

# GeoProg3D: Compositional Visual Reasoning for City-Scale 3D Language Fields

Shunsuke Yasuki<sup>1</sup> Taiki Miyanishi<sup>2,3</sup> Nakamasa Inoue<sup>4</sup> Shuhei Kurita<sup>5,4</sup>  
 Koya Sakamoto<sup>2</sup> Daichi Azuma<sup>2</sup> Masato Taki<sup>1</sup> Yutaka Matsuo<sup>2</sup>

<sup>1</sup>Rikkyo University <sup>2</sup>The University of Tokyo <sup>3</sup>ATR <sup>4</sup>Institute of Science Tokyo  
<sup>5</sup>National Institute of Informatics

Project Page: <https://snskysk.github.io/GeoProg3D/>

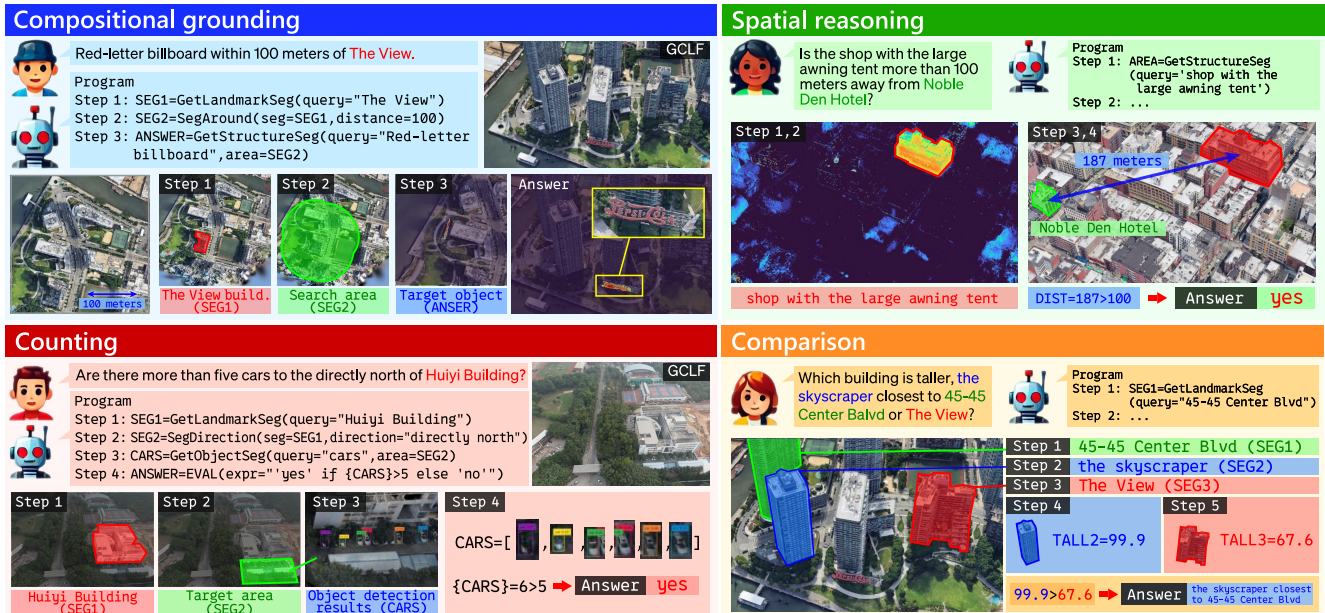


Figure 1. Overview of the proposed compositional geographic reasoning task. This task enables natural language interaction with city-scale 3D scenes, supporting diverse geographic reasoning scenarios. Given a natural language query, our GeoProg3D decomposes the task into modular steps, executes structured visual programs using geographic APIs, and integrates results for interpretable and accurate reasoning.

## Abstract

The advancement of 3D language fields has enabled intuitive interactions with 3D scenes via natural language. However, existing approaches are typically limited to small-scale environments, lacking the scalability and compositional reasoning capabilities necessary for large, complex urban settings. To overcome these limitations, we propose GeoProg3D, a visual programming framework that enables natural language-driven interactions with city-scale high-fidelity 3D scenes. GeoProg3D consists of two key components: (i) a Geography-aware City-scale 3D Language

Field (GCLF) that leverages a memory-efficient hierarchical 3D model to handle large-scale data, integrated with geographic information for efficiently filtering vast urban spaces using directional cues, distance measurements, elevation data, and landmark references; and (ii) Geographical Vision APIs (GV-APIs), specialized geographic vision tools such as area segmentation and object detection. Our framework employs large language models (LLMs) as reasoning engines to dynamically combine GV-APIs and operate GCLF, effectively supporting diverse geographic vision tasks. To assess performance in city-scale reasoning, we introduce GeoEval3D, a comprehensive benchmark

dataset containing 952 query-answer pairs across five challenging tasks: grounding, spatial reasoning, comparison, counting, and measurement. Experiments demonstrate that GeoProg3D significantly outperforms existing 3D language fields and vision-language models across multiple tasks. To our knowledge, GeoProg3D is the first framework enabling compositional geographic reasoning in high-fidelity city-scale 3D environments via natural language.

## 1. Introduction

Large-scale 3D scene reconstruction has emerged as a pivotal technology enabling a wide spectrum of real-world applications. These include the creation of 3D digital worlds [14, 35, 75], urban scene editing [9, 75, 76], and autonomous driving simulation [15, 80, 83]. In particular, recent advances in radiance fields such as Neural Radiance Fields (NeRF) [48] and 3D Gaussian Splatting (3D-GS) [24] have revolutionized the ability to reconstruct city-scale 3D models with unprecedented fidelity. These methods leverage diverse data sources, from street-level imagery from vehicles [13, 67, 93] to aerial and satellite photography [40, 42, 60, 61, 77], enabling the high-fidelity representations of entire cities. However, intuitive and efficient interaction with these detailed 3D city models using natural language remains largely unexplored.

Recently, a promising development in this direction has been the emergence of 3D language fields, which facilitate natural language interaction with high-resolution 3D scenes, enabling precise localization of specific objects or regions within these scenes [3, 23, 26, 53, 59]. However, when extending conventional 3D language fields for large-scale urban 3D scenes, two fundamental difficulties emerge: (1) *Scalability for city-scale 3D data*: Since existing methods primarily focus on indoor scenes, the high-fidelity reconstruction and efficient handling of large-scale urban scenes exceeding  $1\text{km}^2$  [37] remain challenging. (2) *Enhanced task versatility for urban applications*: Current 3D language fields predominantly focus on localizing discrete objects through language queries, yet they fail to adequately address the multifaceted demands intrinsic to urban applications such as interpreting spatial relationships, quantifying object numbers and sizes, or recognizing landmark identifiers. Therefore, extending 3D language fields to urban-scale environments necessitates a more scalable and compositional framework capable of comprehending the intricate complexities of vast citywide landscapes.

In this work, we address these difficulties by proposing GeoProg3D, a visual programming framework which consists of two key components: (i) a Geography-aware City-scale 3D Language Field (GCLF), and (ii) Geographical Vision APIs (GV-APIs) combined with LLMs for code generation. To address the scalability issue (1), GCLF utilizes memory-efficient hierarchical 3D Gaussians to repre-

sent 3D language fields. The structure is useful for high-fidelity reconstruction and fast inference of vast urban environments. It also enables localization of objects based on landmark names by aligning the linguistic and geographic information of real-world 2D maps with 3D Gaussians. In response to issue (2) and to enhance versatility, GV-APIs provide a set of image and geographical processing APIs, including object detection, area segmentation, and distance measurement on GCLF. Figure 1 provides an overview of the response process of GeoProg3D across four distinct scenarios. By leveraging large language models (LLMs) as reasoning engines to dynamically integrate GV-APIs, GeoProg3D effectively decomposes complex queries into simpler subtasks that can be processed on GCLF, facilitating compositional reasoning over city-scale 3D data.

To validate the effectiveness, we introduce novel tasks designed to assess urban-scale geographic visual reasoning capabilities and present GeoEval3D, a benchmark dataset specifically developed for this task. GeoEval3D encompasses 952 carefully designed query-answer pairs across five essential within realistic urban environments derived from city-scale 3D reconstructions [37, 75]. The dataset covers an area exceeding  $3\text{km}^2$  across diverse urban settings in New York (U.S.) and Shenzhen (China).

In summary, our contributions are threefold:

- We propose GeoProg3D, a framework for compositional reasoning over city-scale 3D language fields, where visual programming can perform various 3D geographic vision tasks via image and geographic APIs.
- We introduce five geographic compositional reasoning tasks in city-scale 3D scenes through natural language queries, and present a new benchmark dataset, GeoEval3D, that encompasses these challenges.
- We demonstrate the superior performance of GeoProg3D on GeoEval3D in comparison to existing 3D language fields and state-of-the-art VLMs by a significant margin.

## 2. Related Work

**City-scale 3D scene reconstruction.** 3D reconstruction from large-scale image collections has garnered considerable attention and achieved significant advancements in recent years. NeRF have been widely employed to produce high-fidelity 3D representations from 2D image inputs by modeling volumetric scenes [34, 67, 70, 74, 77, 89, 90]. However, the unique nature of volume rendering in NeRF leads to substantial resource requirements, particularly for high-resolution outputs. More recently, city-scale 3D scene reconstruction using 3D-GS has emerged as a highly efficient alternative to NeRF-based approaches, offering improvements in rendering speed, scalability, and adaptability to real-time applications [18, 25, 36, 40, 41, 60, 61, 82, 85]. Despite these advancements in precise 3D reconstruction, the development of technologies to ground geolocation text

within urban scenes and retrieve geolocation information remains underexplored. We propose GeoProg3D, a novel method designed to ground textual queries and instructions onto precise city-scale scenes for comprehensive analyses.

**3D language field.** Following the rapid advancement of 3D representations, grounding 3D data to text has become an intensive research topic in recent years. Classical approaches to text grounding in point clouds include ScanRefer [7], which uses textual descriptions for 3D localization to generate 3D bounding boxes, and ScanQA [1], which addresses 3D question answering using point clouds from the ScanNet dataset [11, 81]. Neural semantic fields [5, 58], NeRF decomposition [31, 56, 62] and 3D LLMs [19, 69, 78, 95] have been proposed to enhance spatial information representation. LERF [26] embeds CLIP features into a 3D scene representation constructed with NeRF, enabling 3D spatial search using open vocabulary queries. However, as NeRF-based approaches have limitations in both speed and accuracy, 3D-GS based language embed models of LangSplat [53] and LEGaussians [59] and their applications emerge [8, 22, 86]. Specifically, LangSplat avoids the rendering cost problem by adopting 3D-GS for scene representation, and constructs 3D language fields with compressed CLIP features into 3D Gaussians. CLIP features are embedded as features with 3D clear boundaries, guided by the 2D object masks obtained by applying SAM [29] to multiple training views. These 3D language fields are primarily designed for finite-scale indoor scenes. To the best of our knowledge, there are no city-scale 3D language fields.

**LLMs for visual reasoning tasks.** Recent advancements in LLMs and Vision-Language Models (VLMs) have enabled the composition of program codes for extracting relevant features from texts and images, summarizing them, and executing symbolic reasoning tasks [12, 16, 43, 57, 63, 66]. Visual Programming [16] facilitates the resolution of compositional vision tasks by receiving natural language queries generating executable Python programs that invoke external functions. ViperGPT [66] enables Python code generation and execution for visual reasoning and question answering. CodeVQA [63] also introduces a code generation approach tailored for positional question answering from images. In the domain of 3D visual grounding, systems have been developed for zero-shot open-vocabulary 3D visual grounding, primarily within narrow indoor scenes [63], intending for narrow indoor scenes, and the versatility of visual programming is only utilized for the grounding of 3D point cloud space. In contrast, our work introduces GeoProg3D, a visual programming system designed to handle complex queries within city-scale 3D spaces.

**Geography-aware vision and language.** VLMs also have significantly propelled research in geography-aware vision and language modelings, enabling to understand and rea-

son about visual data within a geographical context [94]. A diverse array of geographical tasks has been proposed, encompassing geo-localization [17, 30, 38, 71, 79], visual grounding [65, 84], image captioning [44, 46, 54, 88], question answering [6, 92], and text-to-image generation [27, 68]. Recently, Vision-Language Grounding Foundation Models (VLGFMs) have been proposed as a framework for solving these multiple challenges simultaneously by fine-tuning VLMs [20, 21, 33, 45, 50, 52, 72, 87, 91], enhancing their capacity to understand and analyze complex geospatial information. However, VLGFMs are limited to handling only 2D top-down view images and are not capable of processing city-scale 3D data.

More closely related to our work, recent studies have introduced 3D vision-language tasks that focus on 3D point cloud data of urban environments, moving beyond the reliance on top-down 2D city images. Notable examples include 3D visual grounding [32, 49, 73] and 3D question answering [64]. Early work explores the use of VLMs for predicting urban attributes from 3D city models [4]. However, these frameworks are typically limited to single 3D vision tasks or have not been extended to high-resolution, city-scale 3D reconstruction models. In contrast, our approach is designed for high-fidelity, city-scale 3D data and supports a diverse array of tasks, including compositional grounding, counting, and spatial reasoning, all of which require compositional reasoning. This versatility enables comprehensive interaction with complex urban environments through natural language.

### 3. GeoProg3D Framework

This section introduces GeoProg3D, a framework that enables users to interact with large-scale 3D scenes using natural language queries, consisting of two important components: (i) Geography-aware City-scale 3D Language Field (GCLF), and (ii) Geographical Vision APIs (GV-APIs), specialized modules for geographic visual reasoning tasks. GCLF extends 3D language fields to city-scale and localizes objects, regions, and landmarks using natural language queries. GeoProg3D utilizes visual programming via LLMs to dynamically combine GV-APIs for operating GCLF, and is adaptable to a variety of 3D geographic vision tasks.

#### 3.1. Overview of GeoProg3D

We present an overview of GeoProg3D framework in Figure 2, which consists of two steps: the generation and execution of visual programs. In the first step, a program  $z \in \mathcal{Z}$  that answers the query  $q$  is generated as  $z = \Pi(q, R)$ , where  $\Pi$  is an LLM,  $R$  is in-context examples using GV-APIs, and  $\mathcal{Z}$  is the set of executable Python programs (GV-APIs). In the second step, the generated program  $z$  is executed to obtain the answer  $a$  as  $a = \Lambda(z; \mathcal{T})$ , where  $\Lambda$  is the Python execution engine and  $\mathcal{T}$  is GCLF. The key dif-

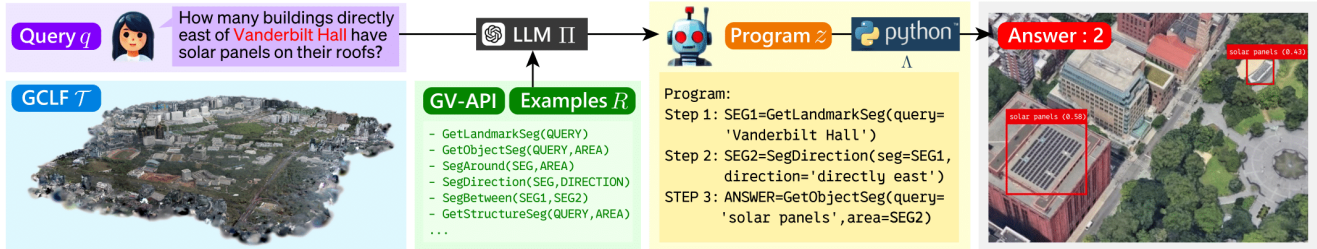


Figure 2. **Framework overview.** Given a user query, GeoProg3D generates a visual program via LLM in-context learning. The program operates GCLF by combining Geographical Vision APIs (GV-APIs) and answers the query.

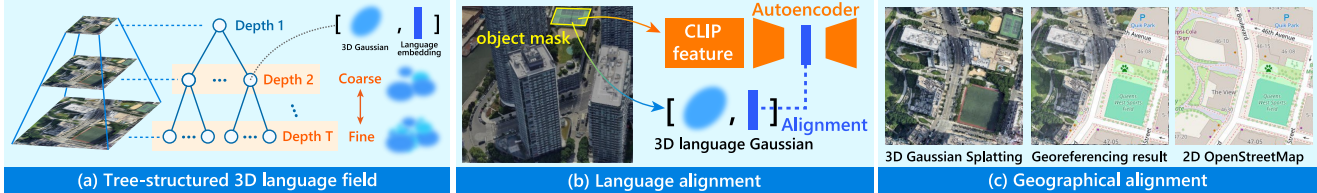


Figure 3. **GCLF structure.** (a) Coarse-to-fine tree structure to represent 3D scenes. Each node represents a pair of a 3D Gaussian and a language embedding. (b) Language alignment using CLIP features. (c) Geographical alignment using OpenStreetMap.

ferences between GeoProg3D and existing visual programming methods [16, 66] are that GeoProg3D can handle 3D scene queries rather than 2D images, and this enables operations and control over 3D language fields. Below, we describe the details of GCLF, GV-APIs, visual programming in Secs. 3.2 to 3.4.

### 3.2. Geography-aware City-scale 3D Language Fields

**Scene representation.** As a first step toward constructing GeoProg3D, we design a city-scale 3D language field (GCLF) with two key features: high fidelity and fast inference. City-scale reconstructions often produce rough details, which can negatively impact localization and image processing performance. Additionally, fast rendering is essential for practical use of large-scale 3D city models. To meet these requirements, GCLF represents urban scenes by embedding language into tree-structured 3D Gaussians. As described in [60], this tree structure learns the nested relationships between Gaussians for detailed representation and those for an overview across multiple layers. Figure 3a shows this structure. The rendering algorithm dynamically selects a hierarchical level in which the 3D Gaussian has a diameter of less than one pixel in image space, and efficiently renders the space that is far from the viewpoint.

**Why GCLF?** To demonstrate the advantages of GCLF over conventional 3D Gaussian Splatting (3D-GS), we compare its efficiency and reconstruction quality. As shown in Table 1, GCLF requires tens of times more 3D Gaussians than LangSplat (vanilla 3D-GS) due to its hierarchical reconstruction approach, yet the rendering speed increases only by a few times, ensuring high-speed rendering remains feasible [60]. Furthermore, as illustrated in Figure 4, the GCLF tree structure reconstructs city scenes with higher fidelity than both point clouds and vanilla 3D-GS, enabling successful inferences such as object detection on rendered images.

Scene	# Gaussians		Rendering speed (ms)	
	LangSplat	GCLF	LangSplat	GCLF
Center Blvd	37,212	1,136,015	2.73	14.46
World Fin Ctr	30,278	763,432	2.21	7.99
Mott St	33,846	1,253,668	2.97	14.48
Washington Sq	31,757	950,932	2.34	12.35
UrbanScene3D	OOM	37,813,418	OOM	20.83

Table 1. Number of Gaussians and inference speed.

**Geo-visual integration.** To effectively integrate visual and geographic data, we train our language embeddings to align with CLIP image features [55] (see Figure 3b), thereby extending the LangSplat approach [53] into a tree-structured 3D language field. Additionally, we georeference the Gaussian coordinates to real-world coordinates (see Figure 3c), which enables our system to generate precise responses incorporating geographic details such as landmark names and measurements in real-world units.

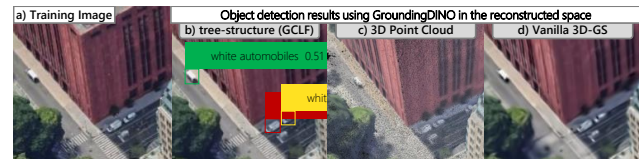


Figure 4. The results of object detection using GroundingDINO.

**Training.** Given a set of multi-view images  $\mathcal{D} = \{x_i\}_{i=1}^N$  for training, a GCLF is trained in the following three steps. First, coarse vanilla 3D Gaussian primitives  $G = \{g_j\}_{j=1}^K$  are trained on  $\mathcal{D}$ , where  $g_j$  is a vector representing the position, colors, scale and rotation of the  $j$ -th Gaussian and  $K$  is the number of Gaussians. Next, following the learning method in [60], the tree-structure  $G'$  is trained based on  $G$ . Third, language embedding into 3D Gaussian is trained. Unlike LangSplat [53], which is based on vanilla 3D-GS, the embedding is trained on the tree structure  $G'$  using our unique implementation. As shown in Figure 3b, the lan-

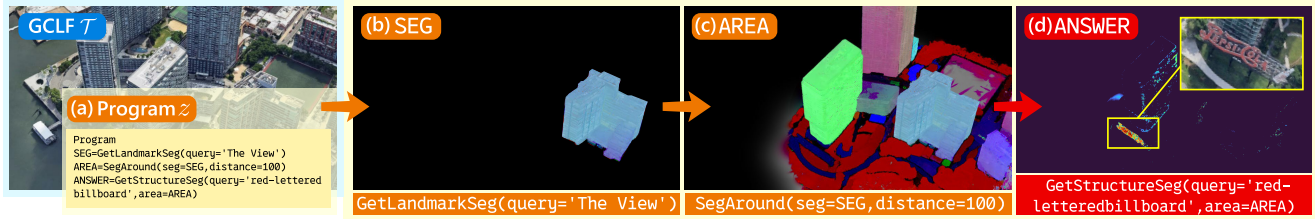


Figure 5. **Execution example.** (a) Program code generated from the query “Red-letter billboard within 100 meters of *The View*.” that consists of three steps. (b) `GetLandmarkSeg` identifies the building *The View*. (c) `SegAround` retrieves the area around *The View* within a 100-meter radius. (d) `GetStructureSeg` produces the segment where the confidence map from GCLF is visualized.

guage embedding teacher data  $c^{(m)}$  is obtained by applying a trained autoencoder for each scene to CLIP features, in the same way as LangSplat[53]. Here,  $c$  is the CLIP image encoder and  $m$  is the index of the object mask obtained from the training image using the Segment Anything model [28].

**Georeferencing.** After training GCLF  $\mathcal{T}$ , georeferencing is performed to align the 3D scene with a 2D map in a semi-automatic manner. First, top-down view images of four small areas are rendered, each with an image size of  $1024 \times 1024$  pixels. Second, more than 20 landmark points are manually chosen as points of interest that are visible in OpenStreetMap. Finally, the geometric transformation is computed between the real-world coordinates on the map and the Gaussian coordinates using the transformation function from the scikit-image library. An example of georeferencing is shown in Figure 3c.

**Inference.** Given a query  $q$ , GCLF localizes the target object by computing the cosine similarity between the CLIP text feature  $T(q)$  and the decoded language embedding  $D(\hat{l}(v))$  at each pixel  $v$  of a rendered 2D image. Here,  $T$  is the CLIP text encoder,  $D$  is the decoder of the autoencoder, and  $\hat{l}(v)$  is the aggregated language embedding at pixel  $v$ . To ensure efficiency, GCLF leverages its hierarchy to render each pixel using only Gaussians projected smaller than one pixel. This avoids exhaustive similarity computations across the entire scene and enables real-time interaction.

### 3.3. Geographical Vision APIs

The functions of the 3D language fields, including GCLF, are primarily limited to the localization for objects corresponding to word-level natural language queries. To leverage GCLF for a variety of tasks, GV-APIs provide a suite of visual and geographic processing functions. Table 2 shows how each function is called and its role. APIs 1) to 6) retrieve the relevant 3D Gaussians within GCLF, effectively narrowing down the region of interest in the vast city space (as shown in Figure 7 GRD SEG2). Specifically, 2) selects 3D Gaussians that closely match the query based on cosine similarity (e.g., Figure 7 GRD Answer). APIs 7) and 8) compute distances within GCLF, while 9) applies GroundingDINO on RGB renderings from GCLF. All these APIs work on trained GCLF without requiring additional training. Its technical contribution is to allow operations over the trained 3D Gaussian space. For example, 8) estimates land-

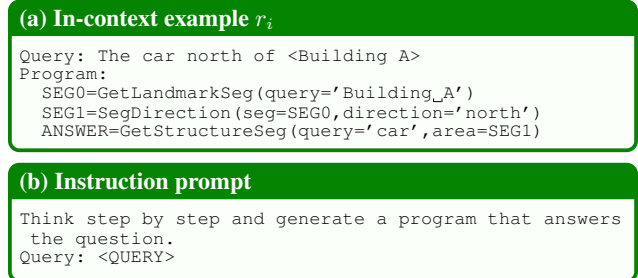


Figure 6. ICE and instruction prompt.

form height by identifying horizontal planes from Gaussian variance directions, while 6) applies clustering to filter out noisy activations and improve segmentation accuracy.

### 3.4. Visual programming

To enable compositional reasoning through dynamic combinations of GV-APIs, we utilize visual programming. Following previous studies on visual programming [16, 66], an LLM  $\Pi$  is utilized to generate a Python program  $z$  given an input query  $q$  with in-context examples (ICEs)  $R$ . Specifically, the GPT-3.5 model (gpt-3.5-turbo-instruct) is used as  $\Pi$ . To instruct  $\Pi$  on how to use GV-APIs  $\mathcal{Z}$ , we provide ten  $R$ . Each example consists of a short query paired with an example program. Figure 6 shows an example of  $R$  and instruction prompts for the grounding task (task details are provided in Section 4.1). The generated program  $z$  is executed by the Python execution engine  $\Lambda$  with pre-trained GCLF  $\mathcal{T}$  to obtain the answer  $a$ .  $z$  is executed within a try block and returns None if an error is encountered during execution. Figure 5 shows an example of the entire process of localizing a specific object (“Red-letter billboard”) by combining GV-APIs. Three functions contribute to the compositional reasoning procedure.

Our framework requires only a small number of ICEs for the LLM to effectively utilize the GV-APIs. Empirically, providing 10-15 examples enables the LLM to generate programs for diverse queries with a high success rate of over 90%. Crucially, the LLM does not merely imitate the provided examples. It combines its understanding of the API definitions with its pre-trained knowledge (e.g., concepts of counting and comparison) to generate programs with novel structures not seen in the examples, demonstrating strong structural-level generalization. This allows the framework to respond robustly even to variations in query phrasing (see

Function	Roles
1) GetLandmarkSeg(query=QUERY)	Get segment by landmark name (e.g., “Trinity Church”).
2) GetStructureSeg(query=QUERY, area=SEG)	Get segment by structure name (e.g., “bridge” and “tower”).
3) SegAround(area=SEG, distance=DIST)	Get segment around the input segment according to DIST.
4) SegDirection(area=SEG, direction=DIR)	Get segment with the specified direction for the input segment according to DIR.
5) SegBetween(seg1=SEG1, seg2=SEG2)	Get segment located between two input segments.
6) LargestSeg(segs=SEG)	Get the largest contiguous segment of the input segments using clustering.
7) MeasureDist(from=SEG1, to=SEG2)	Get real-world distance (meters) between two input segments.
8) MeasureHeight(area=SEG)	Get real-world height (meters) of input segment.
9) GetObjectSeg(query=QUERY, area=SEG)	Execute GroundingDINO object detection by object name (e.g., “car”).

Table 2. How to call the nine functions of the GV-APIs and their roles.

Task	Query examples
GRD	U-shaped building to the west of <i>Liberty Luxe</i> Red canopy shop within 150 meters of <i>Chase Bank</i> . Church with blue copper domes
CNT	How many sports fields are there? How many ships are there northwest of <i>200 Vesey Street</i> ? How many cars are there between <i>Little Stadium</i> and <i>Huiyi building</i> ?
MES	How tall is the skyscraper near <i>Quik Park</i> . How tall is the building with a gray curved roof? How far is the fountain that is closest to <i>Washington Square Arch</i> from <i>Amity Hall</i> ?
CMP	Which is taller, the cubic building or yellow building? Which is taller, <i>The View</i> or <i>Quik Park</i> ? Which is taller, the tallest object around <i>Quik Park</i> or <i>The Avalon</i> ?
SPR	There is a light-blue pointed roof. There is one or more tennis courts around <i>Amity Hall</i> . The skyscraper that is closest to <i>Quik Park</i> is closer to <i>The View</i> than <i>Quik Park</i> .

Table 3. Query examples for each task.

Appendix A.1 for a detailed analysis).

## 4. GeoEval3D Dataset

In this section, we present five tasks for evaluating understanding of city-scale 3D scenes, and introduce GeoEval3D, a dataset covering these tasks. The dataset  $\mathcal{B} = \{(\mathcal{D}_i, \mathcal{Q}_i)\}_{i=1}^S$  consists of pairs multi-view image sets  $\mathcal{D}_i$  and task sets  $\mathcal{Q}_i$ , where  $S$  is the number of outdoor scenes.

### 4.1. Task Definition

The task set  $\mathcal{Q}_i = \{(q_k, a_k)\}_{k=1}^{K_i}$  consists of pairs of queries  $q_k$  and the corresponding ground truth answers  $a_k$ . Each pair represents one of the five tasks listed in Table 3 with examples. The task definitions are summarized below. Further details are provided in Appendix B.

**1) Grounding (GRD).** Given a query  $q_k$  describing a specific target object, this task requires models to identify and localize the object. Following [26, 53], the ground truth  $a_k$  is the segment of the target object.

**2) Counting (CNT).** Given a query  $q_k$  that involves counting objects in a scene, this task requires models to accurately count the number of specified objects. The ground truth  $a_k$  is provided as an integer.

**3) Measuring (MES).** Given a query  $q_k$  describing a question, this task requires models to accurately measure the height (MES-H) and distance (MES-D) of buildings. The ground truth  $a_k$  is provided as an integer.

**4) Comparison (CMP).** Given a query  $q_k$  that involves comparing the sizes of buildings, this task requires models to return a text  $a_k$  identifying the correct building.

**5) Spatial reasoning (SPR).** Given a query  $q_k$  describing object details or spatial relationships, this task requires models to return “yes” if it is correct, and “no” otherwise. The ground truth is provided as  $a_k \in \{\text{yes, no}\}$ .

### 4.2. Images

GeoEval3D comprises urban scenes sourced from two datasets: GoogleEarth [75] and the UrbanScene3D [37]. Each scene has an image set  $\mathcal{D}_i = \{x_j\}_{j=1}^N$  consists of multi-view images  $x_j$  for training 3D scene representation. From the GoogleEarth dataset, we chose four scenes, including scenes of New York (U.S.) collected from Google Earth Studio. Each scene comprises 60 images captured in an orbital pattern around a central subject. The orbit radius ranges from 125 meters to 813 meters, and the altitude varies from 112 meters to 884 meters. The images have a resolution of approximately  $958 \times 538$  pixels, providing slightly coarse visuals suitable for large-scale urban modeling. From the UrbanScene3D dataset, we chose one scene, a large real city in China (Shenzhen), covering a total area of  $2 \text{ km}^2$ . This scene includes multiple orbital images, offering high-definition visuals at approximately 4K resolution. To enable the evaluation of various geographic vision tasks, we provide annotations including object masks and distance measurements as ground truth for text queries.

### 4.3. Dataset construction and statistics.

**Annotation.** All query-answer pairs were manually created by five annotators who were instructed to create high-quality ground truth. To ensure the reliability of the dataset, annotation was performed using common tools. Specifically, GT masks were created by an annotation tool of LabelMe. Distance and height GTs were prepared by GoogleEarth, counting and yes/no correct labels were manually annotated through visual inspection. Each query contains up to three landmark names.

**Statistics.** GeoEval3D is composed of unique 952 queries. It is scaled up more than 10 times compared to the datasets used in previous works, and contains many more words [26, 53] (Appendix Figure 12). The SPR task requires particularly complex compositional reasoning, thus accounting for

Method	GoogleEarth					UrbanScene3D		
	SPR Acc.↑	CMP Acc.↑	CNT MAE↓	MES-H MAE (m)↓	MES-D MAE (m)↓	SPR Acc.↑	CNT MAE ↓	MES-D MAE (m)↓
GPT-4o Vision [51]	24.77	2.63	3.02	158.16	195.29	15.18	4.29	1583.48
LLaVA-1.5 [39]	50.95	36.96	3.08	607.37	433.15	46.04	4.23	837.11
Llama-3.2 Vision [47]	54.84	28.49	2.54	88.06	133.20	57.34	3.54	427.94
Qwen2.5-VL-7B [2]	53.39	26.95	2.65	59.68	175.44	47.24	4.00	412.86
InternVL2.5-8B [10]	54.27	26.95	2.79	51.30	157.14	52.47	4.23	318.71
GeoChat [33]	57.23	41.99	2.89	84.74	89.34	56.76	3.69	328.68
LHRS-BOT [50]	49.45	27.52	4.85	46.17	104.49	41.94	3.49	438.94
VHM [52]	54.55	39.82	5.28	52.58	135.50	56.92	4.34	354.91
TEOChat [21]	59.04	48.11	2.84	150.39	198.89	57.99	4.06	359.71
<b>GeoProg3D</b>	<b>64.00</b>	<b>59.73</b>	<b>2.00</b>	<b>45.24</b>	<b>49.28</b>	<b>60.87</b>	<b>2.51</b>	<b>139.51</b>

Table 4. Performance of spatial reasoning (SPR), comparison (CMP), counting (CNT), and measurement (MES) tasks.

Test scene	Area ( $m^2$ )	LSeg	LERF	LangSplat	GCLF	GeoProg3D
GoogleEarth	$2.4 \times 10^5$	0.96	11.44	14.15	20.09	<b>45.20</b>
UrbanScene3D	$5.0 \times 10^6$	4.65	OOM	OOM	6.98	<b>30.23</b>

Table 5. Localization accuracy (%) on the GRD task.

Test scene	Area ( $m^2$ )	LSeg	LERF	LangSplat	GCLF	GeoProg3D
GoogleEarth	$2.4 \times 10^5$	1.08	6.38	5.19	6.69	<b>18.15</b>
UrbanScene3D	$5.0 \times 10^6$	1.06	OOM	OOM	3.78	<b>8.74</b>

Table 6. 3D semantic segmentation performance on the GRD task. Average IoU scores (%) are reported.

about half of the total queries.

## 5. Experiments

### 5.1. Evaluation metrics

For the GRD task, we report localization accuracy for object localization and Intersection over Union (IoU) for segmentation following the previous study [53]. Localization accuracy is measured at an IoU threshold of 0.15. For the CNT and MES tasks, we calculate the Mean Absolute Error (MAE) between the predicted and true values. For the SPR and CMP tasks, exact match criteria are applied to determine correctness to compute accuracy. We perform experiments on five scenes across the two datasets: four scenes from GoogleEarth and one scene from UrbanScene3D. Note that MES-H and CMP are not evaluated in UrbanScene3D because Ground Truth for height cannot be obtained.

### 5.2. Experimental results

**Localization performance.** In this experiment, we evaluate the localization performance of the 3D language field alone (GCLF) and when using GV-APIs through visual programming (GeoProg3D) in the GRD task. We compare our methods with baselines, including LangSplat [53], which is the SOTA method for high-resolution 3D scene localization. The localization accuracies are shown in Table 5. We observed that GCLF outperforms baselines on GoogleEarth. This suggests not only that language embedding into the tree structure works correctly, but also that high-fidelity re-

construction with the tree structure may help in more accurate localization at the pixel level. In addition, LangSplat caused a memory error with UrbanScene3D in our setting, which implies the efficiency of the tree structure for learning larger scenes [40]. GeoProg3D further improved accuracy on both GoogleEarth and UrbanScene3D. In terms of 3D semantic segmentation, we observed a similar trend in Table 6. It is worth noting that GeoProg3D exhibited superior performance on UrbanScene3D, which spans more than  $2km^2$ . These results demonstrate the limitations of localization using 3D language fields alone in 3D urban scenes and the effectiveness of GV-APIs and visual programming in improving compositional reasoning.

**Various geographic vision tasks performance.** While GCLF is limited to localization, GeoProg3D supports a wide variety of tasks through visual programming. This experiment evaluates the versatility of GeoProg3D in CNT, MES, SPR, and CMP tasks. As baselines, we selected five VLMs: GPT-4o Vision [51], LLaVA-1.5 [39] and Llama 3.2 Vision [47], Qwen2.5-VL-7B [2], InternVL2.5-8B [10], as well as four VLGFM: GeoChat [33], LHRS-BOT [50], VHM [52], and TEOChat [21]. Since VLMs can only process 2D images, we fed top-down view images instead of 3D data. Table 4 summarizes the results. First, in the CNT and MES tasks, which assess the model’s ability to accurately count and measure, our method achieved the lowest MAE values. Notably, our method excelled in the MES-D (distance measurement) task, demonstrating precise horizontal spatial assessment capabilities. These results underscore the superior performance of GeoProg3D in estimating quantities within large-scale 3D scenes and highlight the effectiveness of the program-based inference procedures. In addition, in the SPR and CMP tasks, which require more complex compositional reasoning, our method outperformed all other methods. These results indicate not only its proficiency in understanding spatial relationship but also its effectiveness in comparing and contrasting them. Furthermore, GeoProg3D showed examples of successful inference with discrimination of structures that are difficult to see in the top-down view, such as the sides of buildings and billboards. This shows that 3D language fields enable

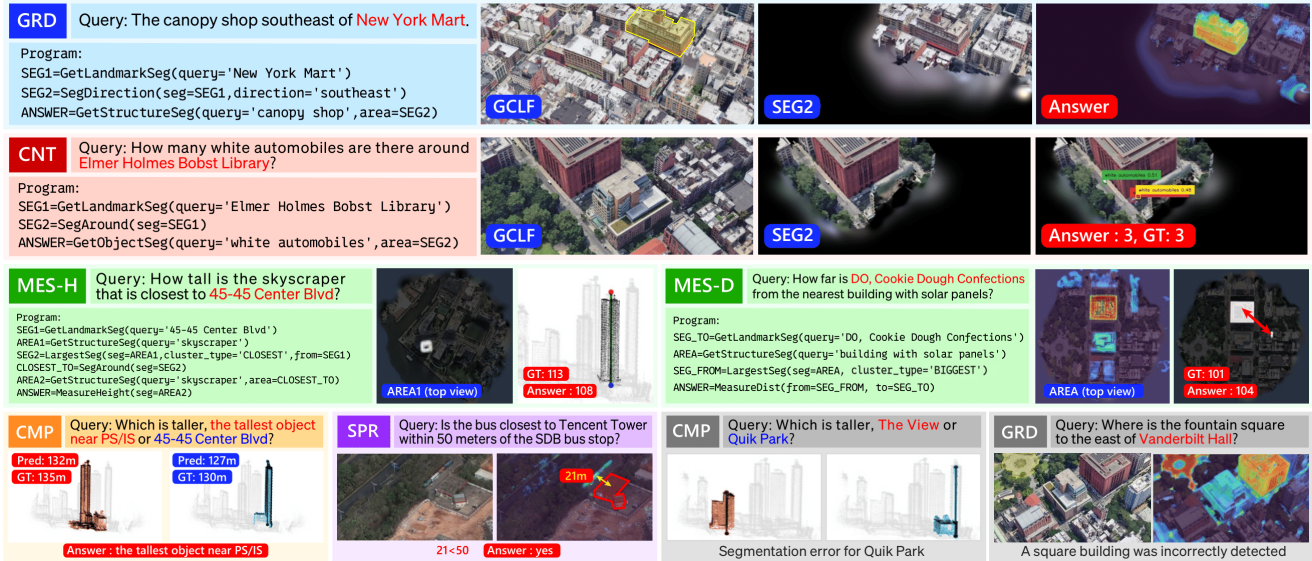


Figure 7. Qualitative results and failure cases. The Ground Truth region for the GRD task is delineated by the yellow frame.

localization that is independent of the viewpoint, taking into account the characteristics of structures which cannot be seen from directly above. See appendix C for details. Note that these comparison VLN-based baselines do not support pixel-level inference, so the GRD task is not evaluated.

**Ablation study.** To assess the impact of each component of GeoProg3D, we conducted an ablation study to investigate the three tasks of GoogleEarth’s GRD, SPR, and CMP. There are also tasks and modules that are not included in the comparison for execution reasons. Specifically, the CNT and MES tasks are not included in the experiment because they cannot be evaluated by accuracy rate, and the dedicated modules 7), 8), and 9) are also not included in the comparison. The module 2) that retrieves similar areas of the query is not ablated because the GRD cannot be executed without it. The results in Table 7 show the significance of each component in maximizing the model’s performance. Among the components, omitting the `GetLandmarkSeg` module led to a significant drop in performance on all tasks, with GRD plummeting to 6.26%, SPR to 15.77%, and CMP to 0.00%. This shows the vital role of segmentation for landmark objects. Similarly, omitting `SegDirection` resulted in a marked decrease in GRD performance to 26.01%, indicating the importance of directional cues in grounding tasks. The omission of `SegAround` impairs performance particularly in the SPR task. This indicates the necessity of identifying objects in close proximity for accurate spatial reasoning. The exclusion of `SegBetween` impacts all tasks, though less drastically than other components. Lastly, omitting the `LargestSeg` module affected the CMP performance, reducing the score to 44.74. This highlights the importance of identifying the largest segment area for precise comparisons. See appendix B for more ablation studies.

**Qualitative results and failure cases.** Figure 7 shows qual-

Method	GRD	SPR	CMP
GeoProg3D	45.20	64.00	59.73
w/o <code>GetLandMarkSeg</code>	6.26	15.77	0.00
w/o <code>SegDirection</code>	26.01	48.95	52.63
w/o <code>SegAround</code>	34.67	36.76	43.42
w/o <code>SegBetween</code>	40.14	58.55	60.53
w/o <code>LargestSeg</code>	45.20	51.51	44.74

Table 7. Ablation study of different Geographical Vision APIs.

itative examples and failure cases. As shown, our approach successfully identifies the specified region to produce answers required for each task. For the example of CNT, when counting automobiles, the specified area is obtained as SEG2, and then the object detection function works to produce the answer that matches the ground truth. However, there are still challenging failure cases to address, as shown in the right bottom of Figure 7. In GRD, there is an over-activation error caused by the square building responding to the square, which means a plaza. In CMP, segmentation errors occur because geographic information is also assigned to parts of adjacent buildings. Possible solutions for future work include redesigning the embedded language features and designing georeferencing to have a margin.

## 6. Conclusion

We introduced GeoProg3D, a novel visual programming framework that enables human-computer interaction with city-scale 3D scenes through natural language queries. We also provided GeoEval3D for benchmarking 3D scene understanding models. Our experiments demonstrated that GeoProg3D significantly improves accuracy across the five visual geographical tasks. We believe our approach and dataset have contributed to the advancement of research in 3D scene understanding and visual programming.

## Acknowledgements

This work was supported by JST PRESTO (Grant Number JPMJPR22P8) and JSPS KAKENHI (Grant Number 25K03177), Japan.

## References

- [1] Daichi Azuma, Taiki Miyayoshi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19129–19139, 2022. 3
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 7, 15
- [3] Yash Sanjay Bhalgat, Iro Laina, João F. Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2
- [4] Valentin Bieri, Marco Zamboni, Nicolas S. Blumer, Qingxuan Chen, and Francis Engelmann. Opacity3d: 3d urban scene understanding with vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025. 3
- [5] Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *Proc. Conference on Robot Learning (CoRL)*, pages 706–717, 2022. 3
- [6] Christel Chappuis, Valérie Zermatten, Sylvain Lobry, Bertrand Le Saux, and Devis Tuia. Prompt-rsvqa: Prompting visual context to a language model for remote sensing visual question answering. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1372–1381, 2022. 3
- [7] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proc. European Conference on Computer Vision (ECCV)*, pages 202–221. Springer, 2020. 3
- [8] Haoran Chen, Kenneth Blomqvist, Francesco Milano, and Roland Siegwart. Panoptic vision-language feature fields. *IEEE Robotics and Automation Letters (RA-L)*, 9(3):2144–2151, 2024. 3
- [9] Yingshu Chen, Huajian Huang, Tuan-Anh Vu, Ka Chun Shum, and Sai-Kit Yeung. Stylecity: Large-scale 3d urban scenes stylization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 7, 15
- [11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 3
- [12] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *arXiv preprint arXiv:2305.15393*, 2023. 3
- [13] Tobias Fischer, Jonas Kulhanek, Samuel Rota Bulò, Lorenzo Porzi, Marc Pollefeys, and Peter Kotschieder. Dynamic 3d gaussian fields for urban areas. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2
- [14] Jiaming Gu, Minchao Jiang, Hongsheng Li, Xiaoyuan Lu, Guangming Zhu, Syed Afaq Ali Shah, Liang Zhang, and Mohammed Bennamoun. Ue4-nerf: neural radiance field for real-time rendering of large-scale scene. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 59124–59136. Curran Associates, Inc., 2023. 2
- [15] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2
- [16] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4, 5
- [17] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12893–12902, 2024. 3
- [18] Yujin Ham, Mateusz Michalkiewicz, and Guha Balakrishnan. Dragon: Drone and ground gaussian splatting for 3d building reconstruction. In *IEEE International Conference on Computational Photography (ICCP)*, 2024. 2
- [19] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *arXiv*, 2023. 3
- [20] Qingqing Hu, Yue Yuan, Jie Mei, Qi Bi, Jinghui Xie, and Qiang Du. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 3
- [21] Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *arXiv preprint arXiv:2410.06234*, 2024. 3, 7, 15
- [22] Mazeyu Ji, Ri-Zhao Qiu, Xueyan Zou, and Xiaolong Wang. Graspplats: Efficient manipulation with 3d feature splatting. *arXiv preprint arXiv:2409.02084*, 2024. 3

- [23] Yuzhou Ji, He Zhu, Junshu Tang, Wuyi Liu, Zhizhong Zhang, Xin Tan, and Yuan Xie. Fastlgs: Speeding up language embedded gaussians with feature grid mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2
- [24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2
- [25] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics (TOG)*, 43(4), 2024. 2
- [26] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19729–19739, 2023. 2, 3, 6
- [27] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *Proc. International Conference on Learning Representations (ICLR)*, 2024. 3
- [28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 5
- [29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3
- [30] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. 3
- [31] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [32] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6687–6696, 2022. 3
- [33] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27831–27840, 2024. 3, 7, 15
- [34] Ruilong Li, Sanja Fidler, Angjoo Kanazawa, and Francis Williams. NeRF-XL: Scaling nerfs with multiple GPUs. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2
- [35] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3205–3215, 2023. 2
- [36] Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, and Wenming Yang. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5166–5175, 2024. 2
- [37] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 2, 6
- [38] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiacong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62:1–16, 2024. 3
- [39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 7, 15
- [40] Yang Liu, He Guan, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2, 7
- [41] Yang Liu, Chuanchen Luo, Zhongkai Mao, Junran Peng, and Zhaoxiang Zhang. Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes, 2024. 2
- [42] Yang Liu, Chuanchen Luo, Zhongkai Mao, Junran Peng, and Zhaoxiang Zhang. Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes. In *ICLR*, 2025. 2
- [43] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3
- [44] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 56:2183–2195, 2018. 3
- [45] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, and Yansheng Li. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024. 3
- [46] Xiangyu Meng, Yue Cao, Bing Zhang, and Liangpei Zhang. A multiscale grouping transformer with clip latents for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62:1–15, 2024. 3
- [47] Meta. Llama 3.2 connect 2024: Vision on the edge and mobile devices. <https://ai.meta.com/blog/llama->

3-2-connect-2024-vision-edge-mobile-devices/, 2024. 7, 15

- [48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 2
- [49] Taiki Miyanishi, Fumihito Kitamori, Shuhei Kurita, Jinyuk Lee, Motoaki Kawanabe, and Naoya Inoue. Cityrefer: Geography-aware 3d visual grounding dataset on city-scale point cloud data. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3
- [50] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 3, 7, 15
- [51] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 7, 15
- [52] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, and Conghui He. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *AAAI*, 2025. 3, 7, 15
- [53] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20051–20060, 2024. 2, 3, 4, 5, 6, 7, 14
- [54] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5, 2016. 3
- [55] 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 *Proc. International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 4
- [56] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14148–14156, 2020. 3
- [57] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. 2023. 3
- [58] Nur Muhammad (Mahi) Shafiqullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *ArXiv*, abs/2210.05663, 2022. 3
- [59] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5333–5343, 2024. 2, 3
- [60] Qing Shuai, Haoyu Guo, Zhen Xu, Haotong Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Real-time view synthesis for large scenes with millions of square meters. 2024. 2, 4, 13, 14
- [61] Kaiwen Song, Xiaoyi Zeng, Chenqu Ren, and Juyong Zhang. City-on-web: Real-time neural rendering of large-scale scenes on the web. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2
- [62] Karl Stelzner, Kristian Kersting, and Adam Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. 2021. 3
- [63] Sanjay Subramanian, Medhini Narasimhan, et al. Modular visual question answering via code generation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 3
- [64] Penglei Sun, Yaoxian Song, Xiang Liu, Xiaofei Yang, Qiang Wang, Tiefeng Li, Yang Yang, and Xiaowen Chu. 3d question answering for city scene understanding. In *Proc. ACM International Conference on Multimedia (ACMMM)*, pages 2156–2165, 2024. 3
- [65] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proc. ACM International Conference on Multimedia (ACMMM)*, page 404–412, New York, NY, USA, 2022. Association for Computing Machinery. 3
- [66] Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 4, 5
- [67] Matthew Tancik, Vincent Casser, Xinchun Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *arXiv*, 2022. 2
- [68] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, and Deyu Meng. Crs-diff: Controllable generative remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2024. 3
- [69] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. *arXiv preprint arXiv:2405.01413*, 2024. 3
- [70] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12931, 2022. 2
- [71] Vicente Vivanco, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3
- [72] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proc. AAAI Conference on Artificial Intelligence*, 2024. 3
- [73] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from

- natural language. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14958–14967, 2024. 3
- [74] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *The European Conference on Computer Vision (ECCV)*, 2022. 2
- [75] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. Citydreamer: Compositional generative model of unbounded 3D cities. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6
- [76] Haozhe Xie, Zhaoxi Chen, Fangzhou Hong, and Ziwei Liu. GaussianCity: Generative gaussian splatting for unbounded 3D city generation. *arXiv 2406.06526*, 2024. 2
- [77] Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, and Dahua Lin. Grid-guided neural radiance fields for large urban scenes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [78] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 3
- [79] Shixiong Xu, Chenghao Zhang, Lubin Fan, and Gaofeng Meng. Addressclip: Empowering vision-language models for city-wide image address localization. In *Proc. European Conference on Computer Vision (ECCV)*, pages 76–92, 2024. 3
- [80] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1389–1399, 2023. 2
- [81] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3
- [82] Gim Hee Lee Yu Chen. Dogs: Distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus. In *arXiv*, 2024. 2
- [83] Tianyuan Yuan, Yucheng Mao, Jiawei Yang, Yicheng Liu, Yue Wang, and Hang Zhao. Presight: Enhancing autonomous vehicle perception with city-scale nerf priors. *arXiv preprint arXiv:2403.09079*, 2024. 2
- [84] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–19, 2022. 3
- [85] Hanyue Zhang, Zhiliu Yang, Xinhe Zuo, Yuxin Tong, Ying Long, and Chen Liu. Garfield++: Reinforced gaussian radiance fields for large-scale 3d scene reconstruction, 2024. 2
- [86] Qihang Zhang, Yinghao Xu, Chaoyang Wang, Hsin-Ying Lee, Gordon Wetzstein, Bolei Zhou, and Ceyuan Yang. 3DitScene: Editing any scene via language-guided disentangled gaussian splatting. In *arXiv*, 2024. 3
- [87] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2024. 3
- [88] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthmarker: A visual prompt learning framework for region-level and point-level remote sensing imagery comprehension. *arXiv preprint arXiv:2407.13596*, 2024. 3
- [89] Yuqi Zhang, Guanying Chen, and Shuguang Cui. Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features. *arXiv preprint arXiv:2303.03003*, 2023. 2
- [90] Yuqi Zhang, Guanying Chen, Jiaying Chen, and Shuguang Cui. Aerial lifting: Neural urban semantic and building instance lifting from aerial imagery. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [91] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62, 2024. 3
- [92] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2021. 3
- [93] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 2
- [94] Yue Zhou, Litong Feng, Yiping Ke, Xue Jiang, Junchi Yan, Xue Yang, and Wayne Zhang. Towards vision-language geo-foundation models: A survey. *arXiv preprint arXiv:2406.09385*, 2024. 3
- [95] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2911–2921, 2023. 3