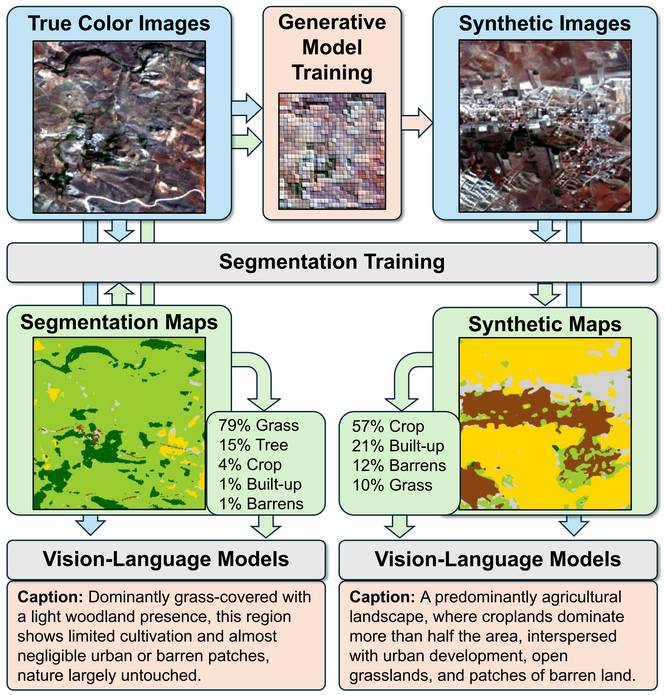
ARAS400k: Grounding Synthetic Data Generation with Vision and Language Models

Ümit Mert Çağlar, Alptekin Temizel
Graduate School of Informatics, Middle East Technical University, Türkiye

ECCV 2026 | MALMÖ | SEPT 8–12

ABSTRACT

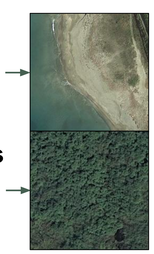
- Framework for remote sensing segmentation, captioning and synthetic data generation
- 100,240 real and 300,000 synthetic images combined with segmentation maps and captions
- A total of 2,001,200 captions generated by different Vision and Language models
- Synthetic captions match human-level CLIPScore
- High volume and less redundant than alternatives
- Image segmentation model training benchmark
- Competitive performance with only synthetic data
- Improved results with synthetic data augmentation



Problem Definition

Problem 1: Redundant captions

Caption (1-5): khaki beach is between green trees and blue ocean

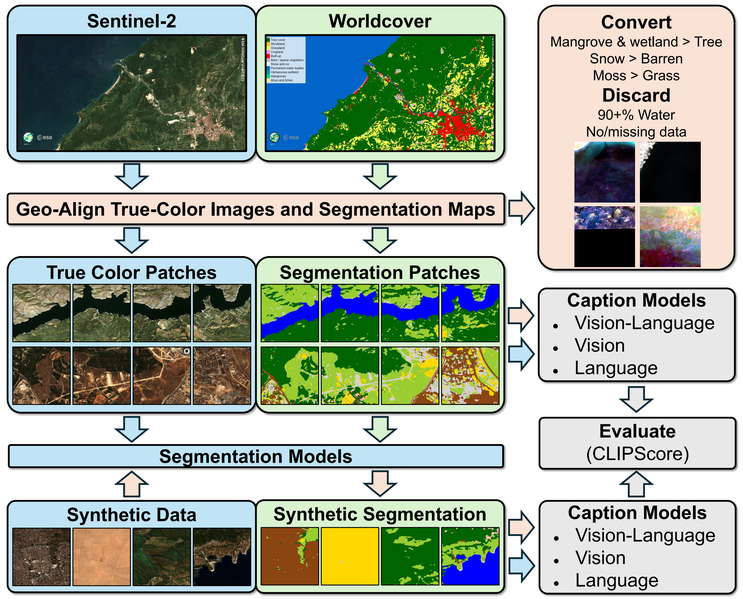


Problem 2: Repetitive captions

This is a dense forest with lots of dark green trees

METHODS AND MATERIALS

- Data collection (Sentinel-2 RGB and Worldcover)
- Synthetic sample generation
- VLM and LLM assisted automated captioning



Data collection, synthetic sample generation and captioning framework

Dataset	Volume	Captions	Redundancy↓	CLIPScore↑
NWPU	31,500	157,500	72.65%	30.25
RSICD	10,921	54,605	67.02%	29.11
UCMC	2,100	10,500	80.88%	30.18
ARAS400k	400,240	2,001,200	12.85%	29.66
-Real	100,240	501,200	8.01%	29.89
-Synth	300,000	1,500,000	13.06%	29.58

Volume, redundancy and CLIPScore of our dataset and literature alternatives
CLIPScore is an automatic, reference-free caption-image matching metric

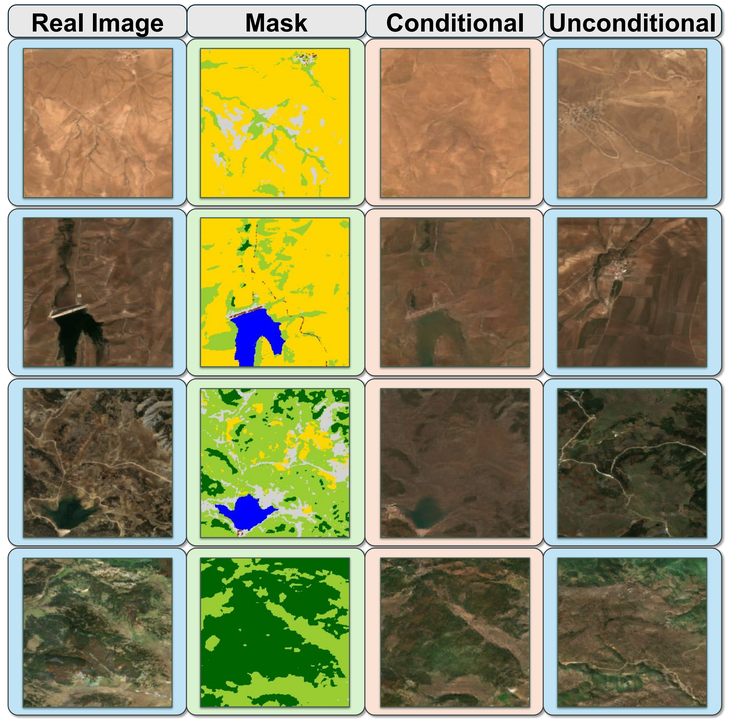
Existing remote sensing caption datasets:

× Limited volume (require labor-intensive captioning)
× Often contain repetitive captions
× Prone to human-error

Our approach:

✓ Larger than existing datasets and scalable volume
✓ Varied captions with low redundancy
✓ Consistent and aligns with human assessment
✓ Synthetic data improves segmentation performance

RESULTS



Benchmark Results

	Real Data				Synthetic Data				Real + Synthetic Data			
Model	**F1**	**IoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Prec.**	**Rec.**	**F1**	**IoU**	**Prec.**	**Rec.**
Unet	76.37	64.49	77.05	**75.86**	74.63	62.31	77.60	72.23	**77.40**	**65.58**	**79.37**	75.64
Unet++	76.38	64.36	76.50	76.39	74.94	62.73	77.81	72.56	**78.23**	**66.67**	**80.05**	**76.65**
PAN	76.20	64.10	75.29	**77.60**	74.65	62.42	77.70	72.22	**77.21**	**65.48**	**79.94**	75.02
DeepLab	76.02	63.91	75.59	**76.77**	74.63	62.39	77.69	72.15	**77.62**	**65.94**	**80.00**	75.65
Segformer	77.09	65.26	78.10	**76.33**	75.10	62.90	77.65	72.95	**77.80**	**66.14**	**79.98**	75.98
FPN	76.85	64.92	77.07	**76.76**	74.81	62.59	77.99	72.26	**77.71**	**66.04**	**80.16**	75.66
Mean	76.49	64.51	76.60	**76.62**	74.79	62.56	77.74	72.40	**77.66**	**65.98**	**79.92**	75.77

Conclusion

- Open source dataset generation pipeline
- Publicly available dataset and results
- Clean and complete image segmentation dataset
- VLM driven automated large-scale captioning
- Results comparable with human captions
- Publicly available synthetic extension data
- Improved segmentation downstream performance