Open-Vocabulary Segmentation of Aerial Photos

Luís Marnoto Lopes

August 17, 2025

Contents

1 Introduction

This is placeholder text for the introduction section.

1.1 Problem Statement

This is placeholder text for the problem statement.

1.2 Contributions

This is placeholder text for the contributions.

2 Related Work

2.1 Aerial Image Segmentation Datasets

2.1.1 Semantic Segmentation Datasets

This is placeholder text about semantic segmentation datasets.

2.1.2 Instance Segmentation Datasets

This is placeholder text about instance segmentation datasets.

2.1.3 Referring Instance Segmentation Datasets

This is placeholder text about referring expression datasets.

2.2 Model Architectures for Aerial Imagery Segmentation

2.2.1 RSRefSeg

2.3 Historic Aerial Imagery

2.4 Multimodal Large Language Models

3 Dataset Construction

Creating a referring segmentation dataset for aerial imagery presents unique challenges, as existing aerial datasets lack the natural language referring expressions required for this task. While high-quality instance and semantic segmentation datasets exist for aerial imagery, they contain only categorical labels and spatial annotations without the descriptive language needed to train referring segmentation models. This chapter describes our systematic approach to transforming these existing datasets into a comprehensive referring segmentation resource through automated rule-based generation of referring expressions.

Our methodology addresses this gap by developing a programmatic system that analyzes spatial relationships, visual characteristics, and positional information to generate natural language descriptions for aerial objects. The process begins with established aerial datasets that provide different annotation paradigms, extracts meaningful spatial and visual features from these annotations, and applies rule-based logic to construct referring expressions that capture object relationships, positions, and visual properties. Additionally, we convert semantic segmentation data into instance-level annotations where appropriate, enabling unified treatment of discrete objects and continuous landscape features within a single referring segmentation framework.

Beyond rule-based expression generation, we further enhance the dataset through multimodal large language model fine-tuning. We leverage the generalization capabilities of open-source multimodal LLMs, which possess both advanced language understanding and vision processing capabilities, to create more natural and diverse referring expressions. Through fine-tuning a multimodal LLM specifically on the task of expression enhancement, we apply this enhanced model to the full extent of our rule-based dataset, more than doubling the number of expressions from the original rule-based generation and significantly increasing the linguistic diversity and naturalness of the referring expressions.

3.1 Source Datasets

The AERIAL-D dataset is constructed from two primary sources of aerial imagery with fundamentally different annotation paradigms. The iSAID dataset is an instance segmentation dataset providing high-resolution aerial images with precise boundaries for individual object instances across fifteen categories including ships, vehicles, planes, buildings, and infrastructure elements such as harbors and bridges. In contrast, the LoveDA dataset is a semantic segmentation dataset that captures land cover and land use patterns, providing pixel-level classification into categories such as buildings, water bodies, agricultural areas, forests, and barren land. These two datasets ensure comprehensive coverage of both discrete objects and continuous landscape features commonly encountered in aerial imagery analysis.

Table 1: Source Dataset Characteristics

iSAID Dataset	
	Contains **2,806** high resolution images at varying widths of 800 to 13,000 pixels, spatial resolution of **0.1m to 4.5m**, with **655,451** instances across **15** object classes including **ships**, **large vehicles**, **small vehicles**, **planes**, etc.
LoveDA Dataset	
	Contains **5,987** images at 1024 pixel width, spatial resolution of **0.3m**, across **6** land cover classes: **building**, **road**, **water**, **barren**, **forest**, and **agriculture**.

3.2 Rule-Based Generation Pipeline

Rule Type	Example Instance
Category	"plane"
Grid Position	"in the top right"
Extreme Position	None
Size Comparison	None
Color Classification	"light"
Directional Relations	"to the bottom right of a plane"
	"to the top right of a plane"

Final Expressions

"the plane in the top right"
"the light plane in the top right"
"the plane in the top right to the bottom right of a plane"
"the light plane in the top right to the bottom right of a plane"
"the plane in the top right to the top right of a plane"
"the light plane in the top right to the top right of a plane"

Figure 1: Example of rule generation for a single instance. The highlighted plane in the top right section demonstrates how the system assigns spatial, visual, and relational rules that will later be combined into referring expressions.

The rule-based generation pipeline transforms raw object annotations into structured linguistic descriptions through a comprehensive analysis of spatial, visual, and relational properties. This systematic approach ensures that every object instance receives detailed characterization across multiple dimensions, creating the foundation for diverse and contextually accurate referring expressions.

The pipeline begins by extracting fundamental metadata from each object instance, including category labels, precise bounding box coordinates, centroid positions, area measurements, and segmentation masks in run-length encoding format. A critical preprocessing step identifies partially visible objects through a cutoff flag, marking instances where less than half the object remains within the patch boundaries. This metadata serves as the foundation for all subsequent rule extraction and linguistic generation processes.

Spatial positioning forms the core structural element of the annotation system. Each image patch undergoes systematic partitioning into a three-by-three grid, establishing nine distinct spatial regions that provide consistent positional references. The system incorporates sophisticated borderline handling through a configurable center zone controlled by parameter alpha set to 0.2, creating a neutral area that helps resolve ambiguous boundary cases. When objects fall near grid cell boundaries, the system records both primary and alternative position labels, ensuring comprehensive coverage of spatial interpretations that later expand into multiple linguistic variants.

Extreme position detection operates independently within each object category, identifying instances that occupy the most prominent spatial positions along each axis. The system assigns topmost, bottommost, leftmost, and rightmost labels when the leading candidate maintains separation from the next closest instance by at least five percent of the total image extent along the corresponding dimension. This threshold ensures that extreme position assignments reflect genuinely distinctive spatial arrangements rather than minor positional variations.

Size-based characterization provides another crucial dimension for object differentiation. The system analyzes area measurements within each category to identify significant size outliers, applying a factor-of-1.5 separation rule to distinguish truly exceptional instances from normal size variations. Largest labels are assigned when an object's area exceeds the second-largest instance by this threshold, while smallest designations apply only to fully visible objects meeting the same separation criteria, preventing misleading size assessments based on partially occluded instances.

Relational analysis captures the complex spatial interactions between nearby objects through pairwise relationship computation. The system employs a dynamic proximity radius that combines a base

distance value with size-dependent adjustments, ensuring that relationship detection scales appropriately with object dimensions. Directional analysis covers eight primary orientations including cardinal directions and diagonals, with sophisticated angular overlap detection using a 15-degree threshold to identify borderline cases where multiple directional interpretations remain valid. The system explicitly excludes containment relationships where one object's bounding box or centroid falls within another, focusing instead on meaningful spatial proximity relationships.

Group formation represents a higher-level organizational strategy that captures collective object arrangements. The system applies DBSCAN clustering algorithms independently within each object category, using minimum bounding box separation as the distance metric to identify natural object groupings. This approach produces meaningful multi-instance clusters while preventing oversized aggregations that would lose spatial coherence. Single-instance groups are created selectively, only when these isolated objects participate in relationships with established multi-instance groups, maintaining relational context across different scales of spatial organization.

Advanced grouping strategies extend beyond simple proximity clustering to capture semantic relationships. Class-level groups aggregate all instances of the same category within a patch, enabling expressions like "all buildings in the image" that refer to complete semantic sets. The system also recognizes special combination patterns, such as the pairing of small and large vehicles when both categories appear together, reflecting common real-world associations in aerial imagery.

Inter-group relationship analysis applies the same directional and proximity principles used for individual instances, but operates at the group level to capture higher-order spatial arrangements. The system computes relationships between multi-instance groups and between groups and individual instances, while avoiding redundant single-to-single relationships that are already captured at the instance level. This hierarchical approach enables complex referring expressions that reference both individual objects and their group affiliations.

Color analysis provides visual characterization through sophisticated pixel sampling and classification. The system extracts HSV color values from object segmentation masks, applying saturation thresholds to distinguish achromatic from chromatic colors and brightness analysis to separate light from dark variants. For chromatic colors, the system requires clear hue dominance before assignment, marking ambiguous cases where multiple color interpretations remain plausible. Domain-specific suppressions prevent inappropriate color assignments for certain categories like buildings and water bodies, where chromatic color variations typically result from imaging artifacts rather than meaningful visual properties.

The expression generation phase synthesizes comprehensive linguistic descriptions by systematically combining the extracted attributes. The system enumerates all valid combinations of category labels, grid positions, spatial relationships, extreme positions, size characteristics, and color properties. Borderline cases identified during attribute extraction expand into multiple expression variants, ensuring comprehensive coverage of alternative linguistic formulations. The system tracks expressions associated with cutoff objects or ambiguous color assignments as candidates for potential removal during final filtering.

Final processing ensures expression uniqueness and linguistic quality through systematic standardization and deduplication. The system normalizes category names and handles plural forms consistently, collapses single-instance groups into singular references, and removes any expression text that appears multiple times within a patch. This strict deduplication policy eliminates all occurrences of duplicated phrases, preventing ambiguous references that would compromise referring expression quality. Additional cleanup removes color expressions for objects with ambiguous color assignments, deletes objects and groups left without valid expressions, and removes entire patches that become empty after filtering. The final step strips intermediate rule fields from the XML annotations, producing clean datasets containing only unique, well-formed referring expressions.

Table 2: Complete Taxonomy of Generated Expression Types

Expression Type	Description	Example
Individual Instance Expressions		
Category Only	Basic object category	"the ship"
Category + Position	Category with grid position	"the ship in the bottom right"
Category + Position + Relationship	Category with position and spatial relationship	"the ship in the bottom right that is to the left of a harbor"
Extreme Position + Category	Extreme spatial position with category	"the topmost ship"
Extreme + Category + Position	Extreme position with grid location	"the topmost ship in the top left"
Extreme + Category + Position + Relationship	Extreme position with relationship	"the topmost ship in the top left that is above a building"
Size + Category + Position	Size attribute with position	"the largest ship in the bottom right"
Size + Category + Position + Relationship	Size attribute with relationship	"the largest ship in the bottom right that is above a harbor"
Size + Extreme + Category + Position	Size with extreme position	"the largest topmost ship in the top left"
Size + Extreme + Category + Position + Relationship	All attributes combined	"the largest topmost ship in the top left that is above a building"
Color + Category	Color attribute with category	"the dark ship"
Color + Category + Position	Color with position	"the dark ship in the bottom right"
Color + Category + Position + Relationship	Color with relationship	"the dark ship in the bottom right that is to the left of a harbor"
Color + Extreme + Category	Color with extreme position	"the dark topmost ship"
Color + Extreme + Category + Position	Color with extreme and position	"the dark topmost ship in the top left"
Color + Extreme + Category + Position + Relationship	Color with all spatial attributes	"the dark topmost ship in the top left that is above a building"
Color + Size + Category + Position	Color with size attribute	"the dark largest ship in the bottom right"
Color + Size + Category + Position + Relationship	Color with size and relationship	"the dark largest ship in the bottom right that is above a harbor"
Color + Size + Extreme + Category + Position	All attributes with color	"the dark largest topmost ship in the top left"
Color + Size + Extreme + Category + Position + Relationship	Maximum complexity expression	"the dark largest topmost ship in the top left that is above a building"
Group Expressions		
Basic Group	Group with size and category	"the group of 3 ships in the center"
Group + Extreme Position	Group with extreme spatial position	"the topmost group of 3 ships in the center"
Group + Relationship	Group with spatial relationship to other groups	"the group of 3 ships in the center that is above a group of 2 buildings"
Single Instance + Group Relationship	Individual object referencing group	"the ship in the bottom right that is to the left of a group of 2 harbors"
Class-Level Groups	Semantic segmentation expressions	"all buildings in the image"
Special Combination Groups	Multi-class semantic groups	"all small and large vehicles in the image"

3.3 LLM Enhancement Component

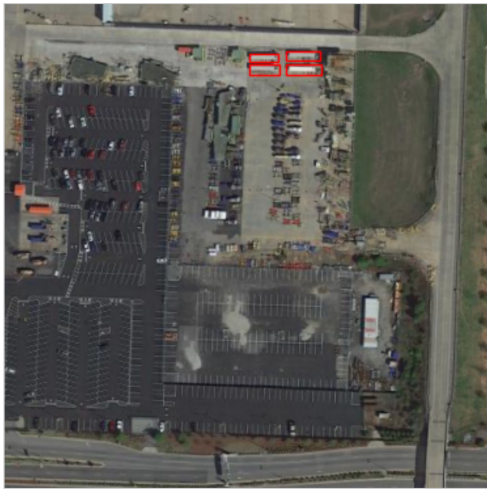

Figure 2: Example aerial image showing a group of four large vehicles (highlighted in red boxes) that demonstrates the LLM enhancement process.

Table 3: Example LLM Enhancement Output for Group Instance

Enhancement Type	**Expression**
Original Expression	the group of 4 large vehicles in the top center
Enhanced Expression	the cluster of four big vehicles near the upper middle
Unique Expressions	the four large vehicles lined up side by side just below the pale paved strip at the very top middle
	the set of four big vehicles parked in a single row in the upper center beside the grassy area to the right

4 Evaluation Setup

4.1 Model Architecture Implementation

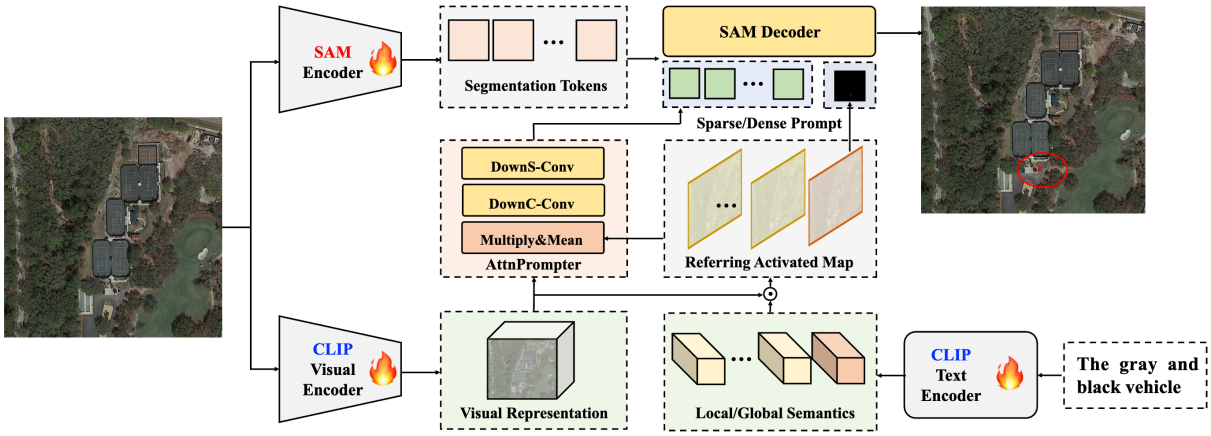

Figure 3: RSRefSeg architecture overview showing the integration of SigLIP2 vision-language encoder with SAM mask decoder through custom prompter networks for text-guided segmentation.

4.2 Dataset Statistics

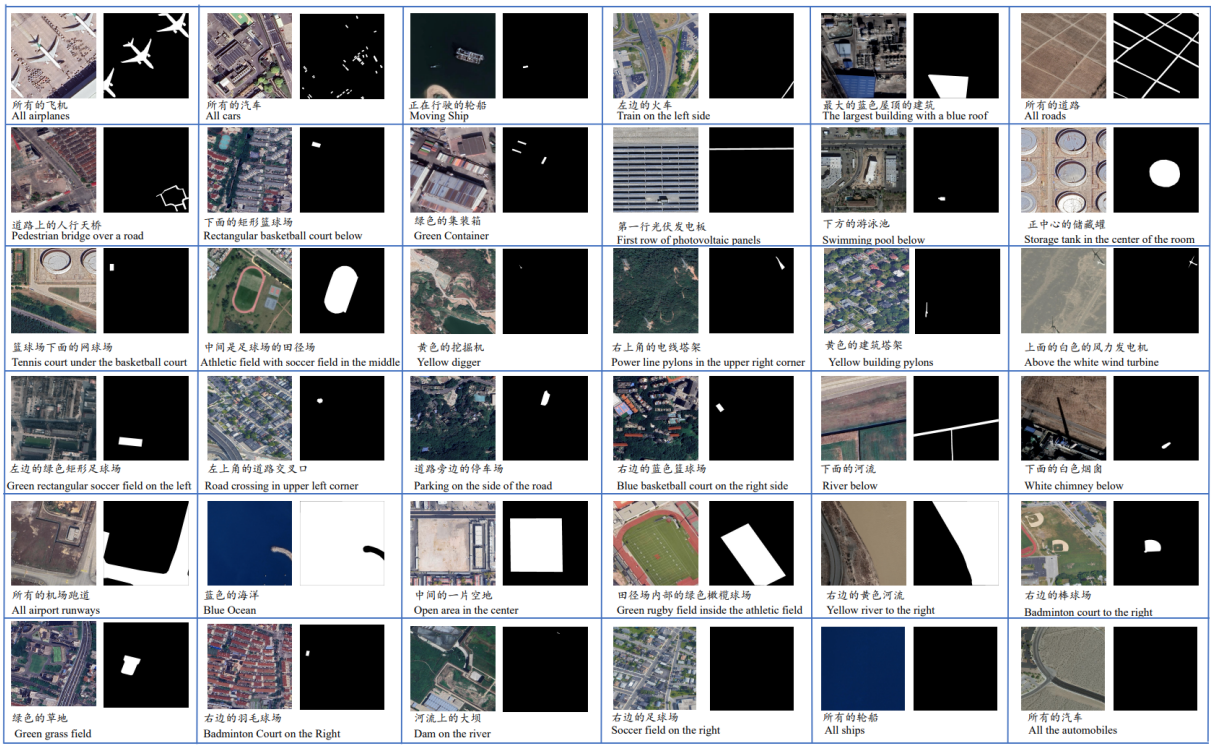

Figure 4: Representative examples from AERIAL-D dataset showing diverse referring expressions with corresponding aerial images and ground truth masks.

Table 4: Dataset Statistics Summary

Metric	Train	Val	Total
Total Patches	32,460	11,054	43,514
Individual Objects with Expressions	94,179	34,536	128,715
Individual Expressions	651,098	244,210	895,308
Groups with Expressions	99,986	34,216	134,202
Group Expressions	487,214	163,472	650,686
Total Samples	1,138,312	407,682	1,545,994
Avg. Expressions per Individual Object	6.91	7.07	6.96
Avg. Expressions per Group	4.87	4.78	4.85

4.3 Category Distribution

Table 5: Object Category Distribution by Instance Type and Source Dataset

Category	Individual Instances	Groups	Instance Expressions	Group Expressions	Source Dataset
Ship	–	–	–	–	iSAID
Large Vehicle	–	–	–	–	iSAID
Small Vehicle	–	–	–	–	iSAID
Building	–	–	–	–	iSAID
Storage Tank	–	–	–	–	iSAID
Harbor	–	–	–	–	iSAID
Swimming Pool	–	–	–	–	iSAID
Tennis Court	–	–	–	–	iSAID
Soccer Ball Field	–	–	–	–	iSAID
Roundabout	–	–	–	–	iSAID
Basketball Court	–	–	–	–	iSAID
Bridge	–	–	–	–	iSAID
Ground Track Field	–	–	–	–	iSAID
Plane	–	–	–	–	iSAID
Helicopter	–	–	–	–	iSAID
Building	–	–	–	–	LoveDA
Water	–	–	–	–	LoveDA
Barren Land	–	–	–	–	LoveDA
Agricultural Area	–	–	–	–	LoveDA
Forest Area	–	–	–	–	LoveDA
Road	–	–	–	–	DeepGlobe

4.4 Expression Type Analysis

Table 6 shows the distribution of different expression types generated across the pipeline, including rule-based expressions and LLM enhancements.

Table 6: Expression Type Distribution

Expression Type	Category	Position	Extreme	Size	Color	Relationship	Total Count
Rule-Based Individual Instance Expressions							
Category Only	✓						–
Category + Position	✓	✓					–
Category + Position + Relationship	✓	✓				✓	–
Extreme + Category	✓		✓				–
Extreme + Category + Position	✓	✓	✓				–
Extreme + Category + Position + Relationship	✓	✓	✓			✓	–
Size + Category + Position	✓	✓		✓			–
Size + Category + Position + Relationship	✓	✓		✓		✓	–
Size + Extreme + Category + Position	✓	✓	✓	✓			–
Size + Extreme + Category + Position + Relationship	✓	✓	✓	✓		✓	–
Color + Category	✓				✓		–
Color + Category + Position	✓	✓			✓		–
Color + Category + Position + Relationship	✓	✓			✓	✓	–
Color + Extreme + Category	✓		✓		✓		–
Color + Extreme + Category + Position	✓	✓	✓		✓		–
Color + Extreme + Category + Position + Relationship	✓	✓	✓		✓	✓	–
Color + Size + Category + Position	✓	✓		✓	✓		–
Color + Size + Category + Position + Relationship	✓	✓		✓	✓	✓	–
Color + Size + Extreme + Category + Position	✓	✓	✓	✓	✓		–
Color + Size + Extreme + Category + Position + Relationship	✓	✓	✓	✓	✓	✓	–
Rule-Based Group Expressions							
Basic Group	✓	✓					–
Group + Extreme Position	✓	✓	✓				–
Group + Relationship	✓	✓				✓	–
Single Instance + Group Relationship	✓	✓				✓	–
Class-Level Groups	✓						–
Special Combination Groups	✓			✓			–

4.5 LLM Enhancement Statistics

Table 7: LLM Enhancement Expression Distribution

Expression Source	Train	Val	Total
Rule-Based Expressions	–	–	–
LLM Enhanced (Language Variations)	–	–	–
LLM Unique (Visual Details)	–	–	–
Total Expressions	–	–	–

4.6 Training Configuration

4.7 Evaluation Methodology

5 Results

5.1 Quantitative Evaluation

Table 8: Cross-Dataset Performance Evaluation on Validation Sets

Dataset	IoU@0.5	IoU@0.7	IoU@0.9	mIoU	oIoU
AERIAL-D	–	–	–	–	–
RefSegRS	–	–	–	–	–
RRSIS-D	–	–	–	–	–
NWPU-Refer	–	–	–	–	–

Table 9: Comparison with Existing RRSIS Datasets

Dataset	Image Resolution	Images	Annotations	Single-object	Multi-object	Resolution	Annotation Generation
RefSegRS	0.13m	4420	4420	✓	×	512	Manual
RRSIS-D	0.5m-30m	17402	17402	✓	×	800	Semi-auto
NWPU-Refer	0.12m-0.5m	15003	49745	✓	✓	1024-2048	Manual
AERIAL-D	**0.3m-2m**	**43,514**	**1,545,994**	✓	✓	**480**	**Automated + LLM**

5.2 Qualitative Analysis

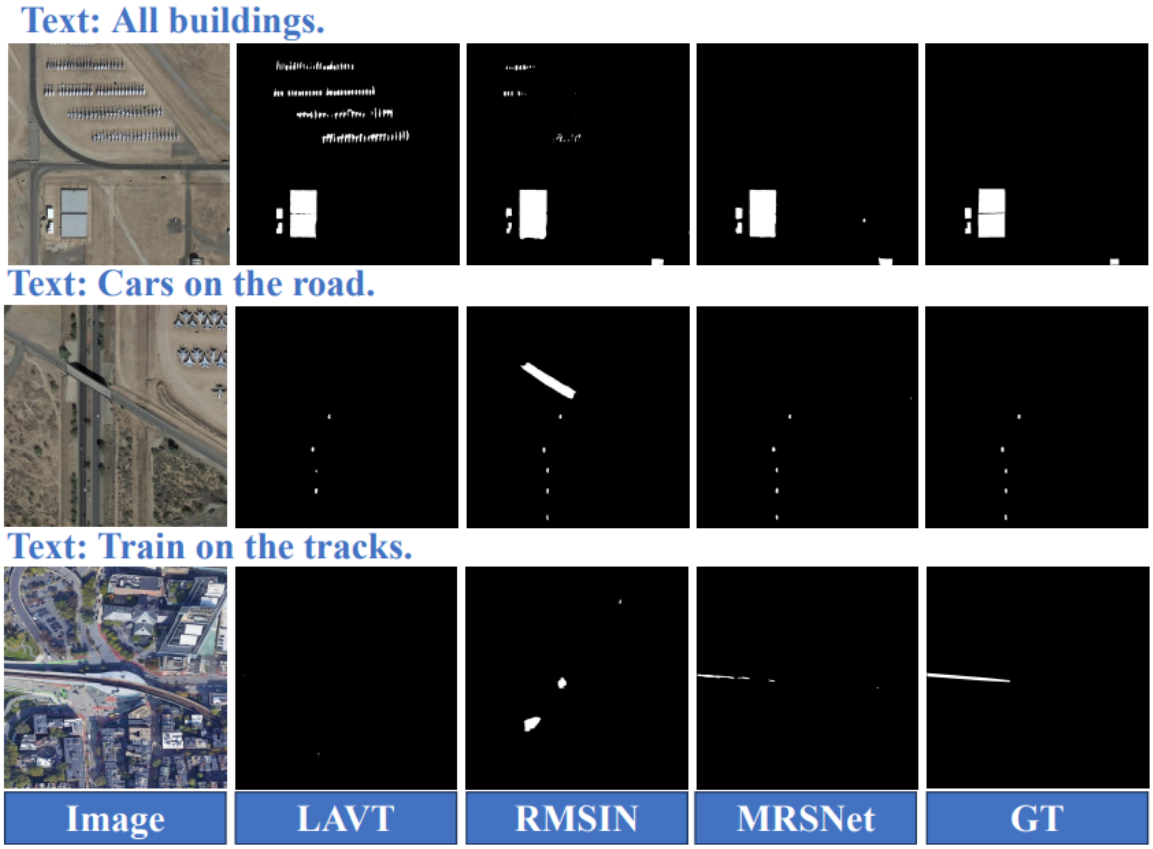

Figure 5: Qualitative segmentation results from RSRefSeg model on AERIAL-D validation set.

Figure 6: Dataset error analysis examples for LLM-generated unique expressions.

5.3 Ablation Studies

Table 10: Ablation Study: Expression Type Training Analysis

Training Configuration	IoU@0.5	IoU@0.7	IoU@0.9	mIoU	oIoU	Training Expressions
Rule-based Only	–	–	–	–	–	–
Language Variations	–	–	–	–	–	–
Unique Expressions	–	–	–	–	–	–
Combined All	–	–	–	–	–	–

6 Conclusion

Acknowledgments

A Pipeline Implementation Details

B LLM Enhancement Prompts