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

Table 1: Dataset Source Contributions to AERIAL-D

Source Dataset	Patches	Individual Instances	Groups
iSAID	–	–	–
LoveDA	–	–	–
DeepGlobe	–	–	–
Total	**43,514**	**128,715**	**134,202**

3.1 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.

Our rule-based component enriches each patch with spatial structure and language-ready annotations before LLM enhancement. It operates on XML annotations produced by the patching steps and outputs three successive datasets: (1) with rules added, (2) with all candidate expressions, and (3) with expressions filtered for uniqueness. Below we summarize the main stages implemented in the pipeline scripts.

1. **Parse instances and metadata**. For every object we read category, bounding box, centroid, area, segmentation (RLE), and a cutoff flag when less than half the object remains inside the patch. This is the starting point for all subsequent rules.

2. **Grid positions with borderline handling**. Each patch is partitioned into a 3×3 grid. A center "no-man's land" controlled by $\alpha = 0.2$ and a small border tolerance mark borderline cases. We assign a primary position label (for example, "in the top left") and, when near cell boundaries, record all alternative positions that might also apply. These alternatives are later expanded into language variants.

3. **Extreme positions**. For each category independently we assign topmost, bottommost, leftmost, and rightmost when the leading candidate is separated from the next by at least five percent of the image extent along the corresponding axis. The extreme tag is stored on the instance.

4. **Size attributes**. Within each category we detect salient size outliers: an instance is tagged as *largest* if its area exceeds the second largest by a factor of 1.5 or more. The *smallest* tag is assigned only among fully visible instances and under the same $1.5\times$ separation rule.

5. **Local relationships between instances**. We compute pairwise spatial relations only to nearby targets using a dynamic radius equal to a base value plus a size-dependent term. Relations cover eight directions (left/right, above/below, and diagonals). A 15-degree angular overlap yields borderline relations that store multiple admissible directions; otherwise a single direction is stored. Containment cases (one box or centroid inside the other) are ignored. We keep distance for optional text rendering and mark if either endpoint is cutoff.

6. **Clustering into groups**. Instances are clustered *per category* using DBSCAN with a distance based on minimum bbox separation, producing multi-instance groups while capping oversized clusters. We create single-instance groups only when they participate in relations with multi-instance

groups. For each group we compute centroid, grid position, size, and a combined segmentation mask.

7. **Higher-level groups**. We add class-level groups that aggregate all instances of the same class present in the patch (semantic "all X in the image" expressions). We also form a special pair group that merges small and large vehicles when both are present.

8. **Group relationships**. We compute relations among groups and between groups and single-instance groups (but not single-to-single), using the same directional scheme and distance gating as for instances while avoiding containment.

9. **Color reasoning with ambiguity**. For each instance we sample HSV pixels from its segmentation and determine a dominant color. Achromatic detection uses a saturation threshold; brightness separates light from dark. Chromatic categories are assigned only when a single hue dominates; otherwise the color is marked ambiguous and we retain a small set of candidate terms. To avoid artifacts, chromatic colors are suppressed for buildings and water; only light/dark are kept for those classes.

10. **Expression generation**. From the attributes above we synthesize comprehensive referring expressions for *instances* and *groups*. We enumerate combinations of category, grid position, relationships, extremes, size, and color. Borderline positions and borderline relationships expand into multiple variants. Expressions associated with cutoff objects or ambiguous color are tracked as "dummy" for potential removal.

11. **Uniqueness filtering and cleanup**. We standardize class names (and plurals), collapse "group of 1" into the singular, and remove any expression text that appears more than once across the patch—if a phrase is duplicated, *all* of its occurrences are dropped. We then discard color expressions for ambiguous objects, delete any object or group with no expressions left, and remove corresponding images when a patch becomes empty. Finally we strip intermediate rule fields from XML, leaving clean annotations with only unique 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"

3.2 LLM Enhancement Component

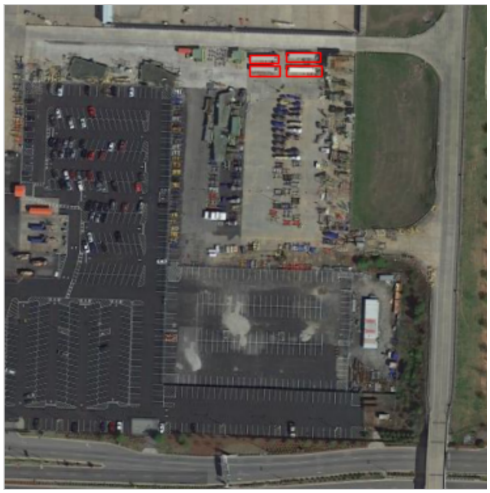

Figure 2: Example aerial image showing a group of four large vehicles (highlighted in red boxes) that demonstrates the LLM enhancement process.

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

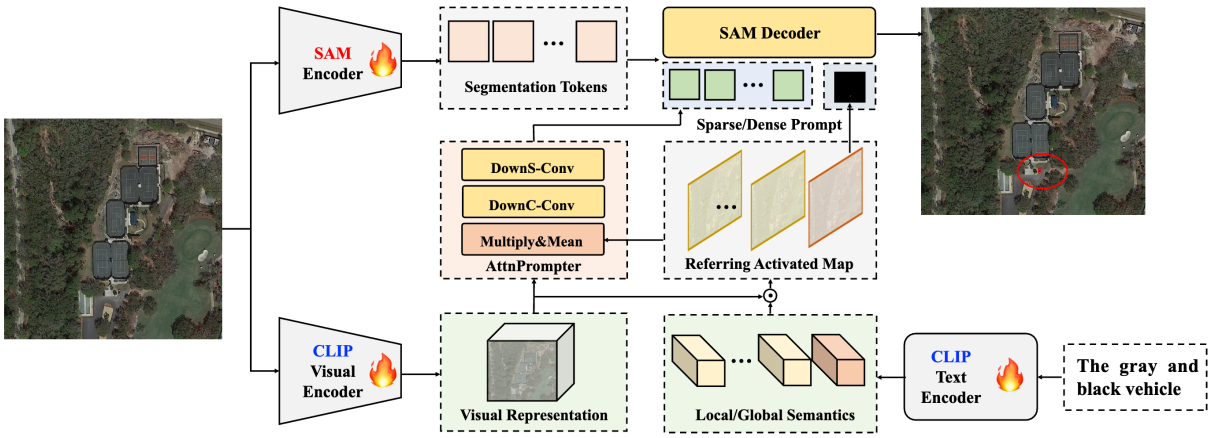

Figure 3: RSRefSeg architecture overview showing the integration of SigLIP2 vision-language encoder with SAM mask decoder through custom prompter networks for text-guided segmentation.

4.2 Dataset Statistics

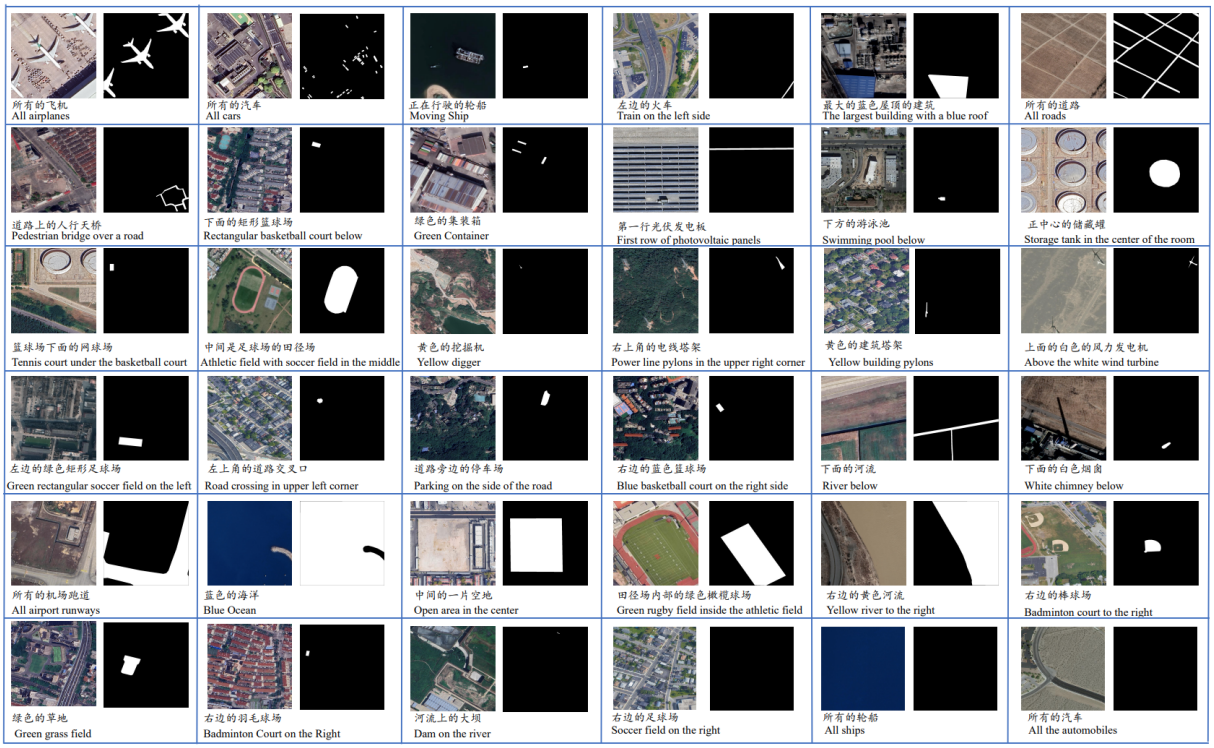

Figure 4: Representative examples from AERIAL-D dataset showing diverse referring expressions with corresponding aerial images and ground truth masks.

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

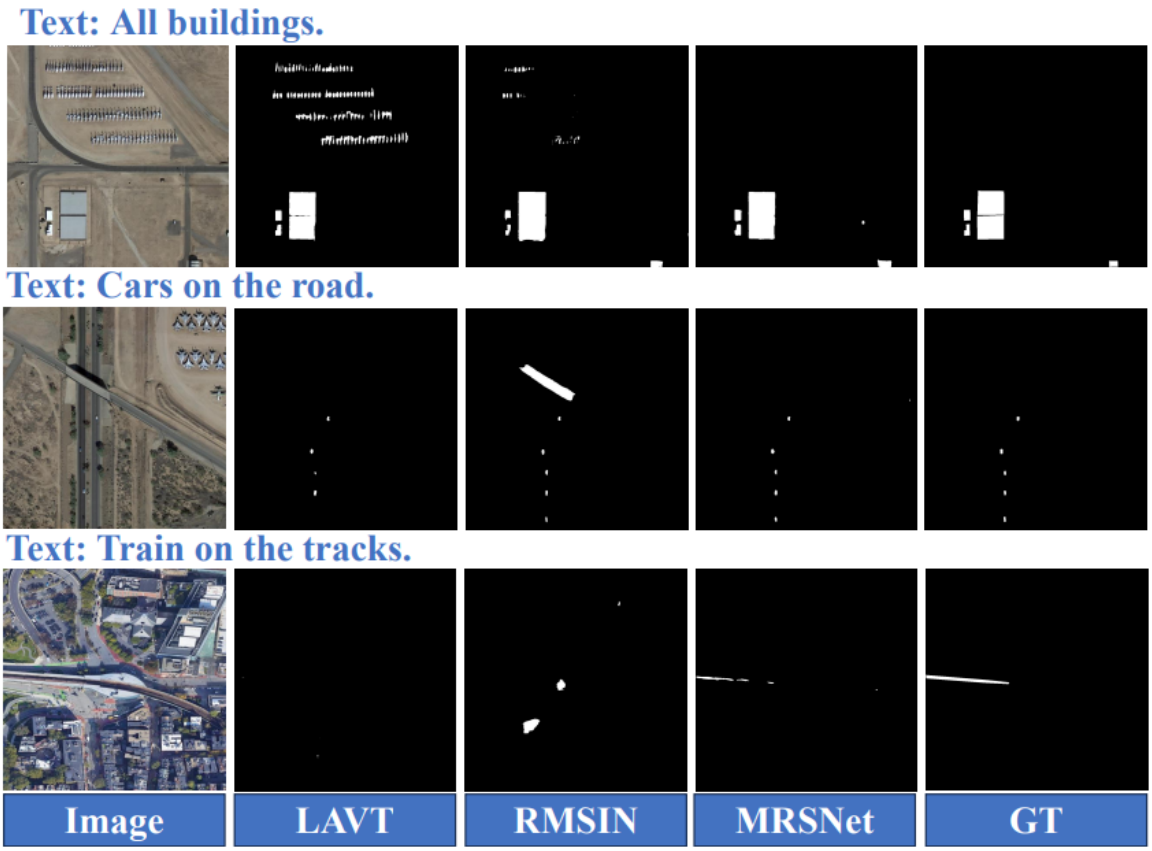

Figure 5: Qualitative segmentation results from RSRefSeg model on AERIAL-D validation set.

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