AEROGROUND: LANGUAGE GROUNDING IN LARGE-SCALE AERIAL GAUSSIAN SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounding natural-language queries in large-scale aerial 3D scenes requires distinguishing visually similar structures and recovering their complete 3D support. Large viewpoint changes and incomplete observations make both tasks challenging. We present AeroGround, a query-time framework that couples active view exploration with evidence-guided structural refinement. Active Query-guided View Exploration (AQVE) maintains competing cross-view target hypotheses and selects informative views to resolve ambiguity. It associates observations across views and accumulates visibility-aware positive and negative evidence to guide subsequent exploration. Evidence-Guided Structural Refinement (ESR) consolidates reliable foreground and background observations into Gaussian-level target posteriors and combines them with local geometric consistency to recover missing support while suppressing foreground leakage. AeroGround operates directly on standard 3D Gaussian Splatting without scene-specific semantic feature optimization or storage. We also introduce AeroRef3D, a benchmark covering spatial–attribute, commonsense, functional, and hypothetical queries with zero-to-many targets across six aerial scenes. Experiments show that AeroGround substantially outperforms the evaluated baselines on AeroRef3D and achieves competitive results on two aerial open-vocabulary segmentation benchmarks.

1 INTRODUCTION

Large-scale 3D reconstructions from UAV imagery support applications such as urban planning, infrastructure inspection, asset management, and digital twins (Hu et al., 2021; Yang et al., 2023; Zhang et al., 2024). Beyond reconstructing geometry and appearance, interacting with these scenes often requires locating specific objects from free-form descriptions and recovering their 3D extent. For example, a user may ask for "the blue-roofed building beside the playground," all structures satisfying a functional description, or whether a specified target exists. We study this problem as language-guided grounding in large-scale aerial Gaussian scenes, where a query may refer to zero, one, or multiple targets.

Language-aware 3D Gaussian representations support semantic retrieval by connecting pretrained vision-language features to scene geometry (Shi et al., 2024; Qin et al., 2024). Instance-aware representations further organize Gaussians into coherent objects (Ye et al., 2024; Zhu et al., 2025b). Recent grounding methods extend these capabilities through spatial-relation modeling and multimodal reasoning over scene observations (He et al., 2025; Xie et al., 2026; Liu et al., 2025; Shi et al., 2026). Together with advances in aerial semantic understanding (Zhang et al., 2024; Wang et al., 2026), these methods provide useful foundations for language interaction. However, aerial grounding still requires determining which observations can distinguish competing targets and how incomplete observations should constrain the target's complete 3D support.

Two challenges are particularly important. *First, target identity can remain ambiguous across views.* Similar buildings and repetitive layouts often produce multiple plausible candidates for the same description. Broad views provide spatial context but may obscure discriminative appearance details, whereas close-up views may omit the reference objects needed to interpret spatial relations. Scene coverage alone therefore does not guarantee that the observed views can distinguish competing candidates. Moreover, a missing detection may result from occlusion rather than the absence of a matching target. *Second, identifying the correct target does not guarantee complete and accurate*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

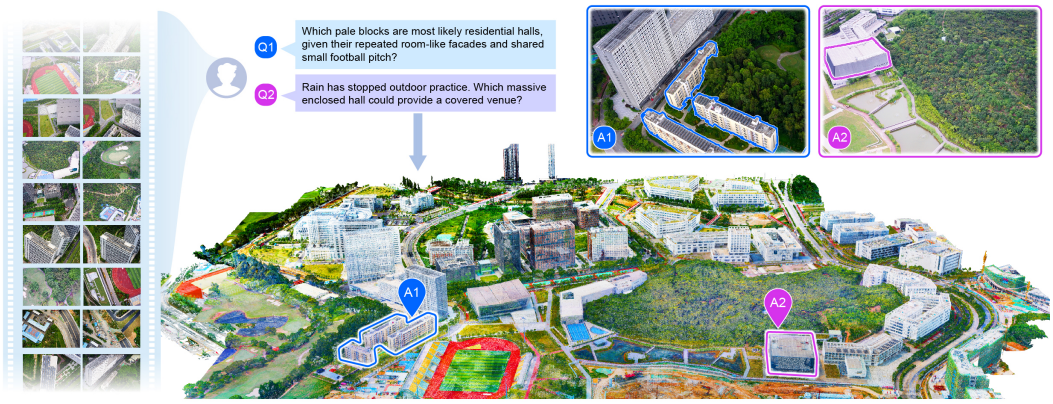

Figure 1: We propose **AeroGround**, an MLLM-agent framework for language-guided 3D grounding in large-scale aerial scenes.

3D support. Occlusion and inconsistent 2D predictions can lead to missing regions and foreground leakage. Uniform expansion may recover gaps in compact buildings, but can also merge nearby objects or distort thin and open structures. Effective refinement must therefore account for both observation reliability and target structure.

We propose **AeroGround**, which couples active target disambiguation with evidence-guided 3D support refinement over a standard 3D Gaussian scene. Active Query-guided View Exploration (AQVE) retrieves and verifies query-relevant observations, lifts accepted masks into 3D, and organizes geometrically consistent observations into competing target hypotheses. The hypotheses accumulate visibility-aware positive and negative evidence, which in turn guides the acquisition of new informative views. Once the evidence stabilizes, evidence-based selection produces the initial Gaussian support. Evidence-Guided Structural Refinement (ESR) then consolidates accepted foreground and background observations into Gaussian-level target posteriors. It further combines the fused positive and negative evidence with local geometric consistency to recover missing support and suppress unsupported foreground regions.

To evaluate language-guided grounding in aerial scenes, we introduce AeroRef3D, comprising UrbanBIS-Ref and GauU-Ref across six large-scale aerial scenes, with queries covering zero, single, and multiple targets.

Our contributions are threefold:

- **Framework.** We present AeroGround, a query-time framework for target disambiguation and 3D support refinement in standard 3DGS scenes.
- **Method.** We propose AQVE to resolve competing target hypotheses through hypothesis-guided, visibility-aware evidence acquisition, and ESR to recover complete target support by fusing multi-view foreground and background evidence with local geometric constraints.
- **Benchmark.** We introduce AeroRef3D, a benchmark covering four reasoning types and zero-to-many targets across six aerial scenes, and demonstrate competitive performance on both aerial grounding and open-vocabulary segmentation.

2 Related Work

Semantic and Instance Gaussian Representations. Recent work extends 3D Gaussian Splatting beyond appearance reconstruction by attaching semantic or instance information to Gaussian primitives. Language-aware methods lift features from pretrained vision-language models into 3D for open-vocabulary retrieval and segmentation (Shi et al., 2024; Qin et al., 2024; Zhou et al., 2024). Instance-oriented approaches further organize Gaussians through mask lifting, identity learning, affinity modeling, or object-level codebooks (Ye et al., 2024; Lyu et al., 2026; Cen et al., 2025a; Wu et al., 2024; Li et al., 2025; Wang et al., 2025; Zhu et al., 2025b). Cross-view descriptor aggregation

and interactions between instance and language fields further improve semantic consistency across observations (Cen et al., 2025b; Zhu et al., 2025a). GaussianCut instead refines coarse multi-view segmentations through graph-cut optimization over Gaussian primitives (Jain et al., 2024). These approaches primarily build semantic or instance representations for open-vocabulary retrieval and segmentation, rather than reasoning over complex language queries at inference time. AeroGround instead constructs query-specific Gaussian support directly from multi-view observations and refines it without scene-specific semantic feature optimization.

3D Language Grounding and Multi-View Reasoning. 3D grounding maps natural-language descriptions to scene objects, from single-target referring in ScanRefer (Chen et al., 2020) to zero-, single-, and multi-target grounding in Multi3DRefer (Zhang et al., 2023). In Gaussian scenes, ReferSplat models spatially aware language–Gaussian interactions (He et al., 2025), while ReLaGS incorporates hierarchical semantics and explicit object relations (Xie et al., 2026). Recent MLLM-based methods further exploit scene observations for complex reasoning: ReasonGrounder infers a target and retrieves a top-down view (Liu et al., 2025); REALM aggregates global observations and refines grounding through local views (Shi et al., 2026); ZeroSplat combines geometric keyframes with language-derived spatial anchors (Ding et al., 2026); and SplatReasoner retrieves and synthesizes query-conditioned views (Kim et al., 2026). These methods advance multi-view reasoning, but do not explicitly maintain competing cross-view target hypotheses to drive sequential evidence acquisition. AQVE instead associates observations into competing hypotheses, accumulates visibility-aware positive and negative evidence, and uses the evolving hypotheses to acquire informative views until the evidence stabilizes or the view budget is exhausted.

Aerial 3D Scene Understanding. Aerial 3D understanding introduces additional challenges from large spatial coverage, oblique viewpoints, scale variation, and repeated urban structures. Aerial Lifting transfers semantic and building-instance labels into urban neural radiance fields while addressing scale variation and cross-view inconsistency (Zhang et al., 2024). ID-Splat associates object identities across aerial images and propagates semantic labels through Gaussian representations (Wang et al., 2026). DiLEGS reduces the cost of large-scene language-field construction by decoupling semantic learning from photometric reconstruction and improving cross-view mask consistency (Li et al., 2026). These works establish strong foundations for semantic and instance understanding in large aerial scenes, but primarily target category-level recognition, object association, or language-field construction. AeroGround instead addresses free-form language grounding, where visually similar candidates require cross-view target disambiguation and incomplete observations can leave the resulting 3D support fragmented or contaminated. These challenges are addressed through active evidence acquisition and evidence-guided 3D support refinement.

3 METHODOLOGY

3.1 OVERVIEW

Given a reconstructed 3D Gaussian scene $\mathcal{G}$ with its calibrated multi-view images $\mathcal{V}$, we take a natural-language query q as input and aim to identify the corresponding regions in the 3D scene. Here, the 3D scene is represented by N Gaussians, $\mathcal{G} = \{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^{N}$ (Kerbl et al., 2023), reconstructed from the calibrated observations $\mathcal{V} = \{(I_v, \pi_v)\}_{v=1}^{V}$, where I_v and π_v denote an aerial image and its camera parameters, respectively. The input q describes the target objects or regions, and the output $\mathbf{s}_q \in [0,1]^N$ assigns a membership score to each Gaussian for language grounding.

Our *AeroGround* decomposes this task into two stages: (1) **target disambiguation**, which determines which candidate regions satisfy the query under ambiguous multi-view observations, and (2) **3D support refinement**, which recovers complete and spatially consistent Gaussian support for the selected targets. In the first stage, the key challenge is to distinguish the intended target from visually similar candidates when relevant evidence is distributed across multiple views. **Active Query-guided View Exploration (AQVE)** first retrieves and verifies query-relevant observations and lifts accepted masks into 3D. Geometrically consistent observations are organized into competing target hypotheses, whose accumulated evidence guides the acquisition of new informative views under a bounded view budget. Evidence-based selection then produces the initial Gaussian support $\mathbf{s}_q^0$. Starting from this support, the second stage addresses residual incompleteness and foreground leakage. **Evidence-Guided Structural Refinement (ESR)** first consolidates multi-view foreground

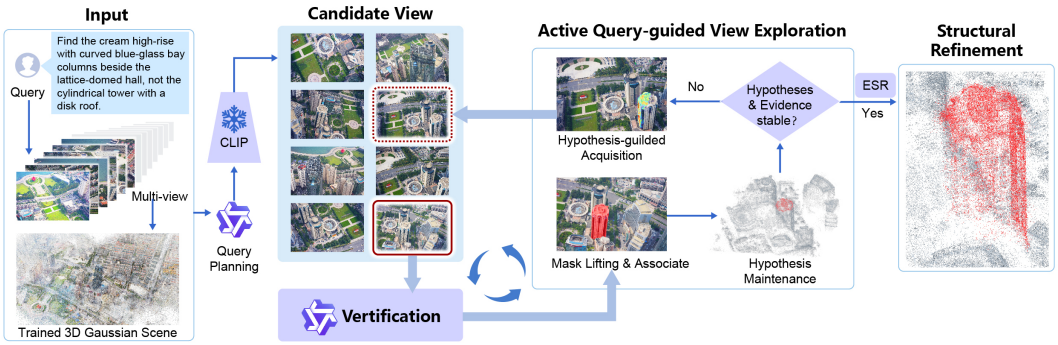

Figure 2: Overview of **AeroGround**. Given a language query and a pretrained aerial 3D Gaussian scene, query planning and CLIP construct a query-conditioned candidate-view pool. AQVE iteratively verifies observations, lifts accepted masks into 3D, maintains competing cross-view target hypotheses, and uses the accumulated evidence to guide subsequent view acquisition. Once the hypotheses and evidence stabilize, the retained hypotheses form the initial Gaussian support, which is further refined by ESR.

and background evidence into Gaussian-level target posteriors, and then incorporates local geometric consistency to recover missing support while suppressing unsupported foreground regions.

3.2 Active Query-guided View Exploration

AQVE resolves target ambiguity through an iterative observation–hypothesis reasoning process. It first retrieves and verifies query-relevant observations and lifts accepted 2D masks to Gaussian support, providing a common 3D basis for cross-view association. Geometrically consistent observations are then organized into competing target hypotheses, which accumulate visibility-aware positive and negative evidence. The current hypotheses guide the acquisition of new informative views, and this hypothesis-update and view-selection loop continues until the evidence stabilizes or the exploration budget is exhausted. Finally, evidence-based selection retains sufficiently supported hypotheses and produces the initial Gaussian support for ESR.

Query-conditioned acquisition. An MLLM interprets q into a canonical target description, target attributes, reference landmarks, spatial constraints, and likely confusers. It also predicts a coarse topology descriptor used during ESR posterior fusion. These complementary cues are used by CLIP (Radford et al., 2021) to score candidate views through multi-scale image-region similarities and construct a query-conditioned view pool. For each acquired view, the MLLM verifies the complete query and produces image-conditioned localization. SAM 3 (Carion et al., 2025) converts each accepted localization into a 2D mask.

Mask lifting. Each accepted 2D mask M_a defines an anchor a in its source view v_a. Let $w_{v_a,i}(\mathbf{x})$ denote the alpha-compositing contribution of Gaussian i to pixel $\mathbf{x}$. We compute its in-mask and total visible contributions as

$$C_{ai}^{\text{in}} = \sum_{\mathbf{x}} M_a(\mathbf{x}) w_{v_a,i}(\mathbf{x}), \qquad C_{ai}^{\text{all}} = \sum_{\mathbf{x}} w_{v_a,i}(\mathbf{x}). \tag{1}$$

We use two complementary quantities. The foreground ratio $r_{ai} = C_{ai}^{\text{in}}/(C_{ai}^{\text{all}} + \epsilon)$ measures how much of the visible Gaussian contribution falls inside the mask. The relative mass $m_{ai} = \min(1, C_{ai}^{\text{in}}/(z_a + \epsilon))$ measures whether this contribution is sufficiently strong, where z_a is the 99.5th percentile of the positive in-mask contributions for anchor a. The lifted support is then

$$s_i^a = \begin{cases} r_{ai}\sqrt{m_{ai}}, & r_{ai} \geq \eta_r \text{ and } m_{ai} \geq \eta_m, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where η_r and η_m are fixed thresholds. The foreground-ratio test suppresses Gaussians that mainly explain pixels outside the target, while the relative-mass test removes negligible or unstable contributions. The Gaussian parameters remain fixed, and no semantic features are trained or stored. The

resulting support $\mathbf{s}^a \in [0,1]^N$ can be rendered into other views to associate observations of the same candidate across viewpoints.

Hypothesis maintenance. To associate target observations across views, we construct a projection graph over the lifted anchors. Two anchors are connected when their masks exhibit sufficient agreement under bidirectional cross-view projection, and geometrically consistent anchors are incrementally grouped into target hypotheses. Each hypothesis h maintains positive and negative evidence (P_h, N_h), initialized to $(1,1)$, with posterior confidence $p_h = P_h/(P_h + N_h)$. After acquiring view v, the evidence is updated as

$$P_h \leftarrow P_h + \Delta^+_{vh}, \qquad N_h \leftarrow N_h + \lambda_- \Delta^-_{vh}. \quad (3)$$

Here, Δ^+_{vh} aggregates confidence-weighted evidence from newly associated anchors. Negative evidence Δ^-_{vh} is admitted only when no matching target is detected and the 3D hypothesis is expected to be sufficiently visible in view v.

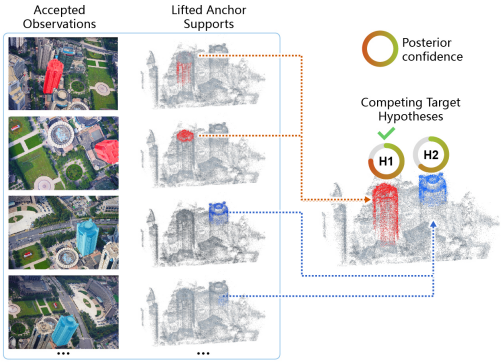

Figure 3: Cross-view hypothesis maintenance. Lifted anchors are associated into competing hypotheses, with accumulated evidence determining posterior confidence.

Hypothesis-guided acquisition. We score each remaining candidate view using up to M maintained hypotheses. The base utility $U_{\text{base}}(v)$ combines retrieval relevance, pose novelty, projected support, and view redundancy. The hypothesis utility $U_{\text{hyp}}(v)$ favors views expected to reduce hypothesis uncertainty or reveal different competing hypotheses. We select

$$v^* = \arg\max_{v \in \mathcal{C}_{\text{rem}}} \left[(1-\lambda_{\text{IG}})U_{\text{base}}(v) + \lambda_{\text{IG}} U_{\text{hyp}}(v)\right], \quad (4)$$

where $\mathcal{C}_{\text{rem}}$ contains the unobserved candidate views. Before any hypothesis is available, selection relies only on U_{base}. AQVE uses a hard upper bound of B verified views and terminates early when the retained hypothesis set and its evidence become stable. It requires one query-planning call and at most B view-verification calls.

Evidence-based selection. At termination, hypotheses whose posterior confidence and associated observation support satisfy the fixed acceptance criteria are retained. Their accepted anchor supports are combined by a pointwise maximum $s^0_{q,i} = \max_{a \in \mathcal{A}_q} s^a_i$. Here $\mathcal{A}_q$ denotes anchors belonging to the retained hypotheses. If no hypothesis is retained, $\mathbf{s}^0_q = \mathbf{0}$. AQVE passes this initial support together with its acquired observations to ESR.

3.3 Evidence-Guided Structural Refinement

ESR refines the initial AQVE support in two successive steps. Gaussian posterior fusion first consolidates the accepted multi-view foreground and background observations into Gaussian-level target posteriors, reducing the effect of noisy or conflicting observations. Signed geometric graph refinement then combines the fused positive and negative evidence with local geometric consistency to recover missing target support and suppress unsupported foreground leakage.

Gaussian posterior fusion. Let $\mathcal{V}_q$ denote the accepted observation views, with masks $\{M_v\}_{v \in \mathcal{V}_q}$. Rendering contributions lift each mask into foreground evidence $e^+_{vi} \in [0,1]$ and reliable background evidence $e^-_{vi} \in [0,1]$ for Gaussian i. Background evidence excludes uncertain mask boundaries and is admitted only when the Gaussian is sufficiently visible.

Each retained AQVE hypothesis defines a target component h for refinement. Let $\mathcal{V}_h \subseteq \mathcal{V}_q$ contain the accepted views used for component h. Fusion uses view-level unions of accepted target masks. For readability, the component index is omitted from e^+_{vi}, e^-_{vi}, and ρ_{vi}. For each Gaussian observed in component h, we compute

$$\bar{p}^h_i = \frac{\sum_{v \in \mathcal{V}_h} \omega_v \rho_{vi} e^+_{vi}}{\max\left(\sum_{v \in \mathcal{V}_h} \omega_v \rho_{vi} \left(e^+_{vi} + e^-_{vi}\right), \epsilon\right)}. \quad (5)$$

Here, $\omega_v \geq 0$ balances observation reliability across camera pose clusters. The robust weight $\rho_{vi} \in [0, 1]$ downweights an observation when its local foreground assignment disagrees with the current multi-view estimate. The constant $\epsilon > 0$ ensures numerical stability.

For each retained target component, we select one representative observation to preserve reliable foreground and background evidence. The remaining uncertain probability mass is then completed using the multi-view estimate $\bar{p}_i^h$:

$$p_i^h = e^+_{r_h i} + \left[1 - e^+_{r_h i} - e^-_{r_h i}\right]_+ \bar{p}_i^h, \tag{6}$$

where r_h is the representative observation of component h, and $[x]_+ = \max(x, 0)$. For non-enclosed targets, additional positive observations may retain visible support absent from the representative view. For multiple retained components, fusion is performed independently within each component and the resulting posteriors are combined as $p_i = \max_{h \in \mathcal{H}_q} p_i^h$, where $\mathcal{H}_q$ is the set of retained target components.

Signed geometric graph refinement. We construct a local k-nearest-neighbor graph from Gaussian positions around the posterior support, where "signed" refers to the use of both positive and negative observation evidence. Spatial proximity defines graph connectivity and pairwise affinity, while opacity and the fused positive and negative observations determine node reliability. Reliable foreground and background Gaussians provide fixed foreground and background constraints. Uncertain and weakly supported Gaussians are grouped into mutable components $\mathcal{C}_G$. The binary labels of these components are inferred by minimizing

$$E(\mathbf{y}) = \sum_{c \in \mathcal{C}_G} \phi_c(y_c) + \lambda_g \sum_{(c,d) \in \mathcal{E}_G} W_{cd} |y_c - y_d|, \tag{7}$$

where $y_c \in \{0, 1\}$ denotes background or foreground. The unary costs are $\phi_c(0) = A_c^+$ and $\phi_c(1) = A_c^-$, where A_c^+ and A_c^- aggregate foreground and background evidence, respectively. $\mathcal{E}_G$ contains neighboring component pairs, $W_{cd} \geq 0$ is their spatial affinity, and λ_g controls pairwise consistency. We minimize this binary energy using graph cut (Kolmogorov & Zabih, 2004).

Graph-cut labels update the local Gaussian support using positive and negative multi-view evidence: supported foreground components are added, whereas weak foreground components contradicted by negative evidence are removed. Gaussians outside the mutable region retain their posterior values p_i, while the Gaussian geometry and appearance remain fixed. The resulting support is denoted by $\mathbf{s}_q$. Its rendered mask in view v is

$$\widehat{M}_v(\mathbf{x}) = \mathbb{I}\left[\sum_{i=1}^{N} w_{v,i}(\mathbf{x}) s_{q,i} \geq \tau_v\right], \tag{8}$$

where $\mathbf{x}$ denotes a pixel, $w_{v,i}(\mathbf{x})$ is the alpha-compositing contribution of Gaussian i, and τ_v is a rendering threshold calibrated from the accepted observation masks. Additional definitions are provided in the appendix.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate language grounding on AeroRef3D, which consists of UrbanBIS-Ref (Qingdao, Lihu, and Wuhu) and GauU-Ref (Campus, SMBU, and SZIIT). The benchmark contains zero-, single-, and multi-target queries covering spatial–attribute, commonsense, functional, and hypothetical reasoning. For open-vocabulary segmentation, we evaluate on two benchmarks. We use four scenes from UrbanBIS (Yang et al., 2023): Yingrenshi, Yuehai, Longhua-1, and Longhua-2. We also evaluate on 3D-AS (Tang et al., 2025), covering City, Country, and Port environments. We provide details of AeroRef3D in the appendix.

Baselines and metrics. We compare against two groups of baselines. The first includes open-vocabulary GS segmentation methods (Shi et al., 2024; Wu et al., 2024; Li et al., 2025; Wang et al., 2025; Zhu et al., 2025b;a): LEGaussians, OpenGaussian, InstanceGaussian, AG^{2}S, Unified-Lift,

Table 1: Quantitative comparison of language grounding on AeroRef3D. The top three results are highlighted with red, orange, and yellow backgrounds, respectively.

Method	UrbanBIS-Ref						GauU-Ref					
	Qingdao		Lihu		Wuhu		Campus		SMBU		SZIIT	
	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑	mIoU↑	mBIoU↑
LEGS	33.59	32.44	23.91	20.69	32.15	25.06	38.42	32.95	25.05	22.70	34.03	25.55
OpenGS	24.49	23.09	27.12	22.01	12.97	11.98	27.61	26.54	19.23	17.57	19.63	18.02
InsGS	10.99	9.61	10.05	6.36	17.14	11.71	9.11	4.86	10.97	5.75	6.50	4.83
AG^{2}S	37.41	29.14	20.71	16.85	20.10	17.55	28.47	18.13	23.33	16.82	29.37	22.81
Unified-Lift	37.23	31.54	35.68	28.52	20.26	17.54	39.81	29.95	21.63	16.74	34.82	25.44
COS3D	31.62	17.66	22.79	17.74	29.85	19.74	23.49	15.09	19.44	14.04	24.77	12.92
ReasonGrounder	22.97	15.68	12.56	8.98	14.94	10.52	24.14	16.27	28.28	21.98	29.80	20.04
ReLaGS	32.11	27.43	14.05	10.94	19.84	16.18	34.45	25.87	41.30	33.84	35.03	26.22
REALM	34.04	28.73	10.75	6.98	19.93	16.13	40.59	35.34	54.85	46.22	19.35	15.75
Ours	81.53	69.63	78.17	68.19	75.79	66.85	86.04	74.49	71.54	62.70	85.10	72.08

Table 2: Open-vocabulary segmentation results on UrbanBIS and 3D-AS. We report mIoU. The top three results are highlighted with red, orange, and yellow backgrounds, respectively.

Method	UrbanBIS					3D-AS			
	Yingrenshi	Yuehai	Longhua-1	Longhua-2	Avg.	City	Country	Port	Avg.
LEGS	33.18	26.37	26.39	26.67	28.15	16.39	13.85	11.02	13.75
OpenGS	32.69	42.78	38.45	33.65	36.89	15.31	13.32	18.32	15.65
InsGS	28.88	30.61	30.91	28.25	29.66	12.81	13.90	14.65	13.79
AG^{2}S	33.33	20.68	43.34	39.64	34.25	21.17	22.79	14.00	19.32
Unified-Lift	33.41	32.44	31.06	30.80	31.93	30.06	19.91	47.80	32.59
COS3D	10.53	21.04	18.86	15.80	16.56	18.40	17.85	15.01	17.09
ReasonGrounder	20.52	12.97	29.87	22.48	21.46	19.42	14.33	19.19	17.65
ReLaGS	9.90	10.81	20.13	11.74	13.15	20.80	17.10	28.10	22.00
REALM	24.54	25.62	27.38	24.29	25.46	48.28	39.75	46.70	44.91
Ours	75.67	69.01	72.23	73.44	72.59	67.90	55.14	65.75	62.93

and COS3D. The second includes language-guided GS grounding methods: ReasonGrounder, ReLaGS, and REALM (Liu et al., 2025; Xie et al., 2026; Shi et al., 2026). We evaluate these methods across both language grounding and open-vocabulary segmentation benchmarks. Following previous works (Kerr et al., 2023; Qin et al., 2024; Shi et al., 2024; Lyu et al., 2026), we report mIoU and mBIoU to quantitatively evaluate the quality of 3D grounding and segmentation results.

Implementation details. We represent each scene using a standard 3DGS optimized for 30k iterations. We use Qwen3.5-35B-A3B (Qwen Team, 2026) as the reasoning model and SAM 3 (Carion et al., 2025) for mask prediction. For contribution-aware lifting, we set the foreground-ratio threshold to $\eta_r = 0.55$ and the relative-mass threshold to $\eta_m = 0.01$. For AQVE, the hard view budget is $B = 24$, and up to $M = 5$ hypotheses are used to evaluate view utility. We set the hypothesis-utility weight to $\lambda_{\text{IG}} = 0.50$ and the negative-evidence weight to $\lambda_- = 0.25$. For ESR, we use $k = 10$ spatial neighbors and set the graph pairwise-consistency weight to $\lambda_g = 1.0$. All experiments are conducted on an NVIDIA L40 GPU.

4.2 Experimental Results

Language grounding. Table 1 shows that AeroGround achieves the highest mIoU and mBIoU across all six AeroRef3D scenes. Open-vocabulary baselines struggle with attribute and relational constraints, while grounding and reasoning baselines still suffer from target confusion and incomplete support. As shown in Fig. 4, AeroGround improves both target localization and support completeness. Following PointGS (Song et al., 2026), we further transfer the Gaussian-level target support to the original scene point clouds and visualize the grounding results in Fig. 5.

Open-vocabulary segmentation. Table 2 shows that AeroGround achieves the highest average mIoU on UrbanBIS (72.59) and 3D-AS (62.93), demonstrating that the framework generalizes beyond language grounding to category-level open-vocabulary segmentation. Qualitative comparisons on both benchmarks are provided in Appendix E.



Figure 4: Qualitative comparison on AeroRef3D. Pink overlays denote the predicted and ground-truth regions for the queries shown on the left. AeroGround yields more complete and cleaner grounding results with less background leakage than the compared methods.

Table 3: Ablation studies on UrbanBIS-Ref. (a) Cumulative contributions of AQVE and ESR, with retrieval-based controls matched to AQVE in per-query view counts. (b) Analysis of AQVE observation acquisition with 12 views per query. Grounding variants share the same retrieved views, whereas retrieval variants share the same downstream MLLM-grounded SAM 3 pipeline.

(a) Component contributions

Variant	mIoU	mBIoU
Fixed retrieval + mask lifting	32.80	30.48
+ Query-conditioned acquisition	64.90	53.66
+ Hypothesis maintenance	71.58	59.83
Full AQVE	**74.47**	**62.56**
+ Gaussian posterior fusion	76.76	66.88
Full AeroGround	**78.49**	**68.22**

(b) AQVE observation analysis

Factor	Variant	mIoU	mBIoU
Grounding	Raw query	29.73	28.07
	Query planning	47.10	41.34
	MLLM grounding	**65.82**	**56.81**
Retrieval	View clustering	40.97	36.19
	Raw-query CLIP	65.51	56.65
	Multi-cue retrieval	**68.40**	**58.47**

4.3 Ablation Studies

Component contributions. Table 3a reports cumulative ablations on UrbanBIS-Ref. For a fair comparison, retrieval-based controls use matched per-query view counts. Starting from fixed retrieval and mask lifting, query-conditioned acquisition provides the largest gain by selecting observations according to the query, while hypothesis maintenance and hypothesis-guided exploration further improve the grounding quality through cross-view association and active evidence acquisition. Starting from the full AQVE output, Gaussian posterior fusion improves the initial 3D support, and geometric graph refinement further recovers missing regions while suppressing unsupported foreground. Together, these components yield the final AeroGround result of 78.49 mIoU and 68.22 mBIoU, confirming the complementary contributions of AQVE and ESR.



Figure 5: Qualitative comparison of language-guided 3D grounding on AeroRef3D. Representative examples from Campus and Wuhu show more accurate target localization and more complete 3D segmentation with reduced background leakage compared with REALM.

AQVE observation analysis. Table 3b separately evaluates per-view grounding and initial view retrieval using 12 views per query, before hypothesis filtering and ESR. For the grounding ablation, we fix the same multi-cue-retrieved views and compare raw-query prompting, planned-text prompting, and image-conditioned MLLM grounding followed by SAM 3. All grounding variants use the same SAM 3 score threshold of 0.5. For the retrieval ablation, we instead fix the downstream MLLM-grounded SAM 3 pipeline and compare view clustering selection, raw-query CLIP retrieval, and planned multi-cue retrieval, with the SAM 3 score threshold fixed to zero for all retrieval variants. MLLM-grounded segmentation substantially outperforms the text-only variants, highlighting the benefit of image-conditioned target verification and localization. Multi-cue retrieval further improves over both view clustering selection and raw-query CLIP retrieval, supporting the use of structured query cues for view selection.

5 Conclusion

We presented AeroGround, a query-time framework for language-guided 3D grounding in large-scale aerial Gaussian scenes. AeroGround addresses target disambiguation and 3D support refinement through two complementary components. AQVE maintains competing cross-view hypotheses and actively acquires informative observations using visibility-aware evidence, while ESR consolidates multi-view foreground and background evidence with local geometric consistency to recover missing support and suppress foreground leakage. We also introduced AeroRef3D, a benchmark covering diverse reasoning types and zero-to-many target queries across large-scale aerial scenes. Experiments on AeroRef3D and two aerial open-vocabulary segmentation benchmarks demonstrate consistent improvements over the evaluated baselines.

AI USE STATEMENT

We used large language models to assist with drafting, revising, and polishing the manuscript. AeroGround uses Qwen3.5-35B-A3B for query planning and view-level target localization and verification. We additionally evaluated alternative Qwen and Gemini models in the model comparison experiments. For AeroRef3D construction, GPT-5.6 assisted in generating candidate queries, reasoning types, target descriptions, and target counts, all of which were subsequently reviewed and corrected by human annotators. The authors reviewed the AI-assisted manuscript content and take responsibility for the final paper, dataset, and reported results.

ETHICS STATEMENT

This work presents AeroGround, a framework for language-guided 3D grounding in large-scale aerial Gaussian scenes. The research focuses on methodological advances in multi-view reasoning and 3D target localization. All experiments use publicly available aerial scene datasets and involve no personal identity information or sensitive personal attributes. The proposed benchmark focuses on scene-level objects, regions, and their spatial or functional relationships. The authors encourage responsible, legal, and privacy-preserving use of this technology, particularly in applications involving aerial imagery.

REFERENCES

Black Forest Labs. FLUX.2 [klein] 4B. Hugging Face model card, 2026. URL `https://huggingface.co/black-forest-labs/FLUX.2-klein-4B`.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3D Gaussians. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (2):1971–1979, 2025a.

Jiazhong Cen, Xudong Zhou, Jiemin Fang, Changsong Wen, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Tackling view-dependent semantics in 3D language Gaussian splatting. In *ICML*, pp. 7013–7034, 2025b.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *ECCV*, pp. 202–221, 2020.

Jiayu Ding, Meilu Song, Xiaoyi Zhang, Hongbo Jin, Yichen Jin, and Xiangtian Si. ZeroSplat: Generalized referring segmentation in 3D Gaussian splatting. In *ECCV*, pp. 147–163, 2026.

Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. ReferSplat: Referring segmentation in 3D Gaussian splatting. In *ICML*, pp. 22456–22467, 2025.

Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges. In *CVPR*, pp. 4977–4987, 2021.

Umangi Jain, Ashkan Mirzaei, and Igor Gilitschenski. GaussianCut: Interactive segmentation via graph cut for 3D Gaussian splatting. In *NeurIPS*, pp. 89184–89212, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *ICCV*, pp. 19729–19739, 2023.

Yu-Ji Kim, Dahye Lee, Jun-Seong Kim, Hyeon-Woo Nam, GeonU Kim, Yongjin Kwon, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. SplatReasoner: Enhancing embodied reasoning and grounding by novel view synthesis. In *ECCV*, pp. 133–151, 2026.

Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. InstanceGaussian: Appearance-semantic joint Gaussian representation for 3D instance-level perception. In *CVPR*, pp. 14078–14088, 2025.

Zhida Li, Jianqiao Zhu, Hejin Huang, Yipeng Qin, Sibei Yang, and Guanbin Li. Direct language embedding enables Gaussian splatting for large scenes. In *CVPR Findings*, pp. 7231–7240, 2026.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. ReasonGrounder: LVLM-guided hierarchical feature splatting for open-vocabulary 3D visual grounding and reasoning. In *CVPR*, pp. 3718–3727, 2025.

Weijie Lyu, Xueting Li, Abhijit Kundu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Gaga: Group any Gaussians via 3D-aware memory bank. *Transactions on Machine Learning Research*, 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI blog, July 2026. URL https://openai.com/index/gpt-5-6/.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3D language Gaussian splatting. In *CVPR*, pp. 20051–20060, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Changyue Shi, Minghao Chen, Yiping Mao, Chuxiao Yang, Xinyuan Hu, Jiajun Ding, and Zhou Yu. REALM: An MLLM-agent framework for open world 3D reasoning segmentation and editing on Gaussian splatting. In *CVPR*, pp. 16779–16788, 2026.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3D Gaussians for open-vocabulary scene understanding. In *CVPR*, pp. 5333–5343, 2024.

Yixiao Song, Qingyong Li, Wen Wang, and Zhicheng Yan. PointGS: Semantic-consistent unsupervised 3D point cloud segmentation with 3D Gaussian splatting. In *CVPR*, pp. 33343–33352, 2026.

Xu Tang, Junan Jia, Yijing Wang, Jingjing Ma, and Xiangrong Zhang. Semantic-aware DropSplat: Adaptive pruning of redundant Gaussians for 3D aerial-view segmentation. arXiv preprint arXiv:2508.09626, 2025.

Yijing Wang, Xu Tang, Xiangrong Zhang, and Jingjing Ma. ID-Splat: Propagating object identities for segmenting 3D aerial-view scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):10261–10269, 2026.

Zhaonan Wang, Manyi Li, and Changhe Tu. AG2aussian: Anchor-graph structured Gaussian splatting for instance-level 3D scene understanding and editing. In *ICCV*, pp. 26806–26816, 2025.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, and Jian Zhang. OpenGaussian: Towards point-level 3D Gaussian-based open vocabulary understanding. In *NeurIPS*, pp. 19114–19138, 2024.

Yaxu Xie, Abdalla Arafa, Alireza Javanmardi, Christen Millerdurai, Jia Cheng Hu, Shaoxiang Wang, Alain Pagani, and Didier Stricker. ReLaGS: Relational language Gaussian splatting. In *CVPR*, pp. 23826–23836, 2026.

Butian Xiong, Nanjun Zheng, Junhua Liu, and Zhen Li. GauU-Scene V2: Assessing the reliability of image-based metrics with expansive LiDAR image dataset using 3DGS and NeRF. *arXiv preprint arXiv:2404.04880*, 2024.

Runjie Yan, Yan-Pei Cao, Peng Wang, Ding Liang, and Yuan-Chen Guo. Generative 3D Gaussians with learned density control. In *SIGGRAPH*, pp. 58:1–58:12, 2026.

Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. UrbanBIS: a large-scale benchmark for fine-grained urban building instance segmentation. In *SIGGRAPH*, pp. 16:1–16:11, 2023.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and edit anything in 3D scenes. In *ECCV*, pp. 162–179, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding text description to multiple 3D objects. In *ICCV*, pp. 15225–15236, 2023.

Yuqi Zhang, Guanying Chen, Jiaxing Chen, and Shuguang Cui. Aerial Lifting: Neural urban semantic and building instance lifting from aerial imagery. In *CVPR*, pp. 21092–21103, 2024.

Qinming Zhou, Chenxi Sun, Deyang Kong, Junhao He, Xiangheng Tang, Peike Yu, Haotian Wu, Leilei Cao, and Linfeng Zhang. OSOR: One-step diffusion inpainting for effect-aware object removal. In *ECCV*, pp. 367–384, 2026.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3DGS: Supercharging 3D Gaussian splatting to enable distilled feature fields. In *CVPR*, pp. 21676–21685, 2024.

Runsong Zhu, Ka-Hei Hui, Zhengzhe Liu, Qianyi Wu, Weiliang Tang, Shi Qiu, Pheng-Ann Heng, and Chi-Wing Fu. COS3D: Collaborative open-vocabulary 3D segmentation. In *NeurIPS*, pp. 11351–11371, 2025a.

Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, and Chi-Wing Fu. Rethinking end-to-end 2D to 3D scene segmentation in Gaussian splatting. In *CVPR*, pp. 3656–3665, 2025b.

A AeroRef3D Benchmark

Task definition and query types. AeroRef3D evaluates referring and reasoning segmentation in aerial 3D scenes. Given a natural-language query, the task is to identify the referred scene elements and recover their spatial extent. Queries cover four reasoning types: *spatial–attribute* queries combine visible attributes with spatial relations; *commonsense* queries require knowledge about objects and their surroundings; *functional* queries identify targets through their uses or task requirements; and *hypothetical* queries identify existing objects under imagined conditions. Hypothetical queries do not require generating new scene content. Reasoning type and scene-level target cardinality are annotated separately: a query may refer to zero, one, or multiple objects.

Dataset composition. AeroRef3D is constructed on six large-scale aerial scenes from UrbanBIS and GauU-Scene, extending category-level segmentation to referring and reasoning segmentation with zero-, single-, and multi-target queries. Spatial–attribute and functional queries account for roughly half and one third of the annotations, respectively, with the remaining queries covering commonsense and hypothetical reasoning. UrbanBIS-Ref is constructed from the Qingdao, Lihu, and Wuhu scenes of UrbanBIS (Yang et al., 2023). For Qingdao, we use a selected subregion of the original scene, referred to as Qingdao throughout the paper. GauU-Ref is constructed from the Lower Campus (abbreviated as Campus), SMBU, and SZIIT scenes of GauU-Scene (Xiong et al., 2024). Each annotated pair consists of an evaluation image, a language query, and a binary target mask. For a multi-target query, each mask contains the visible regions of all referred objects in that image. Table 4 summarizes the task settings and query coverage of related aerial benchmarks. Zero-target queries refer to cases where no object in the scene satisfies the query; an empty mask in an individual view does not by itself indicate a zero-target query.

Annotation and quality control. Figure 6 illustrates the annotation pipeline. We use GPT-5.6 Sol Ultra (OpenAI, 2026) to jointly analyze multiple aerial views of the same scene and generate candidate annotations, including natural-language queries, reasoning types, target descriptions, and target counts. The generated target descriptions are then used to prompt SAM 3 (Carion et al., 2025) for candidate masks. All generated annotations are reviewed and corrected by a team of four human annotators, who verify query meaning, target identity and count, cross-view consistency, and mask boundaries. Ambiguous queries and incorrect target descriptions are revised, while inaccurate masks are manually refined. For hypothetical queries, human verification additionally ensures that the referred targets are existing scene objects consistent with the stated hypothetical condition. Only human-verified queries and masks, with descriptions and mask boundaries corrected where necessary, are retained in the final benchmark.

Evaluation protocol. All methods use the same training/test split, with test images reserved solely for evaluation. AeroGround constructs its grounding predictions from the training images and renders them at the test camera poses for comparison with the human-verified reference masks. For a predicted binary mask P and its ground-truth mask G, we use

$$\mathrm{IoU}(P,G)=\frac{|P\cap G|}{|P\cup G|},\qquad \mathrm{BIoU}(P,G)=\frac{|\mathcal{B}(P)\cap\mathcal{B}(G)|}{|\mathcal{B}(P)\cup\mathcal{B}(G)|}. \tag{9}$$

Table 4: Comparison of aerial 3D segmentation benchmarks under the evaluation settings used in this work. AeroRef3D supports zero-, single-, and multi-target queries across four reasoning types. "Words" denotes category-name prompts, whereas "Inst." denotes natural-language instructions. UrbanBIS statistics refer to the four evaluated regions. *Open-Vocab.*: open-vocabulary segmentation; *Ref.*: explicit referring expressions; *S–A*: spatial–attribute; *Com.*: commonsense; *Func.*: functional; *Hyp.*: hypothetical reasoning.

Dataset	Task	Target Setup	Scenes	Query Scale	Open-Vocab.	Referring and Reasoning				
						Ref.	S–A	Com.	Func.	Hyp.
3D-AS	Semantic Seg.	Category	9	630 Words	✓	–	–	–	–	–
UrbanBIS	Semantic Seg.	Category	4	176 Words	✓	–	–	–	–	–
AeroRef3D (Ours)	**Reasoning Seg.**	**Arbitrary**	**6**	**336 Inst.**	✓	✓	✓	✓	✓	✓

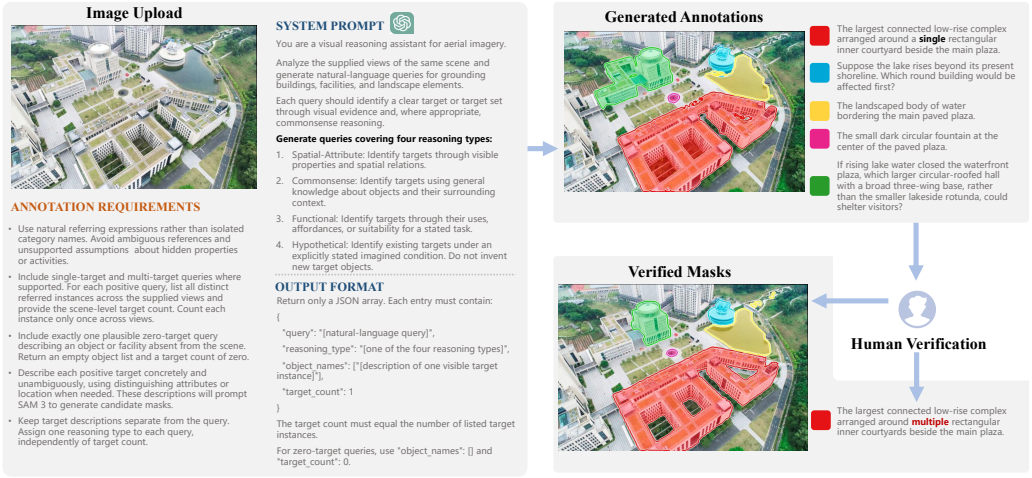

Figure 6: Annotation pipeline of AeroRef3D. GPT-5.6 Sol Ultra first generates candidate language annotations, including queries, reasoning types, target descriptions, and target counts. SAM 3 then generates candidate masks from the target descriptions. Human annotators verify and correct queries, target identities, counts, and mask boundaries to obtain the final annotations.

Here, $\mathcal{B}(M) = M \setminus \mathrm{Erode}(M, d_b)$ denotes the inner boundary band. Following Boundary IoU (Kerr et al., 2023; Qin et al., 2024; Shi et al., 2024; Lyu et al., 2026), we set $d_b = \max(1, \mathrm{round}(0.02\sqrt{H^2 + W^2}))$, where H and W are the image dimensions. Erosion is applied for d_b iterations using a 3×3 neighborhood with zero padding. Both scores are defined as one when both masks are empty, and zero when exactly one mask is empty. We average scores over annotated image–query pairs within each scene to obtain mIoU and mBIoU. Dataset-level results are unweighted averages of the constituent scene scores. Unless stated otherwise, evaluation includes both empty and nonempty ground-truth masks.

B Additional Method Details

We follow the main-paper notation: a indexes observation anchors, h target hypotheses, i Gaussians, and c mutable graph components.

B.1 AQVE

Association and evidence. Let P_a^v be the binary rendering of anchor a in view v, and $\mathcal{A}_h$ be the anchor set of hypothesis h. We associate a new anchor with the best-matching hypothesis using

$$\kappa(a,b) = \sqrt{\mathrm{IoU}(P_a^{v_b}, M_b)\,\mathrm{IoU}(P_b^{v_a}, M_a)}, \qquad \kappa(a,h) = \max_{b \in \mathcal{A}_h} \kappa(a,b). \tag{10}$$

Matches require $\kappa(a,h) \geq 0.05$; connected hypotheses merge with one $(1,1)$ prior, and unmatched anchors initialize hypotheses. A relation-supported fallback requires explicit query relations, at least two functional requirements, exactly two anchors from distinct single-anchor views, confidence ≥ 0.9, satisfied observable relations, and pairwise pose novelty ≥ 0.35. An anchor of confidence c_a contributes c_a, $c_a(1 + \kappa(a,h))/2$, or $0.75c_a$ for initialization, geometric association, or fallback association, respectively.

For rendered support mass $L_a(v)$ over candidate pool $\mathcal{C}$, set $\nu_a(v) = L_a(v)/(\max_{u \in \mathcal{C}} L_a(u) + \epsilon)$ and $\nu_h(v) = 1 - \prod_{a \in \mathcal{A}_h}(1 - \nu_a(v))$. Only a no-target observation contributes negative evidence: $\Delta_{vh}^- = c_v^- \nu_h(v) \mathbb{I}[\nu_h(v) \geq 0.60]$, where c_v^- is absence confidence. The factor λ_- is applied once in Eq. 3; supporting-view counts use distinct cameras.

View utility. Let $\widetilde{r}_v, \widetilde{d}_v$ be min–max normalized retrieval and contrastive scores over $\mathcal{C}$, assigning one to constant vectors. The contrastive score is $d_v = \max(d_v^{\text{target}}, d_v^{\text{alias}}) + 0.5 d_v^{\text{query}} - d_v^{\text{confuser}}$,

using the corresponding text–image similarities. With observed views $\mathcal{O}_t$, camera centers $\mathbf{o}_v$, direction angles θ in degrees, and median positive nearest-camera distance d_{scene}, define

$$\begin{aligned} R(v) &= \widetilde{r}_v + 0.5[\widetilde{d}_v - \widetilde{r}_v]_+ + 5 \times 10^{-7}\widetilde{d}_v, \\ D_t(v) &= 1 - \exp\left[-\min_{u \in \mathcal{O}_t}\left(\frac{\|\mathbf{o}_v - \mathbf{o}_u\|_2}{d_{\text{scene}}} + \frac{\theta(v,u)}{45}\right)\right], \\ S_t(v) &= 1 - \prod_{a \in \mathcal{A}_t}(1 - \nu_a(v)), \\ U_{\text{base}}(v) &= w_R R(v) + w_D D_t(v) + w_S S_t(v) - w_{\text{red}}(1 - D_t(v)). \end{aligned} \tag{11}$$

Here $\mathcal{A}_t$ contains acquired anchors; initially $D_t = 1$, and pairwise novelty uses one comparison view. The weights $(w_R, w_D, w_S, w_{\text{red}})$ are $(0.78, 0.27, 0, 0.06)$ before the first anchor and $(3.12, 0.22, 0.72, 0.12)$ thereafter.

Let $\mathcal{H}_M$ be the top-M hypotheses ranked by p_h, breaking ties by supporting-view count and P_h. For binary entropy $\mathsf{H}(p) = -p\log_2 p - (1-p)\log_2(1-p)$, with $0\log_2 0 = 0$, and its mean $\overline{H}$ over $\mathcal{H}_M$,

$$U_{\text{hyp}}(v) = \frac{0.7}{|\mathcal{H}_M|}\sum_{h \in \mathcal{H}_M}\nu_h(v)\mathsf{H}(p_h) + 0.3\left(\max_h \nu_h(v) - \min_h \nu_h(v)\right)\overline{H}. \tag{12}$$

The extrema use $\mathcal{H}_M$. Eq. 4 uses $M = 5$, $B = 24$, $\lambda_{\text{IG}} = 0.5$, and $\lambda_- = 0.25$; before any hypothesis exists, selection uses only U_{base}.

Acceptance and stopping. With n_h distinct supporting views, retain

$$\mathcal{H}_q = \{h : p_h \geq 0.62 \text{ and } (n_h \geq 2 \text{ or } J_h = 1)\}. \tag{13}$$

$J_h = 1$ requires exactly one anchor, confidence ≥ 0.9, satisfied relational checks when applicable, and mask-area fraction at least $(0.00005, 0.00050, 0.00200, 0.00800)$ for (tiny, small, medium, large) targets; mixed/unspecified scale uses 0.00050.

For stopping, $\mathcal{S}_t$ lists currently eligible hypotheses in $\mathcal{H}_M$, retaining only the highest-ranked one for single-target scope. Stability means a nonempty list unchanged across two consecutive iterations. Let H_b be its mean binary entropy and $H_c = -\sum_h \widehat{p}_h \log_2 \widehat{p}_h / \log_2 |\mathcal{S}_t|$, where $\widehat{p}_h = p_h / \sum_g p_g$ and $H_c = 0$ for a singleton. Use $H_{\text{stop}} = (H_b + H_c)/2$ for single-target scope, and H_b otherwise. Stop at budget B or pool exhaustion; early stopping additionally requires stability and either: (i) at least three verified views with $H_{\text{stop}} \leq 0.3$; or (ii) remaining utility below 0.16, with single-target scope or no unresolved hypothesis in $\mathcal{H}_M$. An unresolved hypothesis has $p_h \geq 0.5$, $n_h < 2$, and $J_h = 0$; remaining utility uses the ranking that selected the current view.

B.2 ESR

Observation evidence and fusion. M_v unions accepted target masks in view v. For hypothesis h, the representative r_h maximizes union-mask area over its associated views. Normally $\mathcal{V}_h = \mathcal{V}_q$; all-target queries with multiple hypotheses instead use each hypothesis's views and restrict its posterior to its lifted anchor-support union before taking the maximum over h.

Evidence covers initial support and its one-hop neighborhood, requiring opacity ≥ 0.1 for expanded nodes. Let $t_{vi} = \sum_{\mathbf{x}} w_{v,i}(\mathbf{x})$; f_{vi} and b_{vi} are contribution fractions inside M_v and the background outside M_v and its two-pixel boundary band. With σ_v the median positive t_{vi} in this domain,

$$\begin{aligned} \gamma_{vi} &= 1 - \exp(-t_{vi}/\sigma_v), \\ e^+_{vi} &= \gamma_{vi} f_{vi}, \qquad e^-_{vi} = \gamma_{vi} b_{vi} \mathbb{I}[\gamma_{vi} \geq 0.15]\mathbb{I}[b_{vi} \geq 0.70]. \end{aligned} \tag{14}$$

Both values are zero for $t_{vi} \leq \epsilon$. Representative views form separate unit-weight pose clusters. Other views are greedily clustered within four median nearest-camera distances and 12 degrees, using directions toward the initial-support median; minimize normalized spatial-plus-angular distance and update cluster centers after assignment. For non-reference views, let $c_v = \sqrt{c_v^{\text{seg}} c_v^{\text{assoc}}}$, using maximum retained SAM 3 confidence and maximum positive anchor-to-mask cross-view IoU.

Fusion uses

$$\omega_v = \frac{c_v \max_{u\in\mathcal{P}(v)} c_u}{\max(\sum_{u\in\mathcal{P}(v)} c_u, \epsilon)}, \qquad \rho_{vi}^{(t+1)} = \left[1 + \left(\frac{o_{vi} - \bar{p}_i^{h,(t)}}{0.25}\right)^2\right]^{-1}, \tag{15}$$

where $o_{vi} = e_{vi}^+ / \max(e_{vi}^+ + e_{vi}^-, \epsilon)$. Alternate two reweighting iterations with Eq. 5, starting from unit weights; zero-evidence observations receive zero weight. Unobserved initial-support nodes retain foreground support, while unobserved expanded nodes start at 0.5.

Set $z_i = \sum_v \omega_v e_{vi}^+ / \max(\sum_v \omega_v \gamma_{vi}, \epsilon)$ and $u_i = \mathsf{H}(z_i)$, using $z_i = 0.5$ when unobserved. If at least two pose clusters observe i and $u_i \geq 0.9$, replace $\bar{p}_i^h$ in Eq. 6 by $\max(\bar{p}_i^h, 0.5)$ for $\alpha_i > 0.1$, or by 0.5 otherwise. Non-enclosed targets additionally retain $\max_{v\in\mathcal{V}_h} e_{vi}^+$ through a pointwise maximum.

Local graph and components. Within each pose cluster, normalized ω_v gives mean evidence $E_{\ell i}^{\pm}$. A cluster observes i when their sum exceeds ϵ, and votes for a sign when its fraction is at least 0.6. Let n_i^{vis} count observing clusters and $q_i^{\pm}$ be their vote fractions. The graph data probability d_i uses robust fusion, admitting negative evidence only with at least $\max(2, \lceil 0.25 n_i^{\text{vis}} \rceil)$ negative votes. Reliability is $r_i^G = 1 - u_i$; outside the evidence domain, $(d_i, r_i^G) = (0.5, 0)$.

Construct a four-hop k-nearest-neighbor graph around $\{i : p_i \geq 0.5\}$, with $k = 10$. Let ℓ_i be the kth-neighbor distance, floored at the first percentile of positive distances. Retain mutual-neighbor edges with normalized distance $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 / \sqrt{\ell_i \ell_j} \leq 0.8$ and affinity $a_{ij} = \exp(-\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 / (2\ell_i \ell_j))$. Fix foreground nodes satisfying $(p_i, d_i, r_i^G) \geq (0.5, 0.9, 0.5)$ and background nodes in the evidence domain with $p_i \leq 0.2, u_i \leq 0.5$. Mutable nodes comprise unfixed foreground nodes, plus non-foreground nodes of opacity ≥ 0.1 outside the fixed-background set. Their edges satisfying $a_{ij} / \sqrt{a_i^{\max} a_j^{\max}} \geq 0.5$ define connected components, including singletons; $a_i^{\max}$ is maximum incident affinity. Between components, $W_{cd} = \max_{(i,j)\in\mathcal{E}_{cd}} a_{ij} \log(1 + |\mathcal{E}_{cd}|)$. We use $\lambda_g = 1$.

Component evidence. For $O_c = \{i \in c : n_i^{\text{vis}} > 0\}$, define

$$\begin{aligned} Q_c^+ &= |c|^{-1} \sum_{i\in c} r_i^G d_i, \quad Q_c^- = |c|^{-1} \sum_{i\in c} r_i^G (1 - d_i), \\ V_c^{\pm} &= \frac{\sum_{i\in O_c} q_i^{\pm}}{\max(|O_c|, 1)}, \qquad F_c = |c|^{-1} |\{i \in c : p_i \geq 0.5\}|. \end{aligned} \tag{16}$$

For each adjacent fixed node j, retain $b_{cj} = \max_{i\in c} a_{ij}$. Let $n_c^{\pm}$ count foreground/background contacts, $\bar{b}_c^{\pm}$ be their mean weights, and $\mathbf{t}_{cj}$ be the unit direction from the component centroid toward j. The foreground directional balance is $D_c^G = 1 - \|\sum_{j\in\mathcal{N}_c^+} b_{cj} \mathbf{t}_{cj} / \max(\sum_{j\in\mathcal{N}_c^+} b_{cj}, \epsilon)\|_2$, or zero without foreground contacts. Frontier nodes have a k-nearest neighbor outside the local graph; n_c^{∂} counts them, and $\bar{b}_c^{\partial}$ averages their maximum incident affinities. Empty contact terms are zero. Then

$$\begin{aligned} T_c^+ &= \bar{b}_c^+ \log(1 + n_c^+)(0.25 + 1.75 D_c^G), \\ T_c^- &= \bar{b}_c^- \log(1 + n_c^-), \qquad T_c^{\partial} = \bar{b}_c^{\partial} \log(1 + n_c^{\partial}), \\ A_c^+ &= T_c^+ + 0.25 Q_c^+ + 0.15 F_c + 0.50 V_c^+, \\ A_c^- &= T_c^- + 0.50 T_c^{\partial} + 0.25 Q_c^- + 0.50 V_c^-. \end{aligned} \tag{17}$$

Support updates and rendering. Positive multi-view support requires $V_c^+ \geq \max(0.30, V_c^-)$, $|O_c|/|c| \geq 0.20$, and mean observing-cluster count over O_c at least one. Without this support and with fewer than three foreground contacts, set $A_c^+ = 0$. Geometric enclosure requires at least three foreground contacts, $D_c^G \geq 0.1$, planar coverage ≥ 0.1, and $n_c^{\partial}/|c| \leq 0.1$. Coverage is the middle-to-largest eigenvalue ratio of the contact-weighted second moment of $\mathbf{t}_{cj}$.

After graph cut, foreground-labeled additions require positive support or enclosure, at least half the nodes with opacity ≥ 0.1, and $A_c^+ \geq A_c^-$. Commit additions only if at least 40% of observed

SYSTEM INSTRUCTION

You are an expert in visual grounding in aerial imagery.

Given one acquired image and a query plan, verify and localize every individually visible instance matching the intended target.

CORE INSTRUCTIONS

1. **TARGET IDENTITY**: Check the original query and target criteria against visible evidence. Do not accept a related facility or lookalike merely because its appearance or location is similar.
2. **DIRECT EVIDENCE:** Accept an instance only when its listed target criteria are directly verifiable. Hidden, off-frame, or unresolved target properties do not count as satisfied.
3. **RELATIONAL CONTEXT:** Check landmarks and relations when visible. An off-frame reference landmark alone does not invalidate an otherwise matching target observation; context is accumulated across views.
4. **INSTANCE LOCALIZATION:** Return one tight box per visible matching instance. Exclude surrounding context and never merge distinct objects into one enclosing box. Do not impose a fixed visible instance count from the query wording.
5. **VIEW-LEVEL RESULT:** Report whether this image provides a verified target observation. A negative result does not establish that the target is absent from the entire scene.

USER PROMPT

Image: [acquired aerial image]

Original query: [original query]

Canonical target: [target description]

Target criteria: [identity criteria]

Context criteria: [landmark and relation criteria]

Likely confusers: [negative descriptions]

OUTPUT

Return JSON only, with target present (0 or 1), confidence, bboxes_2d, checks, evidence, and rejection reason.

Use [x1, y1, x2, y2] boxes with coordinates normalized to [0, 1000].

Boolean checks follow target criteria, then context criteria.

Return an empty box list when no target is verified.

SYSTEM INSTRUCTION

You are an expert in language-guided understanding of aerial scenes.

Given a natural-language query, construct a structured plan for image retrieval and target verification. No scene images are provided.

CORE INSTRUCTIONS

1. **TARGET IDENTITY**: Identify the intended target while preserving its attributes, functional role, and requested scope. Do not substitute a related object or introduce unsupported requirements.
2. **RETRIEVAL CUES:** Produce complementary short phrases describing the same target. Separately identify reference landmarks and plausible lookalikes. A target part is not a negative category.
3. **VERIFICATION CRITERIA:** Separate intrinsic target properties, reference-landmark properties, and target-to-landmark relations. Hard requirements must follow from the query, not speculation.
4. **TARGET SCOPE:** Distinguish a single target, a specified subset, and all instances of a category. This describes the requested scope; it does not establish how many targets exist in the scene.
5. **STRUCTURAL PRIOR:** Infer the expected target morphology from the query. Treat this as a prior, not an observed scene property.

USER PROMPT

Analyze the following aerial grounding query: [original query]

OUTPUT

Return one JSON object containing:

- canonical_target and query_scope;
- sam3_prompts: complementary target phrases;
- landmark_prompts and negative_prompts;
- functional_requirements: intrinsic target conditions;
- reference_constraints and relation_constraints;
- topology_prior and a brief rationale.

Figure 7: Prompt design of AeroGround. Left: the query-planning prompt decomposes the original query into structured retrieval cues, verification criteria, target scope, and structural priors. Right: the view-verification prompt evaluates these criteria against each acquired aerial image and returns instance-level localization and supporting evidence.

mutable components are both eligible and positively supported; no observed components means no addition. Background-labeled removals require $Q_c^- \geq Q_c^+ + 0.05$ and $V_c^- \geq V_c^+ + 0.05$. Commit removals only when eligible foreground nodes collectively comprise at least 0.5% of current foreground. Committed additions/removals receive one/zero; other support values remain unchanged.

Calibrate absolute then relative rendering thresholds by mean IoU against accepted masks: $\tau_v = \max(0.1, \min(\tau_{\text{abs}}, \tau_{\text{rel}}\zeta_v))$, where ζ_v averages up to 64 largest support values ≥ 0.08; fewer than 16 such pixels yields an empty mask. Independently fused all-target components use their associated views for calibration. Retain an addition-only graph proposal if agreement does not decrease, and a signed proposal if it improves agreement by at least 0.0025. Starting from $\mathbf{s}_q^0$, posterior and graph stages are accepted sequentially only with improvement ≥ 0.0025, retaining support and rendering parameters together.

C Implementation Details

C.1 Models

CLIP. We use pretrained CLIP ViT-L/14 (Radford et al., 2021) to encode multi-scale image regions and textual retrieval cues. Image features are computed once per scene. At query time, the planner's target descriptions, reference-landmark cues, and confuser descriptions are encoded and used to rank candidate views. The original query also contributes to retrieval according to the query-dependent retrieval policy.

Table 5: Accuracy–cost trade-off of AQVE on GauU-Ref. Results are averaged equally over Campus, SMBU, and SZIIT. Fixed ranking processes views according to a predetermined order without early stopping, whereas AQVE adaptively selects views and terminates when sufficient evidence is obtained. Qwen calls include one query-planning call and one verification call per acquired view.

Budget B	**Policy**	**Views**	**Qwen Calls**	**mIoU**	**mBIoU**
6	Fixed ranking	6.00	7.00	73.42	63.74
	AQVE	5.86	6.86	73.18	63.41
12	Fixed ranking	12.00	13.00	79.44	68.55
	AQVE	9.37	10.37	79.31	68.47
18	Fixed ranking	17.93	18.93	80.36	69.51
	AQVE	11.74	12.74	80.21	69.44
24	Fixed ranking	23.59	24.59	80.47	69.59
	AQVE (default)	13.28	14.28	80.89	69.76

SAM 3. We use the pretrained SAM 3 image model (Carion et al., 2025). Both CLIP and SAM 3 remain frozen, without dataset-specific fine-tuning. For each accepted localization, SAM 3 receives the source image, a planner-produced target text prompt, and the MLLM-predicted boxes. Box coordinates are converted from the normalized $[0, 1000]$ coordinate system to image coordinates. For multiple boxes, predicted masks are assigned to input boxes using a greedy one-to-one matching based on box IoU. We retain localizations with MLLM-reported confidence of at least 0.9 and require a box IoU of at least 0.10 between the input MLLM box and its matched SAM 3 box. Empty masks are discarded. Each retained instance mask defines a separate anchor for the contribution-aware lifting described in the main paper. The direct-query SAM 3 baseline is an ablation setting and does not use this MLLM-assisted localization procedure. For the observation ablations in Table 3b, the grounding variants share the default SAM 3 score threshold of 0.5. The retrieval variants retain the existing box-prompted segmentation configuration, which uses MLLM verification and box matching without an additional SAM 3 score cutoff (0.0). Each configuration remains fixed within its respective ablation.

MLLM. Unless otherwise specified, AeroGround and all evaluated MLLM-based 3D grounding and segmentation baselines use Qwen3.5-35B-A3B (Qwen Team, 2026) in non-thinking mode as the shared reasoning backbone. This setting controls for differences in the underlying MLLM when comparing methods.

C.2 Preprocessing and Training

Image resolution. We use image downsampling factors of 1, 8, and 6 for 3D-AS, UrbanBIS, and GauU, respectively. A factor of 1 preserves the original image resolution, while factors of 8 and 6 reduce both image width and height by the corresponding factor.

Baseline preprocessing. For each baseline, we perform the method-specific mask generation and feature extraction required by its implementation, including SAM-based masks and CLIP or DINO (Caron et al., 2021) features where applicable. These preprocessing components remain method-specific and are not replaced by AeroGround's CLIP and SAM 3 configuration.

Scene optimization. Gaussian-based baselines follow their respective optimization procedures, with geometric reconstruction and semantic feature learning performed either sequentially or jointly, depending on the method. AeroGround uses a standard 3DGS reconstruction and requires no additional scene-specific semantic feature optimization. Its query-specific Gaussian support is constructed at inference time.

Table 6: Stage-wise hyperparameter sensitivity on SZIIT. AQVE parameters are evaluated before ESR, whereas ESR parameters are evaluated after refinement with fixed AQVE outputs. One parameter is varied at a time within each stage.

Stage	Parameter	Default	Variations → mIoU
AQVE	B	24	12 (80.28), 18 (80.38), **24 (83.35)**, 30 (83.33)
	M	5	1 (83.08), 3 (83.35), **5 (83.35)**, 7 (83.35)
	λ_{IG}	0.50	0.25 (81.79), **0.50 (83.35)**, 0.75 (84.44), 1.00 (78.25)
	λ_{-}	0.25	0 (79.60), 0.125 (83.10), **0.25 (83.35)**, 0.50 (83.94)
ESR	k	10	5 (85.21), **10 (85.10)**, 20 (84.55)
	λ_g	1.00	0.50 (85.10), **1.00 (85.10)**, 2.00 (85.10)

C.3 Prompt Design

Figure 7 shows the prompt templates used for query planning and per-view verification in AQVE. The two stages are deliberately separated: planning organizes the query into structured retrieval and verification cues without accessing scene images, whereas verification evaluates the resulting target criteria against individual acquired observations.

Query planning. The planning prompt receives only the original query, without scene images. It decomposes the query into a canonical target, retrieval cues, reference landmarks, target attributes and relations, likely confusers, query scope, and a coarse structural prior. The prompt explicitly preserves the intended target semantics: related objects or nearby facilities cannot substitute for the queried target, and structural information is treated as a prior rather than an observed scene property. The resulting JSON representation provides complementary prompts for retrieval and subsequent target verification, but does not establish whether a matching target actually exists in the scene.

View verification. For each acquired view, the verification prompt receives the image together with the original query and the structured criteria produced during query planning. It independently verifies every visible candidate against the target identity criteria and, when observable, the associated relational context. Target properties must be supported by direct visual evidence, while an off-frame reference landmark does not by itself invalidate an otherwise consistent observation. Conversely, a relation that cannot be observed in the current image does not by itself provide evidence that the relation is satisfied. The model returns a view-level presence judgment, confidence, instance-wise bounding boxes, criterion-wise checks, and supporting or rejection evidence. Each box is restricted to the visible extent of one matching instance, and distinct objects are never merged into a single enclosing box. Importantly, a negative result is interpreted only at the current-view level and does not imply that the target is absent from the entire scene.

D Scene Editing Applications

The target Gaussians identified by AeroGround can be used for object removal and appearance editing. We also demonstrate insertion of generated assets at user-specified locations. Figure 8 illustrates these applications. Let $\mathcal{G}_q \subseteq \mathcal{G}$ denote the selected target Gaussians, E_v their projected binary editing mask, and $\mathcal{R}_v(\mathcal{G})$ the RGB rendering of the scene. Unlike the observation masks M_v used for grounding, E_v specifies the editing region.

Editing supervision. For RGB images X, Y and a binary editing mask E, we define

$$w_\beta(\mathbf{x}) = E(\mathbf{x}) + \beta[1 - E(\mathbf{x})],$$
$$\ell_\beta(X, Y; E) = \frac{\sum_{\mathbf{x}} w_\beta(\mathbf{x}) \|X(\mathbf{x}) - Y(\mathbf{x})\|_1}{\max(\sum_{\mathbf{x}} w_\beta(\mathbf{x}), \epsilon)}. \tag{18}$$

Here, $\mathbf{x}$ indexes pixels, $\|\cdot\|_1$ sums absolute differences over RGB channels, and $\epsilon > 0$ prevents division by zero. The weight $\beta \in [0, 1]$ controls supervision outside the editing region.

Object removal. We remove $\mathcal{G}_q$ to obtain $\mathcal{G}^- = \mathcal{G} \backslash \mathcal{G}_q$. OSOR-SDXL (Zhou et al., 2026) produces an inpainted image J_v^{rm} from each selected original image and its removal mask. We construct the reference

$$T_v^{\mathrm{rm}} = F_v \odot J_v^{\mathrm{rm}} + (1 - F_v) \odot \mathcal{R}_v(\mathcal{G}^-), \tag{19}$$

where $F_v(\mathbf{x}) \in [0, 1]$ is a per-pixel blending weight: a value of one uses the inpainted prediction, whereas zero preserves the rendering after removal. The operator $\odot$ denotes elementwise multiplication. New Gaussians $\mathcal{G}_{\mathrm{new}}$ are initialized from neighboring retained Gaussians to reconstruct the exposed background. We optimize their geometry, opacity, and appearance using

$$\mathcal{L}_{\mathrm{rm}} = \mathbb{E}_{v \in \mathcal{V}_{\mathrm{edit}}} \left[(1 - \lambda_{\mathrm{rm}}) \ell_{\beta_{\mathrm{rm}}}(\widehat{I}_v, T_v^{\mathrm{rm}}; E_v) + \lambda_{\mathrm{rm}} \mathcal{L}_{\mathrm{LPIPS}}^v \right], \tag{20}$$

where $\mathcal{L}_{\mathrm{LPIPS}}^v$ compares patches within the removal-mask bounding box (Zhang et al., 2018), and $\lambda_{\mathrm{rm}} \in [0, 1]$ balances pixel and perceptual supervision. All Gaussians in $\mathcal{G}^-$ remain fixed. The completed scene is $\mathcal{G}^- \cup \mathcal{G}_{\mathrm{new}}$.

Object insertion. Given an object description, FLUX.2 Klein 4B (Black Forest Labs, 2026) generates a reference image, which TripoSplat (Yan et al., 2026) converts into a Gaussian asset. We place the asset at a user-specified location and align it with a local supporting plane:

$$\boldsymbol{\mu}_i' = \mathbf{A}(\boldsymbol{\mu}_i - \mathbf{b}) + \mathbf{t}, \qquad \boldsymbol{\Sigma}_i' = \mathbf{A} \boldsymbol{\Sigma}_i \mathbf{A}^\top. \tag{21}$$

Here, i indexes asset Gaussians, $\mathbf{b}$ is the asset's base anchor, $\mathbf{t}$ is its scene placement, and $\mathbf{A}$ combines rotation and scaling. The transformed asset is appended to $\mathcal{G}$ without modifying the original scene Gaussians.

Appearance editing. FLUX.2 Klein 4B edits selected scene renderings according to the appearance instruction. Let J_v^{sty} denote its prediction and $I_v^0 = \mathcal{R}_v(\mathcal{G})$ the original rendering. We preserve the reference background by constructing

$$T_v^{\mathrm{sty}} = E_v \odot J_v^{\mathrm{sty}} + (1 - E_v) \odot I_v^0. \tag{22}$$

We transfer these changes into the scene by minimizing

$$\mathcal{L}_{\mathrm{sty}} = \lambda_{\mathrm{sty}} \mathbb{E}_{v \in \mathcal{V}_{\mathrm{edit}}} \left[\ell_{\beta_{\mathrm{sty}}}(\widehat{I}_v, T_v^{\mathrm{sty}}; E_v) \right], \tag{23}$$

where $\lambda_{\mathrm{sty}} > 0$ scales the loss. Only the selected Gaussians' zeroth-order spherical harmonic coefficients are optimized. Geometry, opacity, higher-order coefficients, and all unselected Gaussians remain fixed. The resulting scene supports rendering from novel viewpoints.

E Additional Experiments and More Results

In addition to the experiments reported in the main paper, we provide further analyses and results on hyperparameter robustness, MLLM compatibility, accuracy–cost trade-offs, runtime efficiency, generalization to additional aerial open-vocabulary benchmarks, and qualitative comparisons.

Accuracy–Cost Trade-off. Table 5 compares AQVE with fixed view ranking under different exploration budgets. Across all budgets, AQVE reduces the number of acquired views and MLLM calls while maintaining comparable grounding accuracy. The efficiency advantage becomes more pronounced as the available budget increases, indicating that AQVE can terminate once sufficient evidence has been collected instead of exhaustively consuming the view budget. At the default budget $B = 24$, AQVE uses only 13.28 views and 14.28 Qwen calls on average, compared with 23.59 views and 24.59 calls for fixed ranking, corresponding to reductions of approximately 44% and 42%, respectively. These results demonstrate that AQVE effectively avoids redundant observations and provides a favorable accuracy–cost trade-off.

Qualitative Component Ablation. Figure 9 qualitatively illustrates the progressive effect of the main AeroGround components. Starting from the AQVE output, Gaussian posterior fusion recovers additional target support from multi-view evidence, improving mask completeness. Geometric graph refinement further regularizes the recovered regions according to local structure, leading to more complete and better-aligned target masks.

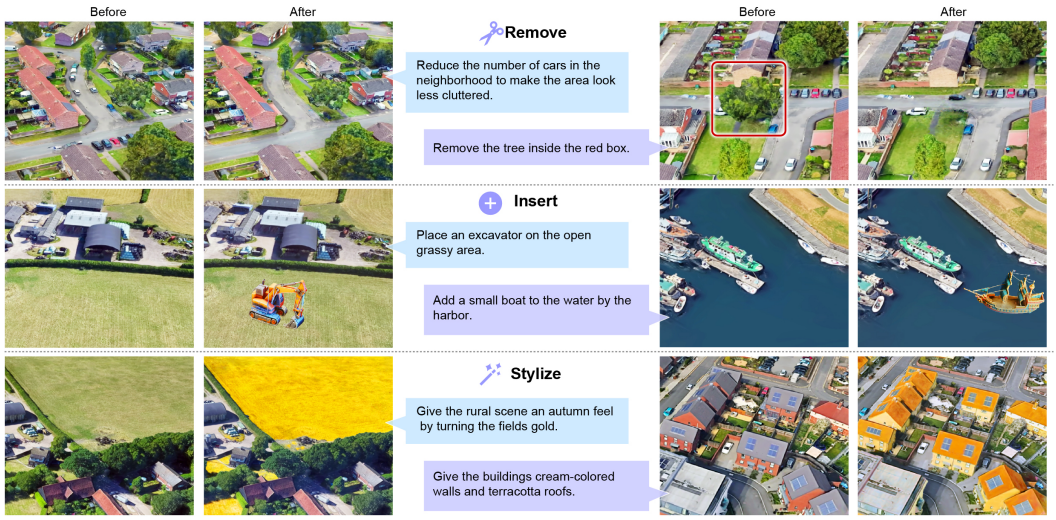

Figure 8: Scene editing applications. Each example shows the scene before and after editing. Top: removing vehicles and a tree. Middle: inserting generated excavator and boat assets at user-specified locations. Bottom: modifying the appearance of fields and buildings.

Table 7: AeroGround with different MLLM configurations on UrbanBIS-Ref.

MLLM	Open-Source	mIoU↑	mBIoU↑
Qwen3.5-9B	Yes	69.57	59.45
Qwen3.5-27B	Yes	74.30	65.37
Qwen3.5-35B-A3B	Yes	78.49	68.22
Qwen3.5-Plus	No	80.56	69.43
Gemini-3.5-Flash-Lite	No	**83.16**	**71.23**

Hyperparameter Sensitivity. Table 6 evaluates the sensitivity of AQVE and ESR parameters on SZIIT. For AQVE, performance is largely stable with respect to M, while increasing the view budget B beyond 24 provides negligible benefit. λ_{IG} performs well over a moderate range but degrades under extreme weighting, whereas introducing nonzero negative evidence through λ_{-} consistently improves performance. For ESR, varying k produces only minor changes, and λ_g remains virtually insensitive over the evaluated range. Overall, the adopted defaults lie in stable high-performing regions and do not require precise hyperparameter tuning.

Effect of MLLM Backbones. Table 7 evaluates AeroGround with different open- and closed-source MLLMs on UrbanBIS-Ref. AeroGround performs consistently across all evaluated models, demonstrating broad compatibility with different MLLM backbones. Performance generally increases across the evaluated configurations, showing that AeroGround benefits from improved multimodal reasoning capabilities.

Computational Efficiency. Table 8 compares offline preparation, retained storage, local query computation, and peak VRAM on SZIIT. Offline measurements exclude separable RGB reconstruction, while local query time excludes model loading and hosted MLLM inference. AeroGround requires no scene-specific feature optimization, prepares its retrieval cache in only 2.19 minutes, and retains only 0.08 GiB of additional scene data, substantially reducing preprocessing and storage costs. This shifts more computation to query time, resulting in 7.47 s of local computation per query and a peak footprint of 11.60 GiB. The MLLM-dependent latency is analyzed separately in the runtime breakdown.

Runtime Breakdown. Table 9 reports module-wise profiling on SZIIT. The summed module time is 52.05 s per query. AQVE accounts for 49.56 s, dominated by view grounding at 42.76 s, whereas ESR contributes 2.49 s. This indicates that view-level reasoning dominates the measured computa-

Table 8: Computational resource comparison on SZIIT. We report profiled offline scene costs, retained scene storage, and query-time computation. Lower is better for all metrics.

Method	Offline Scene Cost				Query Cost	
	Data Prep. (min)	Optimization (min)	Construction/ Postproc. (min)	Storage (GiB)	Compute (s/query)	Peak VRAM (GiB)
LEGS	94.08	251.78†	–	1.92§	0.139	9.25
OpenGS	153.65	91.75	143.23	0.35§	0.019	28.46
InsGS	153.65	137.43†	147.32‡	0.24§	0.019	24.93
AG^{2}S	142.47	60.15†	49.97	0.99§	0.013	19.62
Unified-Lift	58.57	178.40†	–	1.38§	3.475	36.26
COS3D	153.65	105.93	0.47	3.33§	0.368	20.37
ReasonGrounder	154.65	21.39	0.04	11.60	9.734	19.62
ReLaGS	503.39	–	34.59	0.18	0.199	6.97
REALM	58.57	167.01†	0.16	1.43§	15.013	31.38
AeroGround	**2.19**	–	0.08	**0.08**	7.471	11.60

†: joint RGB–feature optimization with RGB cost retained. ‡: additional 10k/5k optimization and language mapping. §: mixed RGB–feature files are retained in full. "–" denotes no separately timed stage.

Table 9: Module-wise runtime on SZIIT, averaged over all queries. Modules are profiled with fixed inputs and loaded models. Total time is the sum of module measurements.

Stage	Module	s/query
AQVE	Query planning and retrieval	6.14
	View grounding	42.76
	Hypothesis update and view selection	0.66
	Subtotal	*49.56*
ESR	Gaussian posterior fusion	1.51
	Geometric graph refinement	0.98
	Subtotal	*2.49*
Total profiled time		**52.05**
Acquired views per query		12.60
Qwen3.5-35B-A3B calls per query		13.60

tion, while posterior fusion and geometric refinement add limited overhead. AQVE acquires 12.60 views per query on average, below the budget of $B = 24$, and makes 13.60 Qwen3.5-35B-A3B calls, including query planning.

Failure Cases. Figure 10 shows representative failure cases of AeroGround. In the first two examples, the correct targets are localized, but complex or elongated structures remain partially incomplete, leading to under-segmented 3D support. In the third example, the pedestrian bridge is correctly identified, but foreground leakage appears on nearby road regions. These failures mainly arise from incomplete support for complex structures and boundary ambiguity around thin targets.

Additional Qualitative Results. We further provide additional qualitative comparisons in Figs. 11–12, 13 and 14. These results complement the quantitative evaluation by illustrating the behavior of different methods across reasoning-based grounding and open-vocabulary semantic segmentation settings. AeroGround generally produces more complete target regions while reducing background leakage, particularly for small or visually ambiguous objects.

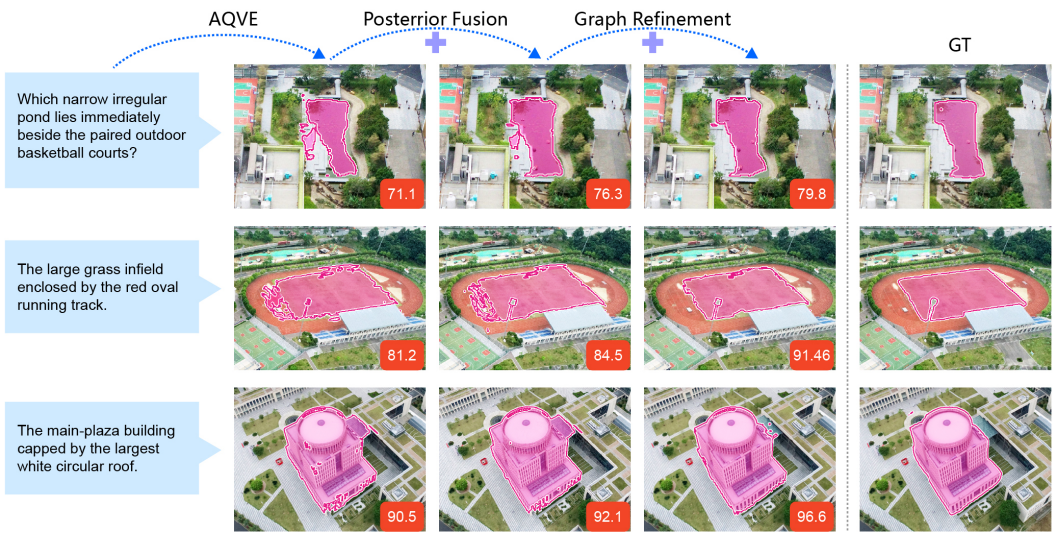

Figure 9: Qualitative component ablation of AeroGround. Starting from AQVE, Gaussian posterior fusion improves target completeness, while geometric graph refinement further regularizes the recovered structure. Numbers denote the per-example IoU.

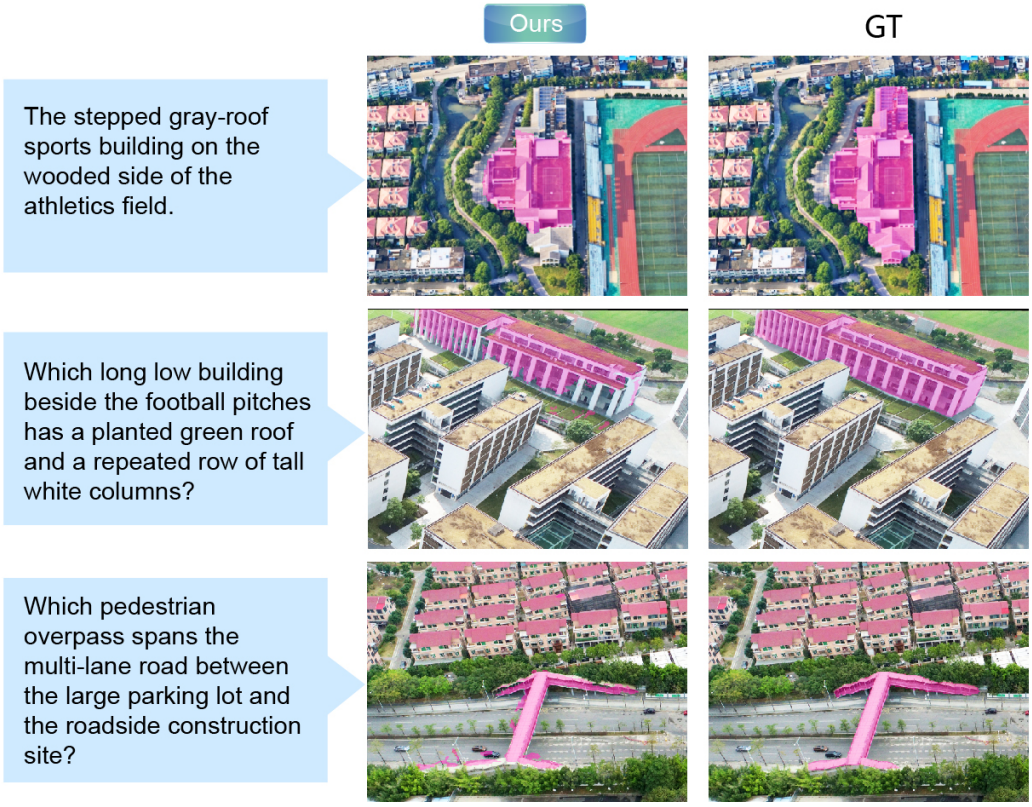

Figure 10: Representative failure cases of AeroGround. The first two examples show incomplete extent recovery for complex or elongated structures, while the last example exhibits foreground leakage around a thin pedestrian bridge. Predictions are shown in pink alongside the corresponding ground truth.

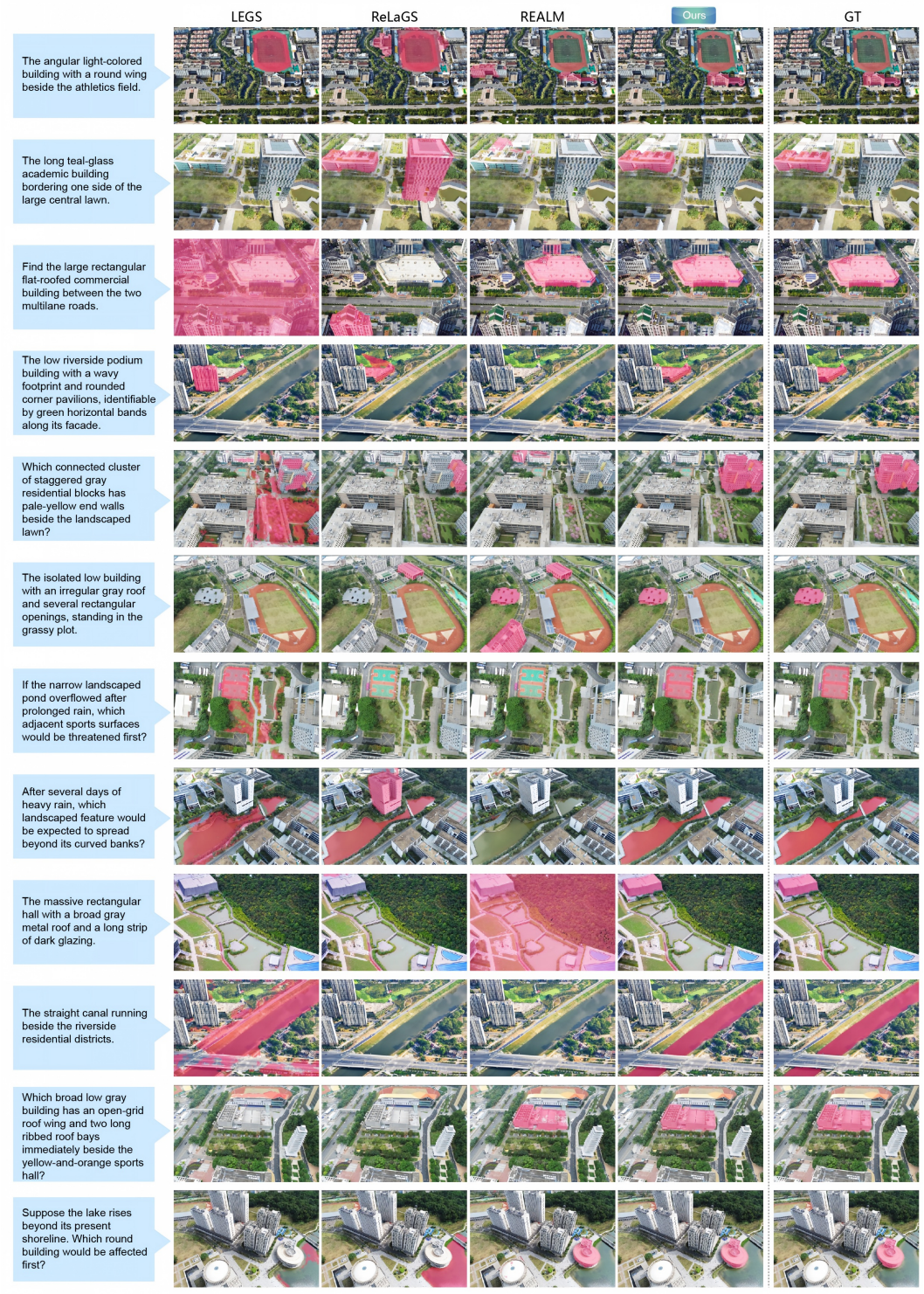

Figure 11: Additional qualitative comparisons of reasoning-based grounding on AeroRef3D. These examples complement the qualitative results in the main paper and cover a broader range of referring expressions and reasoning conditions. Compared with the baselines, AeroGround more consistently identifies the intended targets and produces cleaner, more complete grounding masks with reduced background leakage.

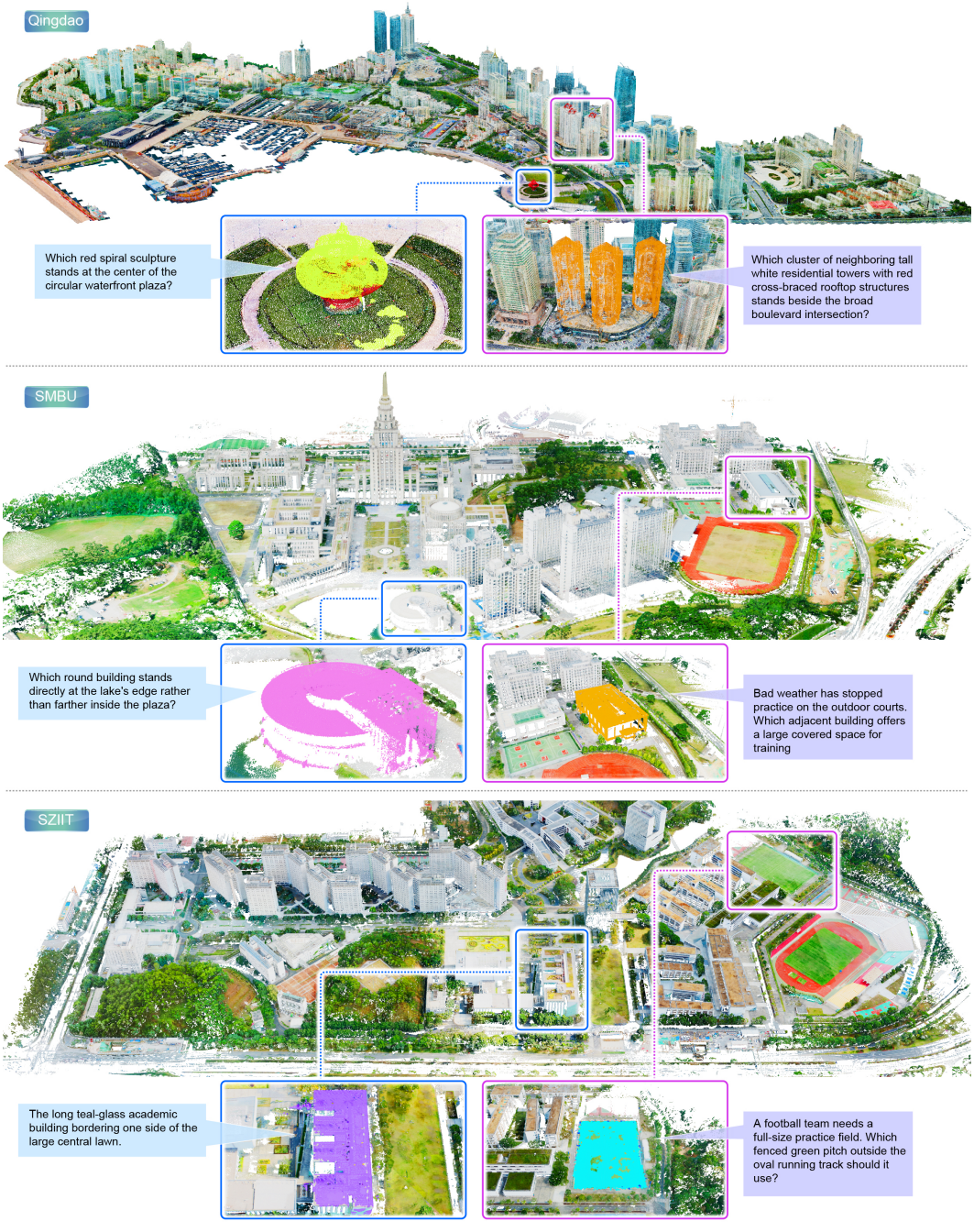

Figure 12: Additional qualitative visualizations of language-guided 3D grounding on AeroRef3D. Examples from Qingdao, SMBU, and SZIIT visualize the recovered target support on the original scene across different object scales, structures, and query types.

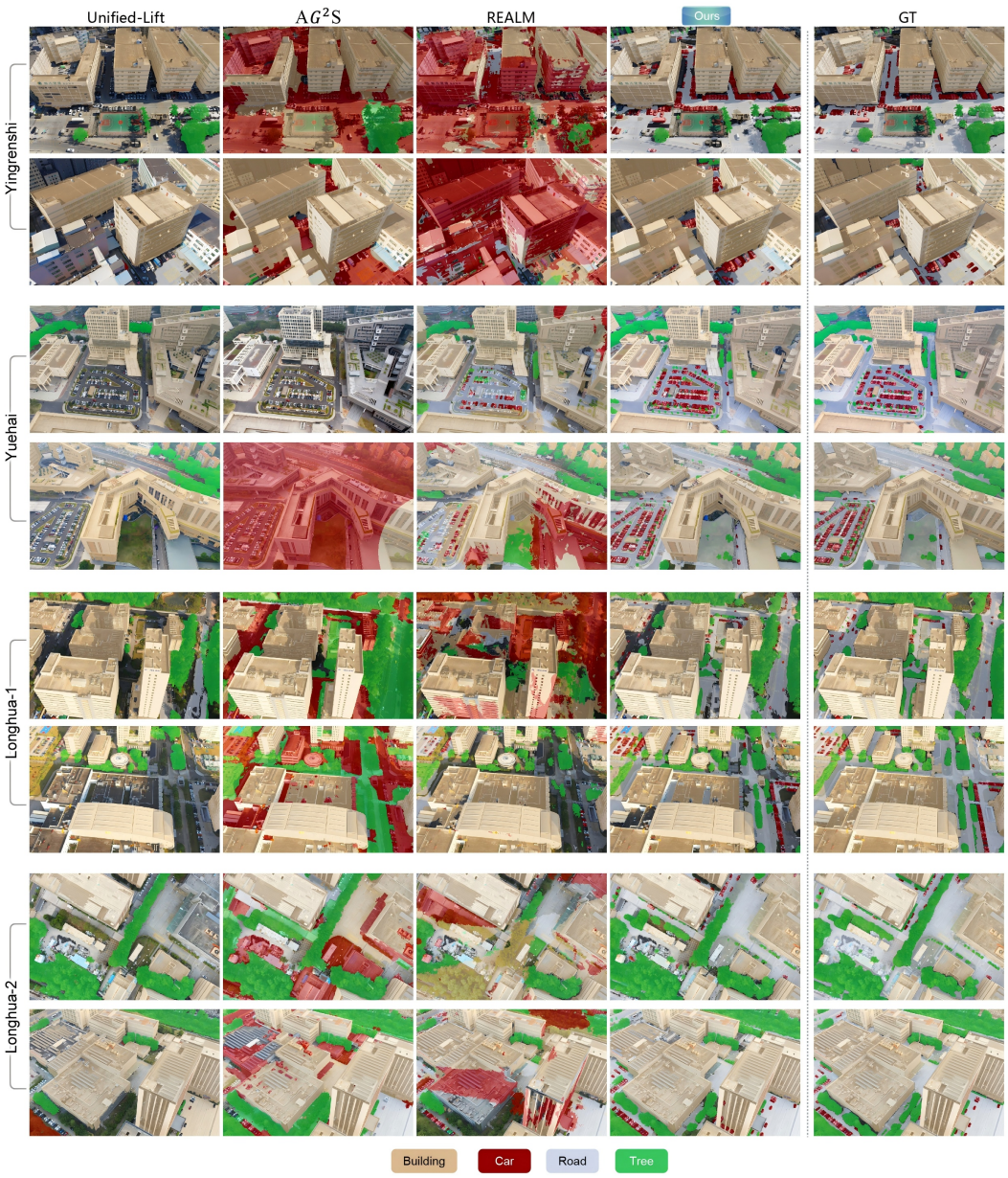

Figure 13: Qualitative comparison of open-vocabulary semantic segmentation on UrbanBIS. Across Yingrenshi, Yuehai, Longhua-1, and Longhua-2, AeroGround produces more complete class regions and fewer spurious predictions than the compared methods, with results that more closely match the ground truth.

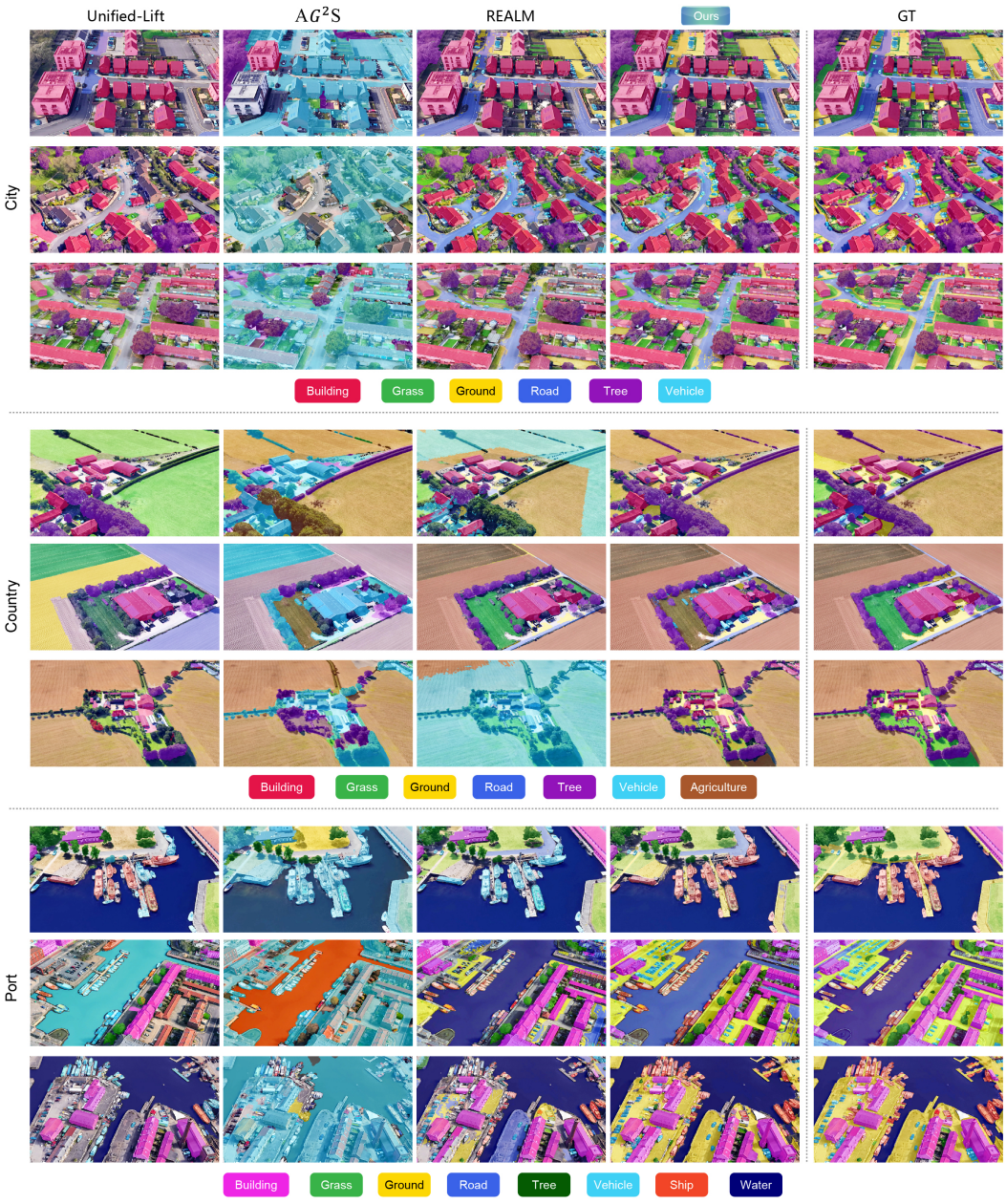

Figure 14: Qualitative comparison of open-vocabulary semantic segmentation on 3D-AS. Examples from City, Country, and Port show that AeroGround yields cleaner and more complete semantic regions that more closely match the ground truth than the compared methods.