AddressVLM: 使用大型视觉语言模型进行图像地址定位的跨视图对齐调整

Shixiong Xu[1,2†], Chenghao Zhang[3†], Lubin Fan[3*], Yuan Zhou[1,2], Bin Fan[4], Shiming Xiang[1,2], Gaofeng Meng[1,2,5*], Jieping Ye[3]

[1*]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, Zhongguancun East Road, Beijing, 100190, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China.
[3]Alibaba Cloud, Beijing, 100020, China.
[4]School of Intelligence Science and Technology, University of Science and Technology, Beijing, 100190, China.
[5]CAIR, HK Institute of Science & Innovation, Chinese Academy of Sciences, HongKong, Country.

*Corresponding author(s). E-mail(s): lubin.flb@alibaba-inc.com; gfmeng@nlpr.ia.ac.cn;
Contributing authors: xushixiong2020@ia.ac.cn; zhangchenghao2018@ia.ac.cn; zhouyuan2023@ia.ac.cn; bin.fan@ieee.org; smxiang@nlpr.ia.ac.cn; yejieping.ye@alibaba-inc.com;
†These authors contributed equally to this work.

Abstract

Large visual language models (LVLMs) have demonstrated impressive performance in coarse-grained geo-localization at the country or city level, but they struggle with fine-grained street-level localization within urban areas. In this paper, we explore integrating city-wide address localization capabilities into LVLMs, facilitating flexible address-related question answering using street-view images. A key challenge is that the street-view visual question-and-answer (VQA) data provides only microscopic visual cues, leading to subpar performance in fine-tuned models. To tackle this issue, we incorporate perspective-invariant satellite images as macro cues and propose cross-view alignment tuning including a satellite-view and street-view image grafting mechanism, along with an automatic label generation mechanism. Then LVLM's global understanding of street

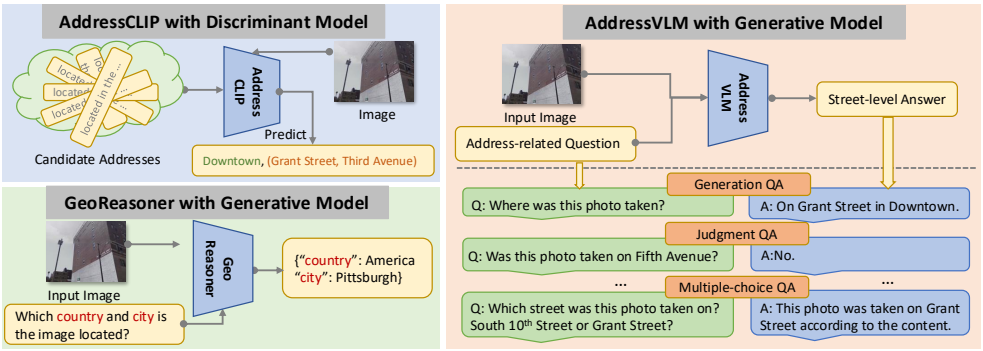


Fig. 1 将我们的 AddressVLM 与 AddressCLIP 和 GeoReasoner 进行比较。我们的方法专注于使用大型视觉语言模型进行城市范围的图像地址定位以及与地址相关的灵活地址问答。

distribution is enhanced through cross-view matching. Our proposed model, named AddressVLM , consists of two-stage training protocols: cross-view alignment tuning and address localization tuning. Furthermore, we have constructed two street-view VQA datasets based on image address localization datasets from Pittsburgh and San Francisco. Qualitative and quantitative evaluations demonstrate that AddressVLM outperforms counterpart LVLMs by over 9 % and 12 % in average address localization accuracy on these two datasets, respectively.

Keywords: Image address localization, Vision-language model, Cross-view alignment

1 介绍

视觉位置识别（VPR）旨在预测给定图像的地理位置，可分为两类：图像地理定位 [1–3] 和图像地址定位 [4] 。大型视觉语言模型（LVLMs）的出现，如 GPT-4V [5] 、Qwen-VL [6] 、和 LLaVA [7] ，显著影响了与图像和语言相关的各种任务。作为能够生成自然语言的生成模型，它们在图像定位任务中展现出更强的适应性和灵活性 [8] 。这种能力源于它们在训练阶段广泛接触街景和地标图像。

最近的工作，GeoReasoner [9] ，将大型视觉-语言模型与人类推理知识相结合，用于具有推理能力的街景地理定位，在国家或城市级别的粗粒度定位中表现出显著优势。然而，当涉及到特定区域（*i.e.* ，市中心）或城市内的街道（*i.e.* ，第五大道）的地址定位时，它可能难以预测准确的文本地址，因为街景图像相似性高且难以区分，而街道级别的地址名称尚未与相应的街景图像充分关联。相比之下，之前的工作 AddressCLIP [4] 通过街景图像和文本地址之间的对比学习研究全市地址定位。然而，这种方法由于依赖于只能在有限候选地址集合中作出区分的判别模型，固有地受到限制。因此，它缺乏提供多样化地址描述和回答其他相关查询的灵活性。

为了结合之前工作的优点，在这项研究中，我们探索如何将街道级地址定位功能整合到一个 LVLM 中。该模型预计能够灵活地响应用户关于地址定位的查询。我们将模型命名为 AddressVLM，旨在处理与地址相关的问题，并提供精确到地区和街道级别的答案。图. 1 展示了所提出的 AddressVLM 与 AddressCLIP 和 GeoReasoner 的比较。我们的方法可以回答各种类型的问题，包括生成、判断和选择题。

为了实现上述目标，一个合理的方法包括使用 LoRA 自适应 [10] 调优一个训练良好的 LVLM，利用街景问答（VQA）数据。然而，这种简单的方法在地址定位调优方面表现不佳。主要原因是街景图像在位置和视角方面采集稀疏，这限制了模型在整个城市中对街道分布的全球理解能力。在测试期间，街景图像被密集采样，因此全球信息对于有效的地址定位至关重要。为了在微调中补充全球信息，我们引入视角不变的卫星图像，以建立稀疏街景图像之间的连接。卫星图像在全球范围内一致并展示重叠特性，使得可以将稀疏街景图像映射到一个全球框架，从而促进图像间的关联。

先前在跨视角地理定位方面的研究 [11] 已显示了将卫星图像与街景图像关联的可行性。基于此，我们提出了一种名为跨视角对齐调整的方法，旨在使 LVLMs 能够将街景图像与标注有街道名称标签的卫星图像上的街道地址进行对齐。此方法将城市环境中街道分布的全局理解整合到 LVLMs 中，包含两个关键组件：卫星视角与街景图像嫁接机制和自动对齐标签生成机制。前者将街景图像放置在其对应区域卫星图像的右上角，作为跨视角对齐调整的输入。后者利用现成的 LVLM 来解释为什么街景图像根据提供的地址提示与卫星图像中的地址匹配，从而自动生成跨视角对齐调整的标签。通过这样做，我们的完整方法包括两个阶段的训练协议：跨视角对齐调整和地址定位调整。

我们介绍了两个城市范围内的街景 VQA 数据集，分别命名为 Pitts-VQA 和 SF-Base-VQA，分别构建于 Pitts-IAL [4, 12] 和 SF-Base-IAL [4, 13] 数据集之上。在 Pitts-VQA 上，AddressVLM 相比于没有跨视角对齐调优的基线提高了 9 %。在 SF-Base-VQA 上，AddressVLM 相比于基线提高了 12 %。此外，与使用 LVLMs 进行图像地址定位的最新方法（SOTA）GeoReasoner [9] 相比，我们的方法在 Pitts-VQA 和 SF-Base-VQA 数据集上分别显示了 11 % 和 14 % 的改进。与通用 LVLMs 相比，所提出的方法表现出卓越的城市范围内地址定位能力。我们进一步提供了定性结果以全面验证所提出的跨视角对齐调优策略的有效性。额外的定量实验显示我们的方法可以扩展到多个城市的地址定位。

本研究的贡献总结如下: (1) 我们探索将全市地址定位功能集成到 LVLMs 中，以实现基于街景图像的灵活地址问答。(2) 我们引入了跨视图对齐调整，结合对城市街道分布的全局理解到 LVLMs 中，其中包括跨视图图像嫁接机制和自动对齐标签生成机制。(3) 我们提出了 AddressVLM，其在基准方法上实现了一致的改进，并且表现优于先进的 GeoReasoner 方法和一般的 LVLMs。

Visual Place Recognition（视觉地点识别）。视觉地点识别旨在预测给定图像的地理位置，且在实际场景中有着广泛的应用。大多数研究人员专注于预测图像的纬度和经度坐标，这被称为通过检索和分类方法进行的图像地理定位。检索方法涉及将给定图像与用 GPS 标记的图像数据库进行匹配，并检索最相似图像的坐标作为预测结果。另一方面，分类方法则是将地球表面或城市划分为地理单元，并预测图像所属的地理单元。最近的趋势涉及利用嵌入视觉-语言模型中的通用文本知识进行地理定位，包括基于 CLIP 的判别模型例如 StreetCLIP，结合区域描述，以及注入 GPS 信息的 GeoCLIP，以及基于 LVLM 的生成模型如具有人工推理知识的 GeoReasoner。然而，这些模型通常仅专注于国家或城市级别的粗粒度定位。AddressCLIP 则关注城市内部的精细街道级定位，但这一判别模型被限制在有限的待选地址集中进行区分，无法提供灵活的地址描述或问答。在本研究中，我们探讨了将精细的全市范围内的地址定位能力整合到 LVLM 中。

大型视觉语言模型。LVLM 已成为一个新兴的研究热点，它利用强大的大型语言模型（LLMs） [24–27] 作为大脑来执行视觉-语言任务。这些通用的 LVLMs 在视觉问答任务中表现出显著的效果 [5–7, 28] ，这暗示了通往人工通用智能的潜在路径。对

于 VPR，LVLMs 可以根据地标、光学字符识别（OCR）信息或其他显著的视觉线索来识别输入图像的位置，通常可以精确到国家甚至城市级别 [8] 。然而，LVLMs 在实现精细的街道级定位方面的利用仍然是一个具有挑战性的问题。本研究利用 LVLMs 的能力来解决街景图像的地址定位问题。

跨视角地理定位。跨视角地理定位的目标与 VPR 类似，只是其数据库由航拍图像组成，而不是地面街景，并且查询可能是全景图像。关键挑战是在特征空间中匹配航拍图像和地面图像之间的特征 [11] 。解决该问题的一个经典方法是实施用于对齐的孪生网络，正如 Vigor [29] 所建议的那样。为了解决地面图像中的时间变化，作者在 [30] 中关注图像的时间不变部分。此外，有些工作 [31, 32] 提出基于部分的图像表示学习来解决方向和局部详细信息匹配问题。总的来说，这些研究证明了将航拍图像与街景图像关联的潜力。受到跨视角匹配精神的启发，我们将这一任务应用于 LVLMs 领域，并调整其引入跨视角对齐微调的方法。

2 方法

2.1 问题陈述

视觉问答中的图像地址定位问题形式化如下：给定一个训练数据集 $D_{train} = \{(I_i, Q_i^j, A_i^j)\}_{i=1}^{M}, j \in [1...N_i]$ ，其中 I_i 代表图像，(Q_i^j, A_i^j) 表示多轮问题和答案，我们的目标是训练一个 LVLM $\mathcal{H}_\theta$ 以根据查询图像和与地址相关的问题预测答案。对于每个图像 I_i ，我们将多轮对话数据组织为一个序列，其中第 $t-$ 轮指令 S_i^t 为：

$$S_i^t = \begin{cases} [I_i, Q_i^1], t = 1 \\ Q_i^t, t > 1 \end{cases}. \tag{1}$$

我们在预测标记上执行 LLM 的地址定位微调。具体来说，对于长度为 N 的序列，我们通过以下公式计算目标答案 A_i 的概率：

$$p(A_i|I_i, S_i) = \prod_{j=1}^{N} p_\theta(x_j|I_i, S_{<j}, A_{<j}), \tag{2}$$

其中 θ 是可训练的模型参数，$S_{<j}$ 和 $A_{<j}$ 分别是在当前预测标记 x_j 之前所有轮次中的指令和答案标记。在测试阶段，给定查询图像 I_k 和一组相关的对话问题 Q_j^k ，模型旨在为每个问题输出相应的答案 A_j^k 。

2.2 跨视图对齐调优

街景图像作为稀疏的微观视觉线索，使得为模型提供全局宏观视角变得具有挑战性，这对于有效的地址定位至关重要，因为在测试过程中，街景图像是密集采样的。相比之下，卫星图像可以被视为补充的宏观信息，它们视角不变且全球稳定，能够在稀疏的街景图像之间建立联系。受以前跨视角匹配工作的启发 [11, 33] ，我们提出了跨视角对齐调整，以将街景图像与卫星图像上的相应街道地址对齐。

卫星视图和街景视图图像嫁接。用于构造跨视图对齐调优输入图像的方法有多种，如图 2 所示。第一种方法是以约 1:1 的比例拼接地图和街景图像。这种方法似乎能够保留地图和街景的最多信息。然而，由于大多数 LVLM 仅接受方形输入图像（例如，

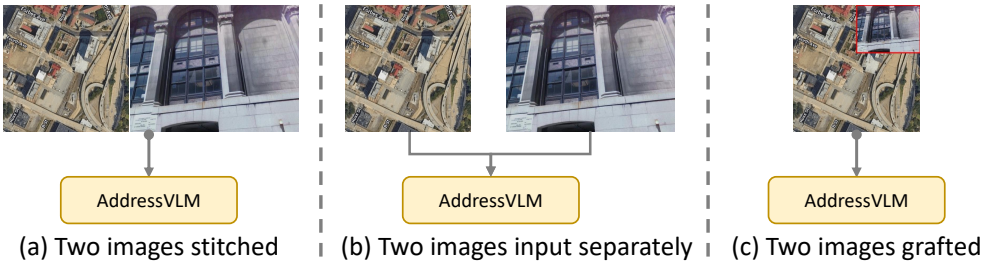


Fig. 2 结合卫星视图和街景图像的三种方式。

336 × 336），因此必要的填充和调整大小操作会导致有效视觉标记的数量减少，这对模型学习不利。第二种方法是分别输入两张图像。这种策略虽然可以保持两张图像的不同特征，但可能会导致模型过度依赖街景内容，忽视地图信息。此外，这种方法实质上使视觉标记的数量翻倍，影响训练效率。为了解决这些问题，并鼓励模型专注于地图中的整体街道分布信息，同时满足 LVLM 的输入大小要求并确保训练效率，我们采用第三种方法进行视觉数据构建。地图的大小被调整为 336 × 336，以适应 LVLM 的输入大小。

为了解决上述前两种方法的问题，我们采用卫星视图和街景视图图像嫁接机制，其中街景图像被缩小后嫁接到卫星图像上。设 I_{sa} 和 I_{st} 分别表示卫星图像和街景图像。嫁接的目标是通过结合两种视图图像生成一个新图像 I_s 。嫁接操作可以表示为：

$$I_s = \boldsymbol{M} \odot I_{sa} + (\mathbf{1} - \boldsymbol{M}) \odot I_{st}, \tag{3}$$

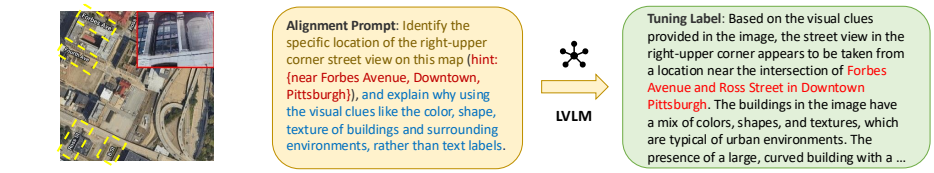
，其中 $\boldsymbol{M}$ 表示一个二进制掩码，指示从哪个视图图像中删除和填充，$\mathbf{1}$ 是一个全为 1 的二进制掩码，$\odot$ 是逐元素相乘运算。我们将街景图像放置在卫星图像的右上角，确保一个较长边的重叠率为 $\delta \in [0, 0.5]$ ，如图 3 所示。每条街道的文本名称都标记在卫星图像上，以便于街景图像和街道地址的对齐，这使得可以使用单个图像作为输入。不同嫁接参数的效果在第 **??** 节中进行了分析。

我们还提供了对于 Table 1 中三种不同嫁接方法的结果比较。 比较这三种不同的

Table 1 三种不同嫁接方法结果的比较。

Method	A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_s^G	A_s^J	A_s^M	$\bar{A}_s$	$\bar{A}$	A_{sd}
Two image stitched	85.73	92.39	93.42	90.95	67.36	91.05	92.17	85.78	87.91	66.45
Two images input separately	86.52	92.90	94.15	91.33	68.29	90.84	92.71	86.67	88.14	67.39
Two images grafted	88.73	93.54	95.16	92.70	72.51	91.70	93.98	87.46	90.02	69.60

嫁接方法，我们的方法产生了最佳结果。这归因于 LLaVA-phi3-mini 是为单图像输入场景设计的，当输入两个图像（两个图像拼接）时，会产生显著的领域差距，偏离其预训练条件。此外，由于 CLIP 适用于方形输入图像，具有高宽高比的图像可能在调整大小和填充操作（两个图像拼接）中丢失标记信息。总之，我们选择了 0.5 的比例，在输入图像构建中嫁接两个图像。



(a) Satellite and Street-view Image Grafting (b) Automatic Alignment Label Generation

Fig. 3 卫星和街景图像嫁接的示意图 (a)，以及对齐提示和生成标签的示例 (b)。(a) 中的红色和黄色框仅用于突出显示，在微调数据中没有标记。

自动对齐标签生成。为了使 LVLMs 能够通过地图建立对城市街道布局的整体理解，我们设计了一项跨视图对齐调整任务。这个任务允许模型通过视觉匹配街景图像与卫星图像从而定位街景图像的地址，其中相应的街道名称已被标记。同时，我们要求模型给出地址预测的理由。在执行跨视图对齐调整任务时，模型可以利用 OCR 能力感知周围的街道信息。

对齐调整的目标依赖于用适当的文本标签训练模型。一种直观的方法是基于人工规则和模板语言构建文本标签，但这种方式无法实现灵活和多样化的描述。为此，我们提出了一种自动对齐标签生成机制。在该机制中，根据规则给出的参考答案是预先给定的，而原因则由一个经过良好训练的 LVLM 生成为文本标签。在这里，我们在对齐提示中提供了一个文本提示作为标准答案，以帮助生成调整标签。图 3 显示了带有标签生成提示的自动对齐标签生成机制的流程。然后，参考答案被隐藏，并使用生成的标签执行对齐调整。 讨论。为了证明所提出的跨视图对齐调整的有效性，我们在图

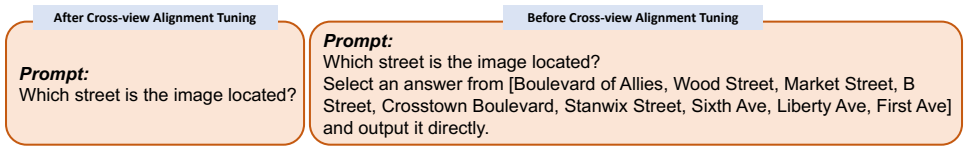


Fig. 4 交叉视图对齐调整前后的模型提示，用于定性结果。

5 中提供了对齐调整前后街道定位概率分布的定性比较。具体来说，我们将 LLM 的温度参数设置为 0.8，以增加推理的可变性。然后，对于每个输入的街景图像，我们使用如图 4 所示的特定提示执行 100 次模型推理。对于每个样本，我们记录在 100 次推理结果中每条街道出现的频率，以近似模型在跨视图对齐调整前后对周围街道分布的理解。道路地图上的红色标记表示输入图像的真实位置，而突出显示的街道是最频繁出现的前三个输出。可以观察到，经过跨视图对齐调整后，预测的街道集中分布在接近真实位置的位置，表明所提出的调整策略成功地将城市街道分布的知识与 LVLMs 整合在一起。

2.3 两阶段训练协议

街景视图视觉问答数据集。为了促进我们的研究，我们构建了两个针对地址相关问答定制的街景视图 VQA 数据集。这些数据集基于来自匹兹堡 [4, 12] 和旧金山 [4, 13] 的图像地址定位数据集。为了丰富问答数据的多样性，我们设计了三种问答模式：生成、判断和多选，如 图 1 所示。问答数据通过语言模板自动生成，并通过一系列多

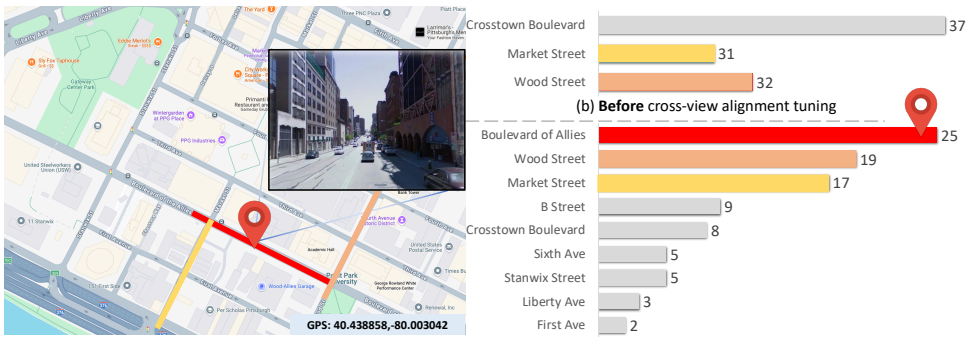


Fig. 5 街道定位概率分布在视角对齐调整前后的定性比较。我们使用红色、橙色和黄色来表示模型在调整对齐后的 100 次重复推断中预测的前三条街道，其中色条的长度表示模型预测相应街道的次数。在跨视角对齐调整后，预测街道聚集并靠近真实位置。可以在 这里 中找到源地图。

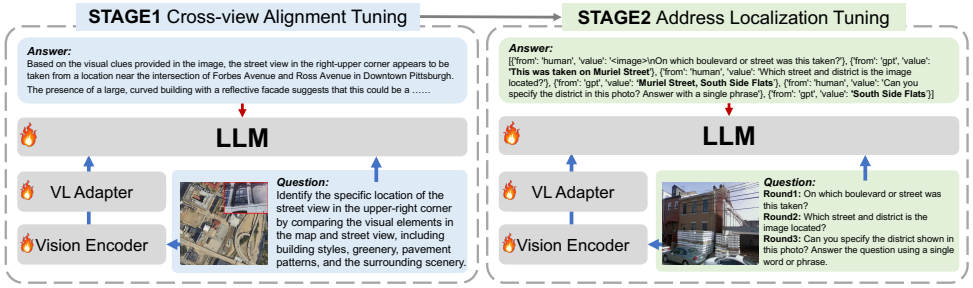


Fig. 6 所提出的框架概述包括两个阶段：跨视图对齐调整和地址定位调整。

轮对话进行组织。我们将这两个城市的 VQA 数据集分别命名为 Pitts-VQA 和 SF-Base-VQA。具体而言，Pitts-VQA 包含 10,586 个地点，每个地点有 24 张来自不同视角的图片，每张图片有 7 轮问答。SF-Base-VQA 包含 17,067 个地点，每个地点有 12 张来自不同视角的图片，每张图片有 7 轮问答。经过人工检查标签后，两个数据集都按 7:2:1 的比例划分为训练集、验证集和测试集。我们将在未来将其发布给社区。更多细节见附录 B 。

模型架构。图 6 展示了 AddressVLM 的架构，该架构基于 LLaVA [7] 所建立的框架设计。该模型由三个模块组成：视觉编码器 g ，视觉-语言 (VL) 适配器 h ，以及预训练的 LLM f 。对于输入的卫星视图或街景视图图像 I ，视觉编码器提供视觉特征 $Z_v = g(I)$ 。VL 适配器将视觉特征映射为语言嵌入标记，以 $H_v = h(Z_v)$ 表示，其中 $H_v \in \mathbb{R}^{N \times D}$ 代表与文本表示兼容的精炼视觉特征。对于另一个输入的文本地址查询 Q ，我们从地址查询中获得嵌入标记，表示为 $T_v = \Theta(Q)$ ，其中 Θ 代表现成的标记器和嵌入模型。最后，压缩的视觉特征序列和文本序列被连接起来以输入到预训练的 LLM 模块中，表示为 $A = f(H_v, T_v)$ 。

AddressVLM 经历了两个阶段的训练：跨视角对齐调优和地址定位调优。在第一阶段，我们的目标是通过卫星视图图像与街景图像的匹配，将整个城市内街道和区域的空间分布整合到 LVLMs 中，以实现地址定位。这个对齐调优过程对于促进第二阶段的地址定位调优至关重要。在第二阶段，我们整合了全局的街道分布信息的先验知

Table 2 与其他地址定位方法在 Pitts-VQA 和 SF-Base-VQA 数据集上的性能比较。

	Method	A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_s^G	A_s^J	A_s^M	$\bar{A}_s$	$\bar{A}$	A_{sd}
Pitts	LLaVA-Phi3-mini	26.64	60.22	37.81	45.52	0.00	56.23	34.62	36.69	41.01	0.00
	Baseline	84.51	92.72	93.23	90.70	64.31	90.25	91.27	84.00	87.27	60.52
	GeoReasoner	83.29	91.65	91.50	89.41	61.89	89.87	89.68	82.80	86.03	57.78
	AddressVLM	88.73	93.54	95.16	92.70	72.51	91.70	93.98	87.46	90.02	69.60
SF-Base	LLaVA-Phi3-mini	3.78	71.73	42.76	46.89	0.15	52.39	30.85	33.85	40.31	0.00
	Baseline	82.19	93.46	93.14	90.49	65.48	88.25	88.57	82.61	86.51	58.62
	GeoReasoner	81.40	91.07	90.81	88.53	62.89	86.46	84.64	80.08	84.26	55.99
	AddressVLM	86.48	93.72	94.50	92.06	76.09	88.92	92.75	86.66	89.33	70.45

识，以推断细粒度的、全市范围的地址位置信息。在这里，我们利用街景（VQA）数据，而不使用卫星视图图像。这两个阶段都使用低秩适应（LoRA）进行微调，从而提升了地址定位的整体性能。这种两阶段的方法使模型能够更好地捕捉图像和地址对之间的复杂关系，通过利用综合的空间知识，提高其准确定位地址的能力。

3 实验

3.1 实验装置

实现细节。AddressVLM 基于 CLIP [21] 和 Phi-3.1-mini [27]，以 LLaVA 方式使用 xtuner [34] 框架构建，该框架是用 PyTorch 实现的。所有图像都调整为 336×336 以适应 CLIP 的输入尺寸。更多细节见附录 A 。

评估指标。为了严格评估模型在不同对话上下文中的地址定位能力，我们采用多种格式和指标来评估不同水平的定位准确性。我们提出了三种类型的问题：生成、判断和选择题。它们分别应用于区域和街道级别。我们将与区域相关的 G 生成、J 判断和 M 多选题的准确性分别记作 A_d^G 、A_d^J 和 A_d^M ，它们的平均准确性记作 $\bar{A}_d$ 。相应地，街道级别评估的准确性分别记作 A_s^G 、A_s^J 和 A_s^M ，平均准确性为 $\bar{A}_s$ 。两个级别的总体定位准确性记作 $\bar{A}$ 。此外，我们研究了模型同时生成街道和区域信息的能力，称为 A_{sd} 。该指标与 AddressCLIP [4] 中街道级别的 top-1 准确性 (SA-1) 有一定相似性。然而，值得注意的是，我们报告的 A_{sd} 适用于生成模型，这使其成为比判别型 SA-1 更具挑战性的度量。

3.2 主要结果

基线。首先，我们评估采用的预训练 LVLM，以展示其在图像地址定位方面的零样本能力，用 LLaVA-Phi3-mini 表示。随后，我们复现 GeoReasoner [9] 在地区和街道级别的结果，作为 SOTA 方法。更多方法细节可以在附录 C 找到。我们仅对 LLaVA-Phi3-mini 进行地址定位微调，称之为两种方法的基线。

比较。表 2 显示了我们的 AddressVLM 模型和前述模型在两个数据集上的结果。我们的方法在所有指标上都取得了最佳结果。具体来说，由于对城市环境的细粒度和多模态理解不足，LLaVA-Phi3-mini 的零样本性能在两个数据集上都表现欠佳。然而，其在 A_d^J 上的表现优于随机猜测（在两个数据集上分别为 60.22 % 对比 50 % 和 71.73 % 对比 50 %），这表明它确实具备一定的城市基础知识。在对 LLaVA-Phi3-mini 应

Table 3 嫁接重叠率 δ 和卫星图像类型对两种数据集上的跨视图对齐调整的消融研究。

Method	δ	Pitts-VQA		SF-Base-VQA	
		$\bar{A}$	A_{sd}	$\bar{A}$	A_{sd}
Satellite w/o road	0.3	88.36	64.05	87.32	65.33
Satellite	0.3	89.32	68.98	88.67	70.42
Satellite w/o road	0.5	88.06	64.63	87.41	65.93
Satellite	0.5	90.02	69.60	89.33	70.45

用我们的方法进行两阶段调优后，与零样本设置相比，AddressVLM 的整体性能显著提高（在两个数据集上 $\bar{A}$ 方面分别提高了 +49.01 % 和 +49.02 %）。对于 SOTA 方法 GeoReasoner，关键在于旨在粗粒度识别和增强推理能力的第一阶段推理调优。虽然这种策略在国家层面上带来了好处，但同一城市内部街景的有限差异可能导致负面影响，导致在两个数据集上 A_{sd} 方面分别下降 2.74 % 和 2.63 %。相比之下，我们的 AddressVLM 在第一阶段调优中构建了卫星图像和街景图像对齐任务，有效地将街道名称和全球街道分布知识整合到模型中。与直接应用定位调优的基线相比，所提出的对齐调优阶段带来了显著且一致的性能提升，例如，在两个数据集上 A_{sd} 方面分别提高了 +9.08 % 和 +11.83 %。此外，我们可以观察到 AddressVLM 和 AddressCLIP 在街道和区域定位性能（A_{sd} ）方面的差距，表明开放集合生成模型在特定任务中要实现与闭集合分类模型相当的结果仍具有挑战性。这是一个有希望的方向，我们希望在未来的工作中进行探索。

跨视图对齐调整的嫁接机制。跨视图对齐调整是 AddressVLM 有效性的关键步骤，对于构建视觉数据有多种选择。第一个关键因素是街景图像较长侧与卫星图像的重叠比例 δ （默认 $\delta = 0.5$ ）。第二个因素是卫星图像的类型 *i.e.* ，即卫星图像是否带有文本街道名称的标注。关于它们的消融结果显示在表 3 。结果表明，将 δ 减少到 0.3 会导致性能下降，表明较小的街景图像无法提供足够的视觉细节。同时，从卫星图像中去除街道标签也会导致性能下降，因为没有 OCR 道路信息的情况下，卫星地图无法充分表示街道布局。因此，我们最终采用带有街道名称的卫星图像并设置 $\delta = 0.5$ 。

Table 4 在不同训练阶段使用不同参数进行训练的消融研究，✔ 表示一个模块是可训练的。

Variants	Stage-1		Stage-2	Pitts-VQA		SF-Base-VQA	
	VE	LLM	VE	$\bar{A}$	A_{sd}	$\bar{A}$	A_{sd}
A				86.58	63.21	86.74	62.94
B			✔	86.42	62.95	86.31	62.78
C		✔		87.48	63.03	85.92	61.21
D		✔	✔	89.53	66.37	89.63	68.95
E	✔	✔		87.37	63.52	87.07	64.68
Ours	✔	✔	✔	90.02	69.60	89.33	70.45

在 LVLM 中的训练组件。LVLMs 中的训练参数是否冻结通常会影响其在特定领域任务上的表现。为此，我们探讨了冻结或解冻 Tab. 4 中的 AddressVLM 组件的影响，其中包括 Vision Encoder（VE）和 LLM。在两个阶段中，VLA 和第二阶段的 LLM

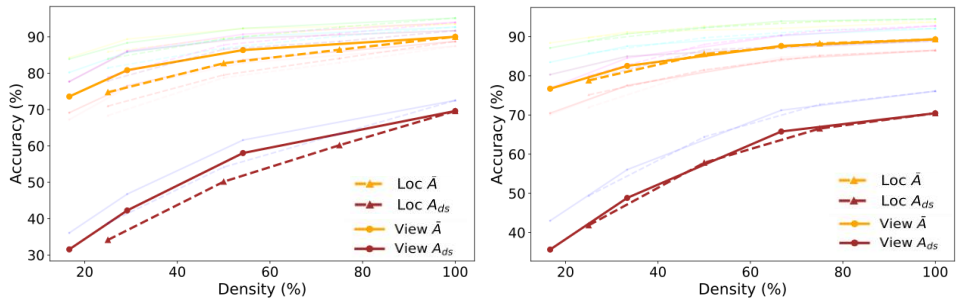


Fig. 7 关于不同密度的街景图像在两个数据集上的地址定位消融实验。背景中的半透明线是 $\bar{A}$ 的子指标，如上所定义。

默认是解冻的。值得注意的是，在第一阶段解冻 LLM 能带来最高的性能提升。同样地，在第二阶段解冻 VE 通常比冻结 VE 取得更好的性能，因为第二阶段训练的输入是街景图片，解冻 VE 能够使模型更好地适应城市街景。最终，解冻所有参数达到了最佳性能。这个结果可以归因于任务的高度特异性以及大型数据集的可用性，这有助于全面的参数优化以实现最佳结果 [35] 。这些发现与社区之前的结论相一致。

街景图像密度研究。我们研究用于地址定位调整的街景图像的不同密度对性能的影响，这可以在两个方面反映出来：i) 视点密度，意指对一个地点可用的街景图数量（例如，100 %，50 %，25 %，12.5 %）。ii) 位置密度，指位置的下采样率（例如，100 %，75 %，50 %，25 %）。我们在图 Fig. 7 中对此进行了独立分析。如观察所示，对于视点密度，当视点下采样到 50 % 时，模型保持超过 88 % 的性能（$\bar{A}$ ）。对于位置密度，当位置下采样到 50 % 时，模型保持超过 71 % 的性能（A_{sd} ）。结果表明，我们的方法在较低数据密度下具有很强的泛化能力。同时，我们注意到，我们的方法对视点和位置密度的敏感性相似，这表明这两个维度的密度对定位性能同等重要。

Table 5 混合训练在 Pitts-IAL 和 SF-IAL-Base 数据集上的效果。

Train / Test	District				Street				$\bar{A}$	A_{sd}
	A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_d^G	A_d^J	A_d^M	$\bar{A}_d$		
Pitts / Pitts	88.73	93.54	95.16	92.70	72.51	91.70	93.98	87.46	90.02	69.60
Pitts + SF / Pitts	89.24	93.17	95.16	92.66	72.90	92.77	94.34	88.18	90.37	70.63
SF / SF	86.48	93.72	94.50	92.06	76.09	88.92	92.75	86.66	89.33	70.45
Pitts + SF / SF	87.40	94.24	94.92	92.66	77.05	91.97	93.00	88.48	90.55	71.36

多城市的扩展性。考虑到图像地址定位在实际中可能涉及多个城市，我们首先在两个数据集上评估 AddressVLM 的扩展性。具体而言，我们合并这些数据集并使用提出的两阶段调优方式训练一个统一的 AddressVLM，然后在两个测试集上进行评估。正如表格 5 所示，令人惊讶的是，这个统一模型在两个数据集上的性能略微超过了每个单独模型的性能。我们推测，更多的同一任务的交叉视角数据有助于模型学习如何使用地图来参考定位街景图像。这一发现进一步证明了我们流程的扩展性，暗示了其在更多城市甚至整个国家扩展能力的潜力。我们还在美国以外的数据集进行了方法性能

Table 6 将不同比例的 llava_v1_5_mix665k 数据纳入我们的阶段 2 训练的结果。

Method	$\bar{A}_d$	$\bar{A}_s$	$\bar{A}$	A_{sd}
w / llava_v1_5_mix665k (5:1)	84.38	75.70	79.95	42.08
w / llava_v1_5_mix665k (1:1)	87.97	79.08	83.43	50.53
wo / llava_v1_5_mix665k	92.70	87.46	90.02	69.60

Table 7 MiniCPM-v2.6 的性能，其中包括 SigLIP 的视觉编码器和 Qwen2-7B 的大型语言模型。

Method	$\bar{A}_d$	$\bar{A}_s$	$\bar{A}$	A_{sd}
CLIP + Phi3-mini	92.70	87.46	90.02	69.60
SigLIP + Qwen2-7B	93.57	88.58	90.97	70.49

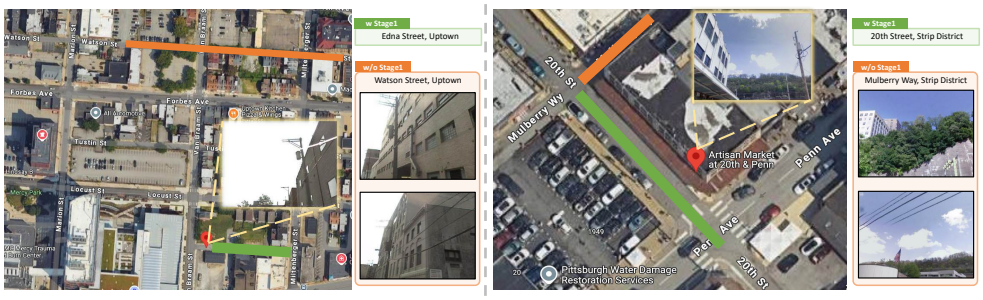


Fig. 8 关于是否使用第一阶段的跨视图对齐调整的影响的定性可视化比较。也描绘了错误预测街道周围的街景图像。

的评估，特别是在东京，详细内容见附录 **??** 。结果表明，即使地址偏差较大，我们的方法也能取得良好的性能。这证明了该方法的扩展性和适应性。

混合一般 VQA 数据的效果。数据混合对于同时保持 LVLM 的多种能力至关重要。为了评估混合一般 VQA 数据对我们方法性能的影响，我们选择了 llava_v1_5_mix665k VQA 数据集作为一般数据。我们分别以 5:1 和 1:1 的比例将其与阶段 2 数据混合。结果如 Table 6 所示。可以观察到，增加地址定位 VQA 数据的比例进一步提升了性能。我们相信，通过仔细调整数据比例，我们可以在提高模型的通用能力和其地址定位能力之间取得平衡。

不同骨干上的性能。为了验证我们方法在不同骨干上的泛化能力，我们采用 MiniCPM-v2.6（由 SigLIP 和 Qwen2-7B 构成）进行了相同的两阶段训练。在 Pitts-VQA 上的结果如 Table 7 所示。可以观察到，基于 SigLIP 和 Qwen2-7B 的 AddressVLM 性能稍高于基于 LLaVA-phi3 的性能，这表明采用更强大的 LVLM 可以取得更好的结果。

在东京 24/7 上的结果。为了验证我们的方法在美国以外城市的有效性，我们在东京 24/7 数据集上进行了评估。在与图像地理定位任务相关的东京数据集中，我们共收集了 52080 张街景图像（来自 4340 个位置，每个位置有 12 张图像），以及用于训练的 52764 组 VQA 对话数据和相应的卫星图像。地址被分为两个级别：丁目和街道。我们完成了两个阶段的微调，测试集的结果（来自 1240 个位置的 7440 张图像）如 Table 8 中所示。

Table 8 东京数据集的结果。

Method	A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_s^G	A_s^J	A_s^M	$\bar{A}_s$	$\bar{A}$	A_{sd}
AddressVLM	73.85	89.49	88.62	84.99	63.52	88.28	86.03	81.57	82.28	65.81
AddressVLM w/o stage1	70.63	86.22	85.93	83.14	58.06	85.39	84.72	78.37	80.76	56.37

如可以看出，即使地址具有高度偏差，我们的方法在东京数据集上依然表现良好，有效利用了东京的地址划分系统。这突出了该方法在现代城市中面对明确定义的地址相关 VQA 数据时的可扩展性和适应性。

关于选择 δ 的更多结果。 Table 9 展示了当 δ 设置为 0.3、0.5 和 0.7 时，Pitts-VQA 上的结果。可以观察到，当 $\delta = 0.5$ 的结果更佳，它优于 $\delta = 0.3$ 和 $\delta = 0.7$ 。这是因为当 δ 太小时，街景图像中没有足够的细粒度信息，而当 delta 太大时，它遮蔽了大部分地图区域，导致缺乏来自地图的信息。

Table 9 在 Pitts-VQA 上的结果具有更多不同的 δ 选择。

δ	A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_s^G	A_s^J	A_s^M	$\bar{A}_s$	$\bar{A}$	A_{sd}
0.3	87.32	92.93	94.59	92.05	71.61	91.04	93.27	86.78	89.32	68.98
0.5	88.73	93.54	95.16	92.70	72.51	91.70	93.98	87.46	90.02	69.60
0.7	86.58	92.33	93.86	91.22	70.35	90.49	92.65	85.81	88.49	68.07

3.3 定性结果

跨视角对齐调整的有效性。为了展示提出的跨视角对齐调整的有效性，我们在图 8 中展示了通过我们的模型进行对齐调整后正确定位的例子。我们展示了在没有第一阶段对齐调整的情况下预测错误的街景图片。可以观察到，这些错误预测的街景相邻街区的街景与真实街区的街景存在高度相似性。这个问题仅通过第二阶段的地址定位调整很难解决。相比之下，第一阶段的对齐调整补充了缺失的全局街道信息，并建立了街景图片之间的连接，从而帮助模型更好地确认街景图片的位置。

案例研究。我们展示了更多例子，其中 AddressVLM 准确定位，而没有跨视图对齐调优的基线模型在定位上出错，如图 9 所示。我们还提供了一些两种模型都无法正确定位的失败案例，如图 9 所示。可以看到，这些图像的视觉提示较少，即便是人类专家也难以识别。

与通用 LVLM 的比较。我们进一步展示了 AddressVLM 在实际推理中的实例，并提供了与 SOTA 通用 LVLM 的比较，例如 GPT-4o [5] 、Sonnet 3.5 [36] 、Qwen2-VL [6, 37] 和 LLaVA-Phi3-mini，见图 10 。AddressVLM 在各种 VQA 场景中始终提供高质量的结果。相反，SOTA 模型的性能受限于输入图像是否包含足够的可识别信息，如街道名称和地标。这表明，经过最少的微调，AddressVLM 可以使用仅 40 亿参数来实现对城市环境的细粒度理解。这确保了其在未来设备上的部署和更新的可行性。

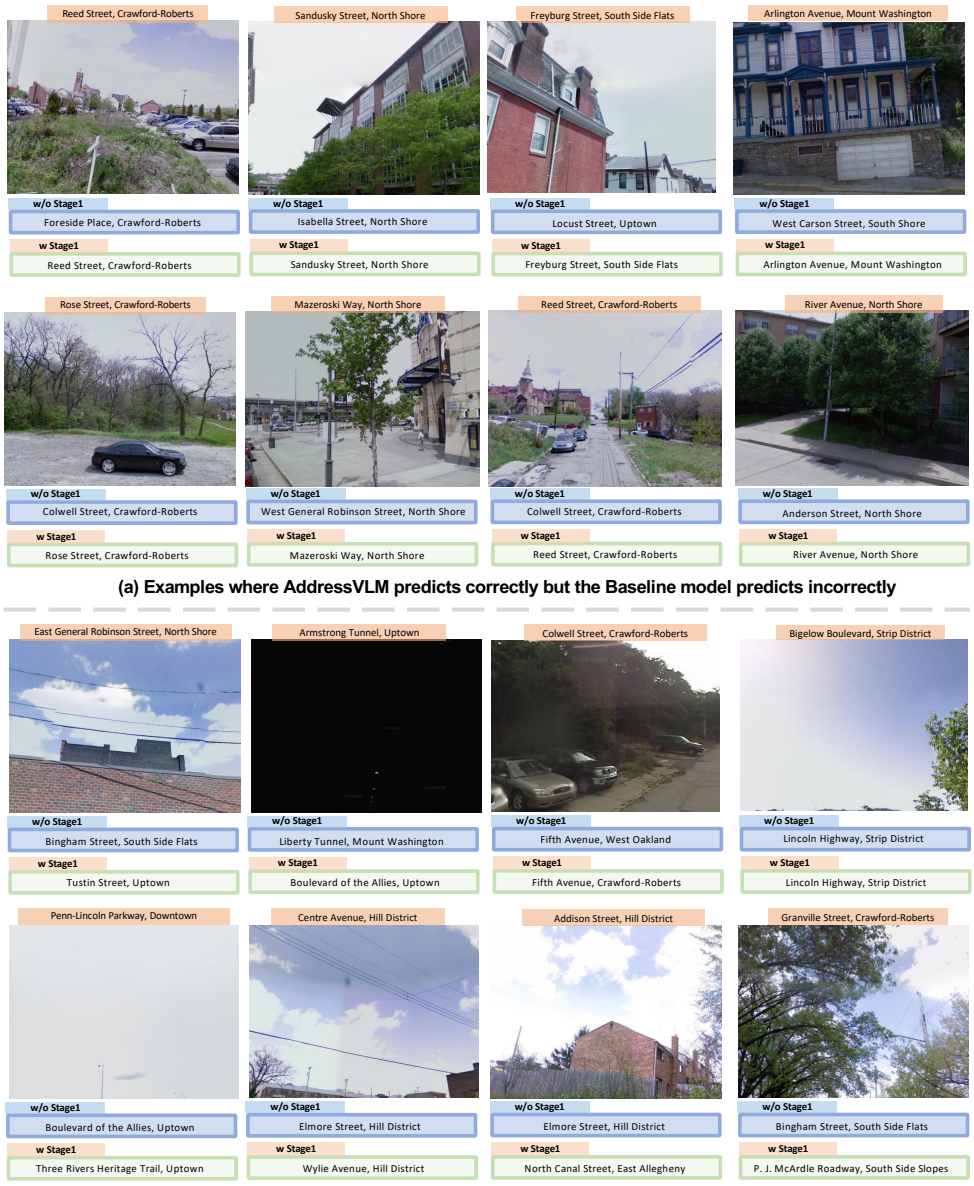


Fig. 9 更多例子显示了 AddressVLM 能够精确定位，而基准模型在定位上产生错误（a），以及失败案例（b）。

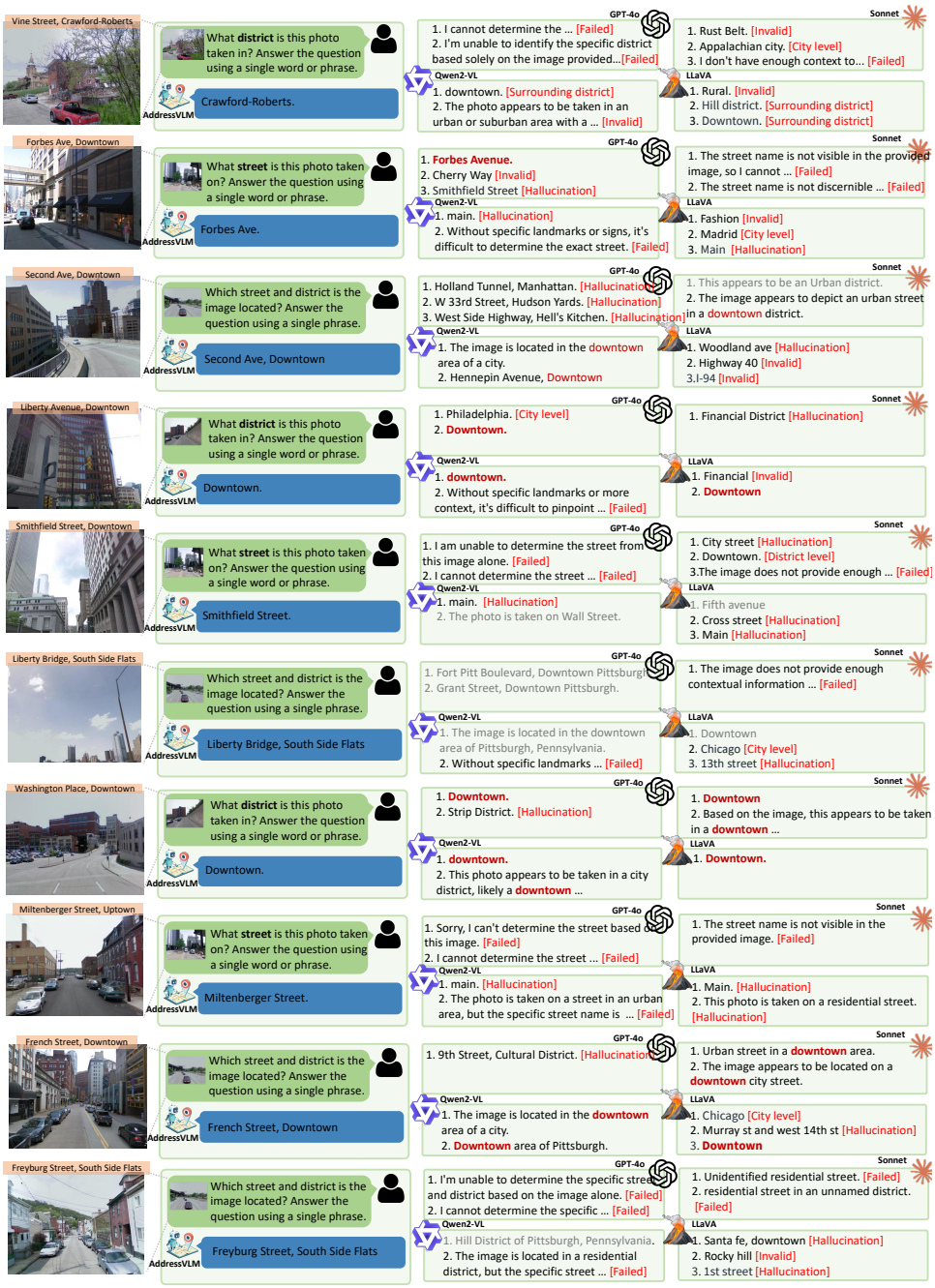


Fig. 10 与通用 LVLMs 在地址问答能力上的定性比较。

在这项工作中，我们提出了用于城市范围内地址定位的 AddressVLM，它能够对街景图像执行灵活的地址问答。其核心思想是利用卫星视图图像与街景图像之间的跨视图对齐调整，将对街道分布的全球理解整合到 LVLM 中。这包含两个关键组件，分别是卫星和街景图像嫁接机制，以及自动对齐标签生成机制。模型经历了两个阶段的微调，包括跨视图对齐调整和地址定位调整。大量实验表明，所提出的 AddressVLM 超过了一般的 LVLMs 和 SOTA 定位 LVLMs，并且可以扩展到多个城市。在未来的工作中，我们希望探索不同大陆的城市并采用更大的 LVLMs。
局限性。多亏了所提出的图像制作机制，AddressVLM 可以执行跨视图对齐调整。当前方法只是一个初步探索，街景图像的相对低分辨率可能会影响 LVLM 理解它们的能力。未来，更复杂的跨视图图像对齐方法值得研究以进一步提高性能。

Appendix A 实现细节

All our experiments are conducted using the xtuner framework on 8 RTX 3090 GPUs. The torch version is 2.4.0, the CUDA version is 12.1, and the transformers version is 4.37.2. The main hyperparameter settings are given in Tab. A1.

Table A1 两个调优阶段的超参数设置。

Hyper-parameter	Values
Batch Size	4×8
Gradient Accumulation	16
Learning Rate	1e-5
Weight Decay	0
Betas	(0.9, 0.999)
Warmup Ratio	0.03
LoRA Rank	128
LoRA Dropout	0.05
Model Max Length	$2048 - (336/14)^2$

Appendix B 数据集详情

We provide detailed information about the two constructed VQA datasets as a supplementary to Sec. 2.3, listed in Tab. B2. The dataset information includes the number of locations, the number of street view images, and the proportions of various dialogue types in the muti-turn conversations for both Pitts-VQA and SF-Base-VQA datasets. Generally, the distribution of address question types in the training set is balanced (1:1:1). In the test set, to accommodate both answer types (Yes/No) in judgment questions, we increased the judgment questions for each district-related and street-related question with answers set as "Yes" or "No", respectively. As a result, the proportion of judgment questions is nearly twice that of the generation and multiple-choice questions. Moreover, we provide comparisons between n the proposed datasets with existing related datasets in Table B3.

Table B2 有关构建的 Pitts-VQA 和 SF-VQA 数据集的更多详细信息。

Statistics	Pitts-VQA		SF-Base-VQA	
	Train	Test	Train	Test
Covered Area	20 km ^{2}	20 km ^{2}	6 km ^{2}	6 km ^{2}
Number of locations	7410	798	11946	1707
Number of Districts	19	19	15	15
Number of Streets	194	165	121	110
Number of images	177840	19152	143352	20484
Number of questions	533520	168409	430056	181943

Table B3 将所提议的数据集与现有相关数据集进行比较。

Statistics	Pitts-VQA	SF-Base-VQA	Pitts-IAL	SF-IAL-Base	Pitts-250K	SF-XL
Images	178K/19K	143K/21K	234K/19K	184K/21K	250K/24K	41.2M/1K
GPS	Yes	Yes	Yes	Yes	Yes	Yes
Address	Yes	Yes	Yes	Yes	No	No
QA	533K/168K	430K/182K	N/A	N/A	N/A	N/A

Additionally, the question templates for different types of questions and address is given in Tab. B4. Each address type includes 10 distinct templates, resulting in 20 templates in total. Subsequently, different question types are generated by appending different prompts for the three question categories, as shown in Tab. B5. We replace the contents in "[]" with the ground truth location names (e.g. street and district) before appending them to the address prompts.

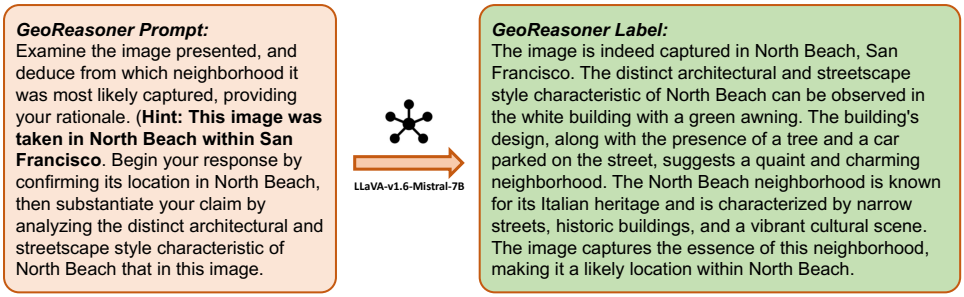


Fig. B1 GeoReasoner 第一阶段的提示和生成的推理标签示例。LLaVA-v1.6-Mistral-7B 模型被用于标签生成。

Appendix C GeoReasoner 的重现

The training process for GeoReasoner [9] consists of two stages. The first stage involves coarse-grained localization at the country level, accompanied by intricate reasoning derived from game data. The second stage is centered on fine-grained localization at

Table B4 问答数据生成的题目模板。

Address Type	Template
District	Tell me the district where this image was captured. I'm curious about the district, where is this? In which urban district was this photo taken? Can you identify which district this is? What district is shown in this photograph? What major district does the photo fall under? I'm looking for the name of the district in this photo, can you help? Can you specify the district shown in this photo? Which district is depicted in the photo? What's the name of the district shown in the photo?
Street	Identify the street in this image, please. What is the street seen in this picture called? On which boulevard or street was this taken? Give me the name of the street that appears in this photograph. Where was this, can you name the street? What's the name of the avenue or street captured in this shot? The street in this image, what is it named? What's the name of this street shown in the photo? Can you tell me which road this is? What thoroughfare is depicted here?

Table B5 附加提示以生成不同类型的问题。

Question Type	Template
Generation	Answer the question using a single word or phrase.
Judgement	Is this image taken [On STREET/IN DISTRICT], Yes or No?
Multiple Choice	Which of the following [STREET/DISTRICT] correctly represents the location shown in the image? (A) [OPTION A] (B) [OPTION B] (C) [OPTION C] (D) [OPTION D]. Please select the correct option (A/B/C/D).

the city level, utilizing Google Street View data. In our study, we replicate this pipeline to achieve district and street-level localization within the same urban area. A primary distinction between GeoReasoner and our AddressVLM lies in the data employed during the first stage. In the original work of GeoReasoner, the first stage data integrates external knowledge sourced from real geo-localization games. For district-level localization, we generate reasoning data by emulating the reasoning generation pipeline utilized for our cross-view tuning data. An example of the prompt and the generated reasoning label for the first stage of GeoReasoner is presented in Fig. B1. To facilitate a comprehensive comparison across various metrics outlined in Sec. 3.1, we employ the same VQA data for training the second stage of GeoReasoner.

Table C6 补充指标消融研究的详细结果。

	Ablations	District				Street				$\bar{A}$	A_{sd}
		A_d^G	A_d^J	A_d^M	$\bar{A}_d$	A_s^G	A_s^J	A_s^M	$\bar{A}_s$		
Pitts-VQA	Satellite w/o road (0.3)	85.93	93.00	93.70	91.33	67.24	91.18	92.55	85.51	88.36	64.05
	Satellite (0.3)	87.32	92.93	94.59	92.05	71.61	91.04	93.27	86.78	89.32	68.98
	Satellite w/o road (0.5)	86.23	92.42	93.54	91.09	67.97	90.50	91.79	85.17	88.06	64.63
	Variant A	85.57	90.73	92.03	89.77	65.60	89.42	90.39	83.90	86.58	63.21
	Variant B	85.48	90.65	92.12	89.59	65.05	89.21	90.02	83.44	86.42	62.95
	Variant C	84.86	91.98	92.85	90.34	66.39	90.63	91.46	84.75	87.48	63.03
	Variant D	87.36	93.23	95.08	92.66	71.19	91.58	93.85	87.02	89.53	66.37
	Variant E	85.00	92.02	92.65	90.34	66.64	90.27	91.05	84.54	87.37	63.52
	View-4/24	69.14	84.29	83.90	80.21	36.08	77.68	77.71	67.25	73.58	31.55
	View-7/24	76.54	89.38	88.33	85.73	46.75	85.90	86.14	76.13	80.83	42.25
	View-13/24	83.67	92.28	92.34	90.04	61.60	89.58	90.69	82.84	86.36	58.04
	Location-1/4	70.95	85.91	83.97	81.47	38.81	77.91	78.92	68.35	74.76	34.14
	Location-2/4	79.53	89.62	89.40	86.91	54.17	86.59	87.79	78.76	82.74	50.16
	Location-3/4	84.06	92.18	92.69	90.18	63.34	88.73	90.88	82.90	86.46	60.19
	AddressVLM	88.73	93.54	95.16	92.70	72.51	91.70	93.98	87.46	90.02	69.60
SF-Base-VQA	Satellite w/o road (0.3)	84.11	91.82	92.64	90.85	73.59	88.38	90.51	84.57	87.32	65.33
	Satellite (0.3)	85.88	93.10	93.92	91.52	75.27	88.04	92.18	85.79	88.67	70.42
	Satellite w/o road (0.5)	84.39	91.85	92.79	90.88	73.87	88.35	90.68	84.32	87.41	65.93
	Variant A	82.92	92.60	92.41	90.07	69.29	88.27	88.13	83.47	86.74	62.94
	Variant B	82.90	92.50	92.03	89.92	68.90	87.14	87.97	82.77	86.31	62.78
	Variant C	82.05	91.99	92.26	89.51	67.90	87.13	87.53	82.39	85.92	61.21
	Variant D	85.87	94.73	95.16	92.57	74.60	90.22	92.05	86.76	89.63	68.95
	Variant E	83.55	92.25	92.75	90.15	71.28	87.91	89.20	84.06	87.07	64.68
	View-2/12	70.47	88.39	87.11	83.47	43.01	80.31	76.87	70.08	76.71	35.65
	View-4/12	77.52	91.18	90.74	87.57	56.05	84.97	84.54	77.59	82.53	48.83
	View-8/12	84.23	92.11	93.94	90.34	71.20	87.12	90.23	84.74	87.63	64.72
	Location-1/4	75.14	89.77	88.84	85.78	49.39	79.06	80.04	71.95	78.80	41.83
	Location-2/4	81.52	92.69	92.27	89.72	64.42	86.62	88.02	81.49	85.56	57.81
	Location-3/4	85.17	93.85	93.92	91.64	72.70	87.80	91.55	84.94	88.26	66.53
	AddressVLM	86.48	93.72	94.50	92.06	76.09	88.92	92.75	86.66	89.33	70.45

Appendix D 定性结果的实现细节

D.1 第 2.2 节中的定性结果

In Fig. 5, we conduct a quantitative analysis of the cross-view alignment tuning by examining the outputs from two distinct models. While the first stage utilizes grafted images as inputs, our principal objective is to establish a connection between street-view images and the street addresses. Consequently, we employ only street-view images as the input for this analytical evaluation.

After Cross-view Alignment Tuning. For discriminative models like CLIP, we can compare the embeddings of street views and address texts to assess whether the model effectively associates street layouts with street views. However, this method is not suitable for the generative models discussed in this study. Instead, we leverage the inherent randomness in the output of generative models. Specifically, we increase the temperature of the model during inference from 0.1 to 0.8 to encourage output variability. By performing inference for 100 times on the same input image, we can

count the number of different valid streets, approximating the output distribution for the model for a given input.
Before Cross-view Alignment Tuning. Since the image address localization task is quite challenging, the model without any downstream fine-tuning (zero-shot model) struggles to produce valid street outputs directly. Therefore, we organize all the street names generated by the model above into options, allowing the zero-shot model to select one street from this given list for output. The difference between the prompts of these two models is given in Fig. 4.

D.2 第 3.3 节中的定性结果

In Sec. 3.3, we demonstrate the results of four current state-of-the-art proprietary and open-source models on several samples in our datasets. Our AddressVLM is capable of generating outputs as the requirement in the prompt. However, the outputs of other LVLMs are more diverse and uncontrollable. Therefore, for each sample, we conduct multiple inferences (5-10 times) for each input, and display several most frequently responses.

Appendix E 消融研究的详细结果

We provide the detailed results of the ablation studies under all the metrics in Tab C6.

References

[1] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR, pp. 5297–5307 (2016)

[2] Wang, R., Shen, Y., Zuo, W., Zhou, S., Zheng, N.: Transvpr: Transformer-based place recognition with multi-level attention aggregation. In: CVPR, pp. 13648–13657 (2022)

[3] Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Mixvpr: Feature mixing for visual place recognition. In: WACV, pp. 2998–3007 (2023)

[4] Xu, S., Zhang, C., Fan, L., Meng, G., Xiang, S., Ye, J.: Addressclip: Empowering vision-language models for city-wide image address localization. In: ECCV (2024)

[5] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[6] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. (2023)

[7] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2024)

[8] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 9 (1), 1 (2023)

[9] Li, L., Ye, Y., Jiang, B., Zeng, W.: Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In: ICML (2024)

[10] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)

[11] Durgam, A., Paheding, S., Dhiman, V., Devabhaktuni, V.: Cross-view geo-localization: a survey. arXiv preprint arXiv:2406.09722 (2024)

[12] Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. In: CVPR, pp. 883–890 (2013)

[13] Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: CVPR, pp. 4878–4888 (2022)

[14] Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. Pattern Recognition 113 , 107760 (2021)

[15] Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: CVPR, pp. 14141–14152 (2021)

[16] Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. IEEE Robotics and Automation Letters (2023)

[17] Seo, P.H., Weyand, T., Sim, J., Han, B.: Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In: ECCV, pp. 536–551 (2018)

[18] Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the world is this image? transformer-based geo-localization in the wild. In: ECCV, pp. 196–215 (2022)

[19] Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In: CVPR, pp. 23182–23190 (2023)

[20] Trivigno, G., Berton, G., Aragon, J., Caputo, B., Masone, C.: Divide & classify: Fine-grained classification for city-wide visual geo-localization. In: ICCV, pp. 11142–11152 (2023)

[21] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry,

G., Askell, A., Mishkin, P., Clark, J., et al. : Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)

[22] Haas, L., Alberti, S., Skreta, M.: Learning generalized zero-shot learners for open-domain image geolocalization. arXiv preprint arXiv:2302.00275 (2023)

[23] Cepeda, V.V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. arXiv preprint arXiv:2309.16020 (2023)

[24] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[25] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

[26] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

[27] Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)

[28] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

[29] Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: CVPR, pp. 3640–3649 (2021)

[30] Ghanem, A., Abdelhay, A., Salah, N.E., Nour Eldeen, A., Elhenawy, M., Masoud, M., Hassan, A.M., Hassan, A.A.: Leveraging cross-view geo-localization with ensemble learning and temporal awareness. PloS one 18 (3), 0283672 (2023)

[31] Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology 32 (2), 867–879 (2021)

[32] Mi, L., Xu, C., Castillo-Navarro, J., Montariol, S., Yang, W., Bosselut, A., Tuia, D.: Congeo: Robust cross-view geo-localization across ground view variations. arXiv preprint arXiv:2403.13965 (2024)

[33] Hao, X., Chen, W., Yan, Y., Zhong, S., Wang, K., Wen, Q., Liang, Y.: Urban-vlp: A multi-granularity vision-language pre-trained foundation model for urban

indicator prediction. arXiv preprint arXiv:2403.16831 (2024)

[34] Contributors, X.: XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner (2023)

[35] Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: CVPR, pp. 26689–26699 (2024)

[36] Claude: The claude 3 model family: Opus, sonnet, haiku. (2024). https://api.semanticscholar.org/CorpusID:268232499

[37] Qwen: Qwen2-vl (2024)