《多模态大模型原理与应用》

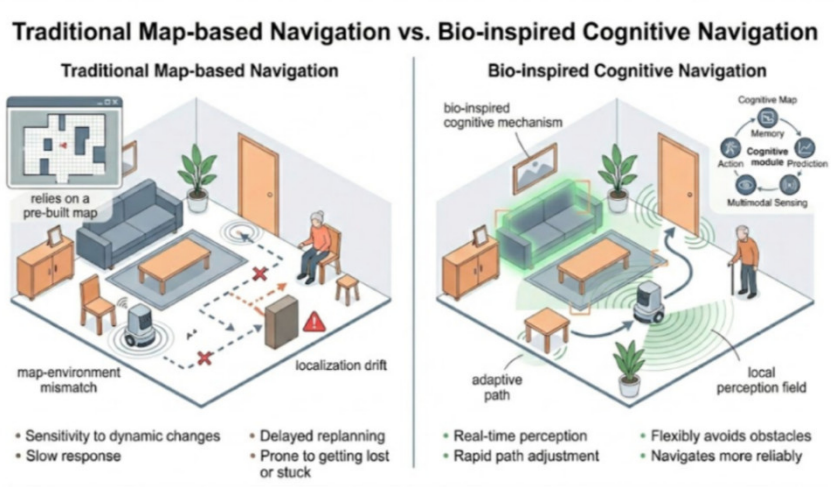
Lecture 11
多模态大模型空间推理

刘阳
中山大学
人机物智能融合实验室（HCP Lab）
liuy856@mail.sysu.edu.cn

本节内容

CONTENTS

什么是空间推理?



AI视角：空间推理 = 三类理解

感知层 L1 "它在哪里？" — 物体识别与定位

关系层 L2 "它跟其他什么关系？" — 方位/距离/拓扑

推理层 L3 "如果……会怎样？" — 运动预测/视角变换

人类视角

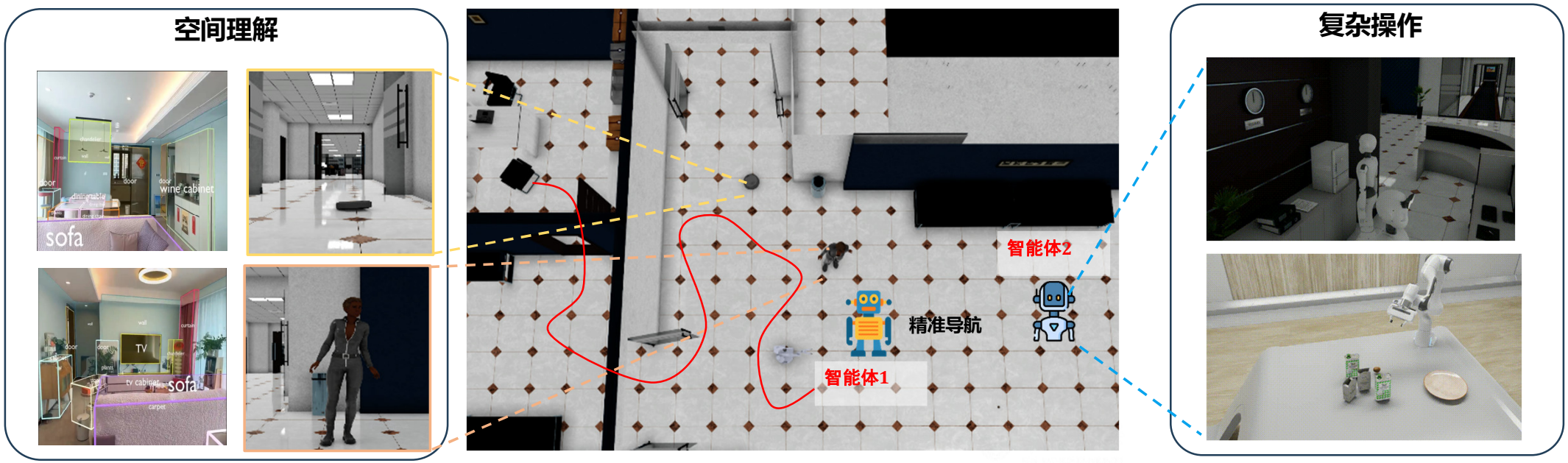
- 遥控器在茶几**上面**（拓扑/接触关系）
- 台灯在沙发**左边**（相对方位）
- 茶几距电视机约**1.5米**（度量/定量距离）

空间推理 = 从多模态输入（图像/点云/视频/音频）中，自动建立三维空间的三类理解能力。

核心问题：如何让AI不仅"看得见"，还能"想得明白"三维空间?

什么是空间推理?

空间智能是指AI系统在三维空间中感知、理解、推演、交互的能力，是具身智能的“核心”，为智能体在真实环境中实现精准导航与复杂操作提供了时空认知基础。



空间推理：从人类认知到机器智能

人类的空间推理能力

感知（出生即有）→ 婴儿伸手抓玩具，利用双眼视差判断距离

表征（经验积累）→ 司机变道，心理模拟周围车辆相对运动

推理（高级认知）→ 建筑师读图纸，从2D构建3D建筑

MLLM的对应能力

视觉感知 → 从图像/点云中识别物体

空间特征提取 → 编码三维几何结构信息

推理与回答 → 回答空间关系问题

核心挑战：空间推理要求感知并操控三维世界中的空间关系，是人类智能的基础，却仍是MLLM的持续挑战。人类的空间推理能力与生俱来却难以形式化，而MLLM试图模仿却困难重重。

为什么空间推理是关键?

具身智能

机器人理解"去厨房拿桌上的杯子"

自动驾驶

判断旁车道车辆与本车的安全距离

医疗影像

分析肿瘤在三维器官中的位置关系

建筑设计

从平面图理解三维空间布局

AR/VR

虚拟物体与真实环境精确融合

科学研究

从X射线衍射图推演分子三维结构

> *"真正的智能，从来不只是'会说话'——而是理解并驾驭物理世界的能力，即空间智能。"*
>
> —— 李飞飞教授

为什么MLLM的空间推理如此困难?

1 数据鸿沟

过去训练VLM的数据集（如COCO、VQAv2）大多是二维图片配文字描述，**深度信息、物体间的精确距离和三维方位关系严重缺失。**

就像只给一个人看世界名画的照片，却不让他走进美术馆亲身感受画作的尺寸与空间布局。

2 表示鸿沟（2D vs 3D）

传统视觉编码器（如CLIP）擅长语义理解，却被优化为处理**实例级语义特征**，对三维空间几何结构的编码能力不足。

模型知道"这是杯子"，却不知道"杯子在桌子的左前方0.5米处"。

3 视角局限

当前LMM在需要**视角依赖理解**的空间推理任务中表现挣扎，主要因为它们被局限于单一、静态的观察视角。

人类会主动移动身体从不同角度观察，而AI只能被动接受给定的图像。

4 缺乏"空间感"

MLLM在语义任务中表现出色，却经常缺乏复杂几何推理所必需的**"空间感"**（spatial sense）。

模型能描述场景中的物体，但无法准确判断物体之间的相对位置、距离和方向关系。

领域发展时间线

2015-2020	2021	2023	2024.01	2024.06	2025.05	2025.06
VQA v1/v2 CLEVR 早期2D空间推理	CLIP 视觉-语言预训练	GPT-4V, LLaVA 多模态对话时代	**SpatialVLM** 首个互联网规模空间推理数据集	**SpatialRGPT** 深度信息插件模块	**SpatialLLM** CVPR 2025 Highlight	**MMSI-Bench** 多图空间智能基准

2025下半年关键进展：

2025.10 多篇综述论文系统化该领域，空间推理研究框架正式确立

2025.11 SpatialThinker（NeurIPS 2025）—— RL+空间奖励实现类人空间感知

2026.05 TwNV, ViSRA, SpatialImaginer —— 视角合成、训练无关推理、想象力增强

从2015年的简单2D VQA到2026年的多视角推理增强，空间推理领域经历了从"能回答"到"能理解三维空间"的范式转变。

多模态空间推理全景图

输入模态 图像 视频 音频 点云 深度图

任务类型

- 空间VQA • 3D定位
- 导航规划 • 场景理解

空间推理
Spatial Reasoning

技术方法

- 数据增强 • 架构改造
- 训练策略 • 推理增强

应用场景 具身智能 自动驾驶 医疗 建筑设计

空间推理需要整合**多种输入模态**，并通过**多种技术方法**来实现从感知到应用的完整闭环

三维世界理解：从图像到行动

行动层 Action

• 导航指令执行 • 操作规划 • 多步任务完成

将空间理解转化为具体行动决策

推理层 Reasoning

• 空间关系推理 • 距离/尺寸估计 • 路径规划 • 遮挡理解 • 视角变换

理解物体之间的空间关系和场景的几何结构

感知层 Perception

• 场景分类 • 物体识别 • 深度估计 • 3D重建 • 语义分割

从认知角度理解空间推理能力

SpatialTree：从认知科学角度组织空间智能，按照推理复杂度划分任务，并映射到现有基准

空间推理任务分类：从2D到3D再到4D

Intrinsic-Static 内在静态

理解物体自身的空间属性，不涉及外部参考系，场景静止。

例： "这个水杯的开口朝向哪个方向？"

Intrinsic-Dynamic 内在动态

推理物体自身的空间变化，如旋转、形变。

例： "如果把这个魔方顺时针旋转90度，红色面会朝哪？"

Extrinsic-Static 外在静态

推理物体之间的空间关系或物体在场景中的位置，场景静止。

例： "桌子左边的椅子是第几把？"

Extrinsic-Dynamic 外在动态

推理物体在场景中的运动轨迹或相对运动。

例： "从A点走到B点需要经过哪些路口？"

空间关系的类型详解

空间关系类型	定义	示例问题
拓扑关系	物体之间的接触/包含/相邻等关系	"杯子在桌子上还是在桌子下？"
方向关系	相对方位（左/右/前/后/上/下）	"从门的角度看，窗户在你的左边还是右边？"
距离关系	定量或定性距离估计	"沙发和电视之间大约有多远？"
尺寸关系	物体的相对大小比较	"图中的三个球，哪个最大？"
运动关系	物体的运动轨迹或速度	"如果这辆车继续直行，它会撞到那个行人吗？"

关键洞察：传统VLM擅长回答"是什么"，但在回答"在哪里/多远/哪个方向"的问题时**频繁出错**。这正是空间推理研究需要解决的核心问题。

2D → 3D：三维空间理解的核心范式

传统VLM看到的是

一只猫在沙发上 —— **这是2D的、定性的**

空间推理需要的是

- 猫在沙发的**左前方**
- 距沙发边缘约**0.3米**
- 距地面垂直距离约**1.2米**
- 猫**面向**窗户方向

三种技术路径对比

路径	输入	代表模型	关键思路
图像基础	2D图像→推导3D	SpatialVLM	从2D图像自动生成空间标注数据
点云基础	直接3D点云	Spatial 3D-LLM	渐进空间感知方案丰富空间嵌入
混合模态	多数据流融合	SpatialLLM	系统整合3D数据、架构与训练

为什么空间推理很难？——关键难题（一）遮挡

问题定义

在特定视角下，物体被其他物体部分或完全遮挡，模型需要**推断被遮挡物体的位置、形状和空间关系。**

为什么对人类简单却对AI困难？

- 人类利用**心理模型**（mental model）进行"想象补全"
- MLLM只有静态的、被动的观察输入

解决思路

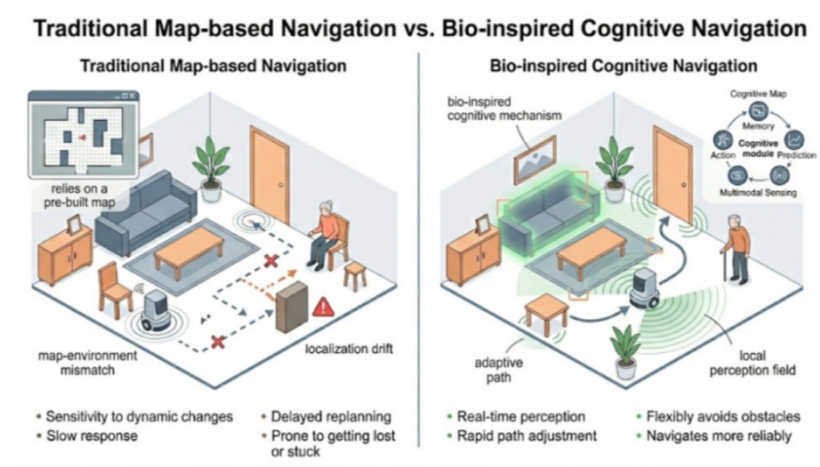
TwNV
多视角推理，生成式新视角合成

SpatialThinker
场景图构建，捕捉空间关系

SpatialImaginer
想象力增强，文本推理+视觉想象

"剪裁和缩放无法揭示被遮挡的几何结构，也无法消解从输入视角无法直接观察到的3D关系结构。"

为什么空间推理很难？——关键难题（二）视角依赖



问题定义

物体在不同视角下呈现出完全不同的外观，但空间关系保持不变。模型需要具备**视角不变性**（viewpoint invariance）

典型表现

从正面看：杯子在花瓶**前面**

从侧面看：杯子在花瓶**右边**

两种说法都对——取决于观察者的视角

实验发现

- 在视角敏感的子任务上，当前LMM表现**最差**
- **TwNV**方法通过多视角合成在视角敏感子任务上获得最大增益（**+1.3至+3.9个百分点**）
- **CAMCUE**框架利用相机姿态作为跨视图融合的显式几何锚点，旋转准确率超**90%**

本节内容

CONTENTS

一、什么是空间推理

二、核心技术方法

三、评估与基准

四、应用与前沿趋势

空间推理技术方法全景图

	训练时（Training-based）	推理时（Reasoning-based）
数据驱动	• 空间数据增强（SpatialVLM） • 3D场景图数据构建（SpatialRGPT） • 合成数据生成（SpaRE）	• 新视角合成（TwNV） • 视觉想象力（SpatialImaginer）
模型驱动	• 双编码器架构（Spatial-MLLM） • 几何-语义融合（SpatialGeo） • 强化学习+空间奖励（SpatialThinker）	• 工具增强推理（LAST） • 结构化场景推理（SSR） • 训练无关空间推理（ViSRA）

核心维度：增强MLLM空间推理的方法可以从"**训练时 vs 推理时**"和"**数据驱动 vs 模型驱动**"两个维度进行分类，形成2×2的方法矩阵。

经典方法一：空间数据增强范式——SpatialVLM

核心思想：VLM空间推理能力有限，不是因为架构缺陷，而是因为**训练数据缺乏3D空间知识。**

数据集构建流程

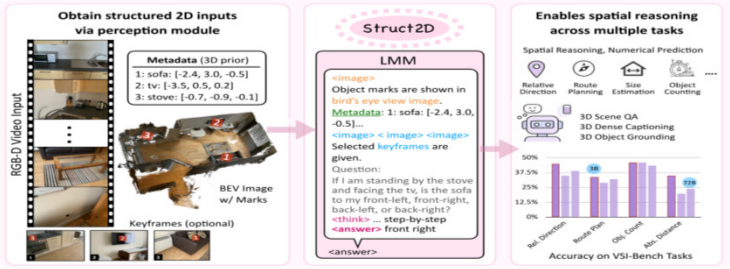
语义过滤
CLIP筛选
→
物体提取
开放词汇检测
→
3D定位
度量深度估计
→
关系抽取
关系三元组
→
VQA生成
模板化/自然语言

关键数据

1000万
张真实世界图像覆盖

20亿
个空间VQA示例生成

业界首个互联网规模的
度量空间推理数据集



Automatic 3D Spatial Data Generation Paradigm for SpatialVLM

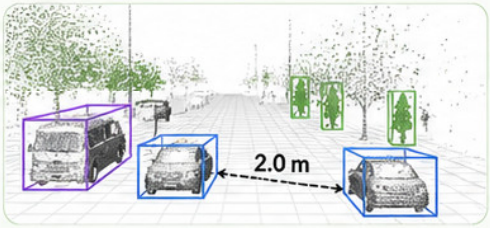
Real-world Images (2D)

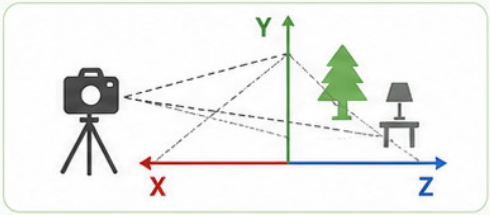
Automatic Processing Pipeline

- Depth Estimation (Far)
- Object Detection
- Semantic Segmentation
- Camera Pose Estimation
- Point Cloud Reconstruction
- ...

Fully Automatic

3D Physical Concepts (relative distance, depth, etc.)





Structured 3D Representation

3D Spatial VQA Data (Massive & Structured)

Q: What is the distance between the car and the bus?
A: 2.0 meters.

Q: Which object is farther away, the tree or the building?
A: The building is farther away.

Q: How high is the lamp from the ground?
A: 1.2 meters.

Q: Is the couch closer to the TV or the table?
A: The couch is closer to the TV.

Massive 3D Spatial Q&A Dataset

SpatialVLM：关键洞察与实验发现

关键洞察

洞察	具体发现
数据质量 > 数量	高质量空间标注数据比大量低质量数据更有效
混合训练最优	95%原始PaLM + 5% Spatial VQA效果最佳
定量推理能力	可对2D输入图像进行度量距离估计

对比GPT-4V

- SpatialVLM在定性空间问答和定量距离估计方面**显著超越**GPT-4V
- 解决了GPT-4V在空间推理方面的"**盲区**"（blind spots）

里程碑意义：从2D图像预测3D距离的能力具有里程碑意义——这意味着模型**不需要深度传感器或3D扫描仪**，仅凭单张RGB照片即可估计物体间距离。

经典方法二：区域级深度增强——SpatialRGPT

核心思想： VLM的空间推理需要**精确的区域级空间感知**——不仅知道"那里有一个杯子"，还要知道"杯子在花瓶的右前方0.8米处"。

两项关键创新

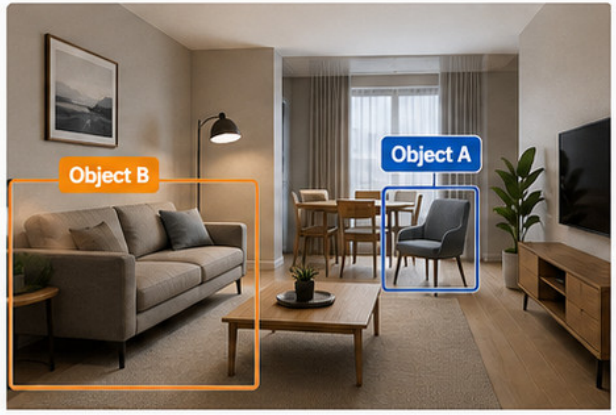
创新点	实现方式	作用
数据管理管道	从3D场景图学习区域表示	提供丰富的空间标注训练数据
深度信息插件模块	灵活集成到现有VLM视觉编码器	增强对三维空间结构的感知

核心能力

- 当提供用户指定的区域提案时，SpatialRGPT可以准确感知它们的**相对方向和距离**
- 提出了**SpatialRGPT-Bench**：涵盖室内、室外和模拟环境的真实3D标注基准

① 2D + Depth INPUT

2D RGB Image (Indoor Scene)

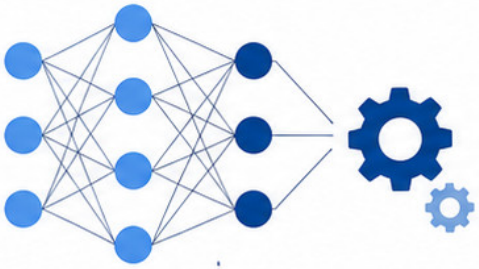


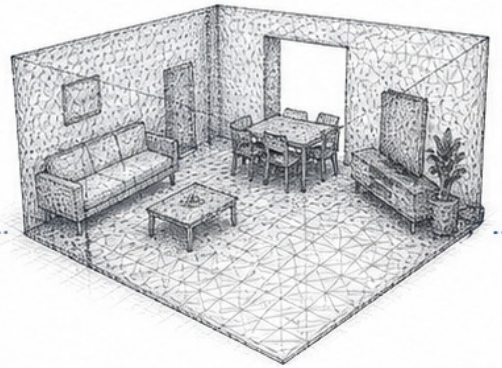
Depth Map

PNG

② SPATIALRGPT PLUGIN

SpatialRGPT
PLUGIN





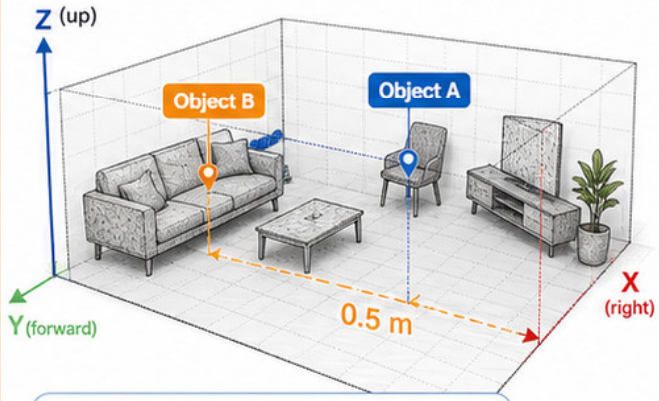
Plugin Integration with
Visual Language Model (VLM)

③ 3D SPATIAL REASONING OUTPUT

Natural Language Output (Example)

Object A is 0.5 meters to the right of Object B.

3D Spatial Scene & Coordinates



- Object A (Chair)
  $(x_A, y_A, z_A) = (0.50, 1.20, 0.45)$ m
- Object B (Sofa)
  $(x_B, y_B, z_B) = (0.00, 1.20, 0.45)$ m
- Relative Position (A w.r.t B)
  $\Delta x = +0.50$ m, $\Delta y = 0.00$ m, $\Delta z = 0.00$ m

两条技术路径对比

维度		SpatialRGPT（架构增强）
设计哲学	"数据即一切"——缺数据就造数据	"表示是关键"——改进模型架构来理解深度
核心贡献	互联网规模的空间VQA数据集	深度信息插件模块+3D场景图学习
输入要求	2D图像（单张即可）	3D场景图或含深度信息的图像
距离估计	从2D图像直接推理度量距离	基于深度信息精确计算距离
优势	数据驱动，泛化性强	区域级精度高，可解释性强
局限	依赖自动标注的准确性	需要深度传感器或3D重建前置步骤
适用范围	通用VLM空间增强	机器人、AR等需要精确空间感知的场景

结论：两条路径并不互斥——当前最优方案往往是**数据+架构+训练策略**的有机融合，这正是后续SpatialLLM等工作的思路。

经典方法三：3D数据×架构×训练——SpatialLLM

CVPR 2025 Highlight

核心贡献：首个系统研究**3D数据、架构和训练三者如何协同**的空间推理框架。

第一步：构建3D感知训练数据（两类）

数据类型	内容	作用
3D感知探测数据	物体的3D位置和方向标注	训练模型理解空间定位

第二步：最优架构与训练方案

通过系统实验探索LMM架构和训练配置的最佳组合。

重要突破：首个在真实图像上构建包含3D方向关系的VQA数据。

超越GPT-4o 8.7%
证明了系统整合3D数据的巨大潜力

SpatialLLM：3D感知训练数据构建

3D探测数据（Probing Data）

Q: "图中红色框标记的椅子面朝什么方向？"

A: "朝向北偏东30度。"

训练目标：学习3D位置和方向的基本概念

3D对话数据（Conversation Data）

Q: "如果左侧来车以当前速度继续行驶，会不会撞到行人？"
A: 综合多物体空间关系进行复杂推理...

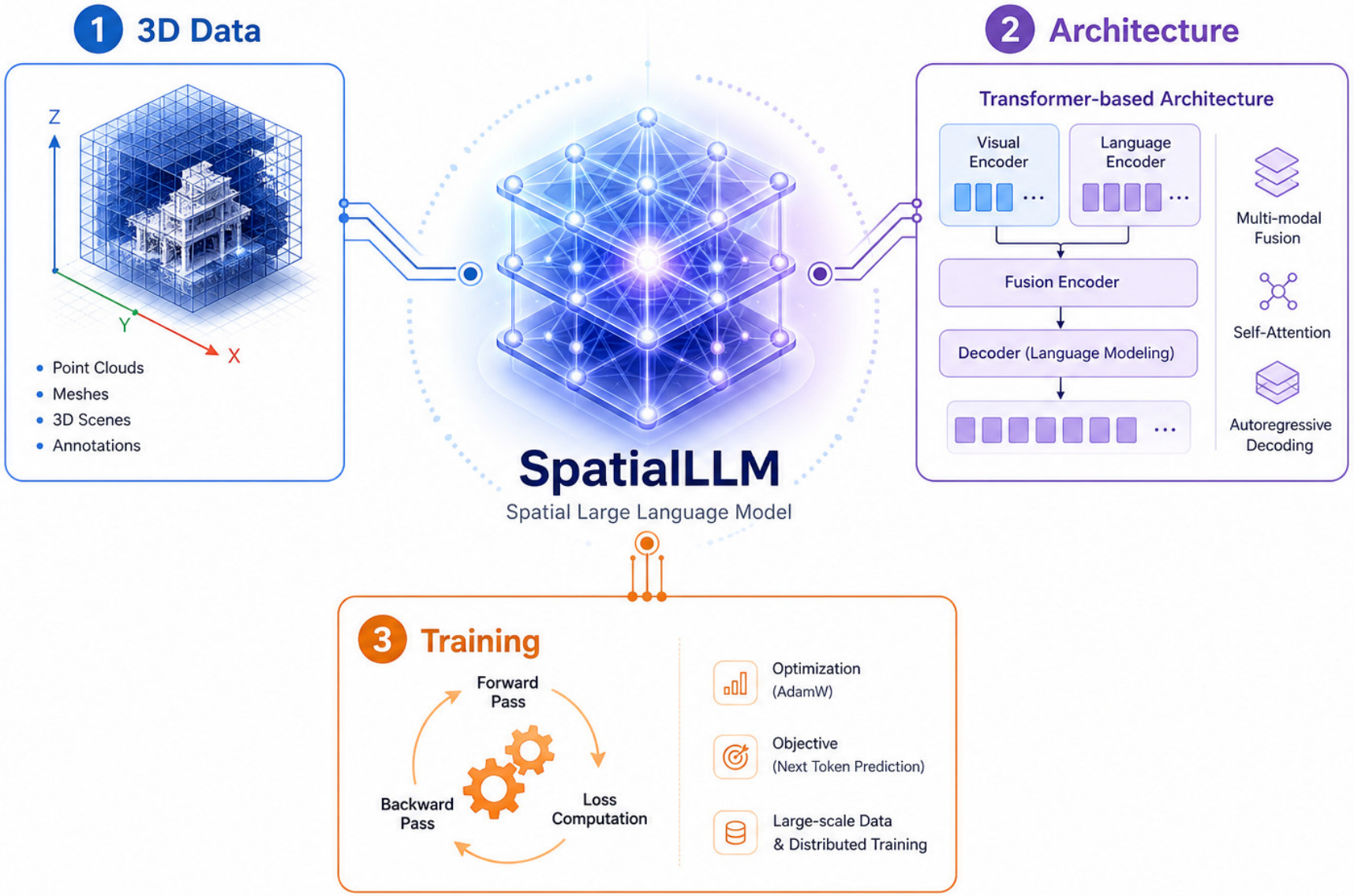
训练目标：综合多物体空间关系推理

设计原则

探测数据
专注单物体的3D属性理解

对话数据
专注多物体的复杂空间关系推理

协同训练
构建完整的3D空间认知能力


1 3D Data
Z
Y
X
Point Clouds
Meshes
3D Scenes
Annotations
SpatialLLM
Spatial Large Language Model
2 Architecture
Transformer-based Architecture
Visual Encoder
Language Encoder
Fusion Encoder
Decoder (Language Modeling)
Multi-modal Fusion
Self-Attention
Autoregressive Decoding
3 Training
Forward Pass
Loss Computation
Backward Pass
Optimization (AdamW)
Objective (Next Token Prediction)
Large-scale Data & Distributed Training

方法创新：纯2D输入实现3D推理——Spatial-MLLM

问题：传统3D MLLM依赖额外的3D或2.5D数据——但在许多实际场景中，我们**只有2D图像或视频**作为输入。

核心思路——双编码器架构

为什么需要两个编码器?

- CLIP擅长语义理解（这是什么？），但丢失了空间几何信息
- 视觉几何模型擅长结构理解（这在哪里？有多远？），需要"唤醒"其结构先验

额外创新

空间感知帧采样策略：在推理时自动从视频中选取空间信息最丰富的帧，使模型在有限token预算下聚焦于对空间推理最关键的时刻。

Spatial-MLLM：从数据到策略的完整训练方案

阶段一
多源数据集构建
- 空间VQA
- 3D场景理解
- 视频定位
- 距离估计

→

阶段二
监督微调 SFT
- 基于标注数据训练
- 学习基本空间推理能力
- 建立空间概念基础

→

阶段三
GRPO优化
- 奖励信号引导优化
- 强化空间推理能力
- 群体相对策略优化

关键特点

✓ 不需要显式的3D数据输入

✓ 不需要深度传感器或3D扫描仪

✓ **真实世界数据集SOTA**

Spatial-MLLM: Pure 2D Input for 3D Reasoning

Direct 2D to 3D Reasoning Without Depth Input

2D Input

Single RGB Image

No Depth, No Point Cloud
Just a Plain 2D Image

Spatial-MLLM

Spatial Multimodal Large Language Model

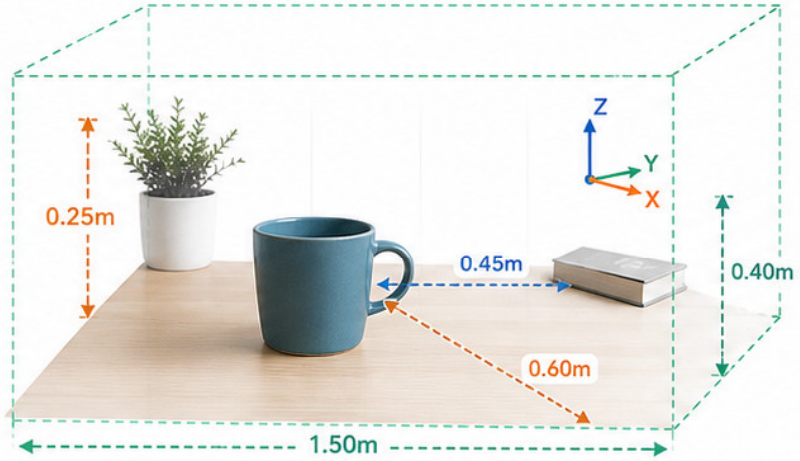
Spatial Understanding

Geometric Reasoning

Language Alignment

3D Reasoning Results

Structured Spatial Understanding



Reasoning Outputs (Examples)

- The mug is 0.45m in front of the book.
- The table length is about 1.50m.
- The plant is 0.25m tall.
- The book is to the right of the mug.
- The mug is 0.60m from the front edge.
- The plant is to the left and behind the mug.

Breakthrough: Achieving 3D Spatial Reasoning Directly from 2D Images

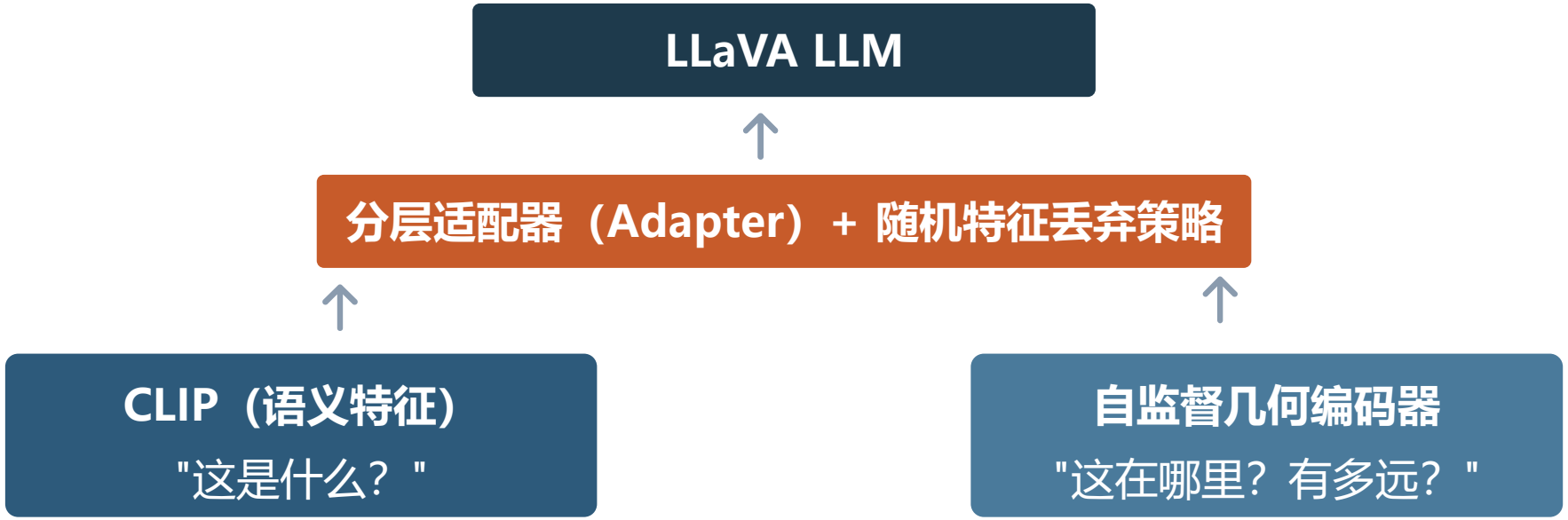
No Depth Sensor

No Point Cloud

No 3D Supervision

方法五：几何-语义融合——SpatialGeo

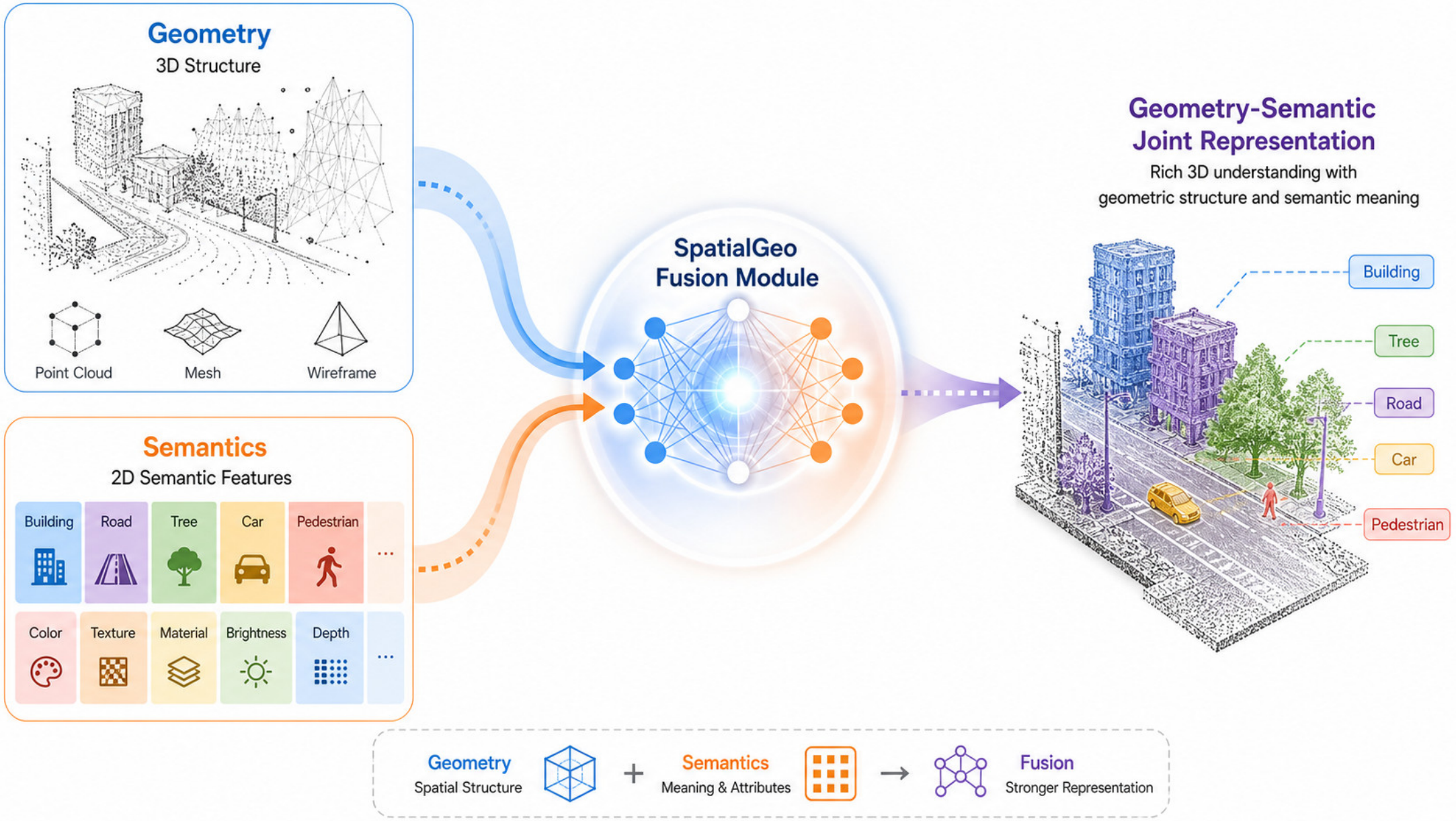
问题根源：大多数MLLM使用的视觉编码器（如CLIP）被限制为**实例级语义特征**，空间模糊问题由此产生。



效果

SpatialRGPT-Bench上提升 **≥8.0%**

推理时内存成本降低约 **50%**


Geometry
3D Structure
Point Cloud
Mesh
Wireframe
Semantics
2D Semantic Features
Building
Road
Tree
Car
Pedestrian
...
Color
Texture
Material
Brightness
Depth
...
SpatialGeo
Fusion Module
Geometry-Semantic
Joint Representation
Rich 3D understanding with
geometric structure and semantic meaning
Building
Tree
Road
Car
Pedestrian
Geometry
Spatial Structure
+
Semantics
Meaning & Attributes
→
Fusion
Stronger Representation

方法六：强化学习+空间奖励——SpatialThinker

问题：现有空间MLLM要么依赖显式的3D输入，要么受限于**大规模数据集和稀疏监督信号。**

三项核心创新

创新	实现方式	优势
场景图构建	自动检测任务相关物体及其空间关系	结构化空间信息
密集空间奖励	多目标RL奖励函数，每步都有空间监督	细粒度学习信号
数据合成管道	自动生成高质量STVQA-7K数据集	减少对人工标注的依赖

核心效果

SpatialThinker-7B **超越SFT**

增益接近翻倍 **+6.5% vs +3.6%**

匹配甚至超越GPT-4o

SpatialThinker：空间奖励机制的深入理解

场景图构建流程

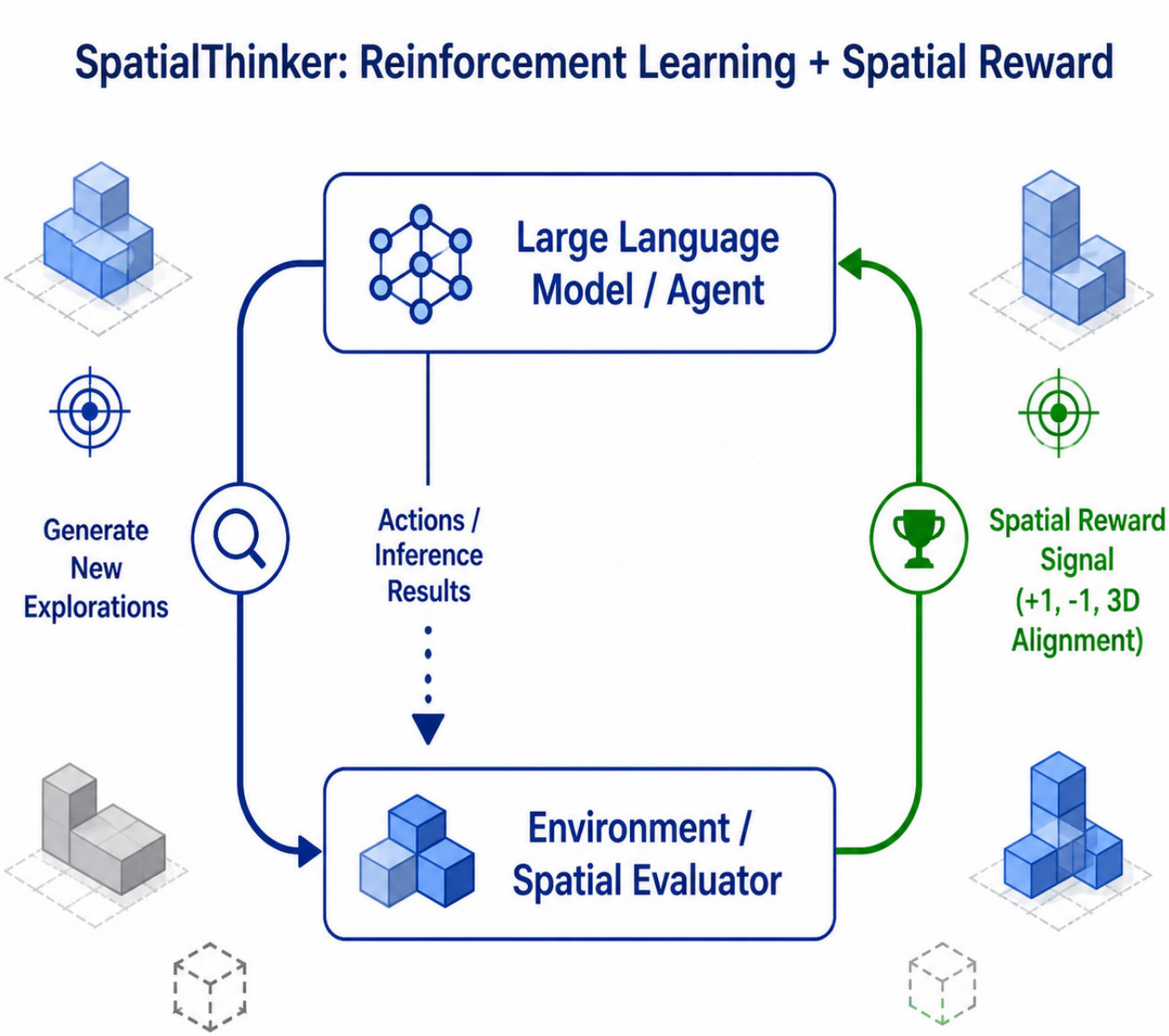
输入图像 → 物体检测 → 空间关系抽取 → 场景图

物体1：杯子 | 关系1：杯子在桌子上面
物体2：桌子 | 关系2：杯子在花瓶右边
物体3：花瓶 | 关系3：桌子在花瓶后面

密集空间奖励的多维度

奖励维度	定义	目的
物体定位奖励	检测位置准确性	确保感知正确
关系准确性奖励	空间关系正确性	确保推理正确
推理步骤奖励	每步推理逻辑性	鼓励多步推理
最终答案奖励	任务完成正确性	目标导向

关键价值：通过结合空间监督与奖励对齐的推理，在**有限数据**下实现鲁棒的3D空间理解，推动MLLM向人类水平的视觉推理迈进。


SpatialThinker: Reinforcement Learning + Spatial Reward
Large Language Model / Agent
Generate New Explorations
Actions / Inference Results
Spatial Reward Signal (+1, -1, 3D Alignment)
Environment / Spatial Evaluator

方法七：结构化场景推理——SSR

问题：模型缺乏复杂几何推理必需的"空间感"；模态对齐成本高昂且缺乏细粒度结构建模。

SSR框架的两大支柱

跨模态对齐
3D几何特征"锚定"到LLM的预对齐2D视觉语义
无需大规模对齐预训练，大幅降低训练开销

场景图生成
全局布局表示为独立局部三元组链
构建"LLM友好"的结构化支架

场景图表示方法

全局场景 → 分解为独立三元组 → 链式结构

例：(杯子，右边0.5m，花瓶) → (花瓶，前面0.3m，盘子) → ...

7B参数在多个空间智能基准达SOTA，VSI-Bench得分73.9，显著超越大得多的模型

Unstructured Scene

(Raw 2D Image)

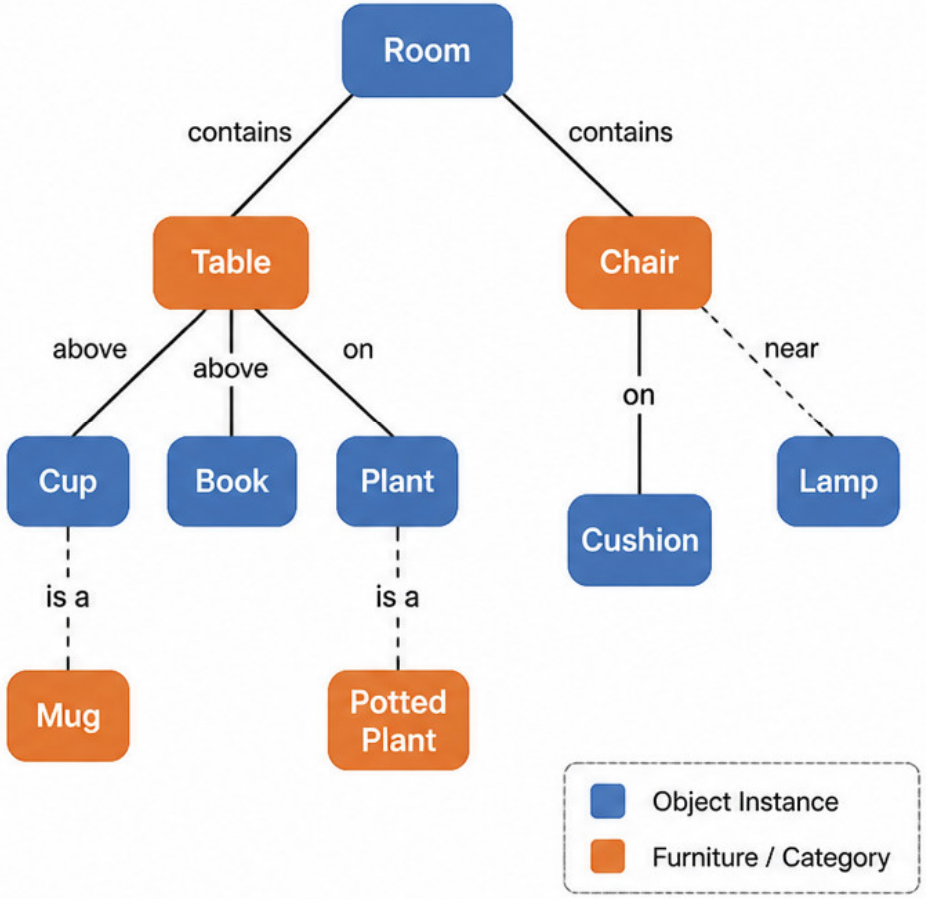
Objects are randomly placed with no explicit structure.

SSR

Structured Scene Reasoning (SSR) Module

Structured Scene Representation

(Scene Graph)



Objects are organized into a structured graph with explicit relationships.

增强空间推理的训练范式总结

数据增强
SpatialVLM
"缺数据
就造数据"

架构改造
SpatialRGPT
"让模型
看懂深度"

训练策略
SpatialThinker
"强化学习
精细奖励"

融合增强
Spatial-MLLM
SpatialGeo
"语义+
几何融合"

推理增强
TwNV
"生成新
视角推理"

范式选择指南

应用场景	推荐范式	代表方法
数据匮乏	数据增强	SpatialVLM
需要精确距离	架构改造	SpatialRGPT
有限标注数据	训练策略	SpatialThinker
纯2D输入	融合增强	Spatial-MLLM / SpatialGeo
复杂场景	推理增强	TwNV / SSR

推理增强——训练方法的有力补充

训练方法的三大局限

数据需求巨大

SFT需要大量高质量空间标注数据，获取成本高

泛化能力有限

训练数据中的空间场景无法覆盖所有真实世界情况

计算成本高昂

全量微调或重新训练大模型的成本往往不可承受

推理增强的优势

即插即用

不需要重新训练模型，在推理时动态增强

灵活组合

可以按需调用不同的工具和能力

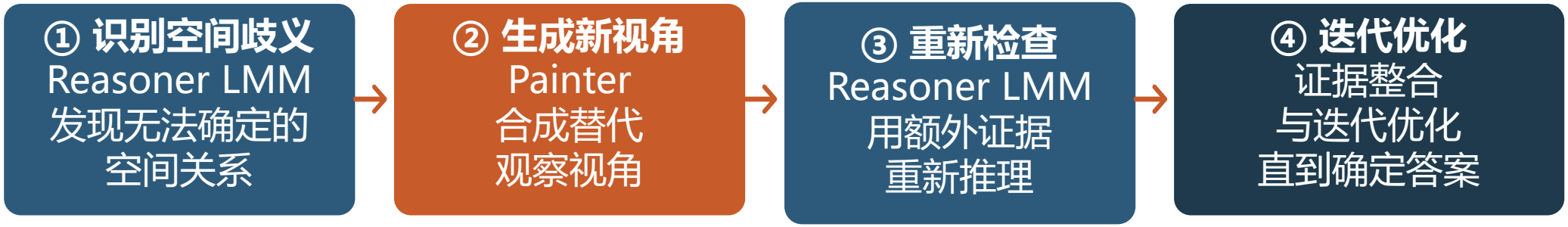
可解释性强

推理过程透明，每一步都可追溯

重要机会：现有方法对推理时方法（inference-time approaches）的探索**相对不足**——这恰恰是推理增强的巨大机会。

推理增强一：视角合成推理——TwNV

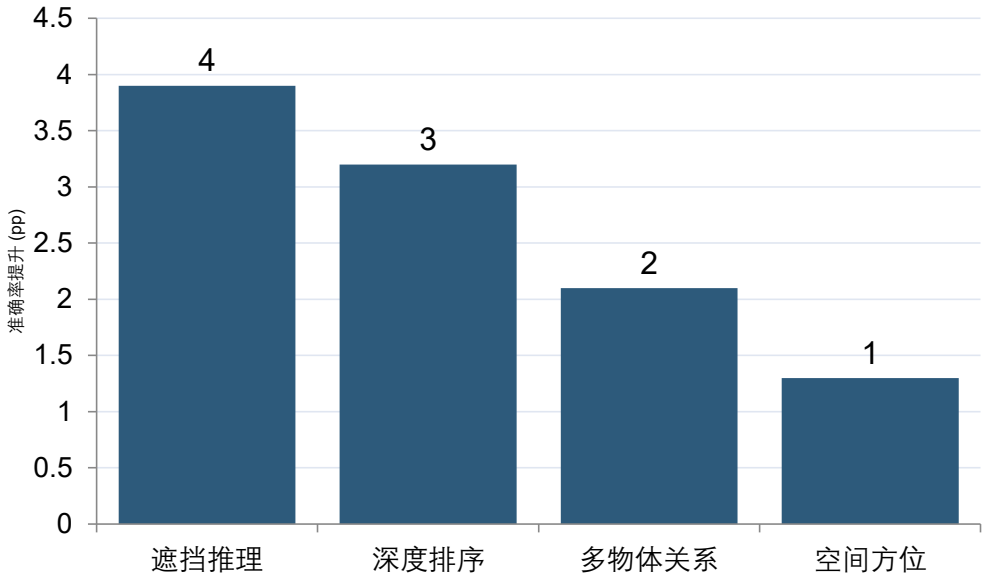
问题： 单一静态观察不足以进行视角依赖的空间推理——人类会主动"想象"不同视角。



核心发现

研究问题	对比方案	结论
指令格式	数值相机姿态 vs 自由语言	数值规格更可靠地控制视角
生成保真度	与下游空间精度紧密耦合	更真实的合成→更准确的推理
视觉缩放	多轮视角迭代优化	迭代进一步提升性能

视角合成推理：TwNV的实验证据



关键实验结果

跨架构验证

4种LMM架构（闭源+开源）均实现正增益

对比GPT-5工具管道

配备裁剪/缩放后准确率**反而下降0.8pp**

多物体关系差距

差距扩大至**2.0个百分点**

核心结论：简单的2D变换（裁剪/缩放）**不足以替代真正的3D视角生成。**TwNV证明了生成式新视角合成在空间推理中的关键价值。

推理增强二：工具增强空间推理——LAST

问题： MLLM在解析复杂几何布局时经常产生**幻觉和不精确性**；数据驱动的缩放方法难以内化结构化几何先验。

LAST四层框架

LLM推理层 — 利用多模态提示进行空间推理

多模态提示生成层 — 将工具输出转化为LLM可理解的格式（标注图像+文本描述）

工具调用层 — 将工具调用抽象为原子指令和可复用技能

LAST-Box沙箱层 — 异质工具的统一抽象（分割模型/深度估计/3D重建）

效果： LAST-7B在4个数据集上相比基线**提升约20%**，**超越强大的闭源LLM**

推理增强三：视觉想象力——SpatialImaginer

问题： MLLM在空间推理中表现出"脆弱的推理痕迹"——空间识别机制与纯文本推理行为之间存在**不匹配**。

文本思维链 CoT
处理高层语义规划
- "我要先看左边"
- "然后判断遮挡"
- "最后确定距离"

⇄

视觉想象力
处理几何敏感状态变换
- 生成新视角图像
- 更新空间状态
- 保留几何细节

数据引擎

难度感知数据引擎 + 闭环验证 —— 训练模型在需要稳定空间状态追踪时**有选择性地调用视觉想象力**，而非每步都生成图像，实现效率与精度的平衡。

推理增强四：训练无关空间推理——ViSRA

问题： 3D空间智能的进展主要由后训练驱动，**推理时方法被相对忽视。**

ViSRA三层架构

专家模型层
深度估计/分割/3D重建

→

空间信息聚合层
结构化提取+多模态提示

→

MLLM推理层
不修改任何权重

核心优势

优势	说明
训练无关	无需任何后训练计算成本
与人类对齐	利用专家模型提取显式空间信息
可迁移3D理解	非任务特定的过拟合
即插即用	模块化、可扩展的灵活框架

实验效果

现有基准最高提升

15.6%

未见过任务提升高达

28.9%

从训练到推理：MLLM空间推理方法发展脉络

趋势总结： 从"数据为王" → "数据+架构+训练协同" → "推理时动态增强"

深入理解视觉编码器：CLIP为什么不够?

CLIP的训练目标

图像 ←→ 文本匹配
"一张猫的照片" ☑
"猫在桌子上" ☑
"桌子上的猫" ☑

CLIP没学到的是

物体间的空间关系
猫距桌子0.5米 ✗
猫在桌子的左前方 ✗
猫面朝窗户 ✗

为什么会这样?

原因	说明
训练数据	互联网图像-文本对，文本很少描述精确的空间关系
训练目标	对比学习优化的是语义对齐，不是空间对齐
表示粒度	全局图像嵌入丢失局部空间信息
缺少3D信号	纯2D训练，没有深度/视角/几何等3D信号

解决方案： SpatialGeo（补充几何编码器）| Spatial-MLLM（增加3D空间编码器）| SpatialRGPT（深度信息作为插件注入CLIP）

MLLM空间推理技术矩阵

方法	类型	输入	需训练?	核心创新	代表性能
SpatialVLM	数据增强	2D图像	SFT	互联网规模空间VQA数据集	超越GPT-4V
SpatialRGPT	架构改造	图像+深度	是	深度插件+3D场景图	区域级精度
SpatialLLM	系统整合	2D+3D标注	SFT	3D数据×架构×训练协同	超GPT-4o 8.7%
Spatial-MLLM	架构创新	2D图像/视频	SFT+GRPO	双编码器，纯2D输入	SOTA
SpatialGeo	编码器增强	2D图像	SFT	几何-语义分层融合	空间基准+8%
SpatialThinker	训练策略	2D图像	RL	场景图+密集空间奖励	匹配GPT-4o
TwNV	推理增强	图像+视角	否	生成式新视角合成	+1.3~3.9pp
ViSRA	推理增强	视频	否	训练无关，即插即用	+15.6%~28.9%

数据是核心：从SpatialVLM看空间数据构建

步骤	技术手段	过程	产出
①	语义过滤	CLIP模型分类	含真实场景的图像（非商品图/GUI截图）
②	物体提取	开放词汇检测器	场景中所有物体的2D边界框
③	3D定位	度量深度估计模型	每个物体的3D位置坐标
④	关系抽取	空间关系三元组生成	（物体A，关系，物体B，距离值）
⑤	VQA生成	模板化+自然语言	定性+定量空间问答对

数据混合策略

95%原始PaLM数据集 + **5%**新构建Spatial VQA数据集 —— 少量高质量空间数据即可显著提升模型空间推理能力

启示：不是所有数据都有同等价值——空间感知数据的边际收益远超通用图像-文本对

视觉编码器的选择与对比

编码器	擅长	不擅长	空间推理中的角色
CLIP	语义理解"这是什么"	空间定位"在哪里"	提供语义基础
SigLIP	细粒度视觉理解	3D几何理解	增强局部细节
DINOv2	自监督视觉特征	语言对齐	补充结构信息
空间几何模型	3D结构先验	语义理解	提供几何理解

关键发现：视觉编码器的选择决定了模型能看到什么——选错编码器就像给模型戴上了"有色眼镜"。**CLIP被优化为实例级语义特征，导致空间模糊。**

最优实践

Spatial-MLLM：CLIP（语义）+ 空间几何模型（结构）

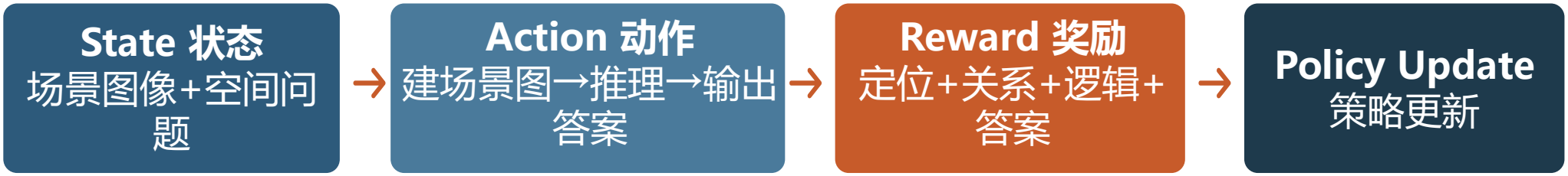
SpatialGeo：同一框架内融合语义和几何特征

强化学习在空间推理中的应用

为什么需要RL

- SFT只学习模仿正确答案，不知道"为什么对"
- RL通过**奖励信号**引导模型学习真正的空间推理能力
- 密集奖励（每步都有反馈）优于稀疏奖励（只看最终结果）

RL训练循环（以SpatialThinker为例）



↻ 循环迭代优化策略

GRPO：群体相对策略优化在空间推理中的应用

GRPO核心思想

① 对同一问题
采样多个回答

→

② 在组内
对回答进行相对排序

→

③ 利用相对排序信号
更新策略

为什么GRPO适用于空间推理?

原因	说明
无需绝对奖励模型	空间推理的"正确答案"难以用单一数值衡量
相对比较更容易	"A比B好"比"A得92分"更容易学习
鲁棒性更强	减少对奖励模型质量的依赖

Spatial-MLLM两阶段训练

阶段一 SFT：学习基本空间推理能力

阶段二 GRPO：进一步优化空间推理质量

范式转变：从单次感知到多视角推理

传统范式："单次感知"

静态图像 → MLLM → 答案

推理完全依赖于单一视角的信息

歧义：无法消解

遮挡：无从判断

新兴范式："多视角推理"

视角1 → 发现问题 → 生成视角2 → 整合证据 → 更可靠的答案

识别空间歧义 → 主动寻求额外视角

代表工作

TwNV
生成式新视角合成+推理循环

SpatialImaginer
视觉想象力+文本思维链协同

SpatialDreamer
主动探索+视觉想象+证据融合

训练方法 vs 推理增强：如何选择?

维度	训练方法（Training）	推理增强（Inference）
计算成本	高（需GPU集群训练）	低（仅推理时额外计算）
数据需求	大量标注数据	无需额外训练数据
部署灵活性	需部署新模型	即插即用，不改权重
泛化能力	受训练数据分布限制	可利用外部专家模型
更新频率	需要重新训练	可随时更新工具链
代表方法	SpatialVLM, SpatialLLM	TwNV, ViSRA, LAST
性能上限	理论上限更高	受基础模型能力限制

适用场景建议

有大规模数据+充足算力 → **训练方法**

快速部署+无需重训 → **推理增强**

最优方案 → 组合使用

模块小结：我们学到的核心技术方法

训练范式（6种核心方法）

SpatialVLM
让数据自己说话

SpatialRGPT
给VLM装上"深度感知"的眼睛

SpatialLLM
数据×架构×训练的系统性协同优化

Spatial-MLLM
纯2D输入也能实现3D空间推理

SpatialGeo
语义理解+几何理解双流合一

SpatialThinker
强化学习让AI像人类一样空间感知

推理增强（4种前沿方法）

TwNV
如果看不清就换个角度看

LAST
调用深度估计工具来帮忙

SpatialImaginer
看不见的地方用想象力补全

ViSRA
零成本空间推理增强

实验分析与关键发现

发现	证据	启示
相对关系优于绝对距离	"A在B左边"比"A距B多少米"更准确	定量空间感知仍不成熟
3D不一定优于2D	3D LLM相比2D LLM未表现出显著优势	有效利用3D信息仍是难题
数据规模≠数据质量	少量高质量空间数据效果远超大量通用数据	空间感知需要"精准打击"
跨场景泛化挑战大	室内表现远超室外/空中场景	场景多样性是关键瓶颈
多步推理仍是痛点	MSR准确率显著低于单步	需要更好的推理链方法

总结：当前MLLM的空间推理能力仍处于"**初级阶段**"——在相对关系判断上表现尚可，但在定量距离估计、多步推理和跨场景泛化方面仍有巨大提升空间。

消融实验：什么因素最重要?

SpatialVLM消融——数据混合比例

- 5%空间数据即可获得绝大部分收益 → 空间数据的关键在于**"质"而非"量"**

Spatial-MLLM消融——训练阶段

- SFT提供基础能力，GRPO进一步精细优化 → **两阶段训练缺一不可**

SpatialGeo消融——编码器选择

- 融合方式比编码器选择更重要 → **好的融合能最大化互补优势**

核心结论： 数据质量 > 数据数量 | 训练阶段协同 > 单阶段 | 融合方式 > 编码器选择

本节内容

CONTENTS

一、什么是空间推理

二、核心技术方法

三、评估与基准

四、应用与前沿趋势

空间推理评估基准分类总览

维度一：按空间维度

2D空间推理： CLEVR, GQA, Spatial-DISE
3D空间推理： ScanQA, SpatialRGPT-Bench, Open3D-VQA
多图/多视角： MMSI-Bench, Spatial-DISE

维度二：按推理类型

感知层（物体识别、定位）→ **关系层**（方位、距离、拓扑）→ **推理层**（多步推理、运动预测、视角变换）

维度三：按场景类型

- 合成场景（CLEVR）
- 真实室内（ScanQA, Matterport3D）
- 真实室外/空中（Open3D-VQA）
- 自动驾驶（SURDS, nuScenes）
- 第一人称（Ego4D, SFI-Bench）

经典基准：CLEVR与GQA

CLEVR（2017）

人工合成的几何场景（球体、立方体、圆柱体）

问题类型：计数、存在、属性比较、空间关系

空间关系：左/右/前/后

优势：完全可控、无歧义、错误可控

局限：合成场景与真实世界差距大

GQA（2019）

基于Visual Genome的真实场景

空间问题子集：涵盖方位、距离、大小比较

优势：真实世界场景

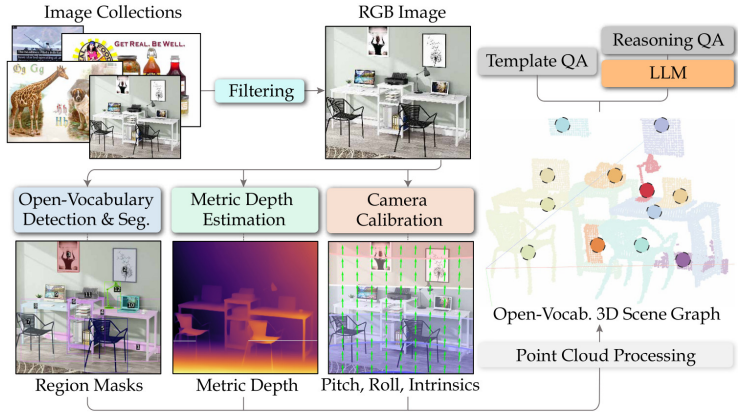
局限：空间推理深度有限，缺乏度量标注

经典基准确立了空间VQA的基本范式，但评估维度较为简单，无法覆盖复杂的三维空间推理需求。

区域级空间认知基准——SpatialRGPT-Bench

核心特点

特点	说明
三维场景覆盖	室内、室外和模拟环境
真实3D标注	使用真实3D标注（ground-truth annotations）
区域级评估	评估对用户指定区域的空间感知



评估维度示例

相对方向

"物体A在物体B的什么方向？"

距离估计

"物体A距离物体B大约多远？"

区域级推理

"区域R1在区域R2的哪个位置？"

空中空间推理——Open3D-VQA

73,000个 QA对，涵盖**7种**通用空间推理任务

多种题型：选择题/判断题/简答题 | 双模态支持：视觉图像+点云数据

7大空间推理任务

物体计数 | 空间关系推理 | 距离比较 | 相对高度判断 | 路径规划 | 遮挡判断 | 场景理解

关键发现

- 模型对**相对空间关系**的回答优于**绝对距离**
- **3D LLM相比2D LLM未表现出显著优势**
- 在模拟数据集上的微调可**显著提升**真实场景的空间推理性能

自动驾驶与6D空间推理——SURDS与Spatial457

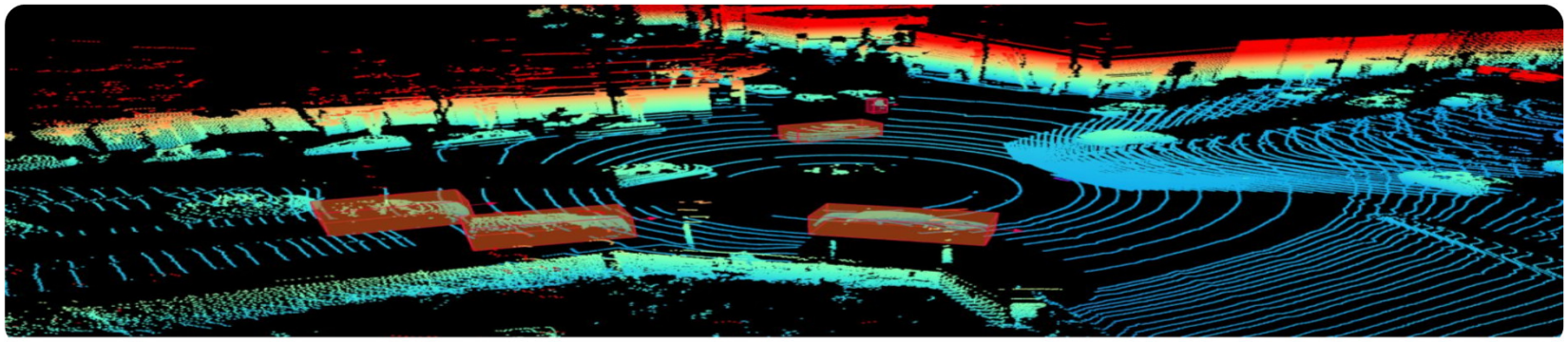
SURDS——自动驾驶空间推理

特征	数值/说明
基础数据	nuScenes自动驾驶数据集
训练样本	41,080个视觉-问答对
评估样本	9,250个
空间类别	6类空间任务
任务示例	物体定位/距离估计/相对运动/路径预测/碰撞判断/场景理解

Spatial457——6D空间推理诊断

核心能力	说明
多物体识别	场景中的物体检测与识别
2D定位	物体在2D图像中的位置
3D定位	物体在3D空间中的位置
3D方向	物体的3D朝向

特色指标：RPDR（相对性能下降率）——揭示模型在3D推理能力方面的关键弱点



MMSI-Bench：多图空间智能的"试金石"

设计理念：现有基准大多仅考察单图像内简单空间关系，真实世界空间理解需要**跨多图像追踪关联实体**。MMSI-Bench专门评估多图像空间推理。

关键数据

1,000	120,000+	300+小时	34个
高质量多选题	张真实图像构建	人工标注（6位研究员）	MLLM被全面评估

10+1种空间推理任务分类

位置关系（6种）： 相机-相机 / 相机-物体 / 相机-区域 / 物体-物体 / 物体-区域 / 区域-区域
属性（2种）： 测量（长度/大小等）/ 外观（形状等）
运动（2种）+ 多步推理（MSR）： 相机运动 / 物体运动 / 复杂顺序推理

(Position) Q: Upon entering from the door in Image 3, facing south, in which direction is the book located relative to the bed?

(Motion) Q: The images were taken continuously. In which direction does the camera rotate from Image 2 to Image 3?

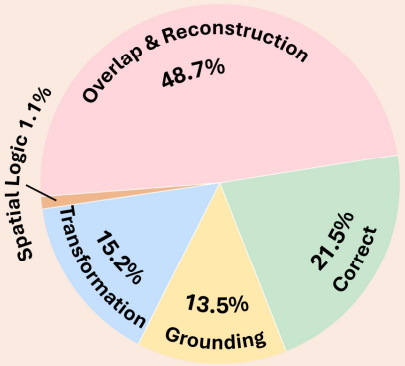
(Attribute) Q: Which is wider: the nightstand in Image 1 or the desk in Image 3?

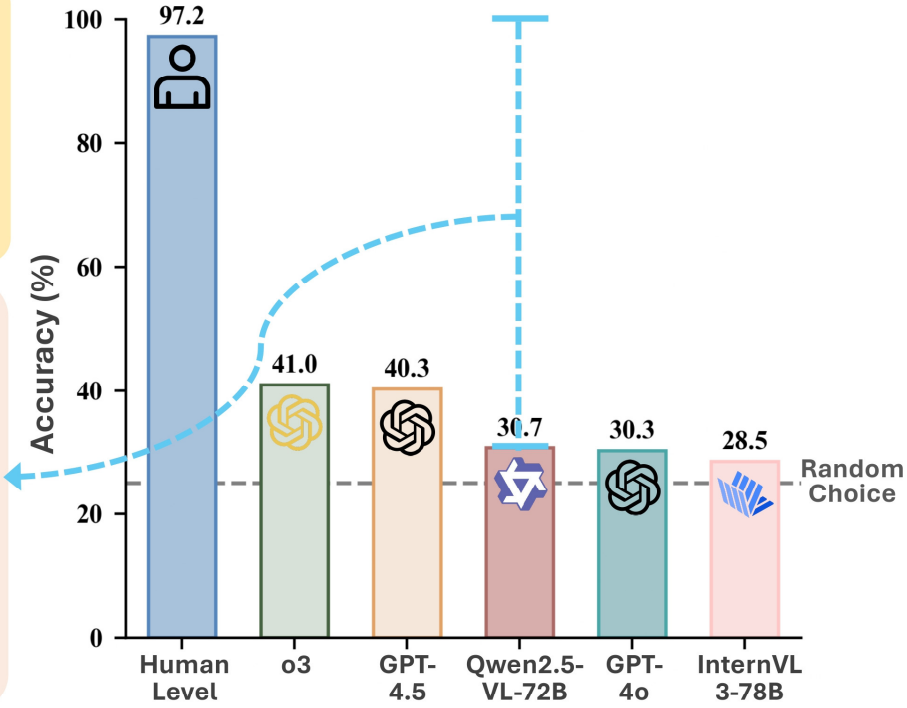
(Multi-Step Reasoning) Q: Upon entering the room through the door in Image 3, one faces south. In which direction is the lamp located relative to the chair?

Answer:
A. East B. West C. North D. South

Reasoning: In Image 3, entering the room while facing south, the book lies ahead and the bed to the front right, indicating that the book is positioned east of the bed.

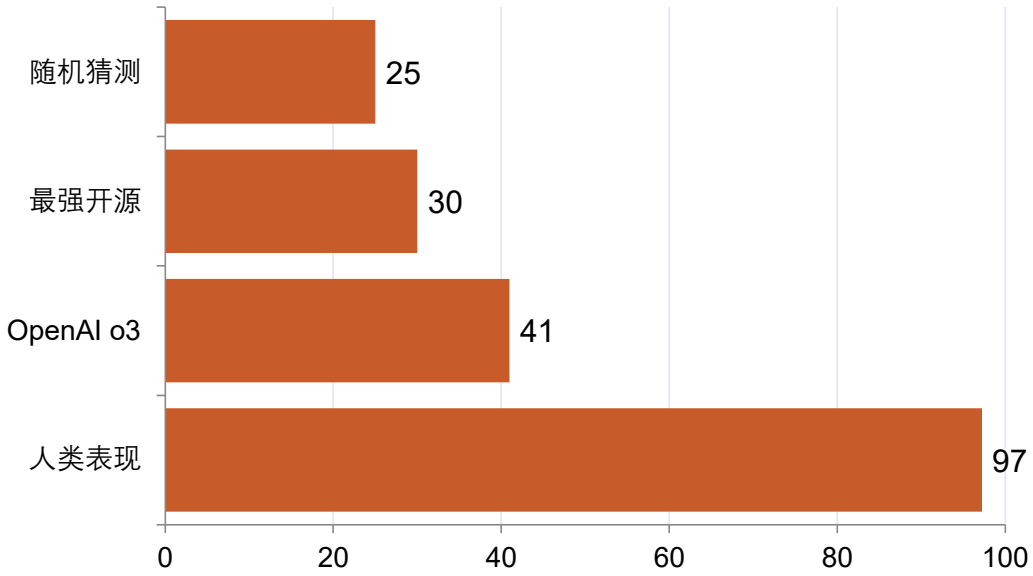
Error Analysis with MLLM





Fully Human Annotated | *1K Data* | *Diverse Scenarios* | *Reasoning Provided* | *Automated Error Analysis*

MMSI-Bench：令人警醒的实验结果



核心发现

- 即使**最先进的MLLM**，在多图像空间推理上仍举步维艰
- 最强开源模型仅约**30%**——大多数模型仅略高于随机猜测
- OpenAI o3准确率仅**41.0%**，而人类高达**97.2%**——差距超56%
- "盲眼GPT-4o"（无图像输入）准确率近乎随机，证明任务对**真实视觉空间推理的依赖**

多图空间推理是人类与AI差距最大的领域

MMSI-Bench：四大错误类型深度分析

Grounding错误

未正确识别图像中的物体——模型"看到"了不存在的物体

场景重建错误

难以解释空间关系和重建整体场景——无法拼接全局理解

情境变换推理错误

无法考虑场景在多张图像间的变化——不能追踪物体运动

空间逻辑错误

空间推理基本逻辑的缺陷——最基础的左右/上下关系判断出错

MMSI-Bench不仅是一把"尺子"，更是一张"诊断书"——精确指出了MLLM空间推理的薄弱环节，为改进指明方向。

Spatial-DISE：从认知角度评估空间推理

Intrinsic-Static 内在静态

"这个杯子的开口朝向哪个方向？" / "这个房间有几个窗户？"

Intrinsic-Dynamic 内在动态

"如果把这个魔方旋转90度，红色面朝哪？" / "盒子打开后是什么形状？"

Extrinsic-Static 外在静态

"桌子左边的椅子是第几把？" / "两个楼哪个更高？"

Extrinsic-Dynamic 外在动态

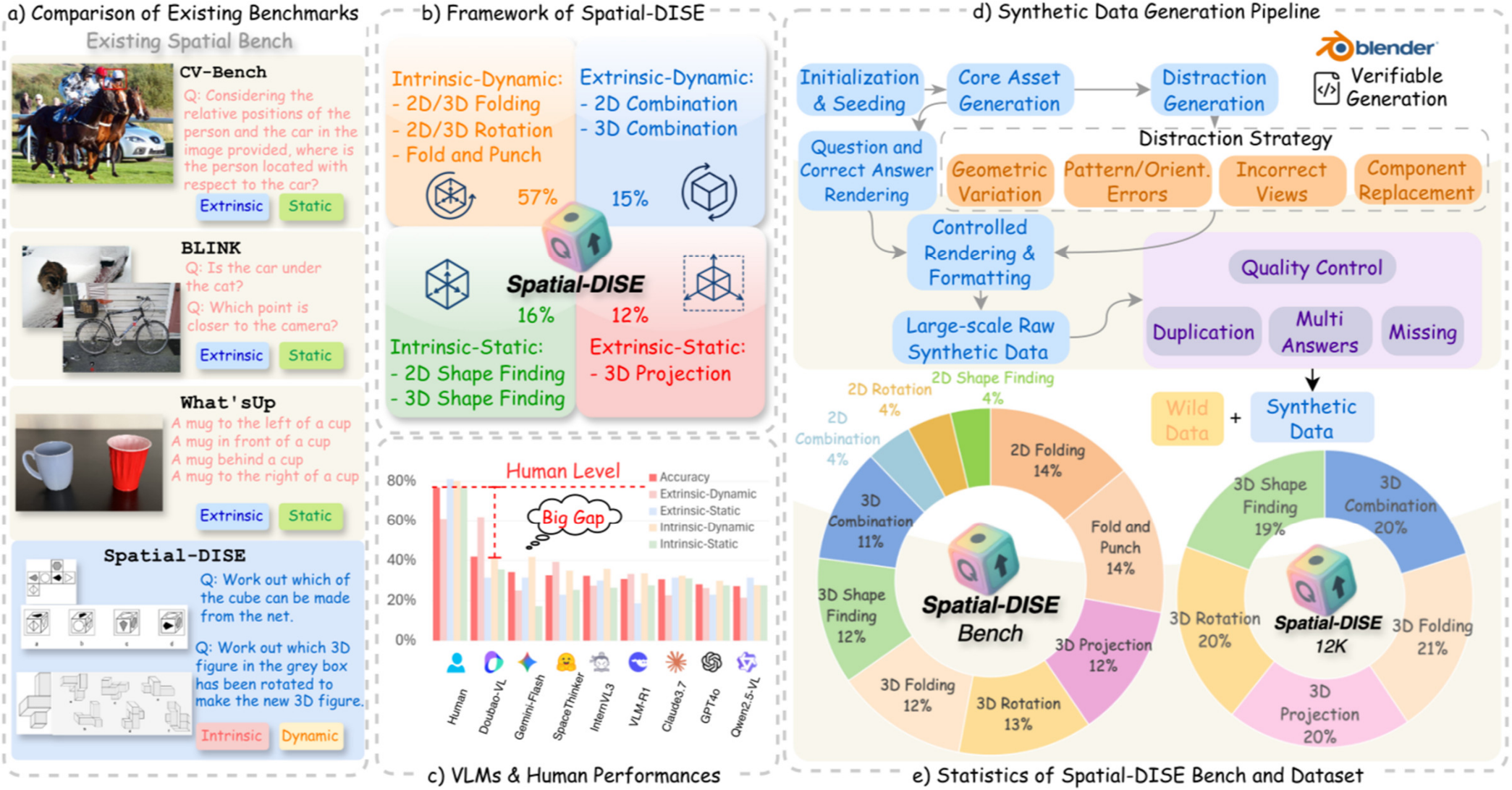
"从A点走到B点要经过哪些路口？" / "两辆车会不会相撞？"

数据集规模

559个高质量评估VQA对

12,000+个训练VQA对

评估**28个**SOTA VLM


a) Comparison of Existing Benchmarks
Existing Spatial Bench
CV-Bench
Q: Considering the relative positions of the person and the car in the image provided, where is the person located with respect to the car?
Extrinsic
Static
BLINK
Q: Is the car under the cat?
Q: Which point is closer to the camera?
Extrinsic
Static
What'sUp
A mug to the left of a cup
A mug in front of a cup
A mug behind a cup
A mug to the right of a cup
Extrinsic
Static
Spatial-DISE
Q: Work out which of the cube can be made from the net.
Q: Work out which 3D figure in the grey box has been rotated to make the new 3D figure.
Intrinsic
Dynamic
b) Framework of Spatial-DISE
Intrinsic-Dynamic:
- 2D/3D Folding
- 2D/3D Rotation
- Fold and Punch
57%
Extrinsic-Dynamic:
- 2D Combination
- 3D Combination
15%
Spatial-DISE
16%
12%
Intrinsic-Static:
- 2D Shape Finding
- 3D Shape Finding
Extrinsic-Static:
- 3D Projection
c) VLMs & Human Performances
Human Level
Big Gap
Accuracy
Extrinsic-Dynamic
Extrinsic-Static
Intrinsic-Dynamic
Intrinsic-Static
80%
60%
40%
20%
0%
Human
Doubao-VL
Gemini-Flash
SpaceThinker
InternVL3
VLM-R1
Claude3.7
GPT4o
Qwen2.5-VL
d) Synthetic Data Generation Pipeline
blender
Verifiable Generation
Initialization & Seeding
Core Asset Generation
Distraction Generation
Question and Correct Answer Rendering
Distraction Strategy
Geometric Variation
Pattern/Orient. Errors
Incorrect Views
Component Replacement
Controlled Rendering & Formatting
Large-scale Raw Synthetic Data
Quality Control
Duplication
Multi Answers
Missing
Wild Data
+
Synthetic Data
e) Statistics of Spatial-DISE Bench and Dataset
2D Rotation 4%
2D Shape Finding 4%
2D Combination 4%
2D Folding 14%
3D Combination 11%
Fold and Punch 14%
3D Shape Finding 12%
Spatial-DISE Bench
3D Projection 12%
3D Folding 12%
3D Rotation 13%
3D Shape Finding 19%
3D Combination 20%
3D Rotation 20%
Spatial-DISE 12K
3D Folding 21%
3D Projection 20%

从空间到功能——SFI-Bench与VSI-Bench

SFI-Bench——空间-功能智能

基于视频的基准，超过**1,500个**专家标注问题

来自多样化的第一人称室内视频扫描

两个互补维度：

① 结构化空间推理

② 功能推理

关键发现：当前MLLM在结合空间记忆与功能推理方面效果较差

VSI-Bench——层次化空间智能

涵盖物体计数、相对方向、路径规划等任务

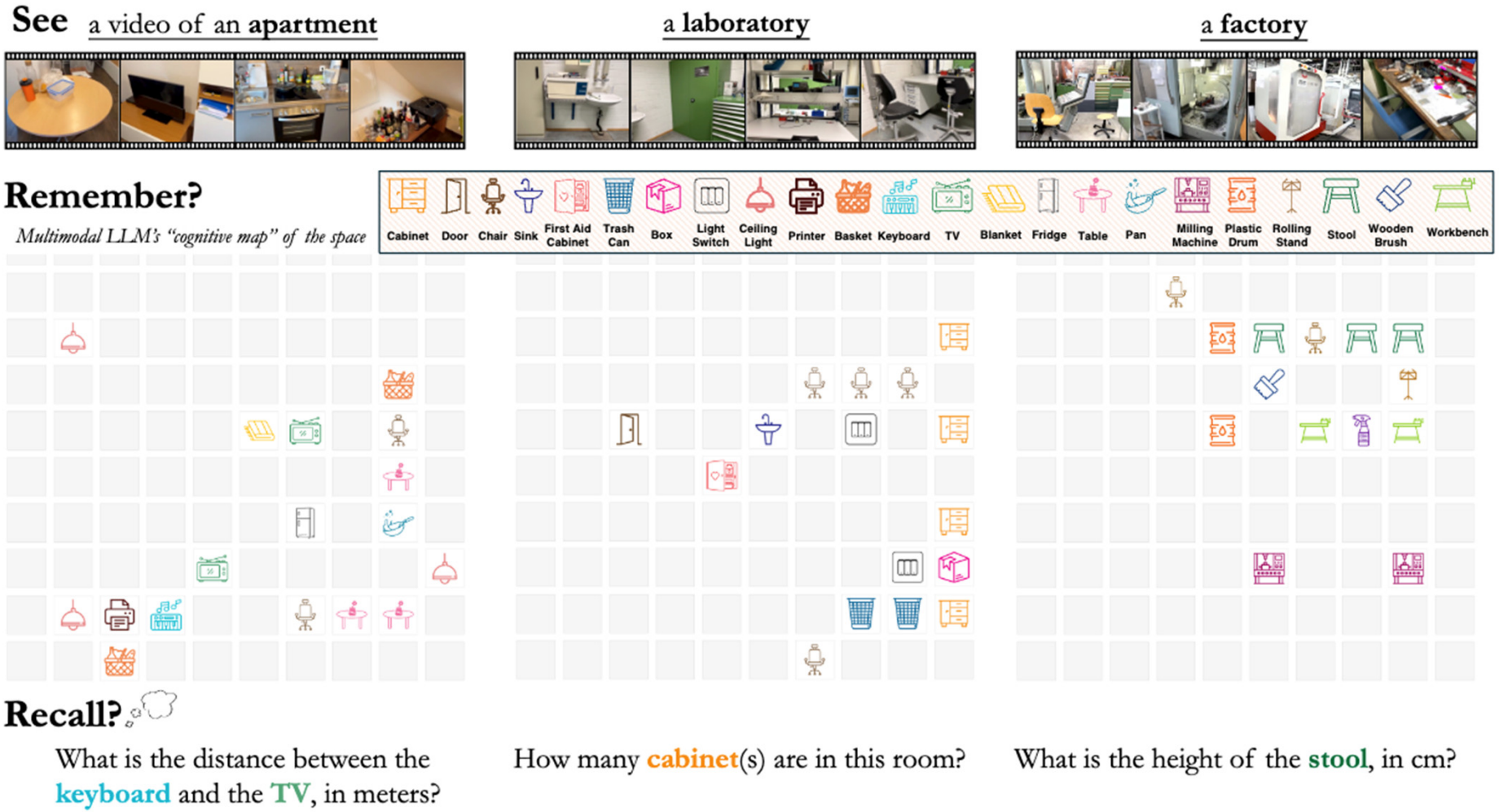
多个代表性方法在此基准上进行对比

领先结果：

SpatialDreamer平均准确率**62.2%**

全面领先其他方法

SFI-Bench


See a video of an apartment
a laboratory
a factory
Remember?
Multimodal LLM's "cognitive map" of the space
Cabinet
Door
Chair
Sink
First Aid Cabinet
Trash Can
Box
Light Switch
Ceiling Light
Printer
Basket
Keyboard
TV
Blanket
Fridge
Table
Pan
Milling Machine
Plastic Drum
Rolling Stand
Stool
Wooden Brush
Workbench
Recall?
What is the distance between the keyboard and the TV, in meters?
How many cabinet(s) are in this room?
What is the height of the stool, in cm?

VSI-Bench

Object Count

How many chairs are there in this room?

Answer: 4

Relative Distance

Measuring from the closest point of each object, which of these objects (refrigerator, sofa, ceiling light, cutting board) is the closest to the printer?

A. refrigerator B. sofa C. ceiling light D. cutting board

Appearance Order

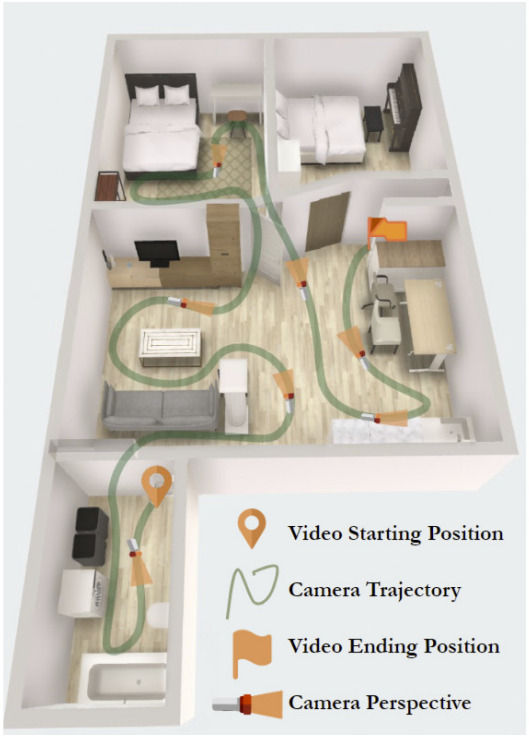
What will be the first-time appearance order of the following categories in the video: basket, printer, refrigerator, kettle?

A. kettle, basket, printer, refrigerator
B. refrigerator, printer, basket, kettle
C. basket, printer, refrigerator, kettle
D. basket, refrigerator, kettle, printer

Relative Direction

If I am standing by the refrigerator and facing the sofa, is the kettle to my left, right, or back?

A. left B. right C. back



Object Size

What is the length of the longest dimension (length, width, or height) of the refrigerator in centimeters?

Answer: 119

Absolute Distance

Measuring from the closest point of each object, what is the distance between the bed and the sofa in meters?

Answer: 3.2

Room Size

What is the size of this room (in square meters)? If multiple rooms are shown, estimate the size of the combined space.

Answer: 57.6

Route Plan

You are a robot beginning at the toilet and facing the washer. Navigate to the pan. Fill in this route: 1. Go forward until the washing machine 2. [?] 3. Go forward until the sofa 4. [?] 5. Go forward until the pan.

A. Turn Left, Turn Left B. Turn Left, Turn Right
C. Turn Back, Turn Right D. Turn Right, Turn Right

空间推理基准全景对比

基准	维度	图像数	场景	数据量	人类表现	模型最优
CLEVR	2D	单图	合成	100万+	~100%	~99%
GQA	2D	单图	真实	2200万+	~90%	~80%
SpatialRGPT	3D	单图+深度	混合	数千	N/A	持续提升
Open3D-VQA	3D	单图+点云	空中	73K	N/A	改进中
SURDS	3D	多图	自动驾驶	50K+	N/A	改进中
MMSI-Bench	3D多图	多图	真实	1K	97.2%	o3 41%
Spatial-DISE	2D/3D	单/多图	合成+真实	12.6K	~95%	显著差距
VSI-Bench	3D多图	多图	真实+合成	数千	N/A	62.2%
SFI-Bench	3D视频	视频	室内	1.5K	N/A	改进中
Spatial457	3D+6D	单图	合成	457	N/A	N/A

当从单图→多图、从简单2D→复杂3D，所有模型的表现都会急剧下降

空间推理的评估方法与指标

主流评估方法

方法	适用任务	优势	局限
准确率	选择题/判断题	简单直观	无法区分部分正确
精确匹配EM	简答题	严格	对表述过于敏感
RPDR	诊断性评估	揭示特定能力弱点	仅适用于合成场景

认知角度的评估方法

多维度评估 — 不只关注最终答案，还关注推理过程
层次化评估 — 按认知层次（感知→映射→模拟→智能体）分别评估
错误类型分析 — 深入分析失败的原因，而非简单算准确率
跨任务对比 — 在统一的认知框架下比较不同任务的难易程度

基准生态系统的挑战与改进方向

当前四大局限

单图像偏见 — 绝大多数基准仅评估单图像内空间关系

合成数据依赖 — 限制问题多样性与真实性

静态评估 — 缺乏对动态场景的评估

粒度不足 — 简单准确率无法揭示失败具体原因

五大改进方向

1. **多图像/多视角评估**（MMSI-Bench范式）
2. **真实世界场景**（Open3D-VQA范式）
3. **认知启发的分类**（Spatial-DISE范式）
4. **细粒度错误分析**（MMSI-Bench范式）
5. **动态+功能推理**（SFI-Bench范式）

本节内容

CONTENTS

一、什么是空间推理

二、核心技术方法

三、评估与基准

四、应用与前沿趋势

空间推理如何赋能现实世界?

智能家居机器人
"从杂乱茶几上递过最近遥控器"
需理解精确相对位置

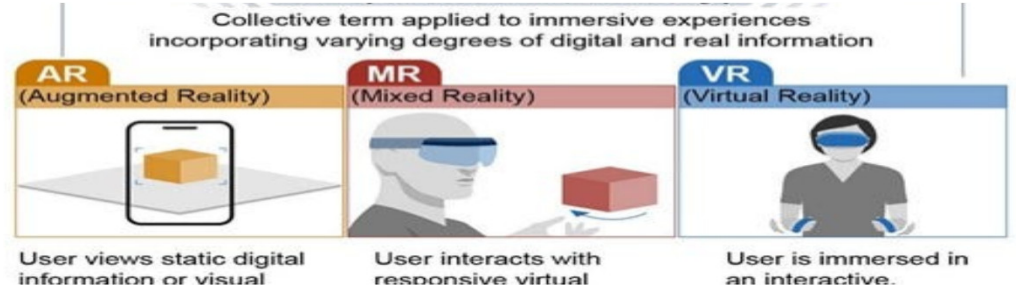
工业机器人
"从货架取出指定零件并按位置组装"
需理解三维操作空间

自动驾驶
"判断旁车道车辆与本车安全距离"
需实时多物体运动预测

医疗影像分析
"分析肿瘤在三维器官中的位置和侵犯范围"
需精确三维空间关系

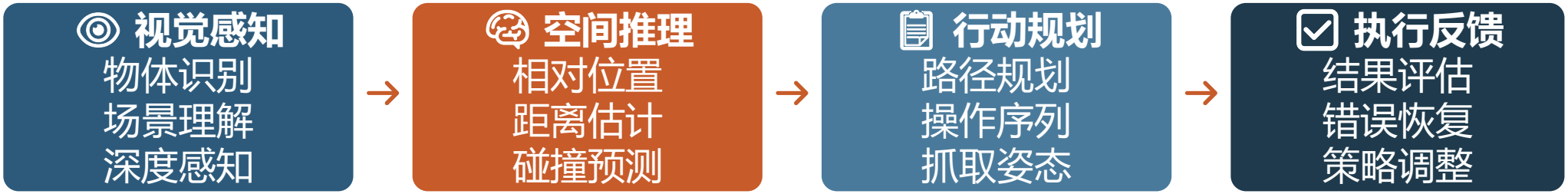
建筑设计
"从平面图纸理解三维空间布局"
需2D→3D心理映射

AR/VR
"虚拟物体与真实环境精确融合"
需实时空间对齐与遮挡处理



具身智能：当空间推理遇上机器人

空间推理是具身智能的"认知引擎"



空间推理缺失 → 机器人将无法理解指令、无法规划路径、无法安全操作

典型任务——视觉语言导航（VLN）

输入： "去厨房，从冰箱里拿一瓶水"
需要： 场景理解 + 空间定位 + 路径规划 + 物体识别 + 操作执行

具身智能中的空间推理技术栈

决策与执行层：视觉语言导航（VLN）/ 机器人操作规划 / 任务分解与重规划 / 安全约束检查

空间推理层（关键枢纽）

空间VQA / 场景图构建 / 距离方向估计 / 碰撞预测 / 路径规划

遮挡推理 / 多视角融合 / 运动预测 / 空间常识推理

多模态感知层：RGB相机 / 深度传感器 / LiDAR / IMU / 触觉传感器

SpatialRGPT在机器人中的应用：充当区域感知密集奖励标注器，为机器人操作任务提供细粒度的空间反馈信号。

世界模型：空间智能的终极载体

概念：一种能对物理环境进行**生成、交互与状态预测**的多模态系统。

李飞飞提出的世界模型三大核心能力

① 感知3D/4D
含时间维度
而非仅处理2D图像

② 理解因果链
动作与结果的
因果关系

③ 主动交互学习
而非被动接受
标注数据

代表性世界模型

模型	机构	核心特点
无界大模型	悠然/码极客	跨空间、跨任务、跨本体泛化
Argus 1.0	如视	从全景图推测空间深度，专注真实复刻
WorldForge	西湖大学	推理时引导——不改权重，即插即用

李飞飞：空间智能——AI的下一战

> "当全球AI竞赛仍聚焦于语言模型的参数与上下文长度时……真正的智能，从来不只是'会说话'——而是**理解并驾驭物理世界的能力**，即**空间智能**。若AI无法掌握空间推理、物体关系与动态预测，所谓的'通用人工智能'终将是空中楼阁。"

空间智能三阶段演进路径

近期 2025-2028
虚拟叙事赋能
赋能电影、游戏
虚拟场景自动生成
沉浸式内容创作

→

中期 2028-2035
服务机器人普及
真正理解家庭环境
自主导航与交互
日常任务执行

→

长期 2035+
科学发现推动
精准医疗与诊断
沉浸式教育普及
科学探索加速

前沿趋势一：想象力增强的空间推理

SpatialDreamer——让AI拥有"空间想象力"

闭环推理流程（模拟人类空间认知）

① 探索 Explore
推理最优自我中心动作
如"前进0.75米""左转45度"

→

② 想象 Imagine
调用世界模型生成执行动作后的新视角图像

→

③ 推理 Reason
整合累积视觉证据生成最终答案

核心突破：从"被动观察"到"主动目标导向的想象"——自主决定"去哪看、看什么、如何推理"

关键实验结果

93.9% SAT真实场景 SOTA	84.9% MindCube-Tiny	62.2% VSI-Bench 平均
合成场景 92.5%	较基线提升超55%	全面领先现有方法

前沿趋势二：推理时空间增强与可插拔范式

ViSRA——训练无关的空间推理增强

三层架构

专家模型层
深度估计 / 分割 / 3D重建
成熟视觉专家模型

空间信息聚合层
结构化空间信息提取
多模态提示生成

MLLM推理层
不修改任何权重
直接利用增强提示推理

核心优势

- **即插即用**——不需要任何后训练
- **专家知识利用**——充分利用成熟视觉专家模型
- **持续升级**——专家模型更新即升级MLLM能力

实验效果

+15.6%
现有基准最高提升

+28.9%
未见过3D任务提升

前沿趋势三：从空间到功能——高阶空间认知

SFI-Bench揭示的新需求

从"**知道在哪里**"到"**理解干什么用**"——几何感知 → 功能推理

双维度评估框架

结构化空间推理

- 理解复杂空间布局
- 形成连贯空间表征
- 条件计数与多跳关系推理
- 从局部到整体的空间整合

功能推理

- 推断物体可供性（affordance）
- 上下文相关的用途理解
- 功能配对与场景适配
- 知识基础的故障排除

⚠ **核心发现**：当前MLLM在结合空间记忆与功能推理方面持续挣扎——这是通往真正具身智能的关键瓶颈

十大开放问题：空间推理的未来研究

1. 如何高效获取高质量3D空间标注数据?
2. 如何设计真正3D原生的视觉编码器?
3. 如何实现实时空间推理?
4. 如何结合空间推理与物理常识?
5. 如何评估多步复杂空间推理?
6. 如何从视频中无监督学习空间知识?
7. 如何实现跨场景的空间推理泛化?
8. 如何利用大语言模型的推理能力?
9. 空间推理的幻觉问题如何解决?
10. 如何实现多智能体协作空间推理?

每一个问题的解决，都将推动空间智能向AGI迈出关键一步

问题和讨论