arXiv:2603.00912v1 [cs.CV] 1 Mar 2026

VGGT-Det：挖掘 VGGT 内部先验，实现无传感器几何信息的多视角室内三维物体检测

Yang Cao[1*] Feize Wu[1, 3*†] Dave ZhenyuChen[2] Yingji Zhong[1] LanqingHong[2] DanXu[1‡]
香港科技大学[2]华为 [3]中山大学

摘要

当前的多视图室内 *3D* 目标检测器依赖成本高昂的传感器几何信息——即经过精确标定的多视角相机位姿——将多视图信息融合为全局场景表示，这限制了其在真实场景中的部署。我们面向一种更实用的设置：无传感器几何信息（***SG-Free***）多视图室内 *3D* 目标检测，即不存在传感器提供的几何输入（多视角位姿或深度）。近期的视觉几何基础变换器（*VGGT*）表明，可以直接从图像中推断出强大的三维线索。基于这一发现，我们提出了 ***VGGT-Det***，这是首个专为无传感器几何信息的多视图室内 *3D* 目标检测设计的框架。我们的方法并非仅仅使用 *VGGT* 的预测结果，而是将 *VGGT* 编码器集成到基于 *Transformer* 的流程中。为了有效利用 *VGGT* 内部的语义先验和几何先验，我们引入了两个新颖的关键组件：（*i*）注意力引导的查询生成（***AG***）：利用 *VGGT* 注意力图作为语义先验来初始化目标查询，通过聚焦物体区域并保留全局空间结构来提升定位效果；（*ii*）查询驱动的特征聚合（***QD***）：一种可学习的See-Query与目标查询交互，以“观察”其所需内容，随后动态聚合 *VGGT* 各层的多层级几何特征，逐步将二维特征提升至三维。实验表明，在无传感器几何信息设置下，*VGGT-Det* 的表现显著优于当前最佳方法，在 *ScanNet* 和 *ARKitScenes* 上的 *mAP@0.25* 分别高出***4.4*** 和***8.6*** 。消融研究表明，我们提出的 *AG* 和 *QD* 能有效利用 *VGGT* 内部学习到的语义先验和几何先验。源代码和预训练模型可在*GitHub* 项目页面获取。

1. 引言

室内多视图三维对象检测是一项基础任务，在机器人学和增强现实领域有着广泛应用。

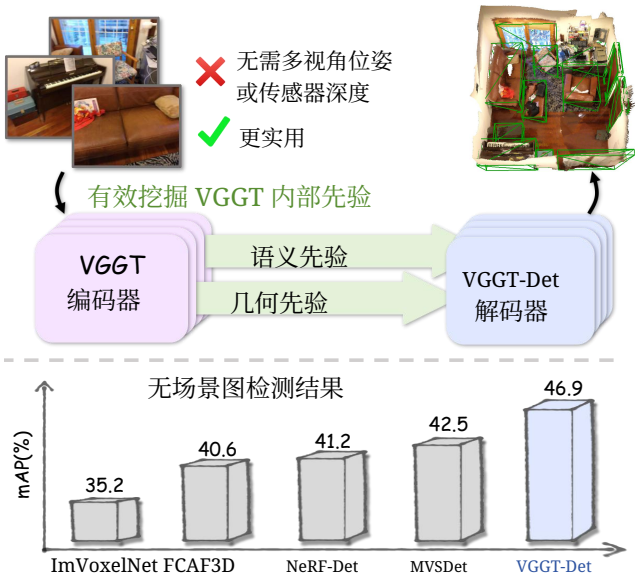


图1. 为实现并提升无传感器几何（SG-Free）多视图室内三维目标检测，VGGT-Det有效利用了VGGT的内部语义与几何先验 [41]，而非仅仅使用其预测结果。VGGT-Det在无传感器几何信息设置下显著优于具有竞争力的方法。

当前方法 [4, 32, 36, 46, 47] 主要依赖传感器导出的几何输入，即经过精确标定的多视图相机位姿或深度。尽管这些方法表现出色，但它们依赖昂贵且往往难以获取的传感器导出的几何输入 [3, 5]，严重限制了其可扩展性和在现实世界中的部署。

相较之下，我们考虑一种更实用的设定：在没有传感器导出的几何输入的情况下，基于多视图图像进行室内三维检测。我们将其称为无传感器几何信息（**SG-Free**）多视图室内三维目标检测。这一设定极具挑战性，因为它排除了传感器导出的多视图相机位姿和深度。近年来，前馈三维重建取得了进展，研究表明可以直接从无位姿的二维图像中推断三维结构 [15, 37, 41, 43, 49, 50]。这些进展

*同等贡献
†本工作是在香港科技大学实习期间完成的
‡通讯作者：danxu@cse.ust.hk

为 SG-Free 多视图室内三维目标检测开辟了新的机遇。前馈模型不仅能够以较低成本从无位姿图像中估计场景几何，还编码了强大的学习到的先验，可提升三维推理能力，而这对于可靠检测至关重要。

本文提出 **VGGT-Det**，这是一种基于视觉几何基础变换器（VGGT）的无传感器几何（SG-Free）多视图室内三维目标检测框架 [41], 。VGGT 是一种具有代表性的前馈三维重建模型。如图 1所示，我们的方法并非仅仅利用 VGGT 的预测结果，而是在基于 Transformer 的流程中集成了预训练的 VGGT 编码器。为了有效利用 VGGT 内部学到的知识，我们深入分析了其中间表征。我们发现，尽管 VGGT 编码器并未针对语义进行显式训练，其注意力图仍能捕捉丰富的语义信息。受这一有趣发现的启发，我们提出了注意力引导的查询生成（**AG**），利用 VGGT 注意力图初始化并引导目标查询，通过聚焦语义区域，同时保留全局空间结构，提升定位效果。此外，鉴于 VGGT 会在各层中逐步将二维特征提升至三维，且每层编码不同层次的几何抽象，我们提出了查询驱动的特征聚合（**QD**）。该模块引入了一个可学习的*See-Query*，它与目标查询交互，以“查看”所需信息，随后动态聚合 VGGT 各层的多层级几何特征，从而有效捕捉层次化表征。这些设计共同释放了 VGGT 在无传感器几何的多视图室内三维检测中的更多潜力。大量实验表明，VGGT-Det 在无传感器几何信息设置下显著优于强基线方法，消融实验也证实，所提出的 AG 和 QD 能有效利用 VGGT 内部学到的先验。我们的贡献总结如下：

• 我们提出了一种更具实用性的设置——无传感器几何信息（SG-Free）多视图室内三维检测，该设置无需传感器导出的几何输入（即多视角相机位姿或深度图）。据我们所知，这是首项工作明确针对这一具有挑战性的设置展开研究。在此设置基础上，我们提出了 VGGT-Det，这是首个基于变换器的无传感器几何信息三维物体检测框架。

• 我们提出注意力引导查询生成（AG），利用 VGGT 编码器注意力中编码的语义先验来初始化目标查询。这些查询聚焦于物体区域，同时保留全局空间结构，从而提升定位效果。

•我们提出查询驱动特征聚合（QD），引入一个可学习的*See-Query*，它与目标查询交互以“了解”其需求，并据此动态聚合多层级几何特征。

• 借助所提出的方法，VGGT-Det取得了显著提升，在无传感器几何信息设置下持续优于表现最佳的方法，在ScanNet和ARKitScenes上的mAP@0.25分别高出**4.4**和**8.6** 点。

2. 相关工作

多视角室外三维目标检测。当前的室外多视角三维目标检测 [7, 10, 42, 45] 大致可分为两类。第一类将三维查询投影到二维图像中，以聚合视觉信息。受 DETR 启发， [6], DETR3D [44] 使用稀疏三维目标查询，通过相机参数索引多视图二维特征。PETR 系列 [24, 25] 引入位置嵌入变换，将三维坐标信息编码到多视图图像特征中。Sparse4D 系列 [19–21]和 SparseBEV [23] 通过探索稀疏策略聚合多帧特征，实现具备时序感知能力的检测。BEVFormer [18] 利用时空变换器引入统一的鸟瞰图表示。第二类通过将二维特征提升到三维空间来构建鸟瞰图表示。LSS [31] 提出一种端到端多视图架构，将每幅图像“提升”为特征视锥，再将其“泼溅”到鸟瞰图网格中，以学习稳健的场景表示。BEVDet [13] 采用标准模块，并通过定制的数据增强和改进的非极大值抑制提升性能。BEVDepth [17] 和 BEVStereo [16] 通过显式深度监督解决深度瓶颈。CorrBEV [48]则更进一步，通过巧妙的设计选择（如辅助视觉原型和语言原型）应对遮挡。我们的工作聚焦于室内检测。与摄像头刚性安装在车辆上的室外场景不同，室内摄像头通常由手持，或经常改变位置。因此，获取可靠的传感器位姿既成本高昂，也往往难以实现，这促使我们为室内场景引入无传感器几何信息设置。

多视图室内三维目标检测。 现有的多视图室内三维目标检测方法根据带有多视角位姿的图像进行3D目标检测 [4, 36, 46, 47]。值得注意的是，多视图立体视觉 [51, 52], 、神经辐射场（NeRF） [29], 和三维高斯泼溅（3DGS） [14] 的最新进展增强了从带位姿多视图图像中进行几何重建的能力，从而能够利用体素 [12, 46] 和泼溅基元 [4] 的三维线索实现有效的三维检测。ImVoxelNet [34] 以端到端方式通过反投影将多视图图像中的二维特征聚合到三维体素体中。NeRF-Det [46] 将来自NeRF模块的多视图几何约束整合到3D目标检测流程中。MVSDet [47] 以平面扫描取代NeRF，并采用概率采样与软加权对深度进行正则化，从而实现精确检测。现有方法依赖于传感器导出的几何输入（例如多视角位姿或深度），这严重限制了其在真实世界中的适用性。为解决这一问题，我们研究无需传感器几何（SG-Free）的多视图室内三维目标检测，并

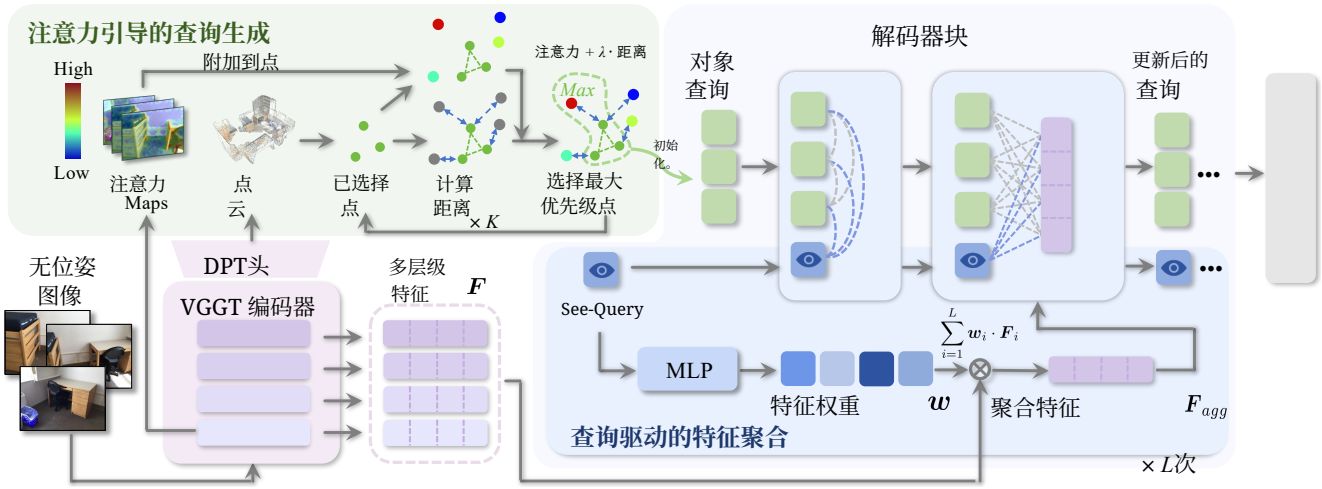


图 2. 所提出的无传感器几何（SG-Free）多视图室内三维目标检测框架概览，*VGGT-Det*。该框架基于 VGGT 编码器构建，[41], 该编码器从多视图图像中提取三维感知特征。解码器处理一组目标查询，这些查询通过交叉注意力关注提取的特征，并迭代更新以完成最终检测。为有效利用 VGGT 内部的语义先验和几何先验，我们精心设计了两个关键组件：注意力引导查询生成（*AG*）和查询驱动特征聚合（*QD*）。AG 利用 VGGT 编码器注意力中的语义先验生成目标查询，使这些查询能够聚焦于物体区域，同时保留全局空间结构。此外，QD 引入了一个可学习的 See-Query，它通过自注意力与目标查询交互，以“了解”它们的需求，并据此动态聚合多层级几何特征。

提出VGGT-Det，该方法利用注意力引导的查询生成和查询驱动的特征聚合，有效挖掘VGGT内部的语义与几何先验，而非仅仅使用其预测结果。

可泛化的三维重建。 给定无位姿的多视图图像，三维重建 [11, 15] 旨在估计图像的几何信息和相机位姿。传统的三维重建流程包含多个子任务，包括关键点检测 [9, 28], 匹配 [2, 22, 35, 38],多视图立体视觉 [40, 51–53], 等。近年来，DUSt3R [43]彻底革新了这一范式，利用单个网络直接估计目标场景的几何信息。MASt3R [15] 继承了这一范式，并为匹配任务配备了一个辅助预测头。DUSt3R和MASt3R都仅支持成对输入，这限制了上下文推理，并导致网络重复进行前向传播以及耗时的全局坐标对齐。Fast3R [50]通过支持对长序列执行单次前向传播并消除坐标对齐的需求，提升了性能。除点图外，VGGT [41]还预测相机位姿及其他三维相关属性。本文提出VGGT-Det，有效利用VGGT内部的先验知识，推动无传感器几何信息的多视角室内三维物体检测的发展。

3. 方法

3.1. 概述

图2展示了我们提出的无传感器几何三维检测框架的概览。VGGT-Det 以多视图图像为输入，输出三维检测结果，从而无需传感器提供几何输入（即多视角位姿和深度）。下文将首先在第 3.2 节介绍我们的基础骨干网络，该网络是一个集成 VGGT 编码器的基于 Transformer 的流程 [41]。为了有效利用 VGGT 内部的语义先验和几何先验，我们引入了两个全新的关键组件：（i）注意力引导查询生成（AG），旨在有效利用 VGGT 编码器注意力中的语义先验，详见第 3.3 节。（ii）查询驱动特征聚合（QD），旨在根据目标查询的需求自适应聚合多层级几何特征，详见第 3.4 节。

3.2. 基础骨干网络

我们设计了一种编码器-解码器变换器 [39] 架构作为骨干网络。给定一组多视图图像$\{I_1, I_2, \ldots, I_V\}$，我们采用 VGGT 编码器 [41] 作为特征提取器，以提取三维感知特征。对于每个视图，VGGT 编码器都会输出一个词元序列，其中每个词元对应图像的一个空间区域或图像块。所有视图提取出的词元表示为 $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_V\}$，其中 $\mathbf{T}_i \in \mathbb{R}^{M \times C}$ 是第 i 个视图的词元序列，每个视图包含 M 个词元，C 为词元维度。沿词元维度拼接这些词元，以形成统一令牌表示：

$$\mathbf{T}_{\text{concat}} = [\mathbf{T}_1; \mathbf{T}_2; \ldots; \mathbf{T}_V] \in \mathbb{R}^{(V \cdot M) \times C}. \quad (1)$$

受 3DETR [30], 启发，我们采用非参数查询采样来初始化目标查询。具体而言，我们对 VGGT 预测的点云 $\mathbf{P}_{\text{pred}}$应用最远点采样（最远点采样），选取 K 个点$\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K\} \subset \mathbf{P}_{\text{pred}}$，其中 K 是目标查询的数量

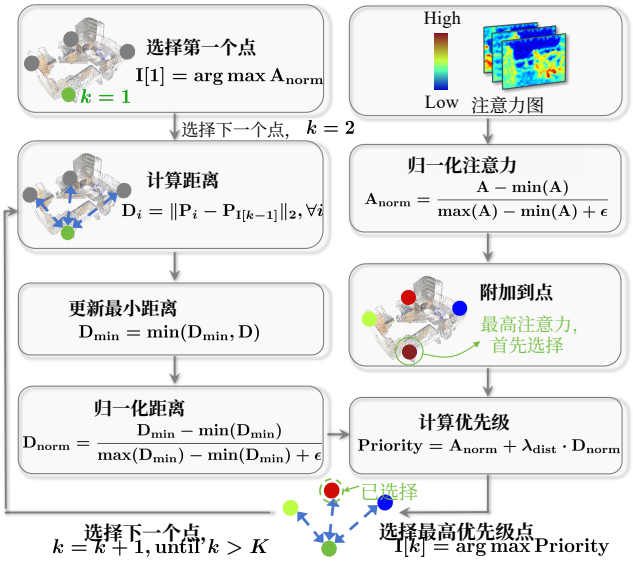


图 3. 注意力引导的查询生成的计算流程。

。这些采样点通过位置嵌入$\mathbf{E}_{\text{pos}} \in \mathbb{R}^{K \times C}$ 进行编码，以初始化目标查询：

$$\mathbf{Q}_0 = \text{Embed}(\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_K\}) + \mathbf{E}_{\text{pos}}. \quad (2)$$

目标查询 $\mathbf{Q}_0$ 随后通过一系列 L Transformer 解码器层。在每个解码器层中，查询首先进行查询间自注意力，以相互交换信息：

$$\mathbf{Q}_l^{\text{self}} = \text{SelfAttention}(\mathbf{Q}_{l-1}), \quad (3)$$

其中，$\mathbf{Q}_{l-1}$ 是第 l 个解码器层的输入，$\mathbf{Q}_l^{\text{self}}$是经过自注意力后的输出。接下来，查询通过交叉注意力关注统一令牌表示 $\mathbf{T}_{\text{concat}}$：

$$\mathbf{Q}_l = \text{CrossAttention}(\mathbf{Q}_l^{\text{self}}, \mathbf{T}_{\text{concat}}). \quad (4)$$

经过所有 L 解码器层后，更新后的目标查询 $\mathbf{Q}_L \in \mathbb{R}^{K \times C}$ 被输入检测头，以生成对象级预测，包括类别标签 $\hat{\mathbf{c}} \in \mathbb{R}^K$ 和边界框 $\hat{\mathbf{b}} \in \mathbb{R}^{K \times 7}$：

$$\{\hat{\mathbf{c}}, \hat{\mathbf{b}}\} = \text{DetectionHead}(\mathbf{Q}_L). \quad (5)$$

通过在不同视图之间拼接词元，模型使目标查询能够直接与多视图令牌表示进行交互。该流程有效利用了 VGGT 编码器提取的三维感知特征，并通过变换器解码器动态优化目标查询，从而实现传感器几何无关三维目标检测。

3.3. 注意力引导的查询生成

由于 VGGT 预测的点云 [41]是场景的稠密重建结果，因此无法区分物体区域和背景区域。因此，仅通过最远点采样（FPS）从预测点云中均匀选取查询，可能会导致大量目标查询落在背景区域，而非实际物体上。这种查询位置偏差会从根本上损害检测效果，因为位于背景区域的查询几乎无法提供有用的检测信息。有趣的是，我们观察到，尽管 VGGT 未针对语义任务进行训练，其编码器中的注意力仍能自然捕获丰富的语义信息，如图4左栏所示。这一观察促使我们提出注意力引导的查询生成（AG），利用 VGGT 编码器的注意力图，引导目标查询定位于具有语义意义的物体区域。所提出的 AG 不仅利用 VGGT 编码器提供的语义引导，还兼顾空间多样性，使采样点能够充分覆盖整个三维空间。

形式化地，VGGT 重建的点云表示为$\mathbf{P} \in \mathbb{R}^{N \times 3}$，其中 N 表示点的数量，每个点均由其三维坐标表示。我们将 VGGT 编码器的注意力权重表示为$\mathbf{A} \in \mathbb{R}^N$，其中每个权重表示一个点的语义重要性。目标是从$\mathbf{P}$中采样 K 查询点$\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K\}$，同时依据注意力权重提供的语义信息和空间分布进行引导。为确保数值稳定性并与基于距离的特征兼容，将注意力权重归一化为：

$$\mathbf{A}_{\text{norm}} = \frac{\mathbf{A} - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A}) + \epsilon}, \quad (6)$$

其中 ϵ 是用于数值稳定性的一个小常数。$\mathbf{A}_{\text{norm}} \in \mathbb{R}^N$。第一个查询点的索引$\mathbf{I}[1]$ 被选为注意力分数最高的点：

$$\mathbf{I}[1] = \arg\max \mathbf{A}_{\text{norm}}. \quad (7)$$

后续点通过融合优先级迭代选择，该优先级平衡语义注意力与空间分散：

$$\text{Priority} = \mathbf{A}_{\text{norm}} + \lambda_{\text{dist}} \cdot \mathbf{D}_{\text{norm}}, \quad (8)$$

其中 $\lambda_{\text{dist}} \in [0, 1]$ 是控制空间分散相对影响的权衡系数，$\mathbf{D}_{\text{norm}}$ 是点$\mathbf{P}$ 与已采样集合 $\{\mathbf{P}^{\mathbf{I}[1]}, \ldots, \mathbf{P}_{\mathbf{I}[k-1]}\}$之间的归一化最小欧几里得距离。具体而言，我们令$\mathbf{D}_{\text{min}}$ 表示$\mathbf{P}$到所有先前已选择点的最小欧几里得距离：

$$\mathbf{D}_{\text{min}} = \min_{j \in \{1, \ldots, k-1\}} \|\mathbf{P} - \mathbf{P}_{\mathbf{I}[j]}\|_2, \quad (9)$$

其中$\mathbf{D}_{\text{min}} \in \mathbb{R}^N$。归一化距离定义为：

$$\mathbf{D}_{\text{norm}} = \frac{\mathbf{D}_{\text{min}} - \min(\mathbf{D}_{\text{min}})}{\max(\mathbf{D}_{\text{min}}) - \min(\mathbf{D}_{\text{min}}) + \epsilon}, \quad (10)$$

其中 $\mathbf{D}_{\text{norm}} \in \mathbb{R}^N$。在每次迭代中，优先级最高的点会被选中：

$$\mathbf{I}[k] = \arg\max_{i \notin \mathcal{S}} \text{Priority}, \quad (11)$$

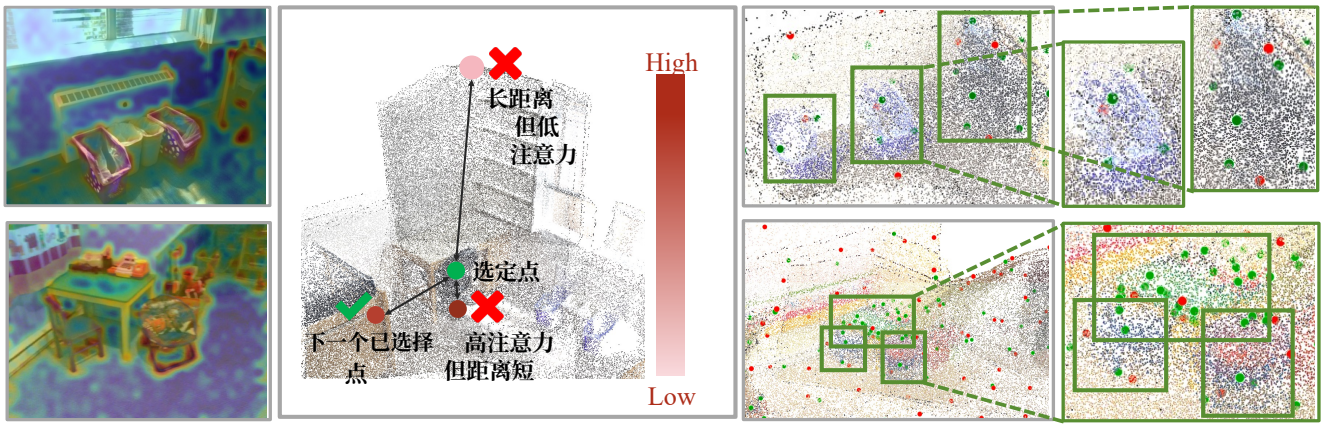


图 4. 我们提出的注意力引导查询生成（AG）受到一项有趣观察的启发：来自 VGGT 编码器的注意力图 [41] 与语义内容高度相关，尽管 VGGT 并未针对语义任务进行显式训练。例如，在左列中，物体区域往往会获得较高的注意力权重。在中间列中，AG 在注意力权重和点分布信息的引导下，从 VGGT 预测的点云中进行采样。在右列中，与无引导的最远点采样（红色点）相比，AG 采样的点（绿色点）更集中于物体区域（绿色方框标出），因此这些区域中的绿色点多于红色点。为清晰起见，建议查看彩色图并放大。

其中 $\mathcal{S} = \{\mathbf{I}[1], \mathbf{I}[2], ..., \mathbf{I}[k-1]\}$ 表示先前迭代中已选择样本的索引集合。在此步骤中，如图 4中间列所示，我们会选择注意力得分较高且距离较远的点。这种设计促使目标查询聚焦于语义上有意义的物体区域，同时保持空间多样性，从而确保覆盖整个三维空间。图 3展示了计算流程，逐步呈现了我们提出的注意力引导查询生成（AG）的实现过程。如图 4右侧列所示，最远点采样往往不加区分地将采样点分布在前景和背景区域，而 AG 则利用 VGGT 的内部语义先验，有策略地将采样查询集中在物体区域，从而促进更有效的三维目标检测。

3.4. 查询驱动的特征聚合

鉴于 VGGT 编码器 [41] 会在其各层中逐步将二维图像特征转换为三维表示，并且每个阶段都会捕获不同层次的几何信息，我们提出了一种查询驱动的特征聚合（QD）策略。该方法引入了一个可学习的*See-Query*，它与目标查询交互，以“了解”目标查询的需求，并据此动态聚合多层级特征。形式化地，设 VGGT 编码器从 L 中间层生成一组特征图，记为 $\{\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_L\}$，其中$\mathbf{F}_i \in \mathbb{R}^{N\times C}$表示第 i 层的特征图，包含 N 个空间元素和 C 个特征维度。为聚合这些多层级几何特征，我们引入了一个可学习的*See-Query* 令牌 $\mathbf{q}_{\text{see}} \in \mathbb{R}^C$，该令牌在所有解码器层之间共享。首先，我们介绍 See-Query 如何聚合多层级特征。See-Query 令牌首先经过一个多层感知机（MLP）随后进行 softmax 归一化，生成作用于$\mathbf{w} \in \mathbb{R}^L$ 各个 L 特征图的注意力权重：

$$\mathbf{w} = \text{Softmax}\left(\text{MLP}(\mathbf{q}_{\text{see}})\right),\ \mathbf{w}_i \geq 0,\ \sum_{i=1}^{L} \mathbf{w}_i = 1. \quad (12)$$

聚合特征$\mathbf{F}_{\text{agg}}$ 随后计算为多级几何编码特征的加权和：

$$\mathbf{F}_{\text{agg}} = \sum_{i=1}^{L} \mathbf{w}_i \cdot \mathbf{F}_i, \quad (13)$$

其中，$\mathbf{F}_{\text{agg}} \in \mathbb{R}^{N\times C}$ 作为键和值输入解码器的交叉注意力块。接下来，我们介绍 See-Query 如何与目标查询交互并更新。See-Query令牌$\mathbf{q}_{\text{see}}$ 在解码器中发挥双重作用。首先，它与目标查询 $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K\}$拼接，形成用于联合解码的统一查询集：

$$\mathbf{Q}_{\text{input}} = [\mathbf{q}_{\text{see}}, \mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K] \in \mathbb{R}^{(K+1)\times C}. \quad (14)$$

这个统一查询集通过自注意力操作在各查询之间交换信息：

$$\mathbf{Q}_{\text{self}} = \text{SelfAttention}(\mathbf{Q}_{\text{input}}), \quad (15)$$

其中，$\mathbf{Q}_{\text{self}} \in \mathbb{R}^{(K+1)\times C}$。通过这一步，See-Query 与目标查询交互，以“了解”它们需要什么，并据此利用$\mathbf{F}_{\text{agg}}$中的相关全局多尺度上下文信息对其进行增强。具体而言，See-Query 和目标查询参与交叉注意力操作，关注聚合几何编码器特征$\mathbf{F}_{\text{agg}}$：

$$\mathbf{Q}_{\text{cross}} = \text{CrossAttention}(\mathbf{Q}_{\text{self}}, \mathbf{F}_{\text{agg}}), \quad (16)$$

其中$\mathbf{Q}_{\text{cross}} \in \mathbb{R}^{(K+1)\times C}$。这会基于以下内容，为 See-Query 和目标查询生成更新后的表示：

方法	cab	bed	椅子	sofa	tabl	door	wind	书架	pic	cntr
ImVoxelNet [34]	20.2	76.3	45.8	60.5	36.0	16.3	09.6	17.9	00.3	23.3
FCAF3D [33]	28.1	81.0	47.2	70.7	40.9	17.1	13.2	22.0	00.7	45.6
NeRF-Det [46]	24.9	85.3	54.4	64.3	42.2	24.3	12.7	31.0	01.0	39.3
MVSDet [47]	28.0	80.6	55.5	65.3	43.2	25.2	09.9	29.4	01.1	43.4
VGGT-Det（基础骨干网络）	20.1	85.5	48.4	78.5	37.0	13.2	09.4	22.1	00.8	40.4
VGGT-Det（基础骨干网络+AG）	25.9	84.7	53.7	73.8	46.3	17.6	13.7	26.4	02.2	40.5
VGGT-Det（基础骨干网络+AG+QD）	27.8	82.0	61.9	74.0	50.3	18.3	17.6	36.9	02.6	40.3

方法	desk	curt	冰箱	淋浴器	toil	sink	bath	烤箱	mAP@0.25
ImVoxelNet [34]	53.1	08.1	29.4	35.0	83.2	43.1	61.3	14.4	35.2
FCAF3D [33]	65.4	08.3	32.3	39.1	83.6	46.9	72.7	15.6	40.6
NeRF-Det [46]	62.6	17.3	41.4	32.8	88.4	39.7	58.7	21.0	41.2
MVSDet [47]	63.8	20.4	39.2	36.7	89.7	44.2	68.4	20.2	42.5
VGGT-Det（基础骨干网络）	68.3	19.3	35.5	41.4	89.5	39.7	76.8	20.0	41.4
VGGT-Det（基础骨干网络+AG）	66.5	18.7	41.5	45.2	84.6	45.3	84.0	24.3	44.2
VGGT-Det（基础骨干网络+AG+QD）	66.9	31.4	49.0	32.9	90.3	50.8	83.2	27.6	**46.9** (+4.4)

表 1. ScanNet 上不同方法的 mAP@0.25 对比。为实现无传感器几何设置（SG-Free）并确保公平比较，使用 VGGT 预测的点云重新训练 FCAF3D。使用 VGGT 预测的多视角相机位姿重新训练 ImVoxelNet、NeRF-Det 和 MVSDet。我们的方法显著优于具有竞争力的方法。

分层编码器特征。经过交叉注意力操作后，更新后的 See-Query 令牌 $\mathbf{q}_{\text{see}}^{(l)}$ 和目标查询将传递到下一解码器层，并从公式（12）开始重复这一过程。See-Query 的迭代细化能够在各解码器层中对多层级几何特征聚合提供动态的上下文感知引导，从而促进细粒度三维检测。

4. 实验

4.1. 实验设置

数据集与设置。 我们在广泛使用的 ScanNet [8]和 ARKitScenes [1]上评估了我们的方法。这些数据集涵盖多种室内环境，如住宅、办公室和教室。ScanNet 包含 18 个标准物体类别，ARKitScenes 则包含 17 个类别。在评估中，我们报告 mAP@0.25，其中平均精度均值（mAP）用于全面评估模型在所有类别上的检测性能。

实现细节。 训练期间冻结预训练的 VGGT 编码器，以保留其原有能力。在实验中，我们使用 256 个目标查询。模型使用 AdamW 优化器 [26] 进行优化，初始学习率为 2.5×10^{-4}，权重衰减为 1×10^{-4}。为确保训练稳定，采用梯度裁剪，最大范数设为 35，范数类型设为 2。学习率调度采用余弦退火策略 [27]，使学习率逐渐衰减至 1×10^{-6}，从而促进平稳收敛，并缓解学习率骤变引发的不稳定。训练损失设置遵循 3DETR [30]。在基础骨干网络中，我们采用一种朴素的编码特征聚合方法，依次查询第 4、11、17 和 23 层（依据 VGGT 架构）的编码特征在解码器中。这些特征按照从浅到深的顺序排列，并逐步进行查询，以逐渐细化对象表示。我们在 8× H800 GPU 上训练模型，完成训练约需两天。

4.2. 方法分析

本节通过消融研究，深入分析所提出设计的贡献。我们逐步将各项设计融入基础骨干网络，以评估其各自的有效性。

注意力引导的查询生成的效果。如表 2a所示，我们的注意力引导的查询生成（第二行）显著提升了基础骨干网络（第一行）的性能，取得了 **+2.8** 个百分点的提升。这表明，与 FPS 采样生成相比，我们的注意力引导的查询生成更具优势。性能提升说明，该方法有效利用了 VGGT 的内部语义先验，引导目标查询聚焦于物体区域，同时保留场景的全局空间结构。

查询驱动的特征聚合效果。对比表2a中的第一行和第二行可以看出，这两行采用了朴素的编码特征聚合策略，来自第4、11、17和23层的编码特征（遵循 VGGT 的默认设置 [41]）会依次在解码器中进行查询。这种直接的策略以简单的方式聚合多层级几何特征。相比之下，我们提出的查询驱动的特征聚合通过与目标查询交互，自适应地聚合多层级几何特征。该方法取得了显著提升，使 mAP@0.25 提高了**2.7**点，展示了 See-Query 机制的有效性。通过在不同解码器层中与目标查询交互，See-Query 学会了识别各阶段最相关的编码特征层级

方法	mAP
BB	41.4
BB+AG	44.2
BB+AG+QD	**46.9**

数量	mAP
40	45.3
60	46.2
80	**46.9**

方法	mAP
Vanilla-v1	37.5
Vanilla-v2	44.2
QD（我们的方法）	**46.9**

方法	时间（秒）	内存（GB）	mAP
VGGT	0.68	11.71	-
使用 MVSDet [47]	0.21	13.81	39.7
与我们的方法	0.23	3.57	**45.3**

（d）表2。消融研究：（a）所提出设计的有效性。“BB”表示我们的基础骨干网络；（b）不同输入帧数量下的性能；（c）不同特征聚合方法的比较；（d）时间和内存分析。VGGT 的计算结果位于第一行，而每种方法额外的时间和内存消耗则显示在后续各行中。

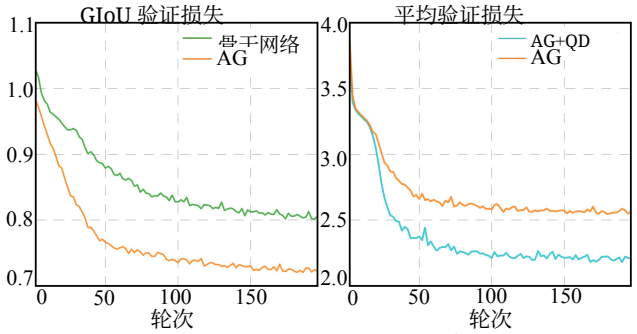


图5。验证损失的可视化。在左侧子图中，应用 AG 后，GIoU损失显著低于基线骨干网络，表明 AG 能够在训练过程中有效改善目标定位。在右侧子图中，随着 See-Query 逐步学习在几个轮次内与目标查询进行交互，并有效聚合编码几何特征，“AG+QD”的损失显著低于“AG”，突出了所提出QD策略的有效性。

，从而实现精确聚合并获得显著的性能提升。

通过损失动态验证动机。为深入验证我们技术贡献背后的动机，我们分析了训练期间的验证损失动态。具体而言，我们评估了注意力引导的查询生成（“AG”）及其与查询驱动的特征聚合（“AG+QD”）结合的效果。如图 5 左侧子图所示，“AG”的 GIoU损失始终低于基线骨干网络。这支持了我们的动机，即注意力图能够有效帮助目标查询聚焦于语义区域，从而提高目标定位的准确性。在右侧子图中，我们进一步考察了查询驱动的特征聚合（“AG+QD”）如何补充“AG”。在最初几个轮次中，“AG”和“AG+QD”的损失相近，因为 See-Query 模块仍在学习与目标查询进行交互。然而，几个轮次后，“AG+QD”的损失显著低于“AG”。这表明 See-Query 模块学会了有效聚合编码几何特征。这些验证损失动态为我们的动机提供了有力证据，并确认了“AG”和“AG+QD”在改善模型训练方面的有效性。

时间和内存分析。我们在表 2d 中进行了时间和内存分析。为确保公平，所有实验均在 1×H800 GPU 上使用 40 个输入帧，并在 Scan Net 上进行评估。该表详细列出了处理一个三维场景所需的时间。对于 MVSDet [47] 等本身并非无传感器几何（SG-Free）的方法，我们通过使用 VGGT 预测的结果（例如多视角位姿）作为训练和评估的输入，将其适配到无传感器几何信息设置。因此，对于 MVSDet，与共享 VGGT 组件相关的时间和内存开销也包含在分析中。VGGT 的计算时间列于表格第一行，而各方法额外的时间和内存消耗则列于后续行中。我们的方法“w/ Ours”与“w/ MVSDet”具有相当的时间消耗，同时显著降低了内存使用量。例如，“w/ Ours”仅消耗 3.57 GB 内存，而“w/ MVSDet”则消耗 13.81 GB。此外，在使用 40 个输入帧时，我们的方法实现了显著的 5.6 点性能提升。

性能随帧数变化。我们在 ScanNet 上进行了消融研究，在评估过程中向模型输入不同数量的帧，如表2b所示。结果表明，更多帧能够提供更丰富的信息，从而使模型取得更高的性能。性能在约 80 帧时趋于饱和，这表明 80 帧足以捕获表示一个 3D场景所需的大部分信息。

不同特征聚合方法下的性能。我们通过与其他特征聚合策略进行比较，分析查询驱动的特征聚合的性能。结果汇总于表 2c。我们考虑了三种特征聚合方法。(i) 在“Vanilla-v1”方法中，仅将 VGGT 编码器最后一层的编码特征 [41] 输入解码器。(ii) 在“Vanilla-v2”方法中，将 VGGT 编码器四个特定层（即第 4、11、17 和 23 层；遵循 VGGT [41]）的编码特征依次输入解码器。(iii) 我们的查询驱动的特征聚合（表 2c中的“QD”）使用 See-Query 在各解码器层与目标查询自适应交互，从而确定并聚合最相关的编码几何特征。如表 2c所示，“Vanilla-v2”的性能优于“Vanilla-v1”，这表明引入多层级特征能够提供更丰富的信息。此外，我们的查询驱动的特征聚合比“Vanilla-v2”高出 **2.7** 个点。



图6。定性　　比较。与 MVSDet 相比 [47], 我们的 VGGT-Det 检测到更多对象，且精度更高　　.

方法	cab	bed	椅子	sofa	tabl	sink	洗衣机	马桶	bath	ovn	refg	架子	炉灶	tvm	stl	fire	dshw	mAP@0.25
ImVoxelNet	06.3	42.4	04.4	26.7	03.4	05.9	23.5	20.9	47.1	09.1	15.0	04.4	00.0	00.0	00.1	0.00	00.8	12.4
NeRF-Det	13.3	51.7	08.8	28.1	08.2	04.5	35.7	42.4	71.7	14.8	19.4	06.6	00.1	00.0	01.3	0.00	00.2	18.1
MVSDet	10.9	53.0	09.0	34.3	07.7	07.4	40.1	43.8	69.8	16.3	25.5	05.7	00.5	00.0	01.2	0.00	03.7	19.4
VGGT-Det	14.6	58.8	15.2	48.2	26.3	09.6	56.4	55.1	64.9	30.9	57.7	09.1	00.0	00.0	10.9	12.5	05.9	**28.0** (+8.6)

表3. ARKitScenes上mAP@0.25的比较。为实现无传感器几何设置（SG-Free）并确保公平比较，ImVoxelNet [34], NeRF-Det [46] 和 MVSDet [47] 均使用VGGT预测的多视角相机位姿重新训练。对ARKitScenes既缺乏结果又缺乏代码支持的方法（例如FCAF3D）未予纳入。我们的方法显著优于这些具有竞争力的方法。

这一改进凸显了我们设计的有效性。

通过与目标查询交互，See-Query 学会自适应地聚合最相关的几何特征，从而显著提升性能。

4.3. 与替代方法的比较

定量比较。表1展示了不同方法在 ScanNet 数据集上的 mAP@0.25 定量比较 [8]。值得注意的是，我们提出的 VGGT-Det 相较于基线方法和具有竞争力的多视图室内三维物体检测方法，取得了显著的性能提升 [32, 46, 47]。具体而言，通过逐步引入注意力引导查询生成（AG）和查询驱动特征聚合（QD），我们的方法整体平均精度均值呈现持续上升趋势，最终达到**46.9**，比 **+4高出 0.4**，超过了 MVSDet [47]。这凸显了我们方法的有效性。请注意，为了将具有竞争力的方法适配到相同的无传感器几何信息设置下进行公平比较，我们使用 VGGT 预测的多视角位姿重新训练 ImVoxelNet、NeRF-Det 和 MVSDet 模型。为提供更全面的分析，我们还与一种具有代表性的基于点云的方法 FCAF3D 进行了比较 [33]。为将其适配到相同的无传感器几何信息设置下进行公平比较，我们使用 VGGT 预测的点云 [41] 重新训练 FCAF3D 模型。VGGT-Det 的表现显著优于 FCAF3D，高出 6.3 个点。总之，VGGT-Det 的优越性能验证了所提出的设计能够有效利用 VGGT 内部学习到的先验。

我们还将其与具有竞争力的多视图室内三维物体检测器进行了比较 [46, 47]在 ARKitScenes 上 [1], , 如表3所示。VGGT-Det 超越了最先进的 MVS-Det [47]，高出 **8.6**个点，进一步证明了我们方法的优越性。我们还发现，少数类别中的物体（例如电视）通常体积较小、较薄，或嵌在橱柜或墙体中。若缺少准确的传感器几何信息作为先验，这些实例极易产生定位误差，导致所有现有模型的性能普遍较低。请注意，为了将具有竞争力的方法适配到相同的无传感器几何信息设置下进行公平比较，我们使用 VGGT 预测的多视角位姿重新训练 ImVoxelNet、NeRF-Det 和 MVSDet 模型。

定性比较。定性比较结果如图6所示。这些结果凸显了我们的VGGT-Det的明显优势：它始终比最先进水平的 MVSDet [47] 检测到更多对象，并取得更高的检测精度，这得益于我们的AG和QD模块对内部VGGT先验的有效利用。

5. 结论

在本文中，我们提出一种更实用的设置：无传感器几何信息的三维目标检测，该设置不再依赖传感器导出的几何输入。为应对这一设置，我们利用 VGGT 编码器构建基于 Transformer的骨干网络。随后，我们提出两项关键创新：注意力引导的查询生成和查询驱动的特征聚合。前者利用 VGGT 注意力机制中的内部

语义先验，生成能够有效聚焦于物体区域、同时保留全局空间结构的目标查询。后者通过与目标查询动态交互，自适应地聚合多层级几何特征。大量实验表明，我们的 VGGT-Det 在 ScanNet 和 ARKitScenes 上的 SG-Free 设置中均显著优于强基线方法。消融实验验证了我们的 AG 和 QD 能够有效利用预训练 VGGT 学习到的先验。

参考文献

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz和Elad Shulman。ARKitScenes：一个使用移动RGB-D数据进行三维室内场景理解的多样化真实世界数据集。NeurIPS，2021年。6, 8[2] Eric Brachmann和Carsten Rother。神经引导RANSAC：学习在何处采样模型假设。ICCV，2019年。3[3] Yang Cao, Yihan Zeng, Hang Xu和Dan Xu。CODA：面向开放词汇三维目标检测的协作式新颖框发现与跨模态对齐。NeurIPS，2023年。1[4] Yang Cao, Yuanliang Jv和Dan Xu。3DGS-Det：以边界引导和框聚焦采样增强3D 高斯泼溅，用于3D目标检测。arXiv预印本 arXiv:2410.01647，2024年。1, 2[5] Yang Cao, Yihan Zeng, Hang Xu和Dan Xu。面向开放词汇三维目标检测的协作式新颖对象发现与框引导跨模态对齐。IEEE 模式分析与机器智能汇刊，2025年。1[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov和Sergey Zagoruyko。基于Transformer的端到端目标检测。ECCV，2020年。2[7] Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus和Adrien Gaidon。面向多视图三维目标检测的视点等变性。CVPR，2023年。2[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Tho mas Funkhouser和Matthias Nießner。ScanNet：室内场景的丰富标注三维重建。CVPR，2017年。6, 8, 2[9] Daniel DeTone, Tomasz Malisiewicz和Andrew Rabinovich。SuperPoint：自监督兴趣点检测与描述。CVPRW，2018年。3[10] Chengjian Feng, Zequn Jie, Yujie Zhong, Xiangxiang Chu和Lin Ma。AEDet：方位角不变的多视图三维目标检测。CVPR，2023年。2[11] Richard Hartley和Andrew Zisserman。计算机视觉中的多视图几何。剑桥大学出版社，2003年。3[12] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai和Chi-Keung Tang。NeRF-RPN：NeRF中目标检测的通用框架。CVPR，2023年。2[13] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye和Dalong Du。BEVDet：鸟瞰图中的高性能多相机3D目标检测。arXiv预印本 arXiv:2112.11790，2021年。2[14] Bernhard Kerbl、Georgios Kopanas、Thomas Leimkühler 和 George Drettakis。用于实时辐射场渲染的3D 高斯泼溅。《ACM 图形学汇刊》，42 14，2023 4 1 2 ( ): – 。[15] Vincent Leroy、Yohann Cabon 和 Jérôme Revaud。借助 MASt3R 将图像匹配建立在三维基础上。载于 ECCV，2024。1，3[16] Yinhao Li、Han Bao、Zheng Ge、Jinrong Yang、Jianjian Sun 和 Zeming Li。BEVStereo：利用时序立体视觉增强多视图3D目标检测中的深度估计。载于人工智能促进协会年会，2023。2[17] Yinhao Li、Zheng Ge、Guanyi Yu、Jinrong Yang、Zengran Wang、Yukang Shi、Jianjian Sun 和 Zeming Li。BEVDepth：为多视图3D目标检测获取可靠深度。载于人工智能促进协会年会，2023。2[18] Zhiqi Li、Wenhai Wang、Hongyang Li、Enze Xie、Chonghao Sima、Tong Lu、Qiao Yu 和 Jifeng Dai。BEVFormer：通过时空变换器从激光雷达-相机数据中学习鸟瞰视角表示。《IEEE 模式分析与机器智能汇刊》，2024。2[19] Xuewu Lin、Tianwei Lin、Zixiang Pei、Lichao Huang 和 Zhizhong Su。Sparse4D：通过稀疏时空融合进行多视图三维目标检测。arXiv 预印本 arXiv:2211.10581，2022。2[20] Xuewu Lin、Tianwei Lin、Zixiang Pei、Lichao Huang 和 Zhizhong Su。Sparse4D v2：基于稀疏模型的循环时序融合。arXiv 预印本 arXiv:2305.14018，2023。[21] Xuewu Lin、Zixiang Pei、Tianwei Lin、Lichao Huang 和 Zhizhong Su。Sparse4D v3：推进端到端三维检测与跟踪。arXiv 预印本 arXiv:2311.11722，2023。2[22] Philipp Lindenberger、Paul-Edouard Sarlin 和 Marc Pollefeys。LightGlue：极速局部特征匹配。载于 ICCV，2023。3[23] Haisong Liu、Yao Teng、Tao Lu、Haiguang Wang 和 Limin Wang。SparseBEV：基于多相机视频的高性能稀疏三维目标检测。载于 ICCV，2023。2[24] Yingfei Liu 、 Tiancai Wang 、Xiangyu Zhang 和 Jian Sun。PETR：用于多视图三维目标检测的位置嵌入变换。载于 ECCV，2022。2[25] Yingfei Liu、Junjie Yan、Fan Jia、Shuailin Li、Aqi Gao、Tiancai Wang 和 Xiangyu Zhang。PETRv2：基于多相机图像进行三维感知的统一框架。载于 ICCV，2023。2[26] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv: *1711 05101*, 2017 6 。[27]

[28] Ilya Loshchilov 和 Frank Hutter。SGDR：带热重启的随机梯度下降。载于国际学习表征会议，2017。6[29]David G Lowe。尺度不变特征点的独特图像特征。《国际计算机视觉杂志》，60：91–110，2004。3[30] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于 ECCV，2020。2[31] Ishan Misra、Rohit Girdhar 和 Armand Joulin。用于三维目标检测的端到端变换器模型。载于 ICCV，2021。3，6Jonah Philion 和 Sanja Fidler。Lift, Splat, Shoot：通过隐式反投影到三维，对来自任意相机系统的图像进行编码。载于 ECCV，2020。2

[32] Charles R Qi、Xinlei Chen、Or Litany 和 Leonidas J Guibas。ImVoteNet：利用图像投票增强点云中的3D目标检测。载于 CVPR，2020。1，8，2[33] Danila Rukhovich、Anna Vorontsova 和 Anton Konushin。FCAF3D：全卷积无锚框三维目标检测。载于 ECCV，2022。6，8，1[34] Danila Rukhovich、Anna Vorontsova 和 Anton Konushin。ImVoxelNet：将图像投影到体素，以进行单目和多视图通用3D目标检测。载于 WACV，2022。2，6，8[35] Paul-Edouard Sarlin、Daniel DeTone、Tomasz Malisiewicz 和 Andrew Rabinovich。SuperGlue：利用图神经网络学习特征匹配。载于 CVPR，2020。3[36] Guanlin Shen、Jingwei Huang、Zhihua Hu 和 Bin Wang。CN-RMA：结合光线步进聚合的网络，用于从多视图图像进行室内3D目标检测。载于 CVPR，2024。1，2[37] Brandon Smart、Chuanxia Zheng、Iro Laina 和 Victor Adrian Prisacariu。Splatt3r：从未标定图像对进行零样本高斯泼溅。arXiv 预印本 arXiv:2408.13912，2024。1[38] Jiaming Sun、Zehong Shen、Yuang Wang、Hujun Bao 和 Xiaowei Zhou。LoFTR：基于变换器的无检测器局部特征匹配。载于 CVPR，2021。3[39] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。载于 NeurIPS，2017。3[40] Fangjinhua Wang、Silvano Galliani、Christoph Vogel、Pablo Speciale 和 Marc Pollefeys。PatchmatchNet：学习式多视图 PatchMatch 立体视觉。载于 CVPR，2021。3[41] Jianyuan Wang、Minghao Chen、Nikita Karaev、Andrea Vedaldi、Christian Rupprecht 和 David Novotny。VGGT：视觉几何基础变换器。载于 CVPR，2025。1，2，3，4，5，6，7，8[42] Shihao Wang、Yingfei Liu、Tiancai Wang、Ying Li 和 Xiangyu Zhang。探索面向高效多视图3D目标检测的以目标为中心的时序建模。载于 ICCV，2023。2[43] Shuzhe Wang、Vincent Leroy、Yohann Cabon、Boris Chidlovskii 和 Jerome Revaud。DUSt3R：让几何3D视觉变得简单。载于 CVPR，2024。1，3[44] Yue Wang、Vitor Campagnolo Guizilini、Tianyuan Zhang、Yilun Wang、Hang Zhao 和 Justin Solomon。DETR3D：通过3D到2D查询从多视图图像进行3D目标检测。载于 CoRL，第180–191页。PMLR，2022。2[45] Kaixin Xiong、Shi Gong、Xiaoqing Ye、Xiao Tan、Ji Wan、Errui Ding、Jingdong Wang 和 Xiang Bai。CAPE：用于多视图3D目标检测的相机视角位置嵌入。载于 CVPR，2023。2[46] Chenfeng Xu、Bichen Wu、Ji Hou、Sam Tsai、Ruilong Li、Jialiang Wang、Wei Zhan、Zijian He、Peter Vajda、Kurt Keutzer 等。NeRF-Det：学习几何感知体积表示，以进行多视图3D目标检测。载于 ICCV，2023。1，2，6，8[47] Yating Xu、Chen Li 和 Gim Hee Lee。MVSDet：通过高效平面扫描进行多视图室内三维目标检测。载于 NeurIPS，2024。1，2，6，7，8，4

[48] 薛梓腾、郭明哲、范恒、张世辉和张志鹏。Corrbev：通过多模态原型的相关性学习实现多视图三维目标检测。发表于 *CVPR*，2025年。2[49] 闫驰、屈德林、Dan Xu、赵斌、王志刚、王栋和李学龙。Gs-slam：利用 3D 高斯泼溅实现稠密视觉同步定位与建图。发表于 *CVPR*，2024年。1[50] 杨嘉宁、Alexander Sax、Kevin J Liang、Mikael Henaff、唐浩、曹昂、Joyce Chai、Franziska Meier 和 Matt Feiszli。Fast3R：朝着单次前向传播实现 1000+ 图像三维重建的方向。*arXiv* 预印本 *arXiv:2501.13928*，2025 年。1，3[51] 姚尧、罗子鑫、李世伟、方天和全龙。Mvsnet：用于非结构化多视图立体视觉的深度推断。发表于 *ECCV*，2018年。2，3[52] 姚尧、罗子鑫、李世伟、沈天伟、方天和全龙。用于高分辨率多视图立体深度推断的循环MVSNet。发表于 *CVPR*，2019年。2[53] 张景阳、姚尧、李世伟、罗子鑫和方天。可见性感知多视图立体网络。 *arXiv* 预印本 *arXiv:2008.07928*，2020年。3

VGGT-Det：挖掘 VGGT 内部先验，实现无传感器几何信息的多视角室内三维目标检测

补充材料

6. 更多消融研究

不同 λ_{dist}下的性能。在本小节中，我们研究改变 λ_{dist} 对我们提出的注意力引导查询生成（AG）性能的影响，其中该变量位于公式（8）中。如表4所示，在我们的评估中，最佳性能约在 $\lambda_{\text{dist}} = 0.8$处取得。直观而言，$\lambda_{\text{dist}}$ 调节了注意力引导与空间分散之间的平衡。当 λ_{dist} 过小时，模型会过度依赖注意力引导，导致过度关注特定的高注意力区域，同时忽略全局空间结构。当 λ_{dist} 过大时，空间分散项占据主导地位，从而减弱注意力引导的影响。总而言之，λ_{dist} 控制着注意力引导与空间分散之间直观的平衡。在消融实验中观察到的约为 $\lambda_{\text{dist}} = 0.8$时的最佳性能，与我们的目标高度一致：初始化查询，使其聚焦于语义区域，同时保留全局空间结构。

方法	mAP@0.25
使用 AG（λ_{dist}=0.9）	44.9
使用 AG（λ_{dist}=0.8）	46.9
使用 AG（λ_{dist}=0.5）	40.7

表 4. 不同条件下的性能 λ_{dist} （公式（8）。

不同噪声水平的影响。在本小节中，我们评估所提方法对 VGGT 预测点云中噪声的鲁棒性。由于预训练 VGGT [41] 输出的质量无法直接控制，我们通过向 VGGT 点云添加高斯噪声来引入受控噪声。具体而言，噪声采样自正态分布 $n \sim \mathcal{N}(0, \sigma^2)$，其中 σ 表示噪声的标准差。为确保噪声强度精确且可扩展，σ 定义为：$\sigma = \text{range} \cdot \text{noise_level}$，其中 $\text{range} = \max(P) - \min(P)$ 表示点云的值范围，而 $\text{noise_level} \in [0, 1]$ 是控制噪声强度的用户自定义参数。含噪点云按如下方式生成：$P' = P + N$，其中 N为噪声矩阵。该方法可实现细粒度噪声控制，便于精确评估噪声影响。我们在 ScanNet 上使用含噪点云评估 VGGT-Det。为进行公平比较，我们还评估了代表性的基于点云的方法 FCAF3D [33], , 该方法同样以含噪点云作为输入。结果汇总于表 5。正如预期，两种方法的性能均随噪声水平升高而下降，证实了噪声对检测的不利影响。然而，FCAF3D 在噪声水平为 0.001 时便开始显著下降，并降至 0.0

噪声水平	FCAF3D	VGGT-Det（我们的方法）
0.30	0.0	34.1
0.20	0.0	39.4
0.10	0.0	44.5
0.01	18.7	47.0
0.005	33.4	47.0
0.001	39.7	46.9
0.000	40.6	46.9

表5. 不同噪声水平下的鲁棒性：FCAF3D [33] 与我们的 VGGT-Det 之间的比较。

在 0.1 时的 mAP@0.25，而 VGGT-Det 仍保持鲁棒性，仅在 0.1 时才开始下降，即使在 0.3 时仍达到 34.1 mAP@0.25。在相同的噪声水平 0.01 下，VGGT-Det 以 **28.3 点**（47.0 对 18.7 ）显著优于 FCAF3D，展现出更强的鲁棒性。

这种鲁棒性源于我们的注意力引导的查询生成机制，它能有效利用 VGGT 的内部语义先验，生成聚焦于语义相关区域的目标查询，如图7所示。该机制降低了对 VGGT 预测点云精确几何细节的依赖，因此能有效减轻噪声的影响。相比之下，FCAF3D 高度依赖输入点云的几何完整性来提取特征，因此更容易受到噪声影响；如表5所示，其性能迅速下降。

不同输入帧数量下的性能。为进行公平比较，我们遵循室内多视角三维目标检测中的标准做法（例如，官方 MVSDet 实现），使用 80 个输入帧。不过，我们的方法并不局限于这一配置。如表6所示，在不同输入帧数量下，我们的表现始终优于最强竞争方法 MVSDet。

帧数	20	40	60	80	100
MVSDet [47]	35.9	39.7	40.9	42.5	42.5
Ours	**42.6**	**45.3**	**46.2**	**46.9**	**47.3**

表6. 不同输入帧数量下的性能

与利用传感器几何信息的方法之间的性能差距。表 8 中，M1、M2 和 M3 分别对应 ImVoxel-Net [32], NeRF-Det [46], 和 MVSDet [47], , 而“+SG”表示利用传感器几何信息的方法。我们的方法在 SG-Free 和基于传感器几何信息两种设置下均取得了最高性能。两者之间仍存在明显差距，这在预料之中，因为传感器提供的几何先验能够为检测提供有力线索。

尽管如此，我们仍认为 SG-Free 设置是一种重要且实用的场景，因为在真实世界中，传感器几何信息往往不可用或难以获取。

方法	M1	M2	M3	Ours
mAP	35.2	41.2	42.5	**46.9**
方法	**M1+场景几何**	**M2+场景几何**	**M3+场景几何**	**我们的+场景几何**
mAP	48.1	49.5	56.2	**58.8**

表 7. 与利用传感器几何信息的方法之间的性能差距。

效率分析中的更多方法。为确保在相同的 SG-Free 设置下进行公平比较，我们训练并评估所有模型的 SG-Free 变体，而不是其原本依赖传感器几何信息的版本。具体而言，首先使用 VGGT 处理多视角 RGB 图像以估计相机位姿，然后将其提供给各个检测器。因此，SG-Free 流程天然包含 VGGT 的运行时间和内存开销，从而确保比较公平。在表 8中，M1、M2 和 M3 分别代表 ImVoxelNet [32], NeRF-Det [46], 和 MVSDet [47], 。如图所示，我们的设计（+Ours）的推理时间与最强竞争方法（+ MVSDet）相当，同时在所有方法中 GPU 内存使用量最低。

方法 VGGT		+M1	+M2	+M3	+我们
T（秒）	0.68	0.09	**0.07**	0.21	0.23
M（GB）	11.71	12.95	6.76	13.81	**3.57**

表8. 在无传感器几何信息设置下，使用 40 个输入帧与不同方法的效率对比。

视图数量对效率的影响。如表9所示，时间和 GPU 内存开销随视图增多而增加，这是多视角 3D 检测中的常见挑战。

帧数	20	40	60	80	100
T（秒）	0.49	0.91	1.48	2.20	3.06
M（GB）	14.46	15.28	17.50	19.48	21.66

表9. 不同输入数量下的效率对比。

可视化分析。为了进一步研究注意力引导的查询生成的有效性，我们对注意力图和生成的目标查询的位置进行了可视化分析。如图7所示，第2列展示了注意力图，其清晰地突出了物体区域。这一观察结果支持了我们的动机：注意力可以为查询生成提供语义引导。在第3列中，我们比较了生成的目标查询位置。红色点表示未使用注意力引导生成的目标查询，而绿色点表示通过我们的注意力引导的查询生成方法生成的查询。值得注意的是，在物体区域（由绿色框突出显示）内，与基线方法（红色点）相比，我们的方法生成了明显更多的目标查询（绿色点）。这表明，我们的注意力引导的查询生成能够有效聚焦于物体区域，从而带来更加准确的查询分布。正如主论文表2a所报告的那样，我们方法的有效性还体现在mAP@0.25提升了**2.8个点**。这一结果凸显了注意力引导的查询生成对整体性能的直接贡献。

7. 与替代方法的比较

定性比较。我们与表现最佳的竞争方法 MVS-Det 进行了定性比较 [47]。为适应无传感器几何设置（SG-Free）并确保公平比较，MVSDet 使用 VGGT 预测的多视角位姿进行训练和测试 [41]。如图8所示，我们的 VGGT-Det 检测到的对象更多，准确率也更高，这与主论文表1中报告的显著性能提升**4.4 个点**一致。性能提升得益于所提出的注意力引导查询生成（AG）和查询驱动特征聚合（QD）模块对 VGGT 内部先验的有效利用。

8. 训练和测试时间

所有消融实验均在八张 H800 GPU 上进行。在 ScanNet 数据集 [8] 上训练我们的模型大约需要 2 天。测试时，整个 ScanNet 测试集约 1 分钟即可处理完毕。

9. 局限性与未来工作

尽管 VGGT-Det 相较于强有力的替代方法取得了显著提升，但仍有若干局限性有待进一步探索。在当前所有 SG-free 流程中，VGGT 都带来了明显的运行时间和内存开销。此外，由于 VGGT 输出的是归一化预测，当前所有 SG-free 流程都会利用数据集提供的尺度对 VGGT 的预测进行反归一化。展望未来，引入一种更轻量、能够生成度量尺度预测的类 VGGT 模型，有望进一步推动这一方向的发展。

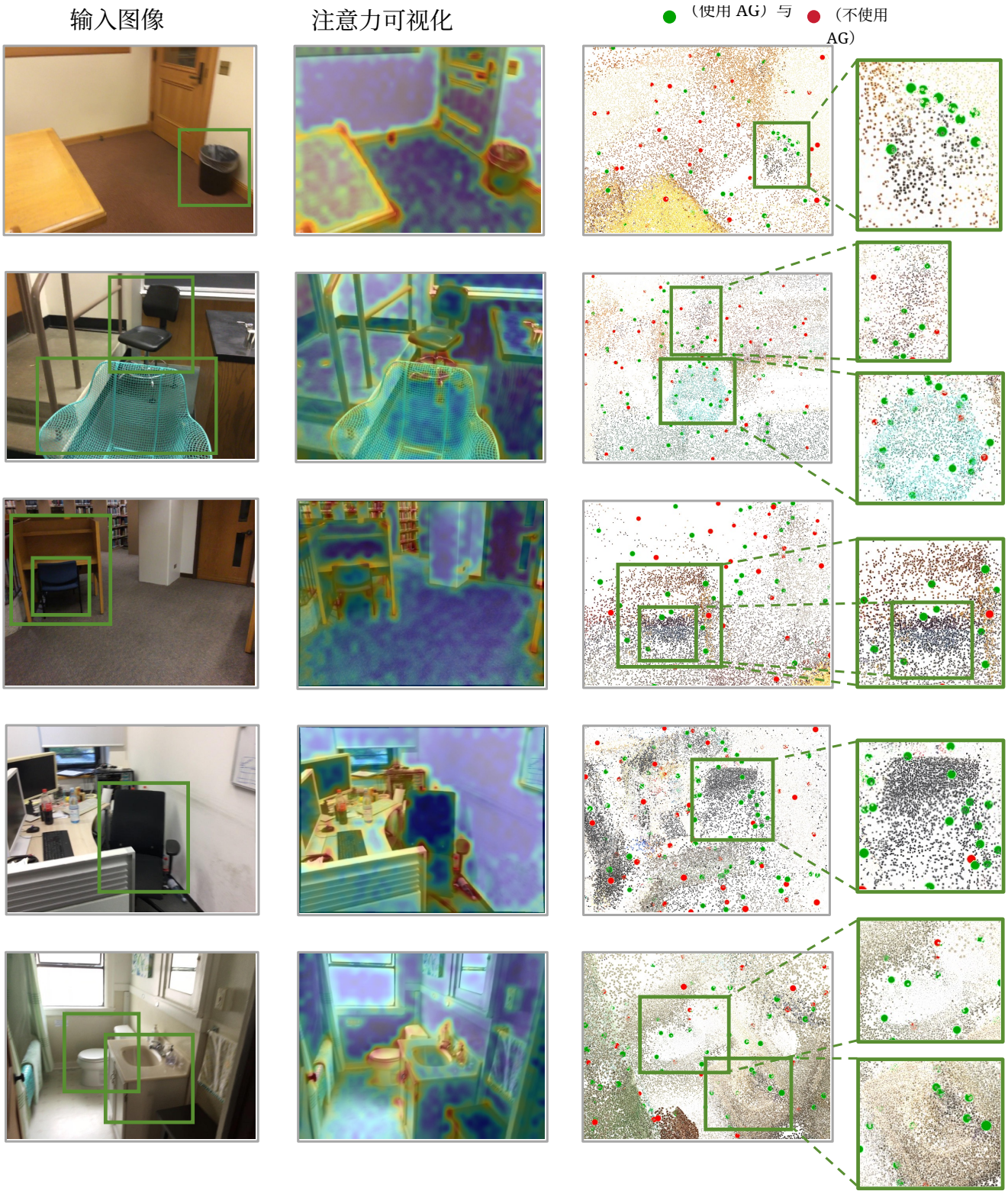


图 7. 注意力和生成的目标查询位置的可视化。与无引导的最远点采样（红色点）相比，AG 采样的点（绿色点）更集中于物体区域（由绿色框标出），因此这些区域内的绿色点比红色点更多。为便于查看，建议以彩色模式查看并放大图像。

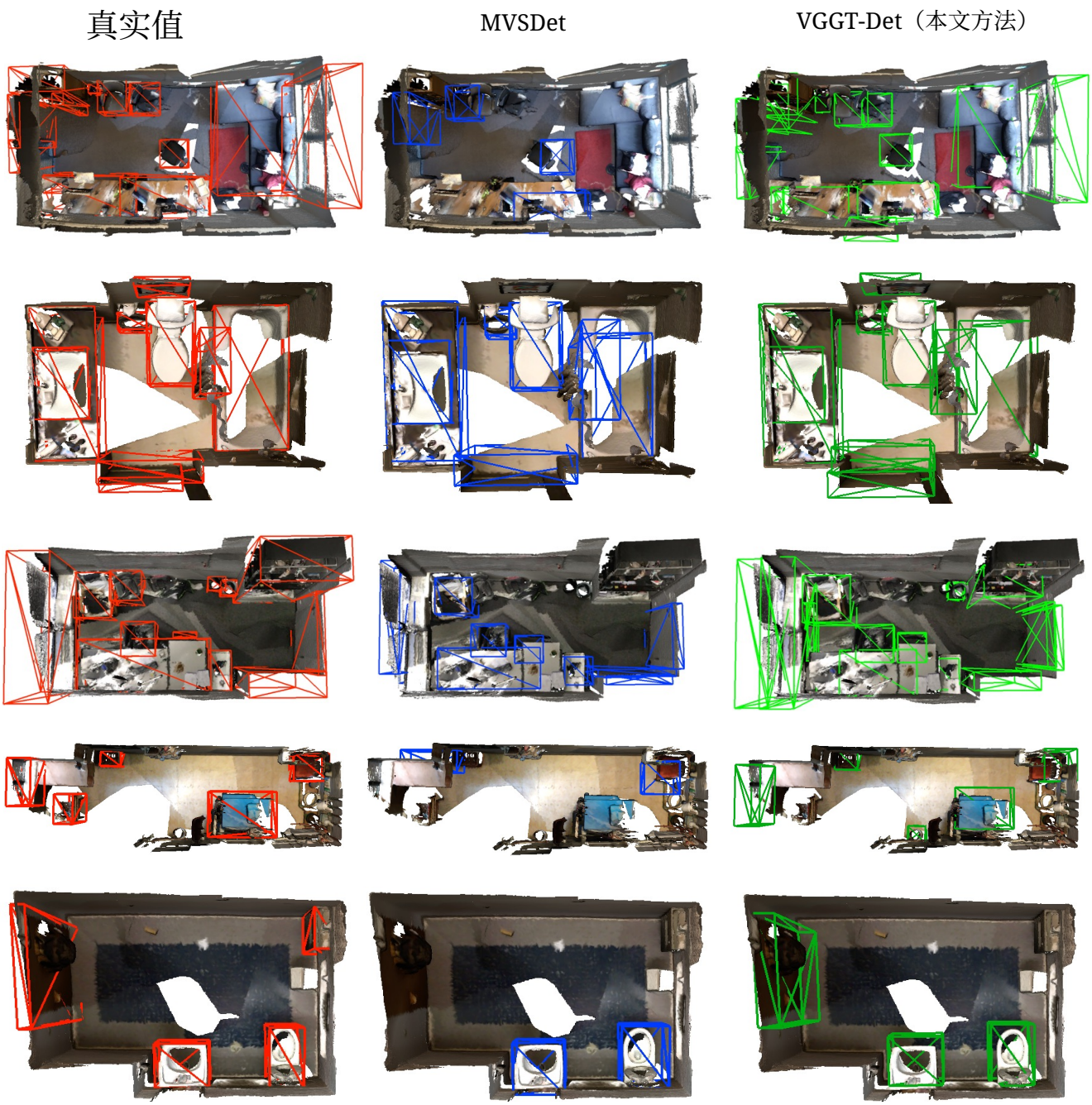


图8. 与MVSDet的定性比较 [47]。为实现无传感器几何设置（SG-Free）并确保公平比较，MVSDet使用由VGGT预测的多视角位姿进行训练 [41]。此处的网格并未在方法中使用，仅用于可视化。