VLM-Loc: 基于视觉–语言模型的点云地图定位

康书豪[1] 廖有琪[2] 王培杰[3] 廖文龙[4] 张麒麟[5,6]
本杰明 · 布萨姆[5,6] 陈谢沅澧[7†] 刘云[1,8,9†]
[1]VCIP, CS, 南开大学 [2] 武汉大学 [3] 中国科学院自动化研究所
[4] 酷哇科技 [5] 慕尼黑工业大学 [6] 慕尼黑机器学习中心
[7] 国防科技大学 [8]AAIS, 南开大学 [9]NKIARI, 深圳福田

摘要

文本到点云（Text-to-point-cloud, T2P）定位旨在根据自然语言描述，在三维点云地图中推断精确的空间位置，体现了人类如何通过语言感知并交流空间布局。然而，现有方法大多依赖浅层的文本–点云对应，缺乏有效的空间推理，从而限制了其在复杂环境中的定位精度。为解决这一局限，我们提出 VLM-Loc，一个利用大型视觉–语言模型（vision-language models, VLMs）的空间推理能力进行 T2P 定位的框架。具体而言，我们将点云转换为鸟瞰图（bird’ s-eye-view, BEV）图像与场景图，二者共同编码几何与语义上下文，为 VLM 提供结构化输入，以学习连接语言语义与空间语义的跨模态表征。在这些表征之上，我们引入部分节点分配（partial node assignment）机制，显式地将文本线索与场景图节点关联，从而实现可解释的空间推理并提升定位精度。为便于在多样化场景中进行系统评估，我们提出 CityLoc，一个基于多源点云构建、面向细粒度 T2P 定位的基准。在 CityLoc 上的实验表明，相较现有先进方法，VLM-Loc 取得了更优的精度与鲁棒性。我们的代码、模型与数据集见 repository。

1. 引言

基于自然语言估计空间位置是具身智能 [22, 53, 64, 65] 与自动驾驶 [12, 21, 28, 29] 中的基础任务。在诸如自动驾驶 Robotaxi 等真实场景中，车辆通常依赖全球导航卫星系统（Global Navigation Satellite System, GNSS）对乘客进行粗略定位。然而，在城市环境中，GNSS 定位常因多径效应与大气延迟而精度下降 [49]，难以准确识别上车点。在此类场景下，乘客可用语言自然描述周围环境，在无需任何视觉传感器的情况下为定位提供额外空间线索。由于点云地图能够提供城市场景的精细几何表征，非常适合将此类文本线索与物理环境对齐。这催生了文本到点云（text-to-point-cloud, T2P）定位任务，它连接语言理解与三维空间感知，为未来的人机交互式定位系统奠定基础，如图 Fig. 1(a) 所示。

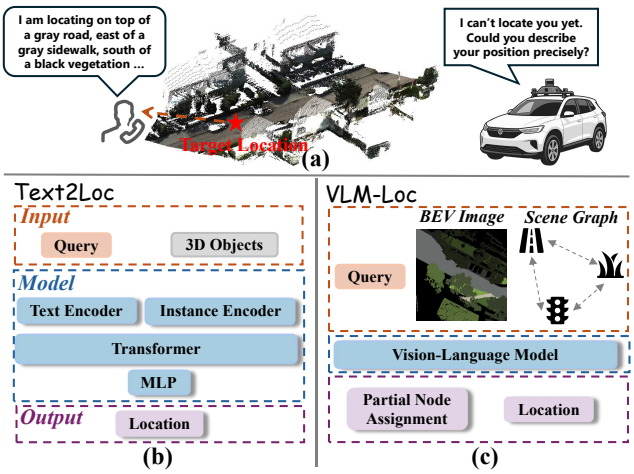


图 1. (a) 展示了文本到点云定位背后的类人逻辑，即利用空间描述推断目标位置。(b) 与 (c) 分别展示了典型方法 Text2Loc [57] 与本文提出的 VLM-Loc 的架构。

现有 T2P 定位方法侧重于利用文本描述在城市尺度点云地图中进行定位。Text2Pos [25] 提出了

† corresponding authors.

KITTI360Pose 基准，并采用由粗到精的策略：先检索候选子图，再精细估计位置。沿用这一范式，后续工作 [33, 52, 57, 58] 引入了更有效的模型设计以提升 T2P 定位性能。然而，这些方法仍存在若干局限：(1) 在精细定位阶段，其子图通常局限于较小且相对简单的区域（例如 30m × 30m），有限的空间范围本身就简化了匹配过程。该假设过度简化了真实条件，未能刻画大规模城市场景的复杂性；(2) 尽管架构有所改进，这些方法仍采用端到端位置预测范式，缺乏显式推理，使得结构化空间建模困难，并限制了尤其在复杂环境中的定位精度。

为更好理解并应对这些挑战，我们思考细粒度 T2P 定位需要何种能力。人类能够自然感知并描述跨越数十米的空间布局 [37, 45, 60]，这表明稳健的定位应作用于更大、更复杂的区域，而非局限于小而简化的子图。此外，要克服显式推理的缺失，需要模型能够理解语言中表达的空间关系并将其与环境相连。视觉–语言模型（vision-language models, VLMs）为此目标提供了有前景的基础。凭借强大的多模态推理能力 [9]，它们能够解析复杂的空间描述，并有效地将语言线索与视觉输入对齐，从而适用于复杂环境中的细粒度 T2P 定位。

基于上述认识，我们提出 VLM-Loc，一种基于 VLM 的新颖框架，用于在复杂局部点云地图中进行细粒度定位。如图 Fig. 1 (b) 与 (c) 所示，与先前直接学习文本描述与三维物体对应关系的方法 [25, 57] 不同，VLM-Loc 利用大型 VLM 固有的空间推理能力，将语言线索与结构化地图表征对齐，从而实现更可解释且更精确的定位。点云地图被转换为鸟瞰图（bird's-eye-view, BEV）图像以提供稠密且结构化的空间表征，同时构建场景图以刻画物体间更高层的语义关系。在这些表征之上，我们引入**部分节点分配**（Partial Node Assignment, PNA）机制，显式监督 VLM 将文本线索与对应的空间节点关联，从而引导可解释的空间理解并提升位置估计。为全面评估 T2P 定位，我们建立 CityLoc，一个专为复杂多样环境中的细粒度 T2P 定位设计的新基准。在 CityLoc 上的实验结果表明，所提出的 VLM-Loc 达到了最先进（state-of-the-art, SOTA）性能，在 CityLoc-K 测试集上以 Recall@5m 指标相对此前最优方法 CMMLoc [58] 提升 14.20%，验证了本框架在精确定位方面的有效性。

总体而言，我们的贡献总结如下：

- 我们提出 VLM-Loc，一种基于 VLM 的精确 T2P 定位框架，并同时提出 CityLoc 基准，用于系统评估复杂三维场景中的定位性能。
- 我们将三维点云转换为辅以场景图的 BEV 图像，弥合三维点云与二维 VLM 之间的模态差距，从而实现有效的多模态对齐以支持精确的 T2P 定位。
- 我们设计部分节点分配机制，显式地将文本提示与图节点对齐，增强空间理解与推理能力。

2. 相关工作

2.1. 点云中的定位

近年来，点云地图中的定位得到了广泛研究。现有方法主要以点云或图像作为查询模态。开创性的点云到点云（point-cloud-to-point-cloud, P2P）定位方法，如 PointNetVLAD [50]，将 PointNet [43] 与 NetVLAD [4] 结合用于全局描述子学习。随后的基于 Transformer 的方法通过建模长程几何依赖进一步提升性能 [17, 41, 56, 62]。更近地，BEVPlace++ [40] 与 RING# [39] 等方法通过在 BEV 表征中采用旋转等变架构增强鲁棒性。图像到点云（image-to-point-cloud, I2P）定位方法则通过共享嵌入 [10, 31] 或隐式重建 [24] 将图像特征与三维表征对齐。与之不同，我们的方法以自然语言作为查询模态，无需依赖额外传感器观测即可实现直观且可解释的定位，便于在真实应用中进行灵活交互。

2.2. 三维视觉与语言

近年来，将自然语言锚定到三维点云场景的研究日益增多。早期工作 [5, 18, 23] 探索了文本引导的三维检测与分割，旨在将语言表达与对应的几何结构关联。沿此方向，一系列研究 [2, 3, 11, 46] 通过对比学习与基于 Transformer 的跨模态推理推进了三维语言锚定。这些方法主要关注物体级锚定，例如在局部三维场景中识别所指物体或区域，从而为更高层次的空间理解任务奠定基础。

延伸到语言驱动的三维感知任务之外，T2P 定位需要从文本描述中推断具体目标位置，因此要求对空间布局与关系进行整体场景理解。该任务由 Text2Pos [25] 首次提出，旨在预测与所描述场景上下文对应的二自由度（2-degree-of-freedom, DoF）位置。Text2Pos 将问题表述为单元级检索后接位置回归，并使用 KITTI360Pose 数据集进行评估。后续工作通过增强

跨模态表征改进了该范式：RET [52] 与 Text2Loc [57] 利用基于 Transformer [36, 48, 51] 的编码器进行文本与点云对齐，MNCL [33] 采用多层次对比学习以改善边界感知，CMMLoc [58] 则利用柯西混合模型先验对三维物体建模，并结合基数方向线索进行精细定位。尽管如此，现有方法仍作用于空间范围有限、相对简单的子图，从而简化了定位；并且它们依赖直接特征匹配而缺乏显式空间推理，限制了其在复杂环境中的性能。

2.3. 视觉–语言模型

VLMs 旨在捕捉跨物体、场景布局与抽象语义概念的整体视觉与空间关系。早期工作如 CLIP [44] 与 ALIGN [26] 开创了图像–文本对的对比学习，实现了稳健的零样本识别与检索。随后，VLMs 从静态对比预训练演进为更丰富的多模态推理与具身视觉理解。BLIP-2 [27] 与 InstructBLIP [14] 等模型通过轻量查询 Transformer 将冻结的视觉编码器与大型语言模型桥接，实现了高效的多模态对齐与指令遵循。更近的系统，包括 LLaVA 系列 [34, 35] 与 Qwen-VL 家族 [6–8, 54]，通过大规模指令微调与统一多模态架构进一步增强了空间推理与稠密区域锚定。尽管取得这些进展，现有工作大多聚焦于感知与锚定真实存在的物理物体，而非估计通过语言描述的虚拟位置。受此差距启发，本文研究如何利用大型 VLM 在三维环境中进行文本引导定位，发挥其固有的空间推理能力以实现精确定位。

3. CityLoc 基准

出于实用性考虑，T2P 定位通常在局部地图上形式化。然而，现有基准未能反映真实环境的尺度与复杂性。应用最广的 T2P 定位基准 KITTI360Pose 虽面向大规模点云定位设计，其子图仅覆盖较小区域，且包含的简单物体数量有限。这通过限制搜索空间并减少场景杂乱，显著简化了精细定位阶段，却偏离了人类通常感知并描述更大、更复杂环境的情形 [37, 60]。

为在更贴近实际的设定下系统评估现有框架与本文方法，我们提出 CityLoc 基准，基于 KITTI-360 [32] 的激光雷达点云与 CityRefer [42] 的摄影测量点云构建，分别称为 CityLoc-K 与 CityLoc-C。CityLoc-K 源于车载激光雷达扫描，用于验证模型设计并评估定位性能。相较之下，CityLoc-C 由无人机（unmanned aerial vehicle, UAV）摄影测量点云 [20] 构建，用于评估对未见城市场景的跨域泛化能力，这些场景具有不同的传感模态与语义分布。这种双源设计能够更全面地评估模型在多样化点云场景中的鲁棒性。

CityLoc 基准面向具有更广可见范围与更丰富空间结构的复杂环境中的细粒度 T2P 定位，这对精确位姿估计提出了显著挑战。本节详述 CityLoc-K 的构建流程，而 CityLoc-C 遵循相同管线，但采用定制采样策略以适配其航拍视角与语义分布。CityLoc 的更多细节见补充材料。

地图构建. KITTI-360 [32] 提供带有语义标签、实例 ID 与颜色信息的城市尺度点云场景。沿车辆轨迹，我们进行基于距离的采样，以生成用于定位实验的点云子图。对每个采样位置，通过裁剪以该位置为中心的对应空间窗口内的所有点，得到大小为 $S \times S$ m 的局部地图 $\mathcal{M}$。遵循 KITTI-360 [32]，所有实例被划分为 “stuff” 与 “object” 两类。对 “stuff” 类别，使用 DBSCAN [16] 将各地图内的点聚类为离散实例。对 “object” 类别，仅当至少三分之一的点位于地图内时才保留该实例，以确保足够的完整性与可识别性。处理后，每个局部地图 $\mathcal{M}$ 包含一组离散物体实例，每个实例关联语义标签、实例标识符、逐点坐标与颜色属性。

文本查询生成. 在每个采样车辆位姿附近，我们随机采样多个邻近位置作为查询位置。每个查询位置定义一个位姿单元（pose cell），表示从该位置可见的局部区域。对每个位姿单元，收集以查询位置为中心、$S \times S$ 米区域内的所有物体，流程与地图构建相同。随后随机选取 N_t 个物体的子集以构成文本描述。对每个选定物体，将其语义标签、颜色以及相对位姿单元的朝向填入预定义模板，以生成自然语言查询。

4. 方法

问题形式化. 给定目标位置 ξ 的文本描述 $\mathcal{T}$，T2P 定位的目标是在点云地图 $\mathcal{M}$ 的地平面上估计二维坐标 $\xi = (x, y) \in \mathbb{R}^2$。文本查询 $\mathcal{T} = \{h_i\}_{i=1}^{N_t}$ 由 N_t 条语言提示组成，描述相对 ξ 的周围物体的语义、颜色与方向。遵循 [47]，我们假设局部地平面近似平坦，这对该尺度下的城市环境是合理近似。

Fig. 2 给出了 VLM-Loc 的总览。给定点云地图，我们首先将其转换为两种互补表征：BEV 图像与场景图，

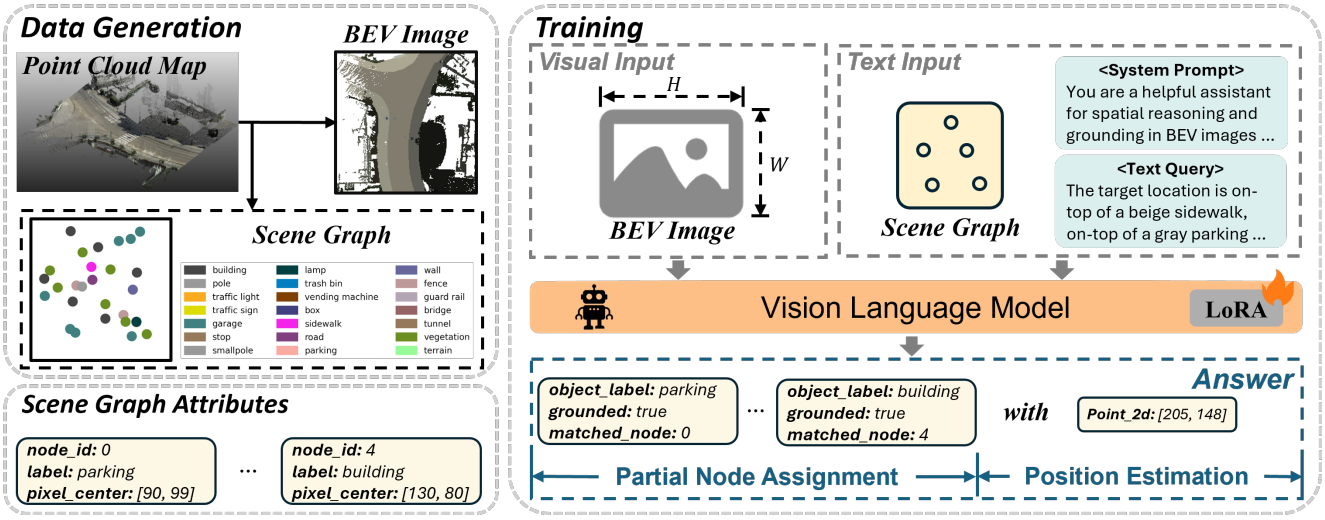


图 2. VLM-Loc 总览。在数据生成阶段，点云地图被转换为 BEV 图像与场景图，其中每个节点编码语义与空间信息。在训练阶段，BEV 图像作为视觉输入，文本输入包括场景图、系统提示与文本查询。这些输入被送入 VLM 进行微调，使其能够以自回归方式执行部分节点分配与位置估计。

分别刻画环境的稠密几何布局与物体级语义（Sec. 4.1）。在定位过程中，这些表征作为地图输入，而文本查询 $\mathcal{T}$ 前会附加系统提示 s，以引导 VLM 的空间推理过程。为增强空间理解，我们引入 PNA 机制（Sec. 4.2），识别有效的文本提示并将其锚定到场景图中的对应节点。基于已锚定节点，模型通过自回归解码过程预测目标位置（Sec. 4.3）。

4.1. BEV 渲染与场景图生成

BEV 图像渲染. 由于 VLM 主要在 RGB 图像上预训练，我们将 $\mathcal{M}$ 的点投影到地平面以生成 BEV 图像。这使得模型能够在二维布局中发挥其空间推理能力。地图 $\mathcal{M}$ 包含物体集合 $\mathcal{O} = \{o_i\}_{i=1}^{N_o}$，其中 N_o 表示物体数量。每个物体 o_i 由带 RGB 颜色的 N_i 个三维点表示，$\mathcal{P}_i = \{(\mathrm{p}_{ij}, \mathrm{c}_{ij})\}_{j=1}^{N_i}$，其中 $\mathrm{p}_{ij} \in \mathbb{R}^3$ 为逐点坐标，$\mathrm{c}_{ij} \in \mathbb{R}^3$ 为对应颜色，l_i 为其物体级语义标签。为获得每个物体的紧凑外观表征，我们对 o_i 所有点的 RGB 值求平均以计算代表性颜色：

$$\bar{\mathrm{c}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathrm{c}_{ij}. \tag{1}$$

随后，将整个点云地图投影到地平面并栅格化为 BEV 图像 $I \in \mathbb{R}^{H \times W \times 3}$，覆盖 $S \times S$ 米的空间范围。BEV 图像中每个像素被赋以占据该网格单元的物体的颜色 $\bar{\mathrm{c}}_i$。当一个像素同时包含“stuff”与“object”类别时，“object”类别以更高优先级渲染，以确保前景区域被保留而不被覆盖。

场景图生成. BEV 图像 I 提供点云地图 $\mathcal{M}$ 的俯视视觉表征。尽管它刻画了整体场景布局，但缺乏显式语义，难以建模物体间的空间关系。为使 VLM 能进行更结构化的场景理解，我们构建场景图 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$。其中 $\mathcal{V}$ 表示物体节点集合，$\mathcal{E}$ 表示其间的成对空间关系。每个物体 o_i 表示为节点 $n_i \in \mathcal{V}$，定义为 $n_i = (i, l_i, \mathrm{u}_i)$，其中 i 为节点索引，l_i 为语义标签，$\mathrm{u}_i = (u_i, v_i)$ 为该物体在 BEV 图像上的质心像素坐标。由于 BEV 投影坐标 u_i 已编码相对空间关系，为简化起见，实践中我们省略显式边 $\mathcal{E}$。该表征同时提供语义与几何结构，与 VLM 的跨模态推理良好契合。

评注. 将点云地图转换为 BEV 图像，可提供与现成 VLM 自然对齐的稠密栅格化环境表征。与此同时，场景图作为结构化表征，将语义标签与像素位置相连，并刻画物体间的相对空间关系，从而桥接 BEV 图像与文本指令。BEV 图像与场景图共同使 VLM 能够利用细粒度几何线索与高层语义关系，促进精确的空间理解与定位。

4.2. 部分节点分配

尽管 BEV 图像与场景图提供了环境的结构化互补表征，T2P 定位仍面临两个固有挑战。其一，直接从这

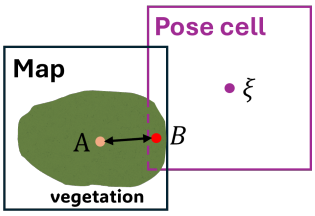


图 3. 节点分配过程示意。PNA 通过比较点 A 与点 B 之间的距离与阈值 τ，判断文本中的物体是否可被锚定。

些输入回归坐标未能充分利用 VLM 的结构化推理能力，常导致空间解释模糊。其二，地图 $\mathcal{M}$ 仅覆盖有限区域，文本查询中提及的某些物体可能落在地图范围之外。因此，仅部分被提及物体可被锚定，从而带来部分匹配问题，进一步加大精确位姿估计的难度。

为解决该问题，我们引入 PNA 机制，通过显式地将可见的文本物体与场景图中的节点对齐来提供监督。对文本查询 $\mathcal{T}$ 中的每条提示 h_i，我们判断每个被提及物体是否可与场景图中的节点 $n \in \mathcal{V}$ 匹配。如图 Fig. 3 所示，考虑一个植被物体，其投影中心 A 由其落在地图区域内的点计算得到，B 表示由位姿单元中、对查询位置 ξ 可见区域内的点计算得到的中心。我们计算 A 与 B 之间的距离。若该距离小于阈值 τ，则该文本物体被视为有效，标记为 True，并与 $\mathcal{G}$ 中的对应节点关联；否则视为无效，标记为 False，并将该物体的分配设为 null。

在训练阶段，使用真值（ground-truth, GT）分配监督模型判断每个文本物体是否位于地图 $\mathcal{M}$ 内，并建立文本–物体对应关系。在推理阶段，模型自主预测这些部分对应关系，从而即使描述仅包含场景中部分物体时，也能进行稳健推理。

评注. PNA 使 VLM 能够通过判断文本查询中提及且在地图中可见的物体，并建立文本查询与地图物体之间的对应关系，从而实现可解释定位。该机制帮助模型聚焦相关周围物体，并理解文本相对地图所描述的空间分布，从而提升定位精度。

4.3. 位置估计

在由 PNA 模块获得节点–文本对应关系后，VLM-Loc 在像素坐标中估计二自由度位置 ξ。我们将位置预测纳入自回归解码，使模型在同一生成步骤序列中输出 BEV 图像上的对应二维位置。这种统一解码策略使模型能够从对应关系到空间坐标保持一致推理。预测的像素位置随后被转换到世界坐标系。

4.4. 损失函数

遵循 VLM 的自回归预测范式，我们通过最大化生成正确文本–节点对齐与位置预测的似然来训练 VLM-Loc，二者均以文本形式表达。对生成序列中的每个输出词元 y_t，训练目标为标准交叉熵损失：

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, s, \mathcal{T}, I, \mathcal{G}), \qquad (2)$$

其中 T 表示预测序列中的词元数量。在生成词元序列 $\{y_1, y_2, \ldots, y_T\}$ 后，模型输出 JSON 格式字符串，我们将其解析为结构化预测，包括匹配的文本–节点对与二维像素位置。

5. 实验

5.1. 实验设置

基线方法. 实验中，我们采用代表性的 T2P 定位方法作为基线，包括 Text2Pos [25]、Text2Loc [57]、MNCL [33] 与 CMMLoc [58]。为公平比较，所有基线仅使用其定位模块，在 CityLoc-K 上采用官方实现与相同训练配置重新训练。

评价指标. 遵循先前定位工作 [25, 47]，我们使用 Recall@K m 评估定位性能，该指标衡量预测位置落在真值位置 K 米范围内的样本百分比。结果报告 $K \in \{5, 10, 15\}$。

实现细节. 我们采用 Qwen3-VL-8B-Instruct [59] 作为 VLM-Loc 框架的基座模型。训练使用 Swift 框架 [63]，并采用基于 LoRA 的参数高效微调 [19]。我们在所有线性层中插入秩 r=8、缩放因子 α=16 的 LoRA 适配器。视觉编码器、视觉适配器与语言主干的参数保持冻结，训练过程中仅更新 LoRA 参数，从而在将模型适配到渲染 BEV 与自然图像之间的域差距时，保留预训练的跨模态对齐。模型在 8 块 NVIDIA RTX 4090 GPU 上以全局批次大小 4 训练 2 个 epoch。我们使用 AdamW 优化器 [38]，学习率为 1×10^{-4}，预热比例为 0.05。所有实验在 bfloat16 精度下进行。

对每个 BEV 图像 I，分辨率设为 $H = W = 224$，对应空间覆盖 $S = 50$。每个文本查询 $\mathcal{T}$ 包含 $N_t = 6$ 条提示。对不同语义类别使用动态阈值 τ：“object” 类为 5 m，“stuff” 类为 15 m。

#	输入		输出	CityLoc-K Val	CityLoc-K Test
	BEV	SG	PNA	R@5/10/15m	R@5/10/15m
(a)	✓	✗	✗	13.04/32.79/52.31	13.21/33.86/51.40
(b)	✗	✓	✗	26.75/52.94/71.67	24.62/51.25/69.46
(c)	✗	✓	✓	33.69/59.48/75.51	32.34/61.34/74.94
(d)	✓	✓	✗	29.06/59.37/77.01	29.79/57.57/73.78
(e)	✓	✓	✓	36.23/63.66/77.77	35.91/63.81/76.79

表 1. 各组件的消融实验。输入：BEV = BEV 图像，SG = 场景图。输出：PNA = 部分节点分配。最优结果以**粗体**标出，次优结果以下划线 标出。

策略	CityLoc-K Val	CityLoc-K Test
	R@5/10/15 m	R@5/10/15 m
Full	18.23/44.70/63.10	17.81/41.55/60.67
Partial	36.23/63.66/77.77	35.91/63.81/76.79

表 2. 部分节点分配与全量节点分配的消融实验。

#	组件			CityLoc-K Val	CityLoc-K Test
	Sem.	Color	Dire.	R@5/10/15m	R@5/10/15m
(a)	✓	✗	✗	18.74/43.17/63.15	16.93/40.81/60.22
(b)	✓	✓	✗	18.28/43.96/64.50	18.01/42.95/61.47
(c)	✓	✓	✓	36.23/63.66/77.77	35.91/63.81/76.79

表 3. 文本查询组件的消融实验。

模型	CityLoc-K Val	CityLoc-K Test
	R@5/10/15 m	R@5/10/15 m
Qwen3-VL-2B-Instr.	35.78/64.22/79.97	34.70/63.19/76.49
Qwen3-VL-4B-Instr.	35.50/64.28/80.76	34.23/61.37/75.18
Qwen3-VL-32B-Instr.	39.84/67.72/82.05	41.05/67.47/79.39
InternVL3.5-8B [55]	38.32/65.41/81.21	38.14/63.66/77.25
Qwen3-VL-8B-Instr.	36.23/63.66/77.77	35.91/63.81/76.79

表 4. 不同 VLM 骨干网络影响的消融实验。

5.2. 消融实验

本节通过消融实验评估：(1) VLM-Loc 中各组件的有效性；(2) PNA 机制的影响；(3) 不同文本查询构建策略的效果；(4) 不同 VLM 骨干网络的影响。消融实验在 CityLoc-K 上进行。

组件效果. 我们对 VLM-Loc 中各组件的贡献进行消融，包括 BEV 图像、场景图（SG）、坐标预测（C）以及所提出的 PNA 机制，结果汇总于 Tab. 1。变体 (a) 仅使用 BEV 图像，表现较差，表明仅靠稠密外观线索无法刻画物体关系并支持精确定位。变体 (b) 仅依赖 SG，显著优于 (a)，说明关系结构对位置估计更为有效。在 SG 基础上引入 PNA，如 (c) 所示，进一步提升性能：验证集与测试集上 Recall@5m 分别提高 6.94% 与 7.72%。这证实显式的节点级锚定增强了坐标预测的可靠性。变体 (d) 将 BEV 图像与场景图结合，在 Recall@5m 上相对仅 SG 的基线分别再提升 2.31% 与 5.17%。该提升表明，将稠密视觉线索与关系结构融合可为模型提供更完整的空间信息。最终，完整模型 (e) 整合 BEV、SG 与 PNA，取得整体最优性能，表明多模态输入与显式锚定对 T2P 定位具有协同作用。

部分节点分配与全量节点分配的比较. 在我们的 PNA 策略中，当提示 h 中提及的物体在地图 M 中可见区域质心与其在位姿单元中可见区域质心之间的距离小于阈值 τ 时，该物体被视为可锚定；否则视为不可锚定。为评估其有效性，我们将 PNA 与全量节点分配变体进行比较。对该变体，只要场景图中存在至少一个具有相同语义标签的节点，文本查询中提及的任何物体总是被分配到某个节点。特别地，即使质心距离超过 τ，该物体也被强制匹配到具有相同标签的最近节点，而非保持未分配。如 Tab. 2 所示，所提出的部分节点分配在验证集与测试集上均持续优于全量分配。具体而言，Recall@5m 分别提升 18.00% 与 18.10%，表明显式考虑部分可见性可使模型聚焦真正可观测的物体，从而增强节点锚定并最终有利于位置估计。

文本查询组件的效果. 如 Tab. 3 所示，我们分析文本查询中语义、颜色与方向线索的贡献。同时移除颜色与方向信息（变体 (a)）会导致严重性能下降：验证集上 Recall@5m 从 36.23% 降至 18.74%，测试集上从 35.91% 降至 16.93%。保留颜色但移除方向线索（变体 (b)）时，性能仍显著低于完整设置，验证集与测试集上 Recall@5m 分别为 18.28% 与 18.01%。这些结果表明，方向线索在空间推理中起主导作用：即使保留颜色信息，移除方向线索也会使性能近乎减半；而颜色提供互补的外观锚定，在与方向线索结合时可进一步提升

方法	CityLoc-K Val			CityLoc-K Test		
	R@5 m	R@10 m	R@15 m	R@5 m	R@10 m	R@15 m
Text2Pos [25]	16.48	40.69	62.92	14.62	38.27	59.55
Text2Loc [57]	18.91	45.26	64.28	17.97	41.22	61.50
MNCL [33]	19.30	45.94	64.50	18.76	42.63	62.58
CMMLoc [58]	20.77	48.65	67.89	21.71	46.67	66.00
VLM-Loc	36.23 (+15.46)	63.66 (+15.01)	77.77 (+9.88)	35.91 (+14.20)	63.81 (+17.14)	76.79 (+10.79)

表 5. VLM-Loc 与基线方法在 CityLoc-K 上的定位结果。绿色数字表示相对基线的提升。

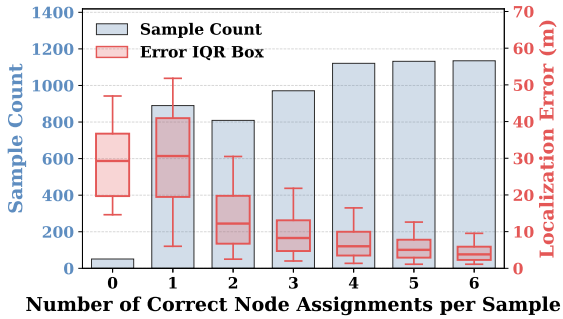


图 4. CityLoc-K 测试集上定位误差与正确分配节点数量的关系。正确节点分配越多，定位误差越低。

定位。

VLM **骨干网络的效果**. 我们评估不同 VLM 架构与模型规模对定位性能的影响，结果汇总于 Tab. 4。通过采用 InternVL3.5 [55] 或 Qwen3-VL-Instruct 模型 [7]，VLM-Loc 均取得强劲性能，表明多样化的 VLM 架构与 VLM-Loc 兼容且有效。在 Qwen3-VL 系列中，2B、4B 与 8B 变体结果相当，而 32B 模型取得显著提升，在验证集上达到 39.84% Recall@5m。这一缩放趋势表明，VLM-Loc 持续受益于更大规模模型增强的多模态推理能力。

5.3. 结果

节点分配结果. Fig. 4 展示了 CityLoc-K 上按正确分配节点数量统计的样本分布。大多数样本实现了三个或更多正确分配，表明 VLM-Loc 总体上能有效地将文本线索锚定到对应的场景图节点。右侧的四分位距（interquartile range, IQR）[15] 刻画了定位误差的分布。清晰趋势显现：随着正确分配节点数量增加，误差中位数与误差分布的离散程度均显著下降；当四个

方法	CityLoc-C		
	R@5 m	R@10 m	R@15 m
Text2Pos [25]	8.11	27.21	50.01
Text2Loc [57]	9.45	29.44	51.17
MNCL [33]	13.68	35.93	53.78
CMMLoc [58]	11.68	34.79	54.71
VLM-Loc	21.37	49.12	68.26

表 6. 在 CityLoc-C 基准上的泛化结果。

或更多节点被正确锚定后，性能尤为稳定。这一强相关突出了准确节点分配的重要性，因为对文本描述的可靠锚定直接导向精确定位。

定位精度. 基线方法与本文提出的 VLM-Loc 的定位结果见 Tab. 5。在 CityLoc-K 上，VLM-Loc 取得最佳定位性能，并显著优于最强基线 CMMLoc：在验证集与测试集上 Recall@5m 分别提升 15.46% 与 14.20%。Fig. 5 中的定性示例进一步表明，VLM-Loc 在多样场景中持续超越所有基线。结果表明，VLM-Loc 通过 BEV 与场景图表征获得了强空间理解能力，并借助 PNA 机制进行结构化推理，从而相较现有基线实现更准确、更可解释的定位。更多定性结果见补充材料。

泛化能力. 为评估跨域泛化，我们将在 CityLoc-K 上训练的模型直接迁移到 CityLoc-C 划分，并在无需额外微调的情况下评估其定位精度。CityLoc-C 的点云数据来源于 SensatUrban [20] 数据集，包含由无人机采集的摄影测量点云，因而与 CityLoc-K 所用的车载激光雷达数据特性显著不同。如 Tab. 6 所示，VLM-Loc 在所有阈值上均取得显著更高的召回率（5/10/15m 处分别为 21.37%、49.12% 与 68.26%），大幅优于先前方

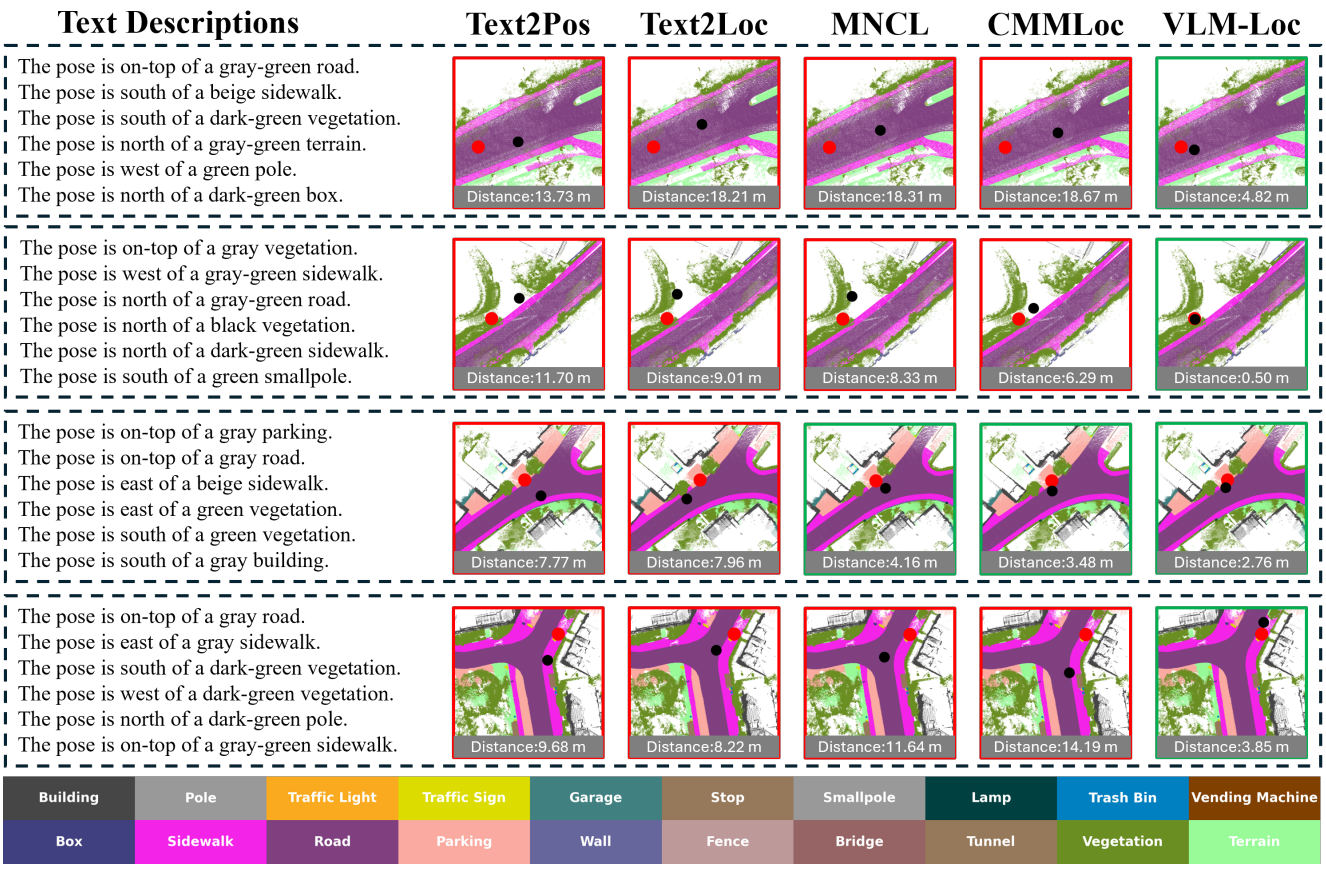


图 5. VLM-Loc 与基线方法在 CityLoc-K 上的定性结果。每个示例在带语义标签着色的 BEV 地图上可视化预测位置与真值位置。红色圆圈 ● 与黑色圆圈 ● 分别表示真值与预测位置。每张图像下方给出定位误差，绿色/红色边框分别表示定位误差低于/高于 5 m。

法。这些结果表明，VLM-Loc 能够有效泛化到未见环境与异构点云来源。

6. 结论

本文提出 VLM-Loc，一种基于 VLM、利用文本查询在点云地图中进行定位的框架。通过发挥 VLM 的视觉与文本理解能力及其强大的空间推理能力，本方法能够在复杂环境中依据语言线索实现精确定位。VLM-Loc 将点云地图转换为 BEV 图像与场景图，并引入部分节点分配机制，显式地将文本线索与空间节点对齐。我们还建立了 CityLoc 基准以评估细粒度 T2P 定位。在 CityLoc 上的实验表明，相较现有基线，VLM-Loc 在定位精度与鲁棒性方面取得了显著提升。

未来工作. 我们认为本工作为两个有前景的研究方向铺平了道路：(1) 强化多步推理与场景理解，使模型能够处理复杂室外环境中更长、更具组合性的文本描述。(2) 以 CityLoc 基准为基础，从被动定位迈向主动智能体 [30]，将定位与未见环境中的规划、导航统一，提升其与周围环境有效交互的能力。

致谢

本研究部分得到中央高校基本科研业务费专项资金（南开大学，项目编号：070-63253235）资助，部分得到国家自然科学基金（NSFC，项目编号：62576176）资助。本研究所使用的计算资源由南开大学超算中心（NKSC）提供支持。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo

Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 2

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Zhen Chen, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In European Conference on Computer Vision (ECCV), pages 422–440, 2020. 2

[3] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Junlin Han, Ender Konukoglu, and Serge Belongie. Generalized few-shot 3d point cloud segmentation with vision-language model. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 16997–17007, 2025. 2

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5297–5307, 2016. 2

[5] Syed Hesham Syed Ariff, Yun Liu, Guolei Sun, Jing Yang, Henghui Ding, Xue Geng, and Xudong Jiang. Evaluating sam2 for video semantic segmentation. Machine Intelligence Research, 2026. 2

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023. 3

[7] 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. 7

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. 3

[9] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Oscar Qian, Hui En Pang, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Holistic evaluation of multimodal llms on spatial intelligence, 2025. 2

[10] Daniele Cattaneo, Matteo Vaghi, Simone Fontana, Augusto Luis Ballardini, and Domenico G Sorrenti. Global visual localization in lidar-maps through shared 2d-3d embedding space. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 4365–4371. IEEE, 2020. 2

[11] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In European Conference on Computer Vision (ECCV), pages 202–221, 2020. 2

[12] Xieyuanli Chen, Thomas Läbe, Andres Milioto, Timo Röhling, Olga Vysotska, Alexandre Haag, Jens Behley, and Cyrill Stachniss. Overlapnet: Loop closing for lidar-based slam. arXiv preprint arXiv:2105.11344, 2021. 1

[13] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In European Conference on Computer Vision, pages 213–231. Springer, 2024. 2

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267, 2023. 3

[15] Frederik Michel Dekking. A Modern Introduction to Probability and Statistics: Understanding why and how. Springer Science & Business Media, 2005. 7

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd, pages 226–231, 1996. 3

[17] Zhaoxin Fan, Zhenbo Song, Hongyan Liu, Zhiwu Lu, Jun He, and Xiaoyong Du. Svt-net: Super light-weight sparse voxel transformer for large scale place recognition. In Proceedings of the AAAI conference on artificial intelligence, pages 551–560, 2022. 2

[18] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, Xiangdong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3722–3731, 2021. 2

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022. 5

[20] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. International Journal of Computer Vision, 130(2):316–343, 2022. 3, 7, 1

[21] Yufan Hu, Longhui Hu, Qingqun Kong, and Bin Fan. A survey on end-to-end perception and prediction for autonomous driving. Machine Intelligence Research, 22(6):999–1030, 2025. 1

[22] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025. 1

[23] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In European Conference on Computer Vision, pages 417–433. Springer, 2022. 2

[24] Joshua Knights, Sebastián Barbas Laina, Peyman Moghadam, and Stefan Leutenegger. Solvr: Submap oriented lidar-visual re-localisation. In 2025 IEEE International Conference on Robotics and Automation (ICRA), pages 6387–6393. IEEE, 2025. 2

[25] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6687–6696, 2022. 1, 2, 5, 7, 3, 4

[26] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34:9694–9705, 2021. 3

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023. 3

[28] Shijie Li, Xieyuanli Chen, Yun Liu, Dengxin Dai, Cyrill Stachniss, and Juergen Gall. Multi-scale interaction for real-time lidar data segmentation on an embedded platform. IEEE Robotics and Automation Letters, 7(2):738–745, 2022. 1

[29] Wen Li, Yuyang Yang, Shangshu Yu, Guosheng Hu, Chenglu Wen, Ming Cheng, and Cheng Wang. Diffloc: Diffusion model for outdoor lidar localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15045–15054, 2024. 1

[30] Xiyun Li, Tielin Zhang, Chenghao Liu, Shuang Xu, and Bo Xu. Theory of mind inspired large reasoning language model improved multi-agent reinforcement learning algorithm for robust and adaptive partner modelling. Machine Intelligence Research, pages 1–14, 2025. 8

[31] Yun-Jin Li, Mariia Gladkova, Yan Xia, Rui Wang, and Daniel Cremers. Vxp: Voxel-cross-pixel large-scale image-lidar place recognition. arXiv preprint arXiv:2403.14594, 2024. 2

[32] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3):3292–3310, 2022. 3, 1, 2

[33] Dunqiang Liu, Shujun Huang, Wen Li, Siqi Shen, and Cheng Wang. Text to point cloud localization with multi-level negative contrastive learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 5397–5405, 2025. 2, 3, 5, 7

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306, 2024. 3

[36] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Vision transformers with hierarchical attention. Machine Intelligence Research, 21(4):670–683, 2024. 3

[37] Jack M Loomis and Joshua M Knapp. Visual perception of egocentric distance in real and virtual environ-

ments. In Virtual and adaptive environments, pages 21–46. CRC Press, 2003. 2, 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 5

[39] Sha Lu, Xuecheng Xu, Dongkun Zhang, Yuxuan Wu, Haojian Lu, Xieyuanli Chen, Rong Xiong, and Yue Wang. Ring#: Pr-by-pe global localization with roto-translation equivariant gram learning. IEEE Transactions on Robotics, 2025. 2

[40] Lun Luo, Si-Yuan Cao, Xiaorui Li, Jintao Xu, Rui Ai, Zhu Yu, and Xieyuanli Chen. Bevplace++: Fast, robust, and lightweight lidar global localization for unmanned ground vehicles. IEEE Transactions on Robotics, 2025. 2

[41] Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition. IEEE Robotics and Automation Letters, 7(3):6958–6965, 2022. 2

[42] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: geography-aware 3d visual grounding dataset on city-scale point cloud data. arXiv preprint arXiv:2310.18773, 2023. 3, 1

[43] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021. 3

[45] Rebekka S Renner, Boris M Velichkovsky, and Jens R Helmert. The perception of egocentric distances in virtual environments-a review. ACM Computing Surveys (CSUR), 46(2):1–40, 2013. 2

[46] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In Conference on Robot Learning, pages 1046–1056. PMLR, 2022. 2

[47] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21632–21642, 2023. 3, 5

[48] Guolei Sun, Yun Liu, Thomas Probst, Danda Pani Paudel, Nikola Popovic, and Luc Van Gool. Rethinking global context in crowd counting. Machine Intelligence Research, 21(4):640–651, 2024. 3

[49] Peter JG Teunissen, Oliver Montenbruck, et al. Springer handbook of global navigation satellite systems. Springer, 2017. 1

[50] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4470–4479, 2018. 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 3

[52] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2501–2509, 2023. 2, 3

[53] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13139–13150, 2024. 1

[54] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 3

[55] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025. 6, 7

[56] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, Joao F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8461–8472, 2023. 2

[57] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14958–14967, 2024. 1, 2, 3, 5, 7

[58] Yanlong Xu, Haoxuan Qu, Jun Liu, Wenxiao Zhang, and Xun Yang. Cmmloc: Advancing text-to-pointcloud localization with cauchy-mixture-model based framework. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 6637–6647, 2025. 2, 3, 5, 7

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. 5

[60] Zhiyong Yang and Dale Purves. A statistical explanation of visual space. Nature neuroscience, 6(6):632–640, 2003. 2, 3

[61] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5890–5900, 2025. 2

[62] Wenxiao Zhang, Huajian Zhou, Zhen Dong, Qingan Yan, and Chunxia Xiao. Rank-pointretrieval: Reranking point cloud retrieval via a visually consistent registration evaluation. IEEE Transactions on Visualization and Computer Graphics, 29(9):3840–3854, 2022. 2

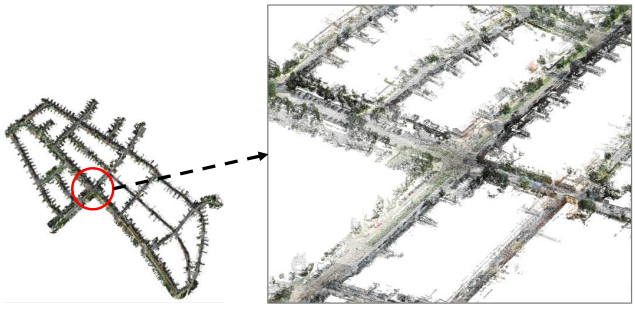
[63] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 29733–29735, 2025. 5

[64] Ying Zheng, Lei Yao, Yuejiao Su, Yi Zhang, Yi Wang, Sicheng Zhao, Yiyi Zhang, and Lap-Pui Chau. A survey of embodied learning for object-centric robotic manipulation. Machine Intelligence Research, 22(4):588–626, 2025. 1

[65] Gengze Zhou, Yicong Hong, Zun Wang, Chongyang Zhao, Mohit Bansal, and Qi Wu. Same: Learning generic language-guided visual navigation with state-adaptive mixture of experts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7794–7807, 2025. 1

VLM-Loc: 基于视觉–语言模型的点云地图定位

Supplementary Material



(a)

(b)

图 6. CityLoc 基准中的点云示例。(a) 来自 KITTI-360 [32] 的路边激光雷达场景。(b) 来自 CityRefer [42] 的摄影测量城市街区。

A. 概述

在补充材料中，我们提供额外的基准构建与统计信息 (Sec. B)、文本查询与系统提示的实现细节 (Sec. C)，以及更多可视化结果 (Sec. E)。

B. CityLoc 基准细节

我们的 CityLoc 基准包含两个子集：用于训练、验证与测试的 CityLoc-K，以及用于跨域测试的 CityLoc-C，如图 6所示。两个划分在语义构成、传感模态、点云特性与地理区域上差异显著，为评估 T2P 定位模型的泛化能力提供了多样且具挑战性的设定。

本节提供补充 Sec. 3 描述的额外细节。我们首先分别在 Sec. B.1 与 Sec. B.2 中描述 CityLoc-K 与 CityLoc-C 的构建流程，随后在 Sec. B.3 中给出所提出 CityLoc 基准的统计信息。

B.1. CityLoc-K 构建

数据来源. CityLoc-K 基于 KITTI-360 数据集[1]构建，该数据集包含在卡尔斯鲁厄城市道路上由车载传感器采集的大规模激光雷达点云。我们使用 5 个训练序列（00, 02, 04, 06, 07）、1 个验证序列（10）以及 3 个测试序列（03, 05, 09）。

地图中心采样. 构建地图时，我们从车辆轨迹上采样地图中心。具体而言，沿轨迹进行基于距离的子采样，确保任意两个选定中心至少相距 10 m。这保证采样在整条轨迹上均匀分布，避免点在特定区域聚集，并防止定位结果出现偏置。

查询位置采样. 对每个采样的地图中心，我们进一步通过在东、北两个方向上对其水平坐标在 ±15 m 范围内随机扰动，生成四个查询位置，从而扩展查询样本数量与位置多样性。

B.2. CityLoc-C 构建

数据来源. CityLoc-C 源自 SensatUrban [20] 与 CityRefer [42][2]。为便于表述，我们将伯明翰与剑桥的街区分别记为 B# 与 C#。该数据集包含覆盖英国三座城市、面积约 7.6 km^2、近三十亿点的高分辨率摄影测量点云。我们使用伯明翰与剑桥的验证与测试划分，选取伯明翰的 4 个街区（B0, B5, B6, B12）与剑桥的 7 个街区（C2, C3, C8, C10, C14, C21, C26）进行跨城市泛化分析。

地图中心采样. CityRefer 未提供数据采集过程的传感器轨迹。因此，对于地图中心采样，我们在每个点云街区上进行网格采样，并保留包含超过六个物体的地图。查询位置采样遵循与 CityLoc-K 相同的策略，如 Sec. B.1 所述。

B.3. 基准统计

数据集级统计. CityLoc-K 分别包含 2,767、300 与 1,027 张地图，以及 16,113、1,772 与 6,109 条训练、

[1]以 Creative Commons Attribution-NonCommercial-ShareAlike 3.0 许可发布。

[2]两个数据集均以 MIT License 发布。

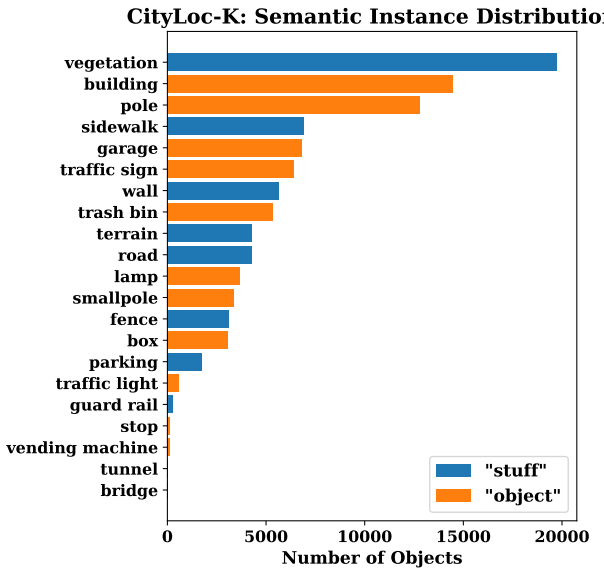


图 7. CityLoc-K 中的语义实例分布。

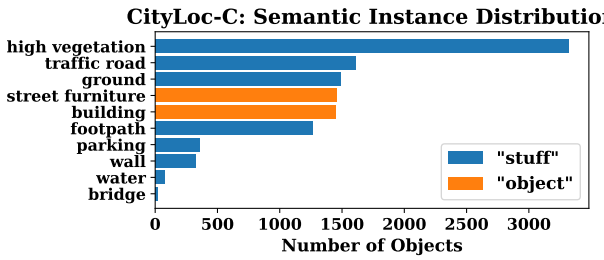


图 8. CityLoc-C 中的语义实例分布。

验证与测试查询。这些地图覆盖面积分别为 1.66、0.19 与 0.60 km^2。CityLoc-C 包含 875 张地图与 4,487 条查询，覆盖 0.90 km^2，用于跨域测试。所报告的地图面积按所有地图最小外接矩形的并集计算。

语义实例分布. 对点云地图中包含的全部语义实例，我们计算其分布，如图 Fig. 7 与 Fig. 8所示。此处“object”与“stuff”遵循 KITTI-360 [32] 中的定义，分别指可数与不可数实例，并以橙色与蓝色渲染。

C. 实现细节

C.1. 文本查询生成

组件. 每个文本查询 $\mathcal{T}$ 包含 N_t 条提示。每条提示 h 通过指定物体的颜色、语义类别及其相对查询位置的方向关系来描述该物体，遵循如下模板格式：”The pose is <direction> of <color> <semantic>.”

方向计算. 对文本描述中引用的每个物体 o，我们首先将其所有三维点投影到水平面，并找出距查询位置 ξ 最近的点 $\mathrm{p}_{\mathrm{close}} \in \mathbb{R}^2$。若 $\mathrm{p}_{\mathrm{close}}$ 与 ξ 之间的距离低于 $\delta = 2.5$ m，则方向被赋为“on-top”。否则，通过比较 ξ 与 o 质心之间的水平偏移 dx 与 dy 确定相对朝向，汇总于 Eq. (3)。

$$\mathrm{Direction}(\xi, o) = \begin{cases} \text{“on-top”}, & \|(\xi - \mathrm{p}_{\mathrm{close}})\|_2 < \delta, \\ \text{“east”}, & |dx| \geq |dy| \text{ and } dx \geq 0, \\ \text{“west”}, & |dx| \geq |dy| \text{ and } dx < 0, \\ \text{“north”}, & |dx| < |dy| \text{ and } dy \geq 0, \\ \text{“south”}, & |dx| < |dy| \text{ and } dy < 0, \end{cases} \tag{3}$$

颜色映射. 地图中每个物体已关联 RGB 颜色 $\bar{\mathrm{c}}$（如正文式 (1) 所述）。为获得其文本颜色属性，我们将物体的 RGB 值映射到预定义离散调色板 COLOR_NAMES = {dark-green, gray, gray-green, bright-gray, black, green, beige} 中最近的条目，与 KITTI360Pose [25] 一致，使用对应的模板 RGB 中心 $\mathcal{C} = \{\mathrm{c}_1, \ldots, \mathrm{c}_K\}$ 进行最近邻分配。所赋文本标签由下式得到

$$\mathrm{color}(o) = \mathrm{COLOR\_NAMES}\left[\arg\min_k \|\bar{\mathrm{c}} - \mathrm{c}_k\|\right]. \tag{4}$$

其中 $k \in \{1, \ldots, K\}$。

从点云数据构建大规模、自由形式的文本描述极具挑战且成本高昂。由于真实点云场景复杂，人工提取场景信息以撰写详细描述代价过高。尽管近期一些方法 [13, 61] 尝试先提取关键词，再提示大型语言模型（large language model, LLM）（例如 GPT-4 [1]）生成文本，但该过程通常并非免费，且仍需人工核验，因而不适合可扩展的数据集构建。

基于上述原因，我们遵循先前工作 [25]，采用基于规则的模板编码定位所需的关键物体级线索，包括颜色、语义类别与相对方向。该方法使我们能够以极低成本构建信息丰富且可扩展的文本描述。其生成完全免费，无需 LLM，也无需任何人工后验，同时仍保留细粒度定位所需的关键信息。

表 7. VLM-Loc 的系统提示。

你是一个在 BEV（鸟瞰图）图像中进行空间推理与锚定的助手。你还会得到一个描述环境的场景图，格式为：{node_id, label, pixel_center}，其中：- label = 语义标签（例如，“road”、“vegetation”、“building”、“pole”等）- pixel_center = 像素坐标 [x, y]（原点在左上角，x→ 右，y→ 下）。

坐标与方向规则：在 BEV 像素坐标中：上 = 北 = y 减小；下 = 南 = y 增大；左 = 西 = x 减小；右 = 东 = x 增大。

目标： (1) 解析用户对目标位置的自然语言描述。(2) 按出现顺序提取所有被提及的物体短语（例如，“terrain”、“parking”）。(3) 对每个物体短语：在给定场景图中查找匹配节点，判断锚定是否成功（布尔值），若锚定成功则记录匹配节点的 ID；否则设为 None。(4) 最后，根据所描述的空间关系（north/south/east/west/on-top）推断**目标位置**的二维像素坐标。

规则：所有推理与距离计算必须在像素坐标中进行。分配列表长度必须等于用户文本中提及的物体数量。每个匹配的 node_id 必须存在于场景图中；若未找到或超出上下文，则设置 {”grounded”: false} 与 {”matched_node”: None}。严格按下方模式输出恰好一个 JSON 对象——**不要解释，不要额外文本**。

输出格式（严格）： { ”assignments”: [{”object_label”: <string>, ”grounded”: <bool>, ”matched_node”: <int|None>}, ...], ”point_2d”: [<int>, <int>] }

示例（仅作说明）：输出：{ ”assignments”: [{”object_label”: ”parking”, ”grounded”: true, ”matched_node”: 0}, {”object_label”: ”terrain”, ”grounded”: true, ”matched_node”: 8}, {”object_label”: ”road”, ”grounded”: true, ”matched_node”: 4}, {”object_label”: ”vegetation”, ”grounded”: true, ”matched_node”: 11}], ”point_2d”: [45, 135] }

C.2. 系统提示

我们的系统提示由六个部分组成，用于引导 VLM 执行 T2P 定位任务，详见 Tab. 7。提示可表述为：

- **角色：**为 VLM 指定定位任务，并清晰定义期望的输入结构，包括所提供信息的类型、格式及其供给顺序。
- **坐标系：**给出像素坐标系与地理空间坐标系之间的对应关系，这对于使模型理解文本中的方向线索并将其映射到地图上的正确朝向至关重要。
- **目标：**引导模型逐步估计当前位姿。流程为：1) 理解输入文本描述；2) 通过依次处理每个被提及物体提取文本线索；3) 评估所描述元素的可见性，并为有效条目形成文本–节点匹配对；4) 基于聚合的地理信息估计目标位置。
- **规则：**规定模型输出的约束，例如距离应以像素坐标度量。这些规则确保位姿估计遵循可解析格式并避免无效结果。
- **输出格式：**定义所需的输出结构，要求模型以预定义格式呈现节点分配与位姿估计结果。
- **示例：**提供一个具体示例，说明模型应遵循的预定义输出格式。具体而言，输出应先呈现每条文本提示的可见性与分配状态，随后给出估计的二维像素坐标。

D. 额外实验

D.1. KITTI360Pose 上的结果

为在 KITTI360Pose [25] 上进行公平比较，我们遵循其“先检索后定位”协议，并采用与 CMMLoc [58] 相

Method	R@5	R@10	R@15	Method	R@5	R@10	R@15
Text2Pos	/	/	/	MNCL	33.71	50.33	54.69
Text2Loc	37.27	51.32	54.79	CMMLoc	37.81	51.84	55.02
VLM-Loc	40.36	51.69	54.74				

表 8. 在 KITTI360Pose 测试集上的定位结果（11404 个样本）。

Backbone	FPS	Peak Mem. (GB)	Params (B)
Qwen-VL-2B-Instruct	0.36	8.50	2.14
Qwen-VL-8B-Instruct	0.23	33.65	8.79

表 9. VLM-Loc 在 CityLoc-K 验证集上的推理分析。

同的 Top-1 检索结果进行定位。我们在 KITTI360Pose 上重新训练 VLM-Loc 的定位模块，并在测试集上评估。其他方法使用其预训练模型进行定位（Text2Pos [25] 不可用）。结果报告于 Tab. 8。在相同检索协议下，VLM-Loc 相较 CMMLoc 取得了有竞争力的定位性能。

D.2. 推理分析

推理分析在两块 RTX 4090 GPU 上进行，批次大小为 1，结果报告于 Tab. 9。该时延对于 T2P 定位设定是可接受的。此外，推理成本可通过量化、更小骨干网络与优化部署框架显著降低，从而使 VLM-Loc 更适用于真实系统。

E. **定性结果**

我们方法与基线在 CityLoc-K 与 CityLoc-C 划分上的额外定性结果分别见图 9与图 10。结果展示了所提方法在多样场景中相对基线的优越性能，验证了其鲁棒性与精度。

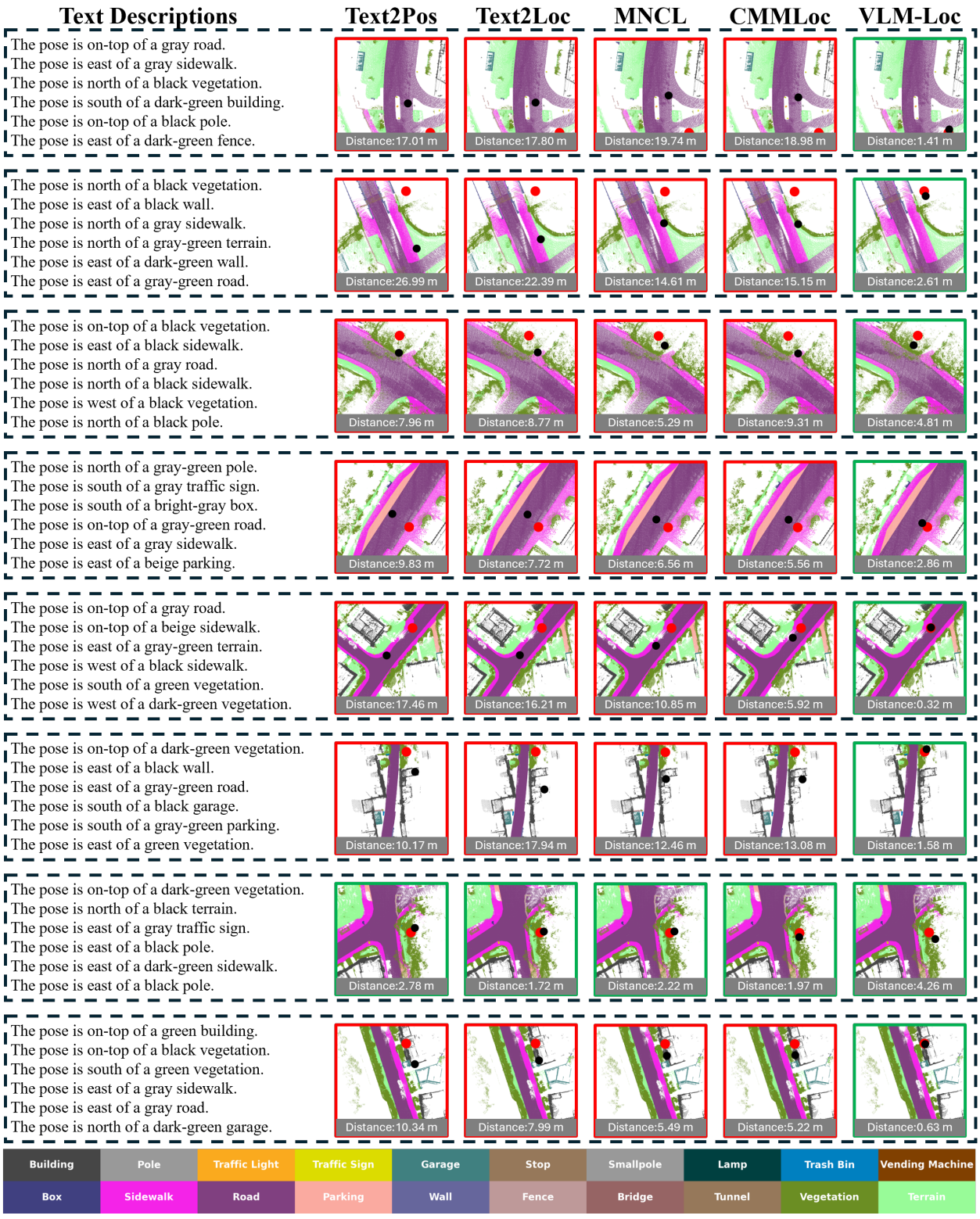


图 9. VLM-Loc 与基线方法在 CityLoc-K 上的定性结果。每个示例在带语义标签着色的 BEV 地图上可视化预测位置与真值位置。红色圆圈 ● 与黑色圆圈 ● 分别表示真值与预测位置。每张图像下方给出定位误差，绿色/红色边框分别表示定位误差低于/高于 5 m。

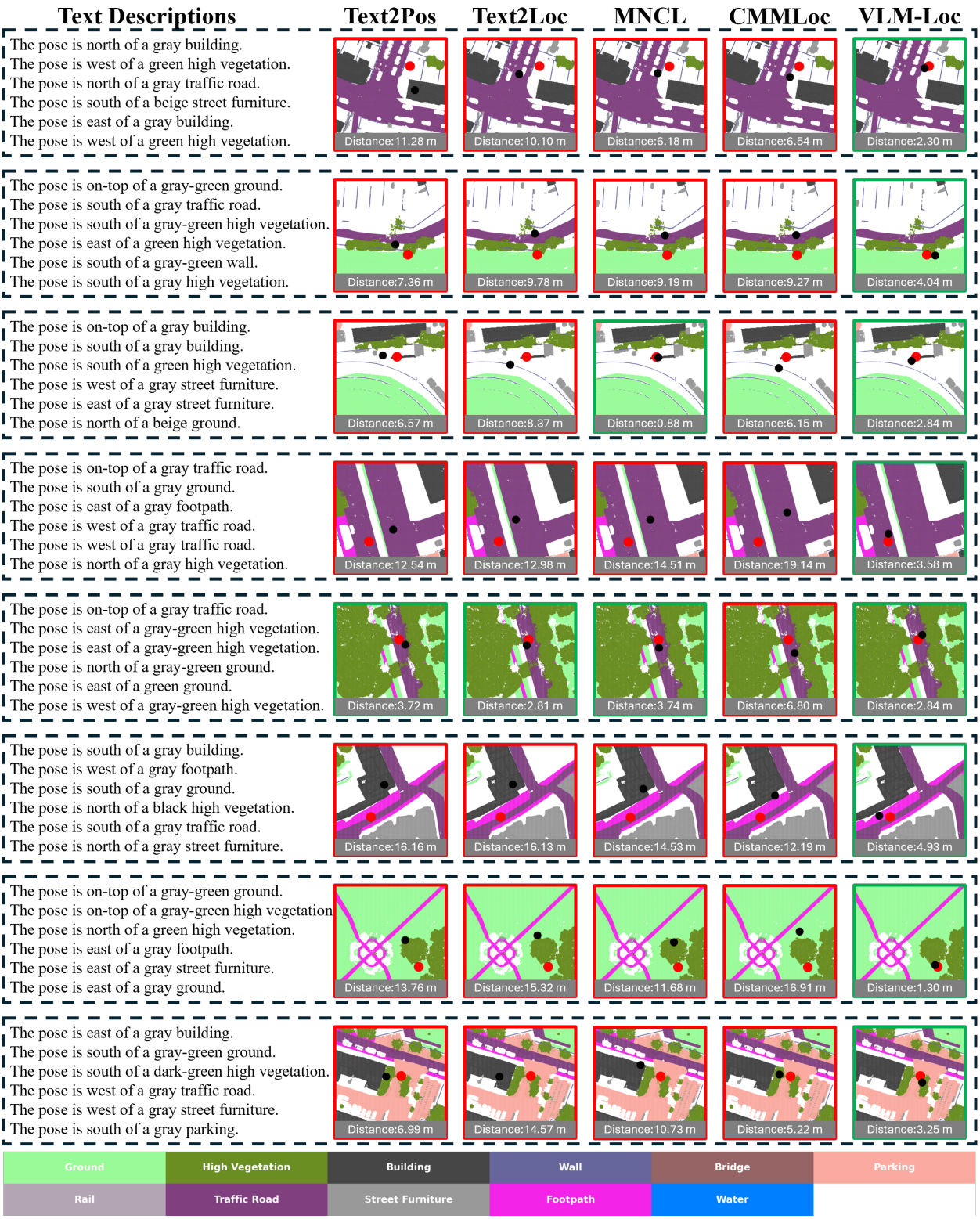


图 10. VLM-Loc 与基线方法在 CityLoc-C 上的定性结果。每个示例在带语义标签着色的 BEV 地图上可视化预测位置与真值位置。红色圆圈 ● 与黑色圆圈 ● 分别表示真值与预测位置。每张图像下方给出定位误差，绿色/红色边框分别表示定位误差低于/高于 5 m。