

OneCanvas: 通过全景重投影实现 3D 场景理解

Bartłomiej Baranowski¹ Dave Zhenyu Chen² Matthias Nießner¹

¹ 慕尼黑工业大学 ² 华为

项目主页: <https://baranowskibr.github.io/onecanvas/>
代码与模型: <https://github.com/baranowskibr/onecanvas>

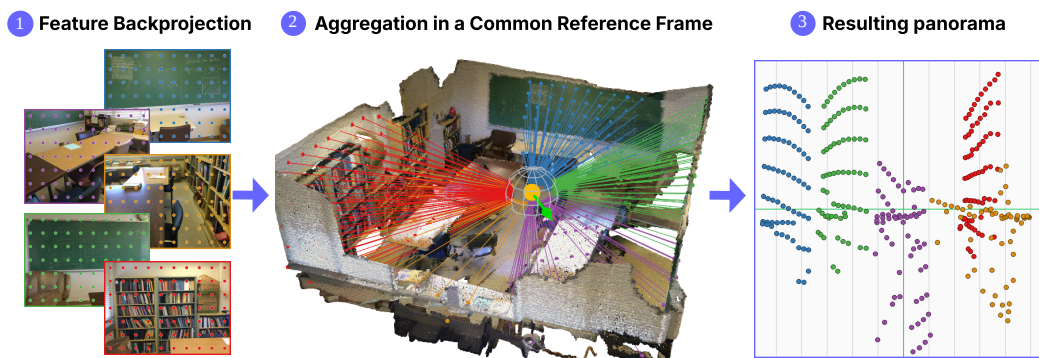


图 1: **OneCanvas** (1) 将逐帧图块特征反投影到 3D; (2) 在以选定视角为中心的公共参考系中聚合这些特征; (3) 按连续角位置把每个图块放置到全景画布上, 由视觉语言模型 (Vision-Language Model, VLM) 作为单张图像消费。

摘要

现有针对视觉语言模型 (Vision-Language Model, VLM) 的 3D 场景理解方法, 要么依赖复杂、与模型绑定的几何编码器, 要么为追求空间推理而付出庞大的训练预算。与之不同, OneCanvas 将所有视角的图块特征 (patch feature) 聚合到单张等距圆柱 (Equirectangular) 全景画布 (panoramic canvas) 上。具体而言, 每个图块由其深度与相机位姿反投影到 3D 世界坐标, 再按该点相对画布原点所见的连续经度与纬度置于画布之上, 既不做栅格化, 也不跨重叠视角聚合。图块度量坐标的三维位置嵌入被加到其特征上, 补回把世界位置压缩为画布角度坐标时所丢失的深度信息。由此, 所有帧的图块共享同一空间坐标系, 既无需融合, 也无需对骨干网络做大改动。预训练的 VLM 把这一表示当作普通图像消费。由于画布可以任意感兴趣的位姿为中心, 同一表示可直接支持从特定视角出发的情境化推理 (situated reasoning), 而这正是机器人学与具身智能 (embodied AI) 中的常见需求。得益于该表示, 我们还能引入空间预训练课程 (spatial pretraining curriculum): 把取自真实图像的物体图块特征以程序化方式放置到原本空白的画布上的指定 3D 世界位置, 即可即时生成覆盖广泛空间推理任务的监督信号, 并控制答案分布以减少空间推理捷径。OneCanvas 在 SQA3D 与 VSI-Bench 上取得最优结果, 并泛化到 SPBench 的分布外 (out-of-distribution) 数据, 而训练算力比基准性能最接近的竞争者低一个数量级。

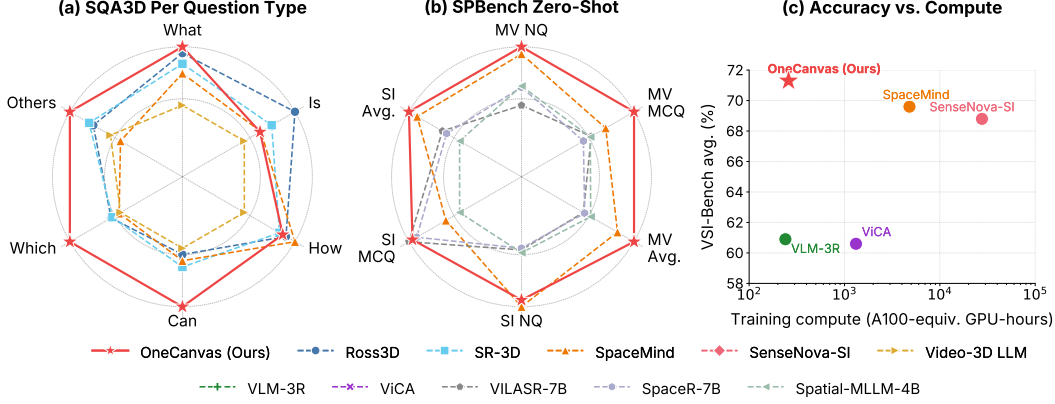


图 2: **OneCanvas** 基准对比。(a) 在 SQA3D 上的分问题类型准确率, 与 (b) 在 SPBench (多视角划分与单图划分) 上的零样本准确率, 均以雷达图形式与基准性能最接近的竞争方法进行对比。(c) 训练算力 (A100 等效 GPU 小时, 对数坐标) 与 VSI-Bench 准确率的对比。

1 引言

视觉语言模型 (Vision-Language Model, VLM) 改变了机器联合理解图像、视频与自然语言的方式, 但许多具有实际价值的理解任务本质上是三维的。在人类环境中行动的机器人、把指令锚定到用户物理环境中的增强现实助手, 以及在陌生空间中规划轨迹的自主智能体, 都需要回答物体在哪里、彼此相距多远、朝哪个方向转才能面对它们、以及从给定视点看什么可见什么不可见。构建能依据场景的普通视频可靠回答这类问题的 VLM, 正是 3D 场景理解的目标。

多数现有具备 3D 感知能力的 VLM 建立在预训练的 2D 视觉语言骨干网络 (Vision-Language Backbone) 之上, 本身没有空间推理机制。要让它们适应 3D, 要么从原始视频中隐式诱导几何信息——这种方式数据低效, 且容易把几何与场景特有的纹理先验混为一谈; 要么为骨干网络附加专门的几何编码器。即便做了这些改造, 所提供的几何信息往往并未被真正使用。近期针对情境化 3D 问答的审计发现, 在广泛引用的基准上, 纯文本基线已能追平甚至接近最先进的几何感知模型 Ma et al. [2026], 说明这些模型依赖的是问题与场景类别的统计规律, 而非对所给几何输入的推理。

尽管如此, 现有 3D VLM 在 3D 推理上仍取得了实质进展。一类方法为 VLM 增加专门的几何模块, 包括点云分词器 (Point-Cloud Tokeniser) Chen et al. [2024], Huang et al. [2024b], Wang et al. [2023], Zhu et al. [2024]、深度条件化的视频位置编码 Zheng et al. [2025b], Zhu et al. [2025], 以及 VLM-3R Fan et al. [2026] 与 SpaceMind Zhao et al. [2025] 等融合几何编码器——它们从单目视频中读取 3D 结构并注入视觉骨干网络。另一类方法扩大空间监督规模: SenseNova-SI Cai et al. [2026] 与 Cambrian-S Yang et al. [2026] 整理了大规模空间问答对, 靠纯粹的数据暴露让几何上毫无特殊设计的 VLM 学会推理距离与方向。两个方向都在各大空间基准上带来切实提升, 共同说明只要愿意在架构或数据上投入, 预训练 VLM 就能获得 3D 推理能力。

两个方向也都要付出高昂代价。架构融合引入的新模块必须在大规模上联合训练, 才能把新的几何编码器与语言骨干网络对齐; 数据扩展则需要整理和筛选数百万条高质量空间问答对。二者都未直接弥合上述审计揭示的缺口: 两个家族的方法仍只能逐帧、零散地观察场景, 也无法明确把握输入中究竟哪些信号编码了它们被要求学习的空间关系。为此, 本文提出 OneCanvas (图 1): 把每个视角的每个图块 (patch) 用其深度与相机位姿提升到 3D 世界坐标, 再按该点相对选定画布原点所见的经度与纬度放到单张等距圆柱 (Equirectangular) 画布上。这些球面坐标填入该图块 MRoPE 位置编码的空间轴, 源帧索引则占据时间轴。于是所有视角的全部图块共享同一空间坐标系, 预训练 VLM 像消费普通单图输入一样消费该结果, 无需对架构做大改。这一表示还支持一种在常规 2D 或视频流程中难以构建的空间预训练课程: 把取自真实图像的物体图块特征按程序放置到空白画布上选定的 3D 世界位置, 即可即时生成度量距离、方向、计数、可观测性与导航等任务的监督信号, 并控制答案分布以抑制近期审计指出的场景统计捷径 Ma et al. [2026]。OneCanvas 在 SQA3D Ma et al. [2023] (65.3 EM@1, 比此前最好结果高 2.3 个点)、VSI-Bench Yang et al. [2025a] (平均 71.3) 与 SPBench Li et al. [2026] (零样本整体 72.1, 高于次优方法 4.8 个点) 上达到最优, 而训练算力比基准表现最接近的竞争者低一个数量级 (图 2, 完整的归一化方法见节 C)。

本文的贡献如下：

- **全景特征重投影。**一种把全部视角的图块特征聚合到单张全景画布上的机制，与 VLM 预训练所用的输入格式一致，并可灵活选择画布原点与朝向以适应不同任务。
- **由该表示支持的空间预训练课程。**把取自真实图像的物体图块特征按程序放置到画布上选定的 3D 位置，即时构建几何监督，并控制答案分布以减少数据集统计捷径。

2 相关工作

通过架构扩展实现的 3D-aware VLM。 现有 3D VLM 大多为预训练语言骨干网络配备辅助模块，以提供几何信息 Ma et al. [2024a], Chen et al. [2024], Huang et al. [2024b], Wang et al. [2023], Zhu et al. [2024], Li et al. [2024], Zheng et al. [2025b], Fan et al. [2026], Zhao et al. [2025], Wang et al. [2025a], Cheng et al. [2025], Qi et al. [2026]。早期基于点云的 3D VLM Hong et al. [2023], Chen et al. [2024], Huang et al. [2024b], Wang et al. [2023], Zhu et al. [2024], Li et al. [2024] 为语言骨干网络搭配专用的点云分词器 (Point-Cloud Tokeniser)，在 3D QA 数据上端到端训练，验证了基于 LLM 进行 3D 场景推理的可行性。基于 RGB-D 的 3D VLM Zheng et al. [2025b], Zhu et al. [2025] 则把输入视作视频序列，将由深度图计算出的 3D 位置编码注入视频表示，从而在无需显式点云编码器的前提下打通视频与 3D 两个领域。HiSpatial Liang et al. [2026] 沿这一路线进一步把度量尺度点图作为辅助输入与 RGB-D 一并送入，并通过从几何感知到抽象空间推理的分层课程对 VLM 进行监督。随着几何基础模型 (Geometry Foundation Model) 的出现 Wang et al. [2024a], Leroy et al. [2024], Wang et al. [2025b], Lin et al. [2026]，较新的 3D VLM 直接从单目视频中提取 3D 词元，并在送入 LLM 之前与视觉特征融合 Fan et al. [2026], Zhao et al. [2025], Wang et al. [2025a], Cheng et al. [2025], Qi et al. [2026], Zheng et al. [2025a]。VLM-3R Fan et al. [2026] 将几何编码器输出的隐式 3D 词元与视觉特征融合。SpaceMind Zhao et al. [2025] 加入 Camera-Guided Modality Fusion 模块，把空间编码器与视觉编码器结合。Ross3D Wang et al. [2025a] 提出重构式视觉指令微调目标，在视频 VLM 的微调阶段提供 3D 感知监督。SR-3D Cheng et al. [2025] 是支持区域提示的 VLM，在共享的视觉词元空间中统一单视图 2D 输入与多视图 3D 输入。GPT4Scene Qi et al. [2026] 在 3D 场景视频序列上微调 Qwen2-VL。SD-VLM Chen et al. [2025] 提出正弦深度位置编码，沿相机 z 轴注入深度信息。Loc3R-VLM Qu et al. [2026] 把预训练几何基础模型的相机先验与 BEV 布局、情境建模等辅助目标相结合。另一条平行路线采用带空间奖励信号的强化学习 Ouyang et al. [2025], Wu et al. [2025b], Batra et al. [2025], Chen et al. [2026]，它与表示选择相互正交，原则上可与本文的全景画布结合。上述方法都在基础 VLM 之上引入了新的架构组件、辅助目标或专用 3D 编码器。相比之下，本文方法不引入任何新架构组件、任何辅助损失项，也不使用专用 3D 编码器，3D 理解能力来自输入的组织方式。

统一场景表示。 全景图像表示提供 360° 视场，天然保留长距离空间关系，这一特性已被用于场景理解 Zheng et al. [2025c]，近期也用于 3D 视觉定位。PanoGrounder Jung et al. [2025] 在预先重建的网格或 3D 高斯泼溅 (3D Gaussian Splatting) 场景周围放置虚拟相机并渲染多张等距圆柱全景图，再用大型 VLM 处理以完成定位。PanoEnv Lin and Zheng [2026] 在合成环境的等距圆柱图像上研究 VLM 的空间智能，表明 ERP 图像的几何畸变仍是现成 VLM 面临的难题。把 2D 特征反投影到 3D、再重投影到规范 2D 视图这一更一般的思路，在自动驾驶中已通过 BEV 方法得到充分确立 Philion and Fidler [2020], Ma et al. [2024b]。Lift3D T et al. [2024] 把各视角特征提升到神经场 (Neural Field) 并渲染目标视图，LiftProj Jia et al. [2025] 则采用类似的先提升后投影范式完成全景拼接。本文将这一原理引入基于 VLM 的 3D 场景理解：各视角特征借助度量深度与相机位姿被提升，一次性投影到全景画布上，并直接由 VLM 消费；画布原点可自由选择，以同时支持以场景为中心和以位姿为中心的情境化推理。与 PanoGrounder、LiftProj 和 PanoEnv 不同，本文不为视觉塔重新渲染等距圆柱图像再编码——视觉塔看到的仍是原始透视图，只有由此得到的图块位置被重新参数化。与渲染或 Lift3D 的光线合成 (Ray Compositing) 也不同，画布还保留被遮挡和位置重合的观测，作为独立词元。受控对比见 小节 B.2。

3 方法

给定一组 K 张带位姿的 RGB 图像 $\{I_k\}_{k=1}^K$ ，配套度量深度图 $\{D_k\}$ 与针孔相机内参 $C_k = (f_x, f_y, c_x, c_y)$ (逐帧的焦距与主点)，目标是构造一种预训练视觉语言模型 (Vision-Language Model, VLM) 可直接接收的单一全景表示。该表示分两步构建 (图 3)：(1) 从每

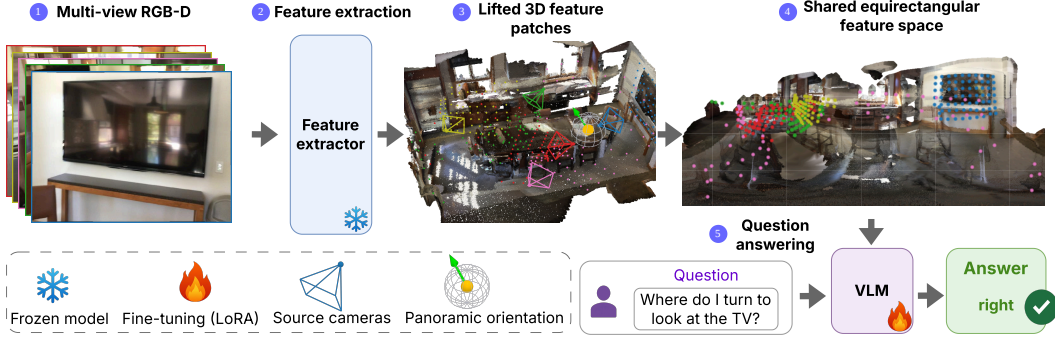


图 3: 方法概览。多视角图像由 Qwen3-VL 的冻结视觉编码器编码，提升后的图块落在共享等距圆柱全景图上的连续位置。VLM 通过其原生注意力层消费该表示。

个视角提取逐图块特征并提升到 3D 世界坐标 (小节 3.1); (2) 把提升后的图块重投影到共享的等距圆柱画布上, 为每个图块的度量坐标加入 3D 位置嵌入, 再用其原生的 3D-RoPE 位置编码拼装 VLM 的输入序列 (小节 3.2)。训练分两阶段 (图 4): 第一阶段为空间预训练 (小节 3.3), 随后是在下游目标数据上的第二阶段适配 (小节 3.4)。

3.1 特征提取与 3D 提升

每张图像 I_k 经 Qwen3-VL 的冻结视觉编码器 Bai et al. [2025] 编码, 得到图块级特征图 $F_k \in \mathbb{R}^{H_f \times W_f \times D}$, 其网格分辨率为编码器的 $H_f \times W_f$, 单图块特征维度为 D 。每个图块 (u, v) 利用其深度 $z = D_k(u, v)$ 、缩放到特征图分辨率的内参以及相机到世界位姿 $T_k \in \text{SE}(3)$ 提升到世界坐标:

$$\mathbf{p}_{u,v}^{\text{world}} = T_k \begin{pmatrix} (u - c_x)z/f_x \\ (v - c_y)z/f_y \\ z \end{pmatrix}, \quad (1)$$

其中 $T_k \in \text{SE}(3)$ 的作用为 $T_k \mathbf{x} = R_k \mathbf{x} + \mathbf{t}_k$, (u, v, z) 遵循 OpenCV 相机约定 (x 向右、 y 向下、 z 向前)。提升全部 K 个视角后, 得到提升图块集合 $\mathcal{P} = \{(\mathbf{p}_i, \mathbf{f}_i, t_i)\}$, 其中世界坐标为 $\mathbf{p}_i \in \mathbb{R}^3$, 特征为 $\mathbf{f}_i \in \mathbb{R}^D$, 逐图块时间索引 $t_i \in \{1, \dots, K\}$ 记录源帧在输入视频中的位置。 t_i 一直保留到全景放置阶段 (小节 3.2), 在那里成为该图块在语言模型时间 MRoPE 轴上的坐标。

3.2 全景画布与位置编码

本文将每个提升图块放置在由原点 $\mathbf{c} \in \mathbb{R}^3$ 与朝向 $R \in \text{SO}(3)$ 定义的等距圆柱画布上。二元组 $(\mathbf{c}, R)$ 可自由设计, 可设为对任务最自然的任意坐标系: 用于情境推理的智能体位姿、使场景覆盖最大的视点, 或其他任意参考。训练与推理所用的具体约定见 小节 4.1。

对每个提升图块, 计算其在画布坐标系下的局部偏移 $\mathbf{q}_i = R^\top (\mathbf{p}_i - \mathbf{c}) = (q_x, q_y, q_z)$ 及连续球坐标: 经度 $\theta_i = \text{atan2}(q_x, q_z)$, 纬度 $\phi_i = \arctan(-q_y / \sqrt{q_x^2 + q_z^2})$ 。图像 y 轴在相机坐标系中向下, 因此 q_y 度量向下方向, 纬度公式中的 $-q_y$ 使正纬度朝上。

连续位置, 而非像素网格。 一种自然的替代做法是把提升图块栅格化到固定像素网格上, 再以人工设计的规则 (平均、基于深度的选择、后写覆盖) 归约格内内容。这会把投影到相同或相近角位置的图块合并 (例如沿同一视线光线的被遮挡表面, 或重叠帧中位置重合的观测), 破坏表面之间的区分, 并用固定归约取代预训练注意力的判断。本文改为让每个提升图块保留为独立词元, 置于其连续的 (ϕ, θ) 位置。位置重合的图块仍是不同词元, 由下文描述的 3D 位置嵌入加以区分, 重叠交予注意力解决, 而非预设规则。栅格化、渲染像素与经度接缝的对照实验见 小节 B.2 与 小节 B.5。

位置编码。 VLM 原样使用其自带的 3D 旋转位置编码 (3D Rotary Position Embedding, 3D-RoPE) Bai et al. [2025]。与标准图像/视频设定相比, 仅位置索引的内容不同。将 W 设为经度 $\theta_i \in [-\pi, \pi]$, H 设为纬度 $\phi_i \in [-\pi/2, \pi/2]$, T 设为源帧索引 $t_i \in \{1, \dots, K\}$, 三者线性重缩放到 $[0, 100]$, 以落在模型预训练时的有效位置分布内。 T 沿用视频中的标准“时间 / 帧序”语义。等距圆柱畸变与经度接缝在 小节 B.5 中讨论。

3D 位置嵌入。 2D VLM 只能从像素中习得依赖类别的尺度先验，而非对实际布局的度量读数。本文通过编码画布坐标系偏移 $\mathbf{q}_i = (q_x, q_y, q_z)$ 显式提供逐图块的度量位置，该编码由三部分拼接而成：(i) 各轴的正弦项 $\{\sin(\omega_k q_a), \cos(\omega_k q_a), q_a\}$, $a \in \{x, y, z\}$ ，取 16 个对数间隔频率 $\omega_k \in [0.1, 100]$ rad/m；(ii) 对径向基量 $\|\mathbf{q}_i\|$ 与 $r_{xz} = \sqrt{q_x^2 + q_z^2}$ 施加相同形式，取 8 个对数间隔频率 $\nu_k \in [0.1, 10]$ rad/m；(iii) 单位光线方向 $\mathbf{q}_i/\|\mathbf{q}_i\|$ 直通。这些径向基量直接暴露出与成对距离、以及占地面积等水平径向读数相关的量，无需模型自行从各轴分量合成。该 136 通道编码经隐层宽度 64 的 2 层 MLP 投影，并通过学习到的标量门控加到图块特征上。在特征空间而非以额外 RoPE 维度注入度量位置，可保持 3D-RoPE 预训练的角向与时间语义不变，并让 VLM 通过与处理其他特征相同的内容注意力路径关注度量位置。

3.3 空间预训练课程

仅靠下游问答难以习得空间推理。其监督是间接的，目标任务分布狭窄，且在几何读数形成之前，场景外观线索主导梯度，因此模型早在学会读取输入之前就已拟合逐模板的答案边缘分布。这一先验一旦被吸收便难以撼动，因为转向几何读数需要经历一次精度暂时下降，而问答梯度不足以奖励这种转变。因此本文希望监督信号中的答案由输入几何因果决定，不存在语言先验或场景记忆的捷径。全景画布正是其自然的输入格式：任意 3D 内容都可放置在世界任意位置，度量坐标由逐图块嵌入携带，几何答案因此可直接从这些放置中读出，走的是模型处理真实场景时的同一路径。

课程物体。 在画布上放置可识别的真实物体可能削弱我们想提供的监督：门的高度往往一致，卫生间尺寸也相差不多，见过足够多场景的模型常能仅凭类别身份回答几何问题。默认方案是在画布上一块框大小的区域内填入从真实场景采样的单个特征图块，并在提示中于类别名出现的位置插入同一图块来指代该放置。这一内联副本与任何其他词元一样，保留其在提示文本中所处槽位的 MRoPE 位置，而非画布上对应物所携带的经纬度画布位置；它同样会被加上 3D 位置嵌入。去掉物体身份这一捷径后，任务主要依赖放置位置及其指代对象的已知几何。该图块取自预计算的激活池，池中激活来自 ScanNet Dai et al. [2017]、ScanNet++ Yeshwanth et al. [2023] 与 ARKitScenes Baruch et al. [2021] 的训练源场景，因此视觉编码器会将其视为普通视觉内容。基于 EmbodiedScan Wang et al. [2024b] 物体特征的另一种类别标注方案，以及与 VSI-Bench 的正面比较，见节 A。

任务类型。 为学习基于几何的空间理解，本文构造一组紧凑的任务，在六个任务族上即时生成：

- **度量测量。** 两个放置物体之间的表面到表面距离；相对距离比较（若干候选中哪个离参考更近，对应 VSI-Bench 的相对距离任务）；以及作为房间尺度信号的占地面积读出。
- **自我中心方向。** 从选定视点出发的“左 / 右 / 前 / 后”查询，分三个难度等级，另有 θ 点钟方向变体与对应下游 SI 类查询的相机视角变体。
- **多轮导航。** 两个放置目标之间的 N 轮路径查询， $N \in \{1, 2, 3, 4\}$ 。
- **可观测性。** 沿帧序列某放置首次可见的时刻（出现顺序），以及从给定位姿能否看到某目标。
- **计数。** 在整除条件（奇偶、能否被三整除）或位置条件（位于智能体左、右、前、后）下对放置物体的计数。
- **多目标包围盒读出。** 放置两到四个物体，提示要求按规范顺序给出每个物体的 3D 包围盒。

3.4 两阶段训练

第一阶段：空间预训练。 仅使用空间预训练课程训练，在刻意抑制真实场景问答中场景先验捷径的监督下，教会模型下游 3D 问答所依赖的几何推理能力。每个课程样本携带的词元远少于完整场景全景图，因此第一阶段运行开销很小，可以承担较长的训练计划。

第二阶段：目标适配。 本文将第一阶段的 LoRA 权重合并回基础语言模型，并在其上训练一个全新的低秩 LoRA 适配器 ($r=64$)，训练数据为下游空间问答的加权混合 (SQA3D、VSI-Bench 风格数据以及 ViCA 系列语料)。合并把第一阶段的更新固化进基础权重，新适配器的较低秩限制了第二阶段偏离第一阶段几何读数的程度，同时仍允许答案格式的适配。

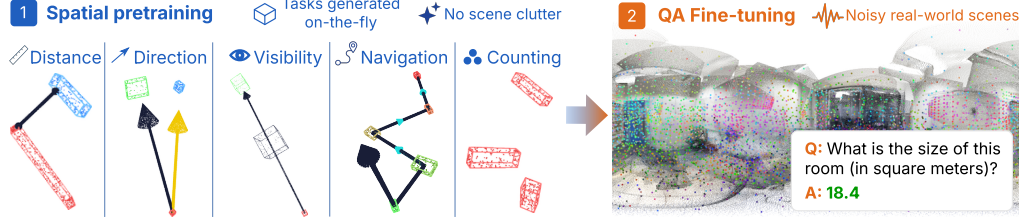


图 4: 两阶段训练。第一阶段在空间预训练课程上训练 LoRA Hu et al. [2022] 适配器与 3D 位置嵌入, 此时放置在空白画布上的物体及其内联指代携带几何信号。第二阶段把第一阶段的适配器合并回基础语言模型, 并在真实场景下游问答上训练一个全新的更小适配器。两个阶段中词元嵌入始终保持冻结。

4 实验

4.1 实现细节

全景投影。 沿视频以均匀间隔的时间戳为每个场景采样 $K=32$ 帧。场景中心的取法: SQA3D 取标注的智能体位置, SPBench 单图查询取被查询的相机位姿, 其余情况取所有相机位置的质心。全部数据源上的训练均采用 320×240 输入分辨率。评测时 SQA3D 使用 320×240 (提高分辨率并未带来整体收益), VSI-Bench 与 SPBench 使用 640×480 , 因为显式的数值测量类查询能从更高的输入分辨率中获益。对 ScanNet 场景, 从每张图像各边裁去 3% (并相应调整内参), 以避免在场景外产生虚假的提升点。数据集提供真值深度与真值相机位姿时即直接使用, 仅在少数已发布真值位姿轨迹损坏的场景上, 退回到前馈重建模型 (feed-forward reconstruction model) 的估计结果。仅在训练期间, 对每个样本随机化全景图的位置与朝向: 从可用相机位置中抽取场景中心, 并把前向轴旋转一个均匀随机的偏航角 (yaw)。

训练。 本文基于 Qwen3-VL-8B Bai et al. [2025] 构建, 该视觉语言模型 (Vision-Language Model, VLM) 采用基于 ViT 的视觉编码器, 以及包含时间、高度、宽度三个维度的 3D RoPE 位置编码方案。训练在 8 张 A6000 GPU 上分两阶段进行, LoRA 适配器作用于语言模型的注意力与 MLP 投影。第 1 阶段在空间预训练课程 (spatial pretraining curriculum) 上训练, LoRA 秩 $r=256$ ($\alpha=512$, dropout 0.05), 有效批大小 16, 学习率 2×10^{-5} , 采用余弦调度与 3% 预热。3D 位置嵌入以基础学习率的 $50 \times$ 联合训练。在第 1 阶段, 它在每次前向传播时被重新缩放到视觉特征平均范数的一半, 因此只学方向而不学幅度。第 2 阶段把其门控 (gate) 初始化为同一尺度, 之后自由学习该尺度。第 1 阶段把每个样本放置物体的图块数上限设为 200, 在放置框之间均匀分配, 且每框不少于 8 个图块以保留形状; 这既让第 1 阶段保持快速, 又不破坏几何信号。样本中还带有 6-12 个未被引用的干扰框 (distractor box), 因此某一放置必须靠其行内图块来识别, 而不能仅凭它是画布上唯一的物体 (节 D)。第 2 阶段的 LoRA 使用 $r=64$, $\alpha=128$, dropout 0.05, 训练 10k 步, 有效批大小 32, 学习率、调度与预热均与第 1 阶段相同。第 2 阶段训练数据是 VLM-3R-VSIBench Fan et al. [2026]、SQA3D Ma et al. [2023] 与 ViCA Feng [2025] 的加权混合, 采样权重约为 70/15/15。

4.2 评测

本文在三个基准上评测: SQA3D Ma et al. [2023] (情境化问答)、VSI-Bench Yang et al. [2025a] (八项空间子任务, 分为数值题与多选题两类), 以及 SPBench Li et al. [2026] (单图与多视角空间推理, 零样本评测)。

由 RGB 预测的几何。 不重新训练, 仅把真值几何替换为由 Depth Anything 3 Lin et al. [2026] 从评测 RGB 帧预测的深度、位姿与内参, SQA3D 得 64.9、VSI-Bench 得 70.0、SPBench 得 71.3, 而使用真值几何时分别为 65.3、71.3 与 72.1。完整的仅推理协议与局限性见小节 B.1。

空间问答。 表 1 报告 SQA3D 上的结果, 将本文方法与基于点云的 3D VLM 以及视频输入方法进行比较。本文取得最佳总体准确率。在最直接考察情境化推理的问题类型 (Which, Can, Others) 上取得最佳结果, 其中 Which 的领先幅度最大。Which 上的大幅领先与画布原点消融 (表 5) 相互印证: 以智能体位姿为中心贡献了其中大部分, 而 Which 正是最

表 1: 在 SQA3D Ma et al. [2023] 测试划分上的评测。方法间跨问题类型的对比。EM@1 与 EM@R1 为主要指标。分问题类型列报告严格 EM@1。最优结果加粗，次优结果加下划线。短横线表示该方法未报告该分数。

方法	分问题类型						EM@1	EM@R1
	What	Is	How	Can	Which	Others		
PQ3D Zhu et al. [2024]	37.1	61.3	44.5	60.9	47.0	45.1	47.1	49.3
3D-VisTA Zhu et al. [2023]	34.8	63.3	45.4	69.8	47.2	48.1	48.5	50.9
LEO Huang et al. [2024b]	39.0	63.9	44.9	66.2	47.7	51.1	50.0	52.4
SIG3D Man et al. [2024]	35.6	67.2	48.5	71.4	49.1	45.8	52.6	54.4
Scene-LLM Fu et al. [2025]	40.9	69.1	45.0	70.8	47.2	52.3	54.2	56.2
ChatScene Huang et al. [2024a]	45.4	67.0	52.0	69.5	49.9	55.0	54.6	57.5
Video-3D LLM Zheng et al. [2025b]	51.1	72.4	55.5	69.8	51.3	56.0	58.6	60.8
GPT4Scene-HDM Qi et al. [2026]	55.9	69.9	50.8	68.7	53.3	60.4	59.4	62.4
VLM-3R Fan et al. [2026]	—	—	—	—	—	—	60.7	63.4
SpaceMind Zhao et al. [2025]	54.1	74.8	61.7	71.0	51.9	53.6	61.1	63.8
SR-3D Cheng et al. [2025]	55.0	<u>76.4</u>	59.8	<u>71.6</u>	54.7	<u>61.1</u>	62.2	—
Loc3R-VLM Qu et al. [2026]	—	—	—	—	—	—	62.8	65.0
Ross3D Wang et al. [2025a]	<u>56.0</u>	79.8	<u>60.6</u>	70.4	<u>55.3</u>	60.1	<u>63.0</u>	<u>65.7</u>
OneCanvas (本文方法)	56.6	74.7	60.2	75.4	72.6	65.7	65.3	68.4

表 2: 在 SPBench Li et al. [2026] 上的评测。所有模型均未使用 SPBench 训练数据。各部分最优结果加粗，次优结果加下划线。

方法	总体	SPBench-MV			SPBench-SI		
		NQ	MCQ	平均	NQ	MCQ	平均
Video-R1 Feng et al. [2025]	43.8	32.5	53.0	42.8	27.7	62.0	44.9
SpaceR-7B Ouyang et al. [2025]	53.5	63.2	53.7	58.5	35.7	61.5	48.6
VILASR-7B Wu et al. [2025b]	54.0	56.2	59.6	57.9	36.6	63.7	50.2
Spatial-MLLM-4B Wu et al. [2025a]	52.5	63.7	58.9	61.3	38.1	49.3	43.7
SpaceMind Zhao et al. [2025]	<u>67.3</u>	<u>76.2</u>	<u>70.5</u>	<u>73.8</u>	66.3	53.2	<u>59.7</u>
OneCanvas (本文方法)	72.1	79.0	91.8	81.5	<u>62.8</u>	<u>62.7</u>	62.8

依赖视角的问题类型。剩余差距集中在 *Is* (是非题，文本先验承载了大部分信号) 与 *How* (以计数为主，而基线 2D VLM 已在该能力上充分预训练)。

表 2 给出 SPBench 上的零样本结果，所有方法均未使用 SPBench 训练数据。本文位居排行榜首位，最大增益集中在多视角多选题上，这与如下直觉一致：单张全景画布为多视角问题提供了统一参考系，而不必迫使 VLM 拼接若干个相互独立的视图。在单图划分的平均准确率上同样位居首位。

表 3 给出 VSI-Bench 上的结果，本文再次位居首位，在路径规划上领先幅度最大，该子任务最直接地考察基于统一全景画布的多步推理。在绝对距离、相对距离与房间尺寸上同样领先，在相对方向上位列第二。

4.3 消融研究

组件消融。 表 4 在 VSI-Bench 上对本文方法的关键组件进行消融。为与两阶段完整模型的 LoRA 总容量对齐 (第 1 阶段 $r=256$ 合并进基础模型，第 2 阶段 $r=64$)，单阶段消融变体均以 $r=256$ ($\alpha=512$) 训练。去掉空间预训练阶段后，路径规划任务上出现大幅下降，印证了该课程对这一能力的价值。在需要估计度量数值的任务上同样表现更好，唯一例外是物体尺寸，它更依赖文本先验。去掉位置编码会在房间尺寸任务上带来进一步的大幅下降，而其余度量类问题相对稳定，因为模型仍可依据场景外观推断尺寸与距离。仅从完整模型中移除 3D 位置嵌入，可在课程保持完整的前提下分离出同一效应：下降集中在度量类子任务上，角度类与类别类子任务不受影响，这与 3D 位置嵌入在全景画布之上充当度量尺度辅助通道的作用相符。与直接送入图像相比，统一全景表示的贡献也得以显现：需要跨场景不同区域关联物体的任务在本文方法下表现明显更好，相对方向上的提升最大。例外是物体计

表 3: 在 VSI-Bench Yang et al. [2025a] 上的评测。在数值题与多选题空间推理子任务上的对比。各部分最优结果加粗，次优结果加下划线。

方法	平均	数值题				多选题			
		物体计数	绝对距离	物体尺寸	房间尺寸	相对距离	相对方向	路径规划	出现顺序
Spacer Ouyang et al. [2025]	45.5	57.8	28.2	59.9	47.1	40.1	45.4	33.5	52.1
ViLaSR Wu et al. [2025b]	45.4	63.5	34.4	60.6	30.9	48.9	45.2	30.4	49.2
Spatial-MLLM Wu et al. [2025a]	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
ViCA Feng [2025]	60.6	68.8	57.0	79.2	<u>75.1</u>	58.5	42.6	34.5	68.8
VLM-3R Fan et al. [2026]	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
VST Yang et al. [2025b]	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
Cambrian-S Yang et al. [2026]	67.5	<u>73.2</u>	50.5	74.9	72.2	<u>71.1</u>	76.2	41.8	80.1
SenseNova-SI Cai et al. [2026]	68.8	72.0	53.5	76.8	72.8	69.6	80.8	48.5	<u>76.4</u>
SpaceMind Zhao et al. [2025]	69.6	73.3	<u>61.4</u>	77.3	74.2	67.2	88.4	44.3	70.6
OneCanvas (本文方法)	71.3	68.9	62.3	75.1	76.8	73.1	<u>88.0</u>	60.8	65.7

表 4: 消融研究 (VSI-Bench)。每行列出该变体所含的组件。完整模型使用全景重投影、3D 位置编码与第 1 阶段空间预训练，而最后一行是不含这些组件、采用多视角输入的基线 VLM。所有变体均在匹配的训练设置下进行 LoRA 微调。每列最优结果以**粗体**标出，次优结果加下划线。

配置	平均	数值题				多选题			
		物体计数	绝对距离	物体尺寸	房间尺寸	相对距离	相对方向	路径规划	出现顺序
完整模型 (本文方法)	71.3	68.9	62.3	75.1	76.8	<u>73.1</u>	88.0	<u>60.8</u>	65.7
完整模型 (去掉 3D PE)	70.4	69.3	58.1	76.1	71.3	70.3	88.6	66.0	63.8
全景 + 3D PE	68.5	68.3	56.6	76.9	<u>74.1</u>	73.2	91.6	45.4	61.3
仅全景	65.3	68.8	54.8	<u>76.8</u>	56.9	71.6	87.4	45.3	60.4
基线 VLM (多视角)	65.5	72.5	53.2	75.8	68.6	68.9	77.4	43.9	63.4

数，基线 VLM 仍保持小幅领先。本文将其归因于重排本身：它使输入偏离了骨干网络预训练时所见的图像布局。计数更多取决于有多少词元呈现被查询物体、这些词元有多集中，而非空间理解本身 Zhang et al. [2018]，因此更好的空间表示无法弥补这一代价（小节 B.2）。

画布原点。 用同一检查点在 SQA3D 上评测四种画布原点策略（表 5）。使用智能体位姿在多数任务上带来明显提升，但 *Is* 与 *How* 两类除外，它们通常与视角无关。随机相机原点从构建画布所用的原始相机位置中采样，因此仍位于已观测空间之内。只有场景外这一条件把原点放在房间之外。其他相机放置策略总体上对结果影响不大，唯有场景外原点使各任务结果一致下降。

5 结论

OneCanvas 把 3D 场景理解重构为全景画布上的 2D 空间推理。多视角 RGB-D 被重投影到单张等距圆柱 (Equirectangular) 全景图上，每个图块通过可加的三维位置嵌入 (3D position embedding) 携带其度量 3D 位置，预训练 VLM 则以其原生注意力机制读取该结果。空间预训练课程把物体放置在原本空白的画布上，并从提示中引用它们，从而在下游微调之前预热几何读取能力。二者的结合在 VSI-Bench 与 SQA3D 上刷新了最优结果，并登顶零样本 (zero-shot) SPBench 排行榜。该表示同样可迁移到其他骨干网络 (backbone) (小节 B.4)。

局限性。 OneCanvas 需要深度与相机位姿，而纯 RGB 方法则无需这两者；位姿估计的失败仍可能使画布质量下降，不过小节 B.1 中仅推理的 Depth Anything 3 评测缩小了这一差距。单一全局全景图以牺牲精细空间精度换取紧凑性，可能限制其在大规模或室外场景中的表现。预训练课程为人工设计，因此要新增一类性质不同的空间技能，就得编写新的任务生成器，而非收集更多数据。

致谢与资助声明

本项目由 ERC Consolidator Grant Gen3D (101171131) 资助。感谢 Angela Dai 为视频配音。

表 5: **SQA3D 上的画布原点消融**。同一检查点在四种画布原点策略下评测。智能体位姿跟随智能体的位置与朝向。其余三种把前向固定为世界 +X，原点分别取场景质心、已观测空间内的一个原始相机位置，以及场景之外。按问题类型划分的各列报告严格 EM@1。每列最优结果以**粗体**标出，次优结果加下划线。

画布原点	按问题类型						EM@1	EM@R1
	What	Is	How	Can	Which	Others		
智能体位姿	56.6	74.7	60.2	75.4	72.6	65.7	65.3	68.4
场景中心	54.6	74.2	59.8	<u>68.9</u>	<u>56.7</u>	<u>58.9</u>	61.2	64.3
随机相机	54.8	<u>74.4</u>	61.3	<u>68.9</u>	55.6	<u>58.9</u>	<u>61.4</u>	<u>64.6</u>
场景外	54.0	73.6	60.0	68.3	51.3	55.7	59.8	62.7

参考文献

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.
- Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatial-thinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*, 2025.
- Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- Jieneng Chen, Wenxin Ma, Ruisheng Yuan, Yunzhi Zhang, Jiajun Wu, and Alan Yuille. Thinking with spatial code for physical-world video reasoning. *arXiv preprint arXiv:2603.05591*, 2026.
- Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024.
- An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, et al. 3d aware region prompted vision language model. *arXiv preprint arXiv:2509.13317*, 2025.
- Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017.
- Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31054–31065, 2026.

- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- Qi Feng. Visuospatial cognitive assistant. *arXiv preprint arXiv:2505.12312*, 2025.
- Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.
- Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024a.
- Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024b.
- Yuan Jia, Ruimin Wu, Rui Song, Jiaojiao Li, and Bin Song. Liftproj: Space lifting and projection-based panorama stitching. *arXiv preprint arXiv:2512.24276*, 2025.
- Seongmin Jung, Seongho Choi, Gunwoo Jeon, Minsu Cho, and Jongwoo Lim. Panogrounder: Bridging 2d and 3d with panoramic scene representations for vlm-based 3d visual grounding. *arXiv preprint arXiv:2512.20907*, 2025.
- Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.
- Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.
- Xiang Li, Jian Ding, Zhaoyang Chen, and Mohamed Elhoseiny. Uni3dl: A unified model for 3d vision-language understanding. In *European Conference on Computer Vision*, pages 74–92. Springer, 2024.
- Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. Hispatial: Taming hierarchical 3d spatial understanding in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.
- Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.
- Zekai Lin and Xu Zheng. Panoenv: Exploring 3d spatial intelligence in panoramic environments with reinforcement learning. *arXiv preprint arXiv:2602.21992*, 2026.
- Xianzheng Ma, Brandon Smart, Yash Bhalgat, Shuai Chen, Xinghui Li, Jian Ding, Jindong Gu, Dave Zhenyu Chen, Songyou Peng, Jia-Wang Bian, et al. When llms step into the 3d world: A survey and meta-analysis of 3d tasks via multi-modal large language models. *arXiv preprint arXiv:2405.10255*, 2024a.

- Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhargat, Jindong Gu, Angel X Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://arxiv.org/abs/2603.23523>.
- Xiaojuan Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023.
- Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10978–10997, 2024b.
- Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2024.
- Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.
- Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pages 194–210. Springer, 2020.
- Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.
- Kevin Qu, Haozhe Qi, Mihai Dusmanu, Mahdi Rad, Rui Wang, and Marc Pollefeys. Loc3r-vm: Language-based localization and 3d reasoning with vision-language models, 2026. URL <https://arxiv.org/abs/2603.18002>.
- Mukund Varma T, Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, and Ravi Ramamoorthi. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21367–21377, 2024.
- Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9275–9286, 2025a.
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025b.
- Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024a.
- Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.
- Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.
- Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

- Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025a.
- Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.
- Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations (ICLR)*, 2018.
- Ruosun Zhao, Zhikang Zhang, Jiale Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv preprint arXiv:2511.23075*, 2025.
- Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.
- Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8995–9006, 2025b.
- Xu Zheng, Chenfei Liao, Ziqiao Weng, Kaiyu Lei, Zihao Dongfang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Lu Qi, Li Chen, et al. Panorama: The rise of omnidirectional vision in the embodied ai era. *arXiv preprint arXiv:2509.12989*, 2025c.
- Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering llms with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.
- Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024.

A 物体指代：类标签替代设计

正文在画布上放置合成轴对齐框，其逐图块特征取自从训练源场景中采集的激活值共享随机池，并在提示中通过同一图块特征的内联副本引用每一次放置。这样做的动机是把物体大小与外观同类别身份解耦，使画布上的几何成为解题的主要信号（参见正文方法中关于课程物体的讨论）。一个自然的替代方案，是用从场景实例标注 Wang et al. [2024b] 中就地采集的真实物体填充画布，连同其原始逐图块特征粘贴到画布上，并在提示中按类名引用。特征

在原始场景内提取，而非从孤立裁剪出的物体重新编码——后者会丢弃每个图块编码时所处的周边场景上下文。该变体顺带引入了文本到物体的接地能力，代价则是重新引入合成框方案本意要抑制的类别先验（门的典型高度、卫生间的典型尺寸）。表 6 在完整 VSI-Bench 测试划分上比较两种课程内容选择，骨干网络、画布构建、MRoPE 调度、任务族、族内扁平加权、阶段 2 微调与评测协议全部固定。两行之间唯一变化的，是阶段 1 中填充画布的内容。

表 6: VSI-Bench 上的课程内容研究（完整测试划分）。两行共享相同的骨干网络、画布构建、任务族、阶段 2 微调与评测协议，只有阶段 1 的课程内容不同。合成框：采用共享随机逐图块特征的轴对齐框，在提示中通过内联图块引用。场景采集的物体：从就地场景裁剪中粘贴、并带原始逐图块特征的真实物体，按类名引用。

课程内容	平均	数值问题				多项选择题			
		物体计数	绝对距离	物体尺寸	房间尺寸	相对距离	相对方向	路径	出现顺序
合成框（正文）	71.3	68.9	62.3	75.1	76.8	73.1	88.0	60.8	65.7
场景采集的物体	70.2	69.8	61.1	76.5	79.6	67.8	86.1	59.8	60.8

B 附加实验与分析

本节汇集若干附加评测。这些协议彼此不可互换：预测几何、栅格化与接缝研究只改变推理；特征与像素全景图的对比以及 ERNIE 对比是匹配的 2,000 步训练；随机种子复现使用完整的 10,000 步调度；Qwen3.5 结果使用完整的两阶段配方。下文在每个结果处说明其边界。

B.1 由 RGB 预测的几何

本文把全部真值深度、位姿与内参替换为 Depth Anything 3 (DA3NESTED-GIANT-LARGE-1.1) Lin et al. [2026] 直接从 RGB 估计的结果。重建恰好使用评测时提供给 VLM 的那些帧：SQA3D 与 VSI-Bench 为 $K=32$ ，SPBench 则使用每道题所附视角的数量。不使用额外视角，也不使用真值几何。这是使用正文检查点、仅改变推理的替换，未围绕几何估计器重新训练。

表 7: 前馈预测几何。在正文完整评测协议下的准确率。此前最优一行分别给出各基准上最强的已有结果：SQA3D 上为 Ross3D，VSI-Bench 与 SPBench 上为 SpaceMind。

几何来源	SQA3D	VSI-Bench	SPBench
此前最优	63.0	69.6	67.3
本文方法，真值几何	65.3	71.3	72.1
本文方法，由 RGB 预测的几何	64.9	70.0	71.3

预测几何使准确率分别下降 0.4、1.3 与 0.8 个点，但在这三个基准上仍高于表中所列已有方法的结果。这只能说明对该特定前馈估计器具有鲁棒性，既不代表对重建失败不敏感，也不代表联合训练的几何编码器的性能。由于该估计器是作为可替换的预处理模块调用，而非与 VLM 联合训练，该结果也无需把正文检查点适配到 DA3。

B.2 连续特征、栅格化与渲染像素

仅推理的栅格化。 本文把提升后的图块分入固定的等距圆柱网格，对每个单元内所有图块的特征、深度与帧索引做均值池化，并把单元中心角度赋给池化后的词元。181×362、90×180 与 45×90 三种网格的角分辨率单元近似为方形，分别约为 1°、2° 与 4°。所有行都使用正文评测协议，以及在连续画布上训练出的同一检查点，因此栅格化得分属于仅推理消融。若为网格重新训练模型，表现可能不同，本文未做评测。

损失主要出现在最细的网格上。在 181×362 下，出现顺序从 62.6 降到 39.8，占 2.9 个点平均差距中的 2.8 个点。把均值池化换成每个被占据单元保留一个原始图块，网格与词元数不变，得分 68.6 而非 68.2，可见平均本身并不能解释该差距。两种归约方式都只保留每个单元的一个词元与一个帧索引：同一表面的重复观测，其帧索引要么被平均，要么只留其一，

表 8: **VSI-Bench 上仅推理的栅格化**。词元数为提供给模型的被占据单元数量。

画布	词元数	VSI-Bench
连续 (本文方法)	9,600	71.1
栅格化 181×362	7,144	68.2
栅格化 90×180	4,009	67.7
栅格化 45×90	1,563	65.8

首次出现因此不再可恢复。尽管如此，对于一维位置编码无法赋予任意角度坐标的 VLM，栅格网格仍可能把画布扩展出去。

匹配的特征与像素对比训练。 本文还用相同的基础检查点、LoRA 秩 256、有效批大小 8、调度、数据混合与留出的 VSI-Bench 问题训练了两组 2,000 步。二者仅场景输入不同：投影后的逐帧特征，与预渲染的等距圆柱 z 缓冲 (Z-Buffer) 全景图。

表 9: **特征画布与渲染像素的对比**。VSI-Bench 上匹配的 2,000 步训练。这是短程对比，并非正文的完整训练调度。

场景输入	VSI-Bench
特征画布 (本文方法)	60.8
像素全景图	52.1

8.7 个点的差距大于两组运行中观察到的约 1 个点的逐步波动。它同时反映两方面差异，不应被解读为单独隔离出其中任一项：z 缓冲渲染把被遮挡与位置重合的表面在每个方向上压成一个表面，而渲染那一组要求一个在透视图像上预训练的视觉塔重新编码等距圆柱图像。本文画布不做可见性测试，把这些观测保留为独立词元，并从原始透视帧中提取其外观特征；它还允许在不重新渲染的情况下更换画布原点。

计数与深度粗化。 在表 4 中，基础 VLM 在物体计数上仍保有微弱领先。深度粗化对照并不支持“重投影误差把实例打碎或合并”这一解释：粗化至 8× 时，图块位移中位数为 0.16 m，29% 的图块落到不同表面，而四个级别的计数得分相近 (68.4、68.8、69.3 与 68.9)。本文转而归因于排列本身——它使输入偏离骨干网络预训练时的图像布局；而计数主要取决于有多少词元显示了被查询物体、以及这些词元有多集中 Zhang et al. [2018]。

B.3 训练随机种子复现

本文在其他设置全部固定的情况下，用三个不同的随机种子额外训练了三组，全部保留，并使用完整的 10,000 步调度，而非答复期的短程运行。与全文一致，检查点按留出集的生成准确率选取。

表 10: **完整 10,000 步训练下的随机种子方差**。标准差基于所报告的三个种子计算。

	VSI-Bench	SQA3D (EM@1)	SPBench
种子 A	71.0	66.2	73.4
种子 B	69.9	64.5	72.8
种子 C	70.3	64.7	73.4
均值 ± 标准差	70.4 ± 0.6	65.1 ± 0.9	73.2 ± 0.3

三组运行都超过了 VSI-Bench 上 69.6、SQA3D 上 63.0 与 SPBench 上 67.3 的此前最优结果。这三个种子刻画的是同一固定配方下的训练种子波动，不涵盖数据、超参数或模型族选择带来的差异。

B.4 其他骨干网络

ERNIE-4.5-VL-28B-A3B。 本文把该表示迁移到 Qwen 家族之外。ERNIE 在真实权重上端到端跑通画布预填充、训练与生成，适配器仅限语言模型，视觉塔不变，架构未做改

表 11: VSI-Bench 上匹配的 2,000 步 ERNIE-4.5-VL-28B-A3B 对比。这是不含空间课程、仅阶段 2 的短程运行，因此有意义的量是受控的行间差异，而非绝对准确率或与完整训练的正文模型的比较。

	物体计数	绝对距离	物体尺寸	房间尺寸	相对距离	相对方向	路径	出现顺序	整体
仅骨干网络	50.6	26.9	70.5	42.2	52.3	44.0	21.3	13.9	40.2
+ 全景画布	65.2	27.9	71.2	55.8	51.3	46.8	28.0	24.6	46.3
Δ	+14.6	+1.0	+0.7	+13.6	-1.0	+2.8	+6.7	+10.7	+6.1

动。下面两组仅在是否使用画布上不同，使用相同的数据、调度、适配器形状与有效批大小训练 2,000 次迭代。

ERNIE 是通用模型而非空间专家，且这些运行省略了阶段 1。因此 +6.1 的整体差异只支持该表示在这一匹配的短程协议下具有可移植性，并不确立 ERNIE 在完整配方下的性能。

Qwen3.5-9B。使用完整两阶段配方与正文完整评测协议，Qwen3.5-9B 在 VSI-Bench 上达到 69.42，SQA3D 上 67.18 EM@1，SPBench 上 69.81。这是一次全规模运行，采用更新一代的 Qwen 与不同的视觉栈，并非所要求的非 Qwen 对照。它与 ERNIE 对比分开报告，也不替代正文检查点及其 71.3、65.3 与 72.1 的主要结果。

B.5 畸变与经度接缝

图块特征来自原始透视帧，因此等距圆柱映射只改变图块放置的位置，不改变其特征。经度是周期的，但为其分配的位置 ID 不是：位于 $+179^\circ$ 与 -179° 的图块在场景中彼此相近，却被赋予经度索引范围的两端。可加的三维位置嵌入仍保持连续，因为它编码的是每个图块相对画布原点的笛卡尔偏移；训练时均匀随机的偏航角会使接缝在各经度间移动，而不是系统性地割裂某一场景区域。尽管如此，局部的角度索引不连续依然存在。

本文用一个仅推理的成对距离探针把它隔离出来。两个物体与答案保持不变，只改变画布朝向，使接缝要么位于两物体之间，要么远离它们。按构造，深度与角度间隔 Δ 共同变化，因此每一行只在行内比较。

表 12: 经度接缝探针。平均绝对距离误差，单位为米。

Δ	接缝不在两者之间	接缝在两者之间
15°	0.14	1.23
30°	0.15	0.50
60°	0.24	0.30
170°	0.69	0.71

误差上升集中在接缝把角度间隔很小的两个物体分开的情形，并随间隔增大而变得可忽略。把两个物体紧邻接缝放置但不分开，在 15° 时误差为 0.23 m，说明该效应源于环绕本身，而非靠近边界。在随机画布朝向下，这一构型影响 0.63% 的 VSI-Bench 绝对距离问题（本次评测 214 题中的 1 题）。推理时可自由设定画布偏航角，把接缝转到远离感兴趣区域的位置。若某些场景中跨越接缝的查询很常见，则可能需要针对性的课程样例或感知环绕的位置编码。

C 训练算力预算

可复现的两阶段 OneCanvas 配方使用 8 块 NVIDIA RTX A6000 GPU（Ampere 架构，每块 48 GB GDDR6），配合 DeepSpeed ZeRO-2 与 Flash Attention 2。阶段 1（空间预训练）实际耗时 ≈ 25.2 h，阶段 2（问答微调）耗时 ≈ 39.6 h，8 块 GPU 上合计 ≈ 64.8 h（518 原始 A6000 GPU 小时）。表 13 通过解析的 A100 等效归一化，把该开销与竞争方法报告的训练成本作比较。这些并非实测的跨硬件运行时间对比。

GPU 归一化。本文把原始 GPU 小时乘以源 GPU 与 A100 SXM4 之间稠密 BF16 Tensor Core 吞吐之比，换算为 A100 等效 GPU 小时。之所以采用稠密 BF16（无结构化 2:4 稀疏），是因为它与训练实际运行的方式一致：BF16 数值精度、不使用稀疏内核。各因子列于

表 13: **训练算力对比**。OneCanvas 与竞争方法的训练实际耗时与总算力, 利用 表 14 中各 GPU 型号的峰值吞吐因子解析归一化为 A100 等效 GPU 小时。OneCanvas 的数值合并了两个训练阶段。

方法	GPU 型号	GPU 数量	实际耗时	原始 GPU 小时	A100 等效 GPU 小时
OneCanvas (本文方法)	A6000	8	≈ 64.8 h	518	259
VLM-3R Fan et al. [2026]	H200	16	5 h	80	240
ViCA Feng [2025]	H100	8	55 h	440	1,320
SpaceMind Zhao et al. [2025]	H100	64	25 h	1,600	4,800
SenseNova-SI Cai et al. [2026]	H100 [†]	128	72 h	9,216	27,648

[†] 论文未说明 GPU 型号, 此处假定为 H100。

表 14, 即 $\text{TFLOPS}_{\text{src}}/\text{TFLOPS}_{\text{A100}}$ 四舍五入到一位有效数字。该解析换算只刻画峰值算力, 应视为归一化上界, 而非实测速度比。实际上, H100/H200 上可获得的加速受内存带宽 (HBM 比值 H100 为 $1.67\times$ 、H200 为 $2.4\times$) 以及 表 13 中 64 与 128 块 GPU 任务的多节点扩展效率所限, 低于 $3\times$ 的比值。在更保守的带宽受限换算下, SpaceMind 与 SenseNova-SI 仍分别使用约 $10\times$ 与 $59\times$ 于本文的归一化算力。

表 14: **表 13 中使用的 GPU 归一化因子**。稠密 BF16 Tensor Core TFLOPS (无 2:4 稀疏) 为 NVIDIA 数据表数值, HBM 带宽仅作参考, 不参与换算。A100 等效因子按 $\text{TFLOPS}_{\text{src}}/\text{TFLOPS}_{\text{A100}} = \text{TFLOPS}_{\text{src}}/312$ 计算, 四舍五入到一位有效数字。

GPU	稠密 BF16	HBM 带宽	A100 等效因子
A6000	154.8 TFLOPS	768 GB/s	$154.8/312 \approx 0.5\times$
A100 SXM4	312 TFLOPS	2,000 GB/s	$1.0\times$
H100 SXM5	989 TFLOPS	3,350 GB/s	$989/312 \approx 3.0\times$
H200 SXM5	989 TFLOPS	4,800 GB/s	$989/312 \approx 3.0\times$

H200 与 H100 共用 GH100 计算核心, 仅在内存容量 (141 GB HBM3e 与 80 GB HBM3 之别) 与带宽上不同, 因而具有相同的稠密 BF16 吞吐与换算因子。989/312 的比值由 3.17 下舍为 3.0, 略微低估了竞争方法的算力, 从而使对比不利于本文方法。

各方法的计算。 表 15 把 表 14 中的因子应用到各方法报告的训练实际耗时上。

表 15: **各方法的 A100 等效算力**。A100 等效 GPU 小时等于 $(\# \text{ GPUs}) \times (\text{wall-clock hours}) \times (\text{A100-equiv. factor})$ 。GPU 数量与实际耗时取自各方法论文, 因子来自 表 14。

方法	GPU 数量	实际耗时 (小时)	因子	A100 等效 GPU 小时	相对本文
OneCanvas (本文方法)	8	64.8	$0.5\times$	259	$1\times$
VLM-3R Fan et al. [2026]	16	5	$3\times$	240	$0.9\times$
ViCA Feng [2025]	8	55	$3\times$	1,320	$5.1\times$
SpaceMind Zhao et al. [2025]	64	25	$3\times$	4,800	$19\times$
SenseNova-SI Cai et al. [2026]	128	72	$3\times$	27,648	$107\times$

该算力论断的范围限于与本文准确率相近的方法, 而非所有基线。SpaceMind (VSI-Bench 69.6) 与 SenseNova-SI (68.8) 是距本文 71.3 最近的两个方法, 所用峰值归一化算力分别为本文的 $19\times$ 与 $107\times$ 。VLM-3R 的归一化算力略低于本文 (240 对 259 A100 等效 GPU 小时), ViCA 为本文的 $5.1\times$, 但二者的 VSI-Bench 得分仅为 60.9 与 60.6。算力相当时, VLM-3R 低 10.4 个点。因此证据支持的是准确率相当时训练成本更低, 而非算力上的普遍优势。

D 空间预训练课程：任务细节

本节详细介绍构成阶段 1 空间预训练课程的各项任务。课程把任务实例组织为六个族, 并采用族内扁平加权: 每个族在每个小批次中占相同份额, 族内每个任务权重相同。六个族分别是度量测量、自我中心方向、多轮导航、可观测性、计数与多目标包围盒读出。

所有任务都在只含所放置物体的画布上训练，这些物体的中心、尺寸与偏航角由程序化采样得到。阶段 1 中从不在画布上加载任何场景，模型因此只看到所放置的物体。

每个放置的框被实现为落入其体积内的一组画布图块，这些图块都携带同一个特征向量，向量从预计算的真实 ViT 图块特征池中采样。该池只需构建一次：用 Qwen3-VL 冻结的视觉编码器在 ScanNet Dai et al. [2017]、ARKitScenes Baruch et al. [2021] 与 ScanNet++ Yeshwanth et al. [2023] 的训练源场景上运行得到，训练时框特征从中独立抽取。由于该池与场景无关，框不会泄漏当前场景的身份，多目标任务也无法靠聚类来自同一源帧的词元走捷径。又因为图块是真实 ViT 输出，其逐词元特征分布与 VLM 在真实场景上所见的一致，阶段 1 习得的读取能力因而可迁移到阶段 2——此时画布由真实场景图块填充。

干扰框。 样本中会加入 6–12 个提示从不引用的框。每个框各自抽取特征，与任务指代物及其他干扰框都不相同，因此既无法靠“唯一不同者”也无法靠“一致的类别”走捷径。占地面积任务采样时不加干扰框，因为额外的框会改变所询问的量。

下面依次介绍各族，每个任务给出简短的描述性名称，随后是其精确的输入与输出格式。

D.1 度量测量

成对距离。 在画布上放置两个支撑不重叠、面到面间隔至少 5 cm 的框。提示通过内联图块引用二者，要求模型以米为单位报告两个表面之间的最短距离，保留一位小数。标签即标量回归目标。

四者中的最近者。 在画布上放置五个框（一个目标、四个候选），并约束最优与次优候选的面到面距离差至少 5 cm。提示通过内联图块引用目标与四个候选，模型以答案为 {A, B, C, D} 的四分类选择题选出距离目标最近的候选。

非矩形占地面积。 在场景中放置一个合成的 L 形或 T 形地面多边形，按随机偏航角旋转，并渲染为薄板状的提升点。模型以平方米读出占地面积（保留一位小数的回归）。

D.2 自我中心方向

世界坐标系方向任务（下文的二分类、三分类、四分类对角、四分类基本方位与钟面变体）放置 3 个框：位于智能体位置的“ref”框、定义朝向的“fwd”框，以及待分类的目标框。提示写作“If you are standing at [ref] and facing [fwd], ...”，方括号中的名称渲染为绑定到相应画布图块的内联图块。相机坐标系方向任务放置 2 个框（目标与枢轴），并把画布本身重新定向到从帧序列中抽取的真实场景相机位姿上，因此无需显式的朝向参照。答案空间随任务而异。

二分类方向。 二分类选择题：{left, right}。问题为“目标是位于左侧还是右侧？”

三分类方向。 三分类选择题：{left, right, back}。当目标与朝向的夹角超过 135° 时，“back”单元生效。

四分类对角方向。 四分类对角选择题：{front-left, front-right, back-left, back-right}。采样时强制与基本方位轴保持角度余量，使对角标签定义明确。

四分类基本方位方向。 四分类基本方位选择题 {front, back, left, right}，单元边界为 $\pm 45^\circ$ 。

相机坐标系方向（易 / 中 / 难）。 答案空间与世界坐标系的二分类、三分类和四分类对角变体相同，但画布被重新居中并定向到从帧序列中抽取的真实场景相机位姿上。问题从相机视角提出：“From the camera’s perspective, is X to Y’s ...?”。

钟面方位。 答案为 {1, 2, ..., 12} 中的钟面方位，按相对朝向的顺时针角度计算（12 = 前，3 = 右，6 = 后，9 = 左）。角分辨率高于基本方位或对角变体。

D.3 多轮导航

导航族包含四个转弯数量的 N 轮路径规划任务（ $N \in \{1, 2, 3, 4\}$ ），各在族内权重相同。

N 轮路径规划 ($N \in \{1, 2, 3, 4\}$)。在画布上按采样位置放置起点框、终点框与 N 个中间途经点框。提示通过内联图块引用每一个。答案为把智能体从起点带到终点的自我中心转向动作序列，每轮一个动作。格式为填空题式的选择题：题目把各动作列为 “[please fill in]” 空位，模型从 4 个字母选项中选一个。每一步路径转向取自 $\{\text{Turn Left}, \text{Turn Right}\}$ 。此外，第一轮额外加入 “Turn Back” 选项，因此初始动作取自 $\{\text{Turn Left}, \text{Turn Right}, \text{Turn Back}\}$ 。

D.4 可观测性

出现顺序。 在画布上按采样位置放置四个框。每个框的逐图块 MRoPE T 值被覆盖为受控分布：对随机选取的 k ，四个框被赋予连续的首现 T 起点 $\{k, k+1, k+2, k+3\}$ ，每个框其余图块的 T 值在 $[t_{\text{start}}, T_{\text{max}}]$ 内均匀采样，并将其中一个图块固定为 t_{start} 。提示通过内联图块引用这四个框，要求以四分类选择题给出完整的首次出现顺序（正确排列加三个随机排列干扰项）。该构造在单帧尺度上隔离出细粒度的 T 轴分辨能力，且没有场景内容可供走捷径。

视线可达性。 一个是 / 否任务：给定观察者框 p_1 、目标框 p_2 及二者之间的合成轴对齐遮挡框， p_1 到 p_2 的视线是否可达？可见性按第一性原理计算为： p_2 的轴对齐包围盒 (Axis-Aligned Bounding Box, AABB) 中可由一条不与遮挡框相交的线段从 p_1 到达的比例；该比例至少为 10% 时标签为 “yes”。遮挡框始终存在，因此框的数量不会泄漏标签。遮挡侧、沿线位置与线段长度在两个标签间取自相同分布，只有遮挡框中心垂直于线段方向的偏移与标签相关。该任务不涉时序。特征与位姿无关；之所以列在可观测性下，是因为它检验与出现顺序相同的“模型能否判断某区域是否可达”这一能力。

D.5 计数与算术读出

物体计数。 在画布上放置 N 个框， N 在较宽的整数范围内均匀抽取。这 N 个框与提示通过内联图块引用的单个参照框共享同一次特征抽取：“How many [ref] are in the room?”。模型输出整数计数。共享特征让这 N 个框具有可通过内容注意力读取的共同身份，而未被引用的干扰框各自携带不同的特征抽取，因此不会被误认为指代物。

计数奇偶性。 设置与物体计数相同，但问题为 “Is the number of [ref] in the room odd or even?”，答案为 “even” 或 “odd”。放置数量经采样使两个标签等概率，模型无法靠猜测边缘分布解题。

计数能否被三整除。 设置相同，但问题为 “Is the number of [ref] in the room a multiple of three?”，答案为 “yes” 或 “no”。放置数量经采样使两个标签等概率。该任务表述为二值整除判断而非三分类余数分类，从而与其他是 / 否任务的答案格式保持一致。

按方位计数。 一项混合任务。给定带前向朝向的原点位姿与一个参照框，模型从智能体视角统计指定基本方位一侧有多少个同特征框。方位从 $\{\text{right}, \text{left}, \text{front}, \text{back}\}$ 均匀抽取，目标计数从 $\{3, 4, 5, 6\}$ 按权重 $[30, 35, 25, 10]$ 抽取。答案为整数。该任务把方向分类与计数结合起来。

D.6 多目标包围盒读出

多目标包围盒读出。 在画布上按采样得到的中心、尺寸与偏航角放置二至四个框。提示通过带字母标记的内联图块引用每一个：“Provide the 3D bounding box of each highlighted region, in the same order: A: [m1], B: [m2], ...”。模型按规范的由近及远顺序，为每个框输出轴对齐 3D 包围盒（中心与度量尺寸）。