面关联的学习，深入理解场景中的因果关系，并能够进行“如果……会怎样”的思辨，从而在复杂、动态的真实世界中做出更合理、更安全的决策。其三，人机协同的空间认知是一个极具潜力的方向，通过研究如何将人类的直觉、常识和实时指导有效融入模型的训练与推理回路，有望实现高效的人机协同空间问题解决。

## 参考文献

[1] Azuma, D., Miyanishi, T., Kurita, S. and Kawanabe, M. (2022) ScanQA: 3D Question Answering for Spatial Scene Understanding. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 19-24 June 2022, 19107-19117. https://doi.org/10.1109/cvpr52688.2022.01854

[2] Team, G., Georgiev, P., Lei, V.I., Burnell, R., *et al.* (2024) Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. https://arxiv.org/abs/2403.05530

[3] Wu, J., Wang, Y., Xue, T., *et al*. (2017) MarrNet: 3D Shape Reconstruction via 2.5 D Sketches. *Advances in Neural Information Processing Systems*, **30**, 1-11.

[4] 严永嘉, 蹇木伟, 刘宏哲, 等. 基于深度学习的视觉 SLAM 研究综述[C]//中国计算机用户协会网络应用分会. 中国计算机用户协会网络应用分会 2023 年第二十七届网络新技术与应用年会论文集. 镇江, 2023: 55-58. https://kns.cnki.net/kcms2/article/abstract?v=Jz-lw5xPjDacNhj1bXaSiYL-pJVHArlVQka4-Jhyj_MDeqg7raOKKh0NFINXQ1P91RSnND436dfb9QWqC8bbi2fdpqdlFRQzHBE9hBFcGXmp3XW7US1p-8jpu-VxRv36_5a5e0YkPqd4NGj1uex8glTGHy1Fm8PhteH9dLZCQrbLCOvh8UhV3anYyMwXXTMpE&uniplatform=NZKPT&language=CHS

[5] Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., *et al*. (2025) SpatialBot: Precise Spatial Understanding with Vision Language Models. 2025 *IEEE International Conference on Robotics and Automation* (*ICRA*), Atlanta, 19-23 May 2025, 9490-9498. https://doi.org/10.1109/icra55743.2025.11128671

[6] Liu, C., Wang, H., Henry, F., *et al*. (2025) MIRAGE: A Multi-Modal Benchmark for Spatial Perception, Reasoning, and Intelligence. https://arxiv.org/abs/2505.10604

[7] Yang, S., Xu, R., Xie, Y., *et al*. (2025) MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. https://arxiv.org/abs/2505.23764

[8] Chen, Z., Wu, J.N., Wang, W.H., *et al*. (2023) InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. https://ieeexplore.ieee.org/document/10656429

[9] Team, G., Anil, R., Borgeaud, S., *et al*. (2023) Gemini: A Family of Highly Capable Multimodal Models. https://arxiv.org/abs/2312.11805

[10] Wang, P., Bai, S., Tan, S., *et al*. (2024) Qwen2-vl: Enhancing Vision-Language Model's Perception of the World at Any Resolution. https://arxiv.org/abs/2409.12191

[11] Newcombe, R.A., Fitzgibbon, A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., *et al*. (2011) Kinectfusion: Real-Time Dense Surface Mapping and Tracking. 2011 10*th IEEE International Symposium on Mixed and Augmented Reality*, Basel, 26-29 October 2011, 127-136. https://doi.org/10.1109/ismar.2011.6092378

[12] Wu, D., Liu, F., Hung, Y.H., *et al*. (2025) Spatial-MLLM: Boosting MLLM Capabilities in Visual-Based Spatial Intelligence. https://arxiv.org/abs/2505.23747

[13] Liao, K., Wu, S., Wu, Z., *et al*. (2025) Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation. https://arxiv.org/abs/2510.08673

[14] Zhang, L., Rao, A. and Agrawala, M. (2023) Adding Conditional Control to Text-to-Image Diffusion Models. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3813-3824. https://doi.org/10.1109/iccv51070.2023.00355

[15] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) High-Resolution Image Synthesis with Latent Diffusion Models. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 10674-10685. https://doi.org/10.1109/cvpr52688.2022.01042

[16] Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L. and Xie, S. (2025) Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. 2025 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 10632-10643. https://doi.org/10.1109/cvpr52734.2025.00994

[17] Hurst, A., Lerer, A., Goucher, A.P., *et al*. (2024) Gpt-4o System Card. https://arxiv.org/abs/2410.21276

[18] Sprague, Z., Yin, F., Rodriguez, J.D., *et al*. (2024) To Cot or not to Cot? Chain-of-Thought Helps Mainly on Math and Symbolic Reasoning. https://arxiv.org/abs/2409.12183

[19] Xu, R., Wang, W., Tang, H., *et al*. (2025) Multi-Spatial MLLM: Multi-Frame Spatial Understanding with Multi-Modal Large Language Models. https://arxiv.org/abs/2505.17015

[20] Zheng, D., Huang, S. and Wang, L. (2025) Video-3d LLM: Learning Position-Aware Video Representation for 3D Scene Understanding. 2025 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 2995-9006. https://doi.org/10.1109/cvpr52734.2025.00841

[21] Mildenhall, B., Srinivasan, P.P., Tancik, M., *et al*. (2021) Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, **65**, 99-106.

[22] Gao, C., Saraf, A., Kopf, J. and Huang, J. (2021) Dynamic View Synthesis from Dynamic Monocular Video. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 5692-5701. https://doi.org/10.1109/iccv48922.2021.00566

[23] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., *et al*. (2019) Habitat: A Platform for Embodied AI Research. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 9338-9346. https://doi.org/10.1109/iccv.2019.00943

[24] Kolve, E., Mottaghi, R., Han, W., *et al*. (2017) AI2-thor. An Interactive 3d Environment for Visual AI. https://arxiv.org/abs/1712.05474