Appendix

In this appendix, we provide additional analyses, implementation details, and qualitative results to complement the main paper. These include the effects of in-context example count, generalization to rephrased queries, technical contributions of GCLF, detailed scene-wise performance, and visualizations that further support the robustness and interpretability of the GeoProg3D framework.

A. Additional Analysis on Visual Programming

This section provides a detailed analysis of the visual programming component in GeoProg3D, focusing on the impact of in-context examples and the framework's generalization capabilities.

A.1. Effect of In-Context Example Count

To evaluate the impact of the number of in-context examples (ICEs) on the performance of our framework, we conducted an experiment by varying the number of ICEs provided to LLM. As shown in Figure 8, we measured the success rate of program generation for each of our five tasks, using 5, 10, and 15 ICEs. The results demonstrate a clear trend: the program generation success rate improves as the number of ICEs increases. With 15 ICEs, the success rate reaches approximately 90% for most tasks and begins to saturate. It is noteworthy that even with only 5 ICEs, GeoProg3D achieves a program generation success rate of around 70%. This level of performance is sufficient to significantly outperform baseline methods. For instance, on the GRD task, GeoProg3D with 5 ICEs scores 38.02% in localization accuracy, whereas LangSplat achieves only 17.07%. This highlights the efficiency of our approach in leveraging LLMs for compositional reasoning with a minimal number of examples.

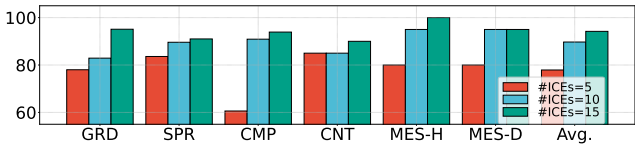


Figure 8. Success rate of program generation for each task

A.2. Generalization to Rephrased Queries

To assess the generalization capability and linguistic robustness of GeoProg3D, we evaluated its performance against rephrased queries. We used GPT-4o to generate paraphrased versions of the original queries in our GeoEval3D benchmark, creating a new test set denoted as "Rephrased Queries (RQ)" (Table 10). We then ran our full framework, GeoProg3D, on this RQ set. The results are presented in Table 8 (for the GRD task) and Table 9 (for other tasks).

Test scene	LangSplat	GCLF	GeoProg3D	GeoProg3D+RQ
GoogleEarth	17.07	26.83	46.34	46.34

Table 8. Localization accuracy (%) on the GRD task.

Method	SPR Acc.↑	CMP Acc.↑	CNT MAE↓	MES-H MAE (m)↓	MES-D MAE (m)↓
InternVL	64.2	30.3	1.8	102.8	115
TEOChat	53.7	63.6	3.8	236.2	205.6
GeoProg3D	71.6	78.8	1.4	13.6	20.1
GeoProg3D+RQ	71.8	78.8	1.3	13.2	18.6

Table 9. Performance of SPR, CMP, CNT, and MES tasks.

As the tables show, the performance of GeoProg3D with rephrased queries (GeoProg3D+RQ) is comparable to its performance with the original queries across all tasks. For instance, the SPR accuracy is 71.6% for the original queries and 71.8% for the rephrased ones, and the MAE for the MES-D task is 20.1m and 18.6m, respectively. This minimal difference in performance indicates that our framework is not reliant on specific keywords or phrasing. Instead, it demonstrates a strong ability to understand the semantic intent of a query and translate it into a correct executable program, highlighting the robust generalization capabilities of our approach.

ID	Original queries
1	How many buildings are there?
2	There are at least two streets facing BldgA.
3	There are 2 or less buildings to the directly west of BldgA.
4	Which is taller, the tallest object around BldgA or BldgB?
5	There are 2 or more grass areas to the directly west of BldgA.
ID	**Rephrased queries**
1	What is the total number of buildings?
2	BldgA faces at least two streets.
3	At most two buildings stand directly to the west of BldgA.
4	Which is taller, the tallest object surrounding BldgA or BldgB?
5	To the direct west of BldgA, there are two or more grassy areas.

Table 10. Examples of test queries rephrased by LLM.

B. Implementation details

To extract object masks, the SAM model using the ViT-H backbone was used. For extracting CLIP features, the OpenCLIP ViT-B/16 model was used. The tree-structure is implemented by utilizing the LoG rendering implementation [60]. Each training stage is repeated 2,000, 15,000, and 30,000 times for the Google Earth dataset, and 300,000, 600,000, and 500,000 times for the UrbanScene3D dataset. Our autoencoder is implemented using an MLP, which compresses 512-dimensional CLIP features into 3-dimensional latent features. For GoogleEarch scenes, each composed of 60 images in with a resolution of 958×538, training took about 15 minutes on an NVIDIA Quadro RTX 8000 GPT using 2GB of memory. For the UrbanScene3D scene composed of 5,871 images with a resolution of 1620×1080, training took about 6 hours using 40GB of memory. The

Test Scene	Area (m^2)	LSeg	LERF	LangSplat	GCLF	GeoProg3D
Center Blvd	2.7×10^5	0.03	0.07	7.69	19.23	**42.31**
World Fin Ctr	4.7×10^5	0	0.05	20.00	16.00	**44.00**
Mott St	1.7×10^5	0	0.05	10.71	17.86	**53.57**
Washington Sq	1.3×10^5	0.01	0.08	18.18	27.27	**40.91**
Campus	5.0×10^6	0.01	OOM	OOM	6.98	**30.23**

Table 11. Localization accuracy (%) for GRN task by scene.

Test Scene	Area (m^2)	LSeg	LERF	LangSplat	GCLF	GeoProg3D
Center Blvd	2.7×10^5	0.04	0.15	4.28	6.10	**19.74**
World Fin Ctr	4.7×10^5	0	0.08	6.03	6.31	**16.12**
Mott St	1.7×10^5	0	0.05	5.21	7.56	**25.60**
Washington Sq	1.3×10^5	0	0.18	5.22	6.80	**11.12**
Campus	5.0×10^6	0.05	OOM	OOM	3.78	**8.74**

Table 12. 3D semantic segmentation performance for GNR task by scene. IoU scores (%) are reporeted.

total number of images used for training is 6,111.

C. Technical contributions of GCLF

Our central contribution in designing GCLF lies in addressing the challenge of geometric distortion that arises from a naive integration of existing methods like LangSplat [53] and hierarchical representations [60]. To tackle this issue, GCLF employs a two-stage training strategy. First, it aligns the 3D Gaussians with real-world coordinates by referencing 2D geographic information from OpenStreetMap, and then freezes their geometric configuration. Subsequently, it learns hierarchical language features upon this fixed structure. This unique approach establishes GCLF as the first large-scale, hierarchical 3D language field for city-scale environments, achieving both high-fidelity scene representation and efficient language query processing.

D. Detailed results and analysis

GRD task. Tables 11 and 12 present the grounding performance by scene. In all scenes, GeoProg3D achieves significantly higher performance compared to GCLF and other baselines. Additionally, GCLF outperforms LangSplat in most scenes in terms of segmentation performance, indicating that its high-quality 3D representation leads to improved performance in the GRD task.

SPR task. Table 13 presents the spatial reasoning performance by scene, demonstrating that GeoProg3D achieves the highest or competitive accuracy in all scenes. Its accuracy remains above 60% in every scene, showcasing strong generalization across diverse environments. The other models including GeoChat and Llama-3.2 Vision perform moderately well but lag behind, while GPT-4o Vision struggles with accuracies below 40% in most cases.

CMP task. Table 14 shows that GeoProg3D outperforms the other models in three out of four scenes for the CMP task. While Llama-3.2 Vision shows strong performance in Center Blvd (73.68%), its accuracy drops significantly in the other scenes. InternVL2.5 and VHM perform moderately well but fall short of GeoProg3D. LLaVA-1.5 and GPT-4o Vision struggle significantly, with the latter showing near-zero performance across most scenes.

CNT task. As shown in Table 15, GeoProg3D also demonstrates strong performance in the CNT task, achieving competitive Mean Absolute Error (MAE) values across all scenes. While LHRS-BOT achieves slightly better MAE in Center Blvd (1.54), GeoProg3D consistently performs well, maintaining low MAEs across all scenes. LLaVA-1.5 and GPT-4o Vision show higher errors, indicating limited reliability in this task. Qwen2.5-VL and Llama-3.2 Vision perform moderately but lack the consistency shown by GeoProg3D. Table 19 and Figure 9 show the average of ground truth counts by scene and a comparison of predicted and ground-truth counts, respectively. The R-squared values in Figure 9 indicate that methods other than GeoProg3D poorly align with the ground truth distribution. These methods often output "1" as the answer to the query, which may explain the smaller MAE observed in some cases for the Center Blvd scene. These results highlight GeoProg3D's capability to balance precision and generalization in diverse counting scenarios.

MES task. Table 16 shows that GeoProg3D demonstrates superior performance in both height (MES-H) and distance (MES-D) measurement tasks across most test scenes, as indicated by its consistently low MAE values. For MES-H, GeoProg3D outperforms the other methods in the scenes except for Mott St. In the Mott St scene, there are 26 queries for MES-H, and GeoProg3D was unable to provide answers for 3 of them. This issue is caused by an error in the program generation by the LLM. Additionally, the Mott St scene contains many low-lying buildings (as shown in the rightmost scene in Figure 11), where baseline method achieves low error by consistently providing monotonous responses with small values, as illustrated in Figure 10. For MES-D, GeoProg3D achieves the lowest MAE across all test scenes, highlighting its precision in distance measurement. Notably, while most methods struggle with higher errors in a large environment like Campus, where GPT-4o Vision has an MAE exceeding 1500 meters, GeoProg3D showed superior performance on this scene.

Ablation study using different LLMs. Table 18 compares the performance of GeoProg3D using different LLM backbones, specifically GPT-3.5 and GPT-4o, across three tasks: GRD, SPR, and CMP. The results indicate that GPT-4o consistently outperforms GPT-3.5 across all tasks, with notable improvements in CMP (64.76% compared to 59.7%). The differences in GRD and SPR are smaller. The improvements in CMP are particularly significant, indicating GPT-

Spatial Reasoning: Accuracy (%) ↑										
Test Scene	LLaVA-1.5	Llama-3.2 Vision	GPT-4o Vision	Qwen2.5-VL	InternVL2.5	GeoChat	TEOChat	LHRS-BOT	VHM	**GeoProg3D**
Center Blvd	47.88	49.23	37.98	56.86	56.45	58.41	58.41	49.43	**63.47**	60.17
World Fin Ctr	47.78	62.65	24.86	52.21	43.23	61.67	61.04	53.19	50.75	**67.14**
Mott St	52.12	53.87	16.08	50.73	57.38	58.41	61.55	50.37	52.78	**67.18**
Washington Sq	56.02	53.62	20.17	53.74	60.01	50.41	55.17	44.83	51.18	**61.52**
Campus	46.04	57.34	15.18	47.24	52.47	56.76	57.99	41.94	56.92	**60.87**

Table 13. SPR performance by scene. LLaVA-1.5 [39], Llama-3.2 Vision [47], GPT-4o Vision [51], Qwen2.5-VL [2], InternVL2.5 [10], GeoChat [33], TEOChat [21], LHRS-BOT [50], VHM [52] and GeoProg3D are evaluated.

Comparison: Accuracy (%) ↑										
Test Scene	LLaVA-1.5	Llama-3.2 Vision	GPT-4o Vision	Qwen2.5-VL	InternVL2.5	GeoChat	TEOChat	LHRS-BOT	VHM	**GeoProg3D**
Center Blvd	21.05	**73.68**	0.00	26.32	57.89	26.32	57.89	5.26	36.84	63.16
World Fin Ctr	42.11	21.05	10.53	52.63	31.58	52.63	31.58	36.84	36.84	**68.42**
Mott St	36.84	10.53	0.00	15.79	36.84	36.84	**42.11**	15.79	**42.11**	**42.11**
Washington Sq	47.83	8.70	0.00	13.04	47.83	52.17	60.87	52.17	43.48	**65.22**

Table 14. CMP performance by scene. LLaVA-1.5 [39], Llama-3.2 Vision [47], GPT-4o Vision [51], Qwen2.5-VL [2], InternVL2.5 [10], GeoChat [33], TEOChat [21], LHRS-BOT [50], VHM [52] and GeoProg3D are evaluated.

Counting: MAE ↓										
Test Scene	LLaVA-1.5	Llama-3.2 Vision	GPT-4o Vision	Qwen2.5-VL	InternVL2.5	GeoChat	TEOChat	LHRS-BOT	VHM	**GeoProg3D**
Center Blvd	2.23	1.77	1.77	2.00	1.85	3.15	2.46	**1.54**	4.62	1.92
World Fin Ctr	2.86	2.86	3.07	2.14	3.00	2.71	2.93	2.71	2.21	**1.93**
Mott St	3.84	2.83	3.68	3.58	3.42	2.95	2.89	10.63	7.68	**1.63**
Washington Sq	3.39	2.70	3.57	2.87	2.91	2.74	3.09	4.52	6.61	**2.52**
Campus	4.23	3.54	4.29	4.00	4.23	3.69	4.06	3.49	4.34	**2.51**

Table 15. CNT performance by scene. LLaVA-1.5 [39], Llama-3.2 Vision [47], GPT-4o Vision [51], Qwen2.5-VL [2], InternVL2.5 [10], GeoChat [33], TEOChat [21], LHRS-BOT [50], VHM [52] and GeoProg3D are evaluated.

Measurement (Height): MAE (m↓)										
Test Scene	LLaVA-1.5	Llama-3.2 Vision	GPT-4o Vision	Qwen2.5-VL	InternVL2.5	GeoChat	TEOChat	LHRS-BOT	VHM	**GeoProg3D**
Center Blvd	418.37	85.18	231.21	71.53	57.79	124.05	193.58	61.05	53.00	**50.99**
World Fin Ctr	349.89	104.78	198.26	90.58	91.26	53.68	184.32	68.21	59.26	**58.16**
Mott St	962.12	43.35	167.73	37.88	**18.58**	62.12	107.62	22.65	69.27	56.28
Washington Sq	699.09	118.91	35.43	38.74	37.57	99.09	116.04	32.78	28.78	**15.52**

Table 16. MES-H performance by scene. LLaVA-1.5 [39], Llama-3.2 Vision [47], GPT-4o Vision [51], Qwen2.5-VL [2], InternVL2.5 [10], GeoChat [33], TEOChat [21], LHRS-BOT [50], VHM [52] and GeoProg3D are evaluated.

Measurement (Distance): MAE (m↓)										
Test Scene	LLaVA-1.5	Llama-3.2 Vision	GPT-4o Vision	Qwen2.5-VL	InternVL2.5	GeoChat	TEOChat	LHRS-BOT	VHM	**GeoProg3D**
Center Blvd	169.95	199.07	196.84	164.47	165.42	82.58	164.42	160.16	141.05	**61.60**
World Fin Ctr	460.21	220.00	285.05	219.79	239.79	129.05	178.47	212.11	162.47	**67.94**
Mott St	306.58	162.47	164.68	162.47	125.21	103.32	264.05	160.37	142.47	**55.00**
Washington Sq	98.95	123.81	283.10	155.05	98.14	60.86	188.62	118.57	96.00	**28.71**
Campus	837.11	427.94	1583.48	412.86	318.71	328.68	359.71	438.94	354.91	**139.51**

Table 17. MES-D performance by scene. LLaVA-1.5 [39], Llama-3.2 Vision [47], GPT-4o Vision [51], Qwen2.5-VL [2], InternVL2.5 [10], GeoChat [33], TEOChat [21], LHRS-BOT [50], VHM [52] and GeoProg3D are evaluated.

LLM	GRD	SPR	CMP
GPT-3.5	45.20	64.00	59.73
GPT-4o	46.09	64.40	64.76

Table 18. Ablation of GeoProg3D using different LLM backbones.

4o's ability to handle more complex tasks. This highlights the potential to enhance GeoProg3D's performance across a wide range of metrics by utilizing an advanced LLM backbone, depending on the available budget.

Reconstruction quality. Figure 11 compares the image reconstruction quality of GCLF and a vanilla 3D-GS in several scenes. As a result of training with the same number of epochs, the vanilla 3D-GS lacks texture details.

Statistics. Table 1 presents statistics for LangSplat and GCLF in terms of the number of Gaussians and inference speed across various scenes. GCLF consistently generates

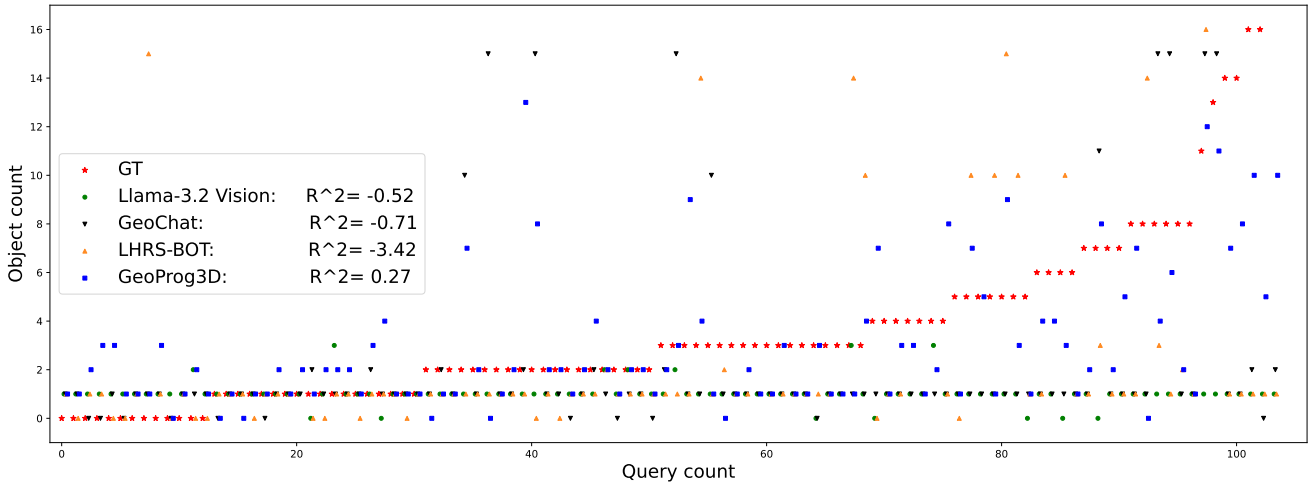


Figure 9. Comparison of predicted counts versus ground truth across different methods.

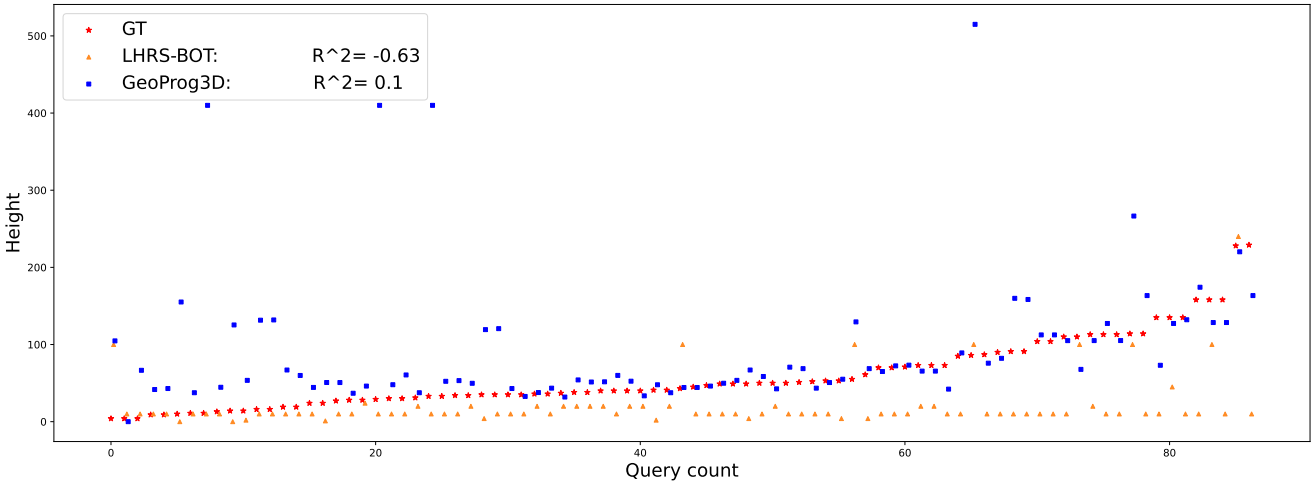


Figure 10. Comparison of predicted height versus ground truth across different methods.

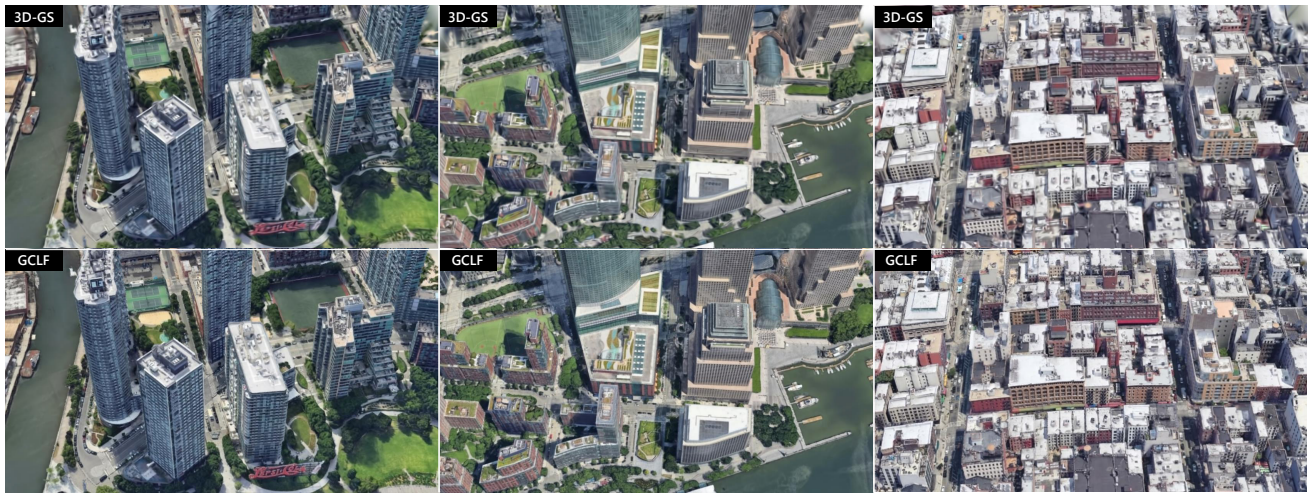


Figure 11. Comparison of 3D scene reconstruction quality between 3D-GS and GCLF.

significantly more number of Gaussians than LangSplat, achieving more detailed representations. However, this reduces the rendering speed, as LangSplat is consistently faster across all cases. These results highlight the trade-off between achieving detailed scene representations and maintaining computational efficiency. Nevertheless, GCLF

Scene	Avg. CNT	Avg. MES-H	Avg. MES-D
Center Blvd	2.31	75.21	161.21
World Fin Ctr	3.07	96.74	220.00
Mott St	3.74	37.88	162.47
Washington Sq	3.61	40.04	123.81
Campus	4.31	-	427.94
Overall	3.63	59.46	284.94

Table 19. Average ground truth values for CNT, MES-H, and MES-D queries.

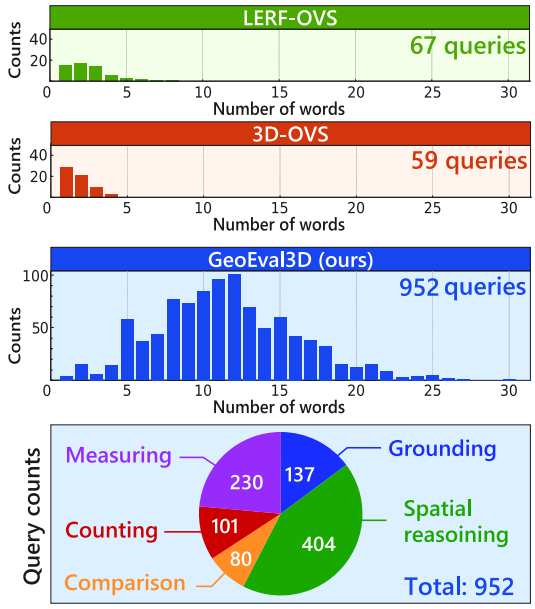


Figure 12. Word count and query distributions. GeoEval3D contains 952 unique queries covering five tasks. Queries include more words than those in the previous evaluation datasets, indicating complexity of the proposed.

demonstrates practical usability, as they are capable of querying even large-scale data within realistic time frames, ensuring their suitability for real-world applications.

E. Qualitative examples

Visualization of CLIP features. Figure 14 visualizes CLIP features projected into a three-dimensional feature space using an autoencoder during the training of the GCLF. As illustrated, buildings and their surroundings are grouped into distinct clusters.

Viewpoint-independent localization by GCLF. Since the VLMs that are compared in the versatility experiment can only process 2D images, all methods, including GeoProg3D, are input with top-down view images. Information from directly above is not sufficient for searches related to the sides of buildings and signboards. However, GCLF boosts the high performance of GeoProg3D by localizing it to take into account the characteristics of structures that cannot be seen from directly above. Figure 13 shows examples of SPR and MES-H that have succeeded in reasoning in GeoProg3D with viewpoint-independent localization. In SPR, the red lettering signboard is not visible in the top-down view used for inference, but it is correctly activated in the localization of the inference process. In MES-H, whether a building is glass-fronted or not cannot be seen in the top-down view, but localization succeeded.

Qualitative examples. Figure 15 shows additional qualitative examples, demonstrating the capability of GeoProg3D across various tasks and environments. Figure 16 illustrates the output obtained by executing a notebook included with our code. Figure 17 shows examples of language-guided 3D Gaussian editing as an additional task. This editing task requires models to localize the object and modify it based on a given query q_k.

F. Dataset details

To further ensure the quality and reliability of the dataset, we evaluated the inter-annotator agreement. For the SPR task, annotations from two independent annotators showed substantial agreement with a Cohen's Kappa of 0.78. For the MES-H and MES-D tasks, we observed high agreement as well, with Pearson correlation coefficients of 0.81 and 0.95, respectively. These results confirm that our annotations are consistent and reliable. Furthermore, for the SPR task, we explicitly instructed annotators to maintain a balanced label distribution, resulting in approximately 54% "yes" and 46% "no" answers to avoid potential bias.

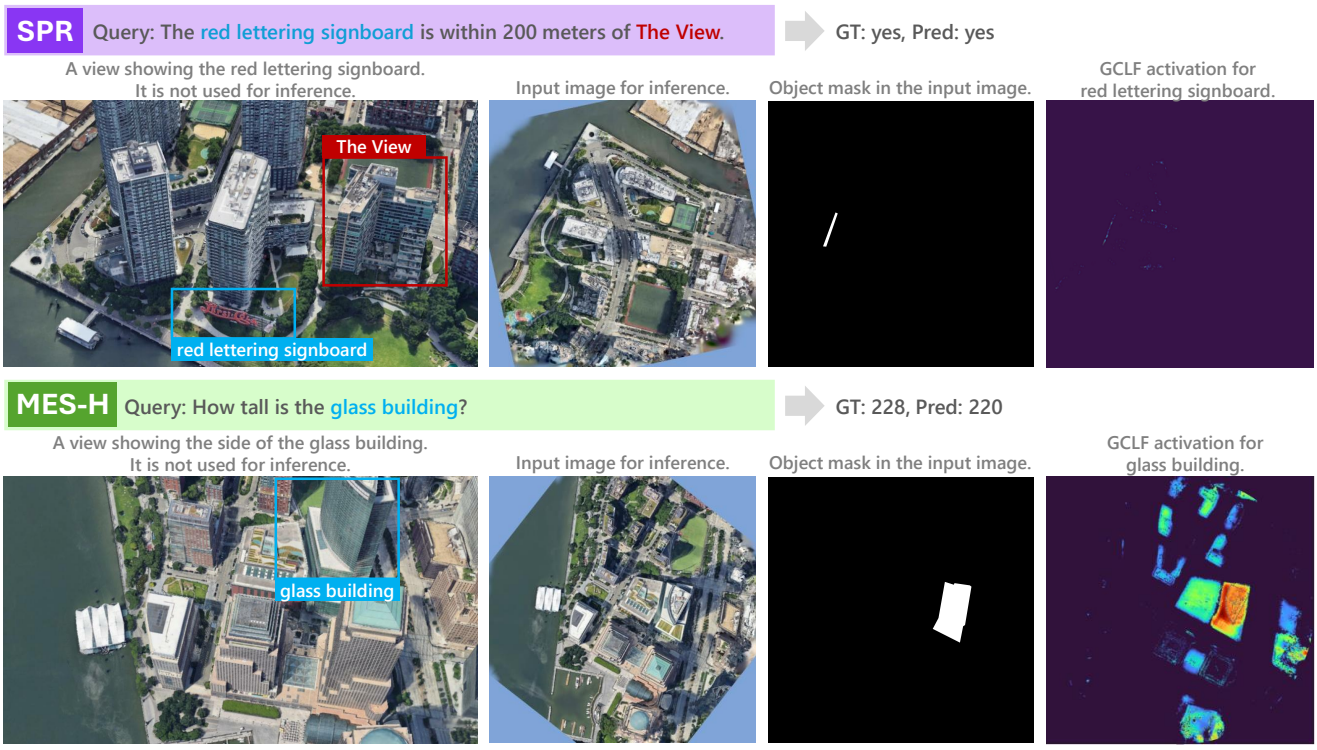


Figure 13. Examples of viewpoint-independent localization by GCLF.

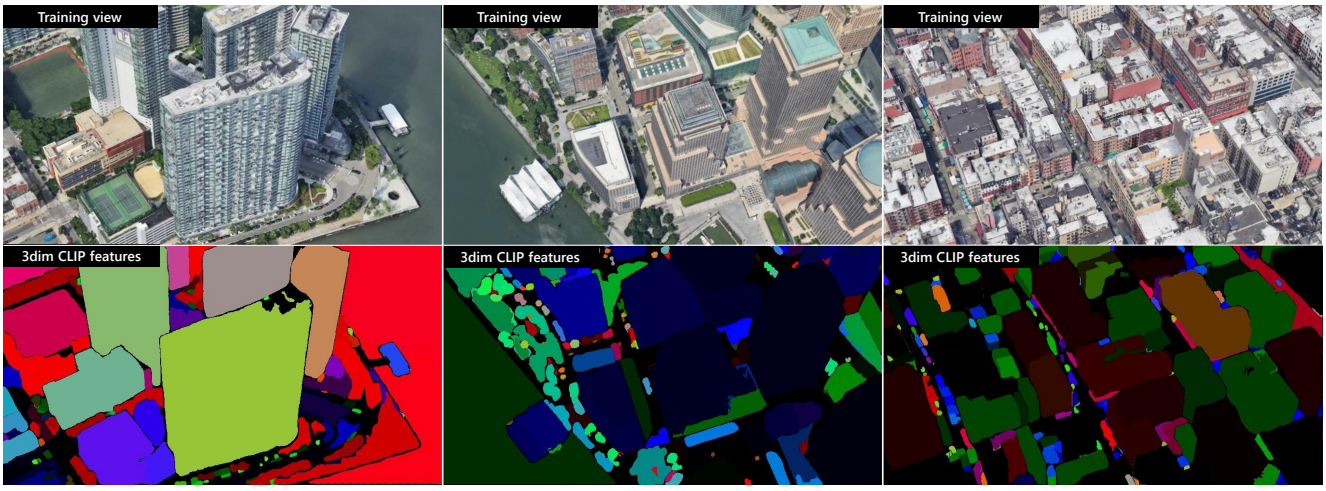


Figure 14. Visualization of CLIP features for training 3D language fields.

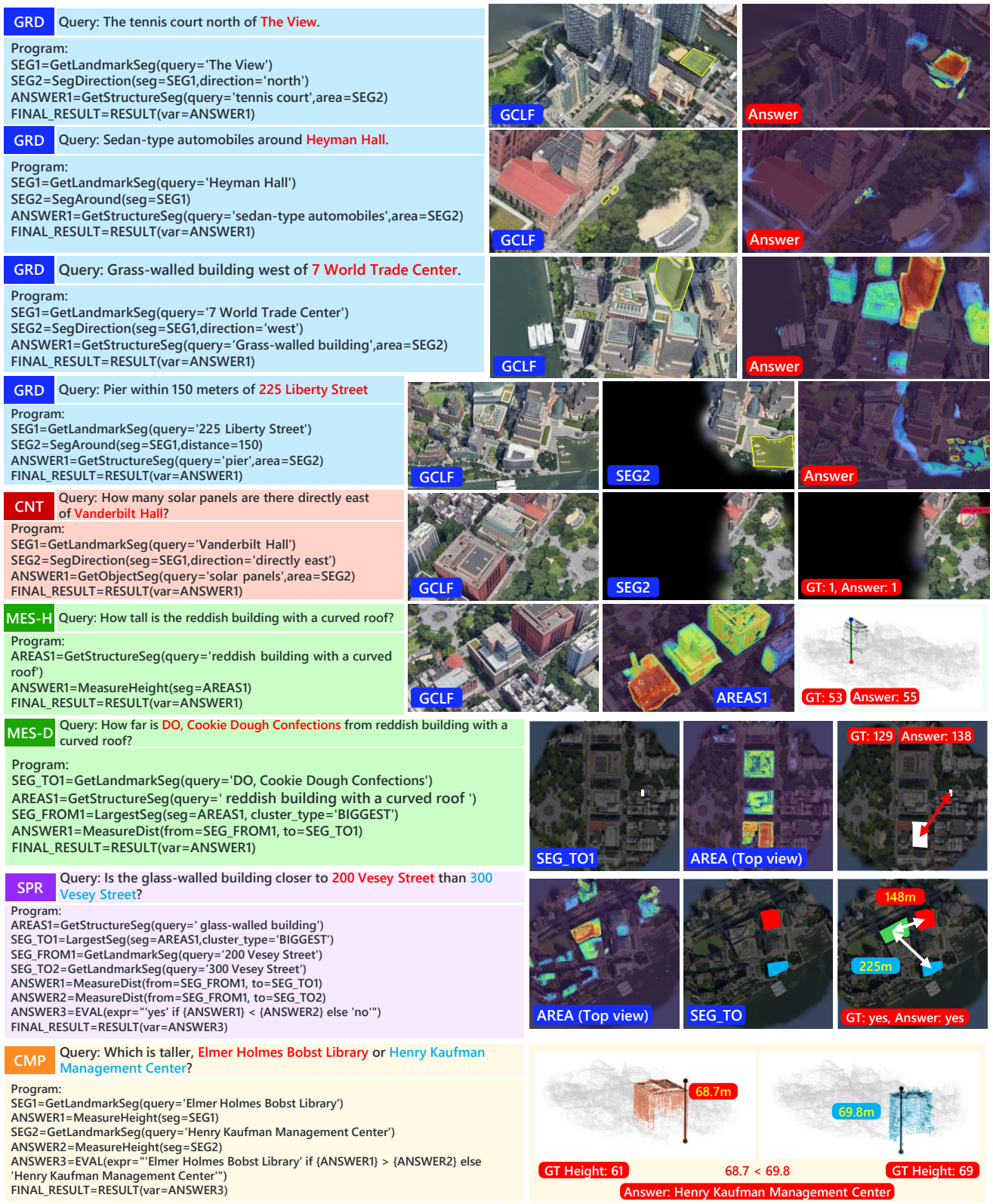
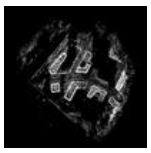
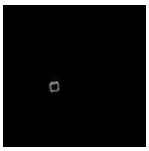


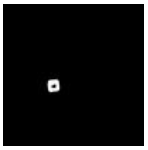
Figure 15. Other qualitative results.

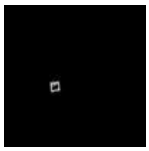
Query: Which is taller, the skyscraper that is closest to 45-45 Center Blvd or The View?

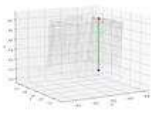
CLOSEST_FROM0=GetLandmarkSeg(query='45-45 Center Blvd')=51802px 

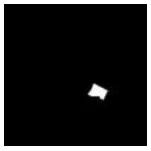
AREAS0=GetStructureSeg(query='skyscraper',area=None)=293548px 

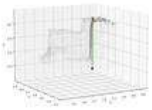
SEG0=LargestSeg(seg=AREAS0,cluster_type='CLOSEST',from=CLOSEST_FROM0)=5634pix 

CLOSEST_TO0=SegAround(seg=SEG0,distance=None)=24966pix 

AREAS1=GetStructureSeg(query='skyscraper',area=CLOSEST_TO0)=9640px 

ANSWER0=MeasureHeight(seg=AREAS1)=99.94 meter -- res: 

SEG1=GetLandmarkSeg(query='The View')=32489px 

ANSWER1=MeasureHeight(seg=SEG1)=67.59 meter -- res: 

ANSWER2=EVAL(expression="'the skyscraper that is closest to <45-45 Center Blvd>' if ANSWER0 > ANSWER1 else 'The View'")=EVAL(expression="'the skyscraper that is closest to <45-45 Center Blvd>' if 99.94 > 67.59 else 'The View'")=**the skyscraper that is closest to <45-45 Center Blvd>**

FINAL_RESULT->RESULT->**the skyscraper that is closest to <45-45 Center Blvd>**

Figure 16. Visual rationales generated by GeoProg3D.

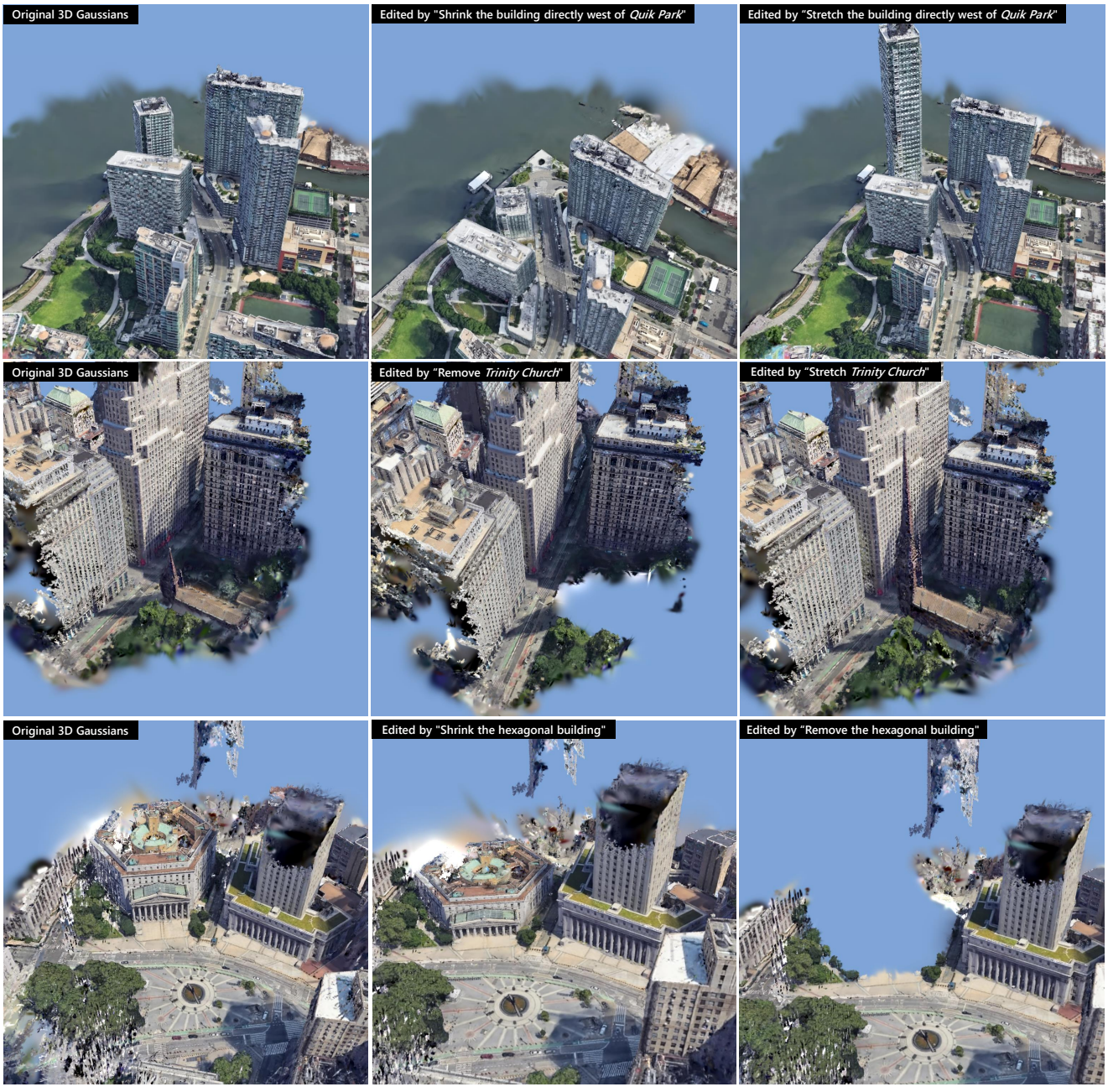


Figure 17. Examples of language guided 3D Gaussian editing.