www.inesc-id.pt

joao.daniel.silva@tecnico.ulisboa.pt

Large Language Models for Captioning and Retrieving Remote Sensing Images

João Daniel Silva[1,2] João Magalhães[3] Devis Tuia[4] Bruno Martins[1,2]

[1]INESC-ID, Portugal - [2]Instituto Superior Técnico, University of Lisbon, Portugal - [3]Faculty of Science and Technology, Universidade NOVA de Lisboa - [4]ECEO, Ecole Polytechnique Fédérale de Lausanne

Motivation

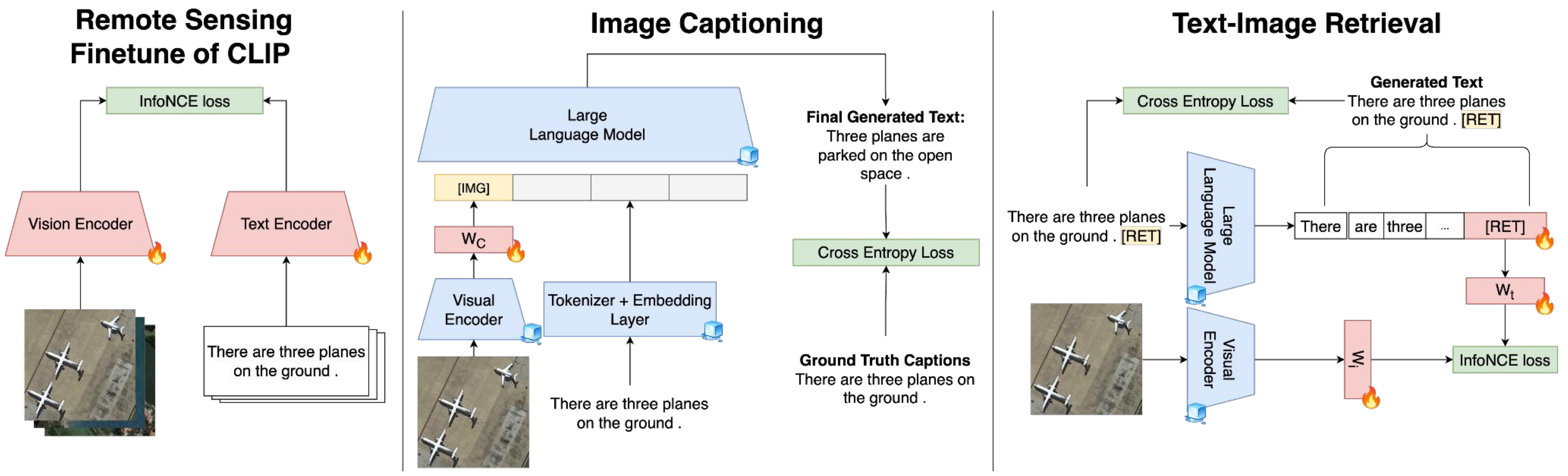
- Image captioning and text-image retrieval tasks can help non-expert users interact with Earth Observation data.
- Move beyond CNN-LSTM framework for text generation based on image inputs.
- Address both generative tasks (image captioning) and embedding tasks (text-image retrieval).

Data

- Aggregation of available image captioning datasets:

Dataset	#Images	Image Size	Spatial Resolution	#Total Captions
NWPU-Captions (Cheng et al., 2022)	31,500	256 × 256	~30-0.2m	157,500
RSICD (Lu et al., 2018)	10,921	224 × 224	different resolutions	54,605
Sydney-Captions (Qu et al., 2016)	613	500 × 500	0.5m	3,065
UCM-Captions (Qu et al., 2016)	2,100	256 × 256	~0.3m	10,500
Cap-4	45,134	224 × 224	different resolutions	225,670
RemoteCLIP	165,745	different sizes	different resolutions	828,725

Method



Remote Sensing Finetune of CLIP
InfoNCE loss
Vision Encoder
Text Encoder
There are three planes on the ground .
Image Captioning
Large Language Model
[IMG]
W_C
Visual Encoder
Tokenizer + Embedding Layer
There are three planes on the ground .
Final Generated Text:
Three planes are parked on the open space .
Cross Entropy Loss
Ground Truth Captions
There are three planes on the ground .
Text-Image Retrieval
Cross Entropy Loss
Generated Text
There are three planes on the ground . [RET]
There are three planes on the ground . [RET]
Large Language Model
There
are
three
...
[RET]
W_t
Visual Encoder
W_i
InfoNCE loss

Main Results

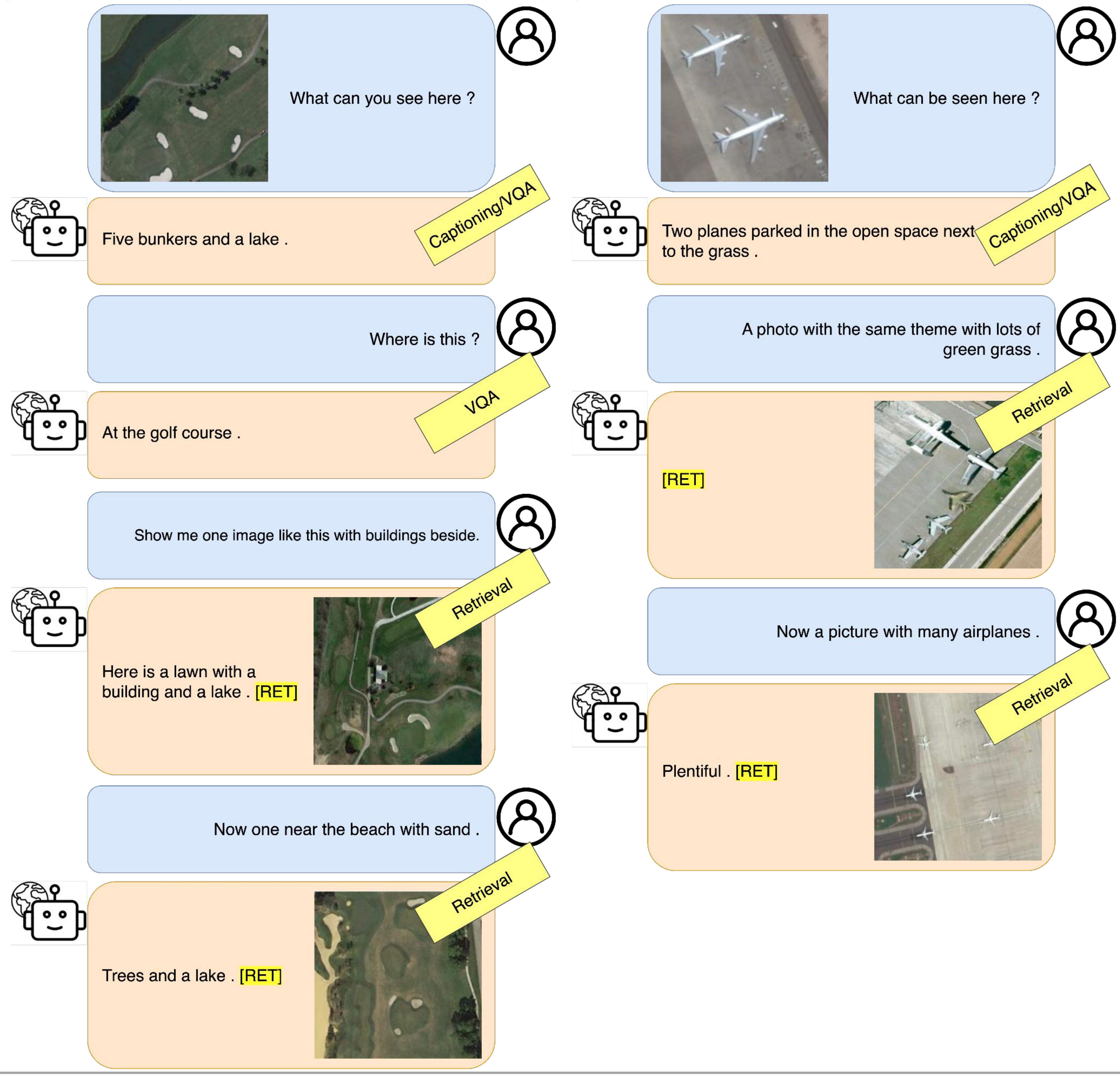
Image Captioning

Evaluation Dataset	Method	Visual Encoder	Text Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
NWPU	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.745	0.624	0.541	0.478	0.337	0.601	1.164	0.285
	RS-CapRet	CLIP-Cap-4	LLamaV2	**0.871**	0.786	0.713	0.650	**0.439**	0.775	1.919	**0.320**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	**0.871**	**0.787**	**0.717**	**0.656**	0.436	**0.776**	**1.929**	0.311
RSICD	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.757	0.634	0.539	0.461	0.351	0.646	2.356	0.444
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.703	0.542	0.440	0.368	0.301	0.533	1.029	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.867**	**0.767**	**0.673**	**0.600**	0.354	0.626	0.837	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.741	0.622	0.529	0.455	**0.376**	**0.649**	**2.605**	**0.484**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.720	0.599	0.506	0.433	0.370	0.633	2.502	0.474
UCM	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.826	0.770	0.717	0.668	0.435	0.772	3.240	0.473
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.861	0.791	0.723	0.657	0.422	0.783	3.332	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.907**	**0.857**	**0.816**	**0.784**	0.462	0.795	2.368	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.833	0.760	0.699	0.645	0.447	0.786	3.429	**0.525**
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.843	0.779	0.722	0.670	**0.472**	**0.817**	**3.548**	**0.525**
Sydney	MLCA-NET (Cheng et al., 2022)	VGG16	LSTM	0.831	0.742	0.659	0.580	0.390	0.711	2.324	0.409
	RSGPT (Hu et al., 2023)	EVA-G	Vicuna	0.823	0.753	0.686	0.622	0.414	0.748	**2.731**	NA
	SkyEyeGPT (Zhan et al., 2024)	EVA-G	LLamaV2-Chat	**0.919**	**0.856**	**0.809**	**0.774**	**0.466**	**0.777**	1.811	NA
	RS-CapRet	CLIP-Cap-4	LLamaV2	0.782	0.688	0.611	0.545	0.383	0.704	2.390	0.423
	RS-CapRet$_{finetuned}$	CLIP-Cap-4	LLamaV2	0.787	0.700	0.628	0.564	0.388	0.707	2.392	**0.434**

Text-Image Retrieval

Dataset	Method	Visual Backbone	Finetune Data	Text-Image Retrieval			
				R@1	R@5	R@10	mR_T2I
RSICD	GaLR (Yuan et al., 2022)	ResNet18	RSICD	4.69	19.48	32.13	18.77
	KCR (Mi et al., 2022)	ResNet101	RSICD	5.40	22.44	37.36	21.73
	CLIP (Radford et al., 2021)†	ViT-B	Zero-shot	5.80	16.85	28.23	16.96
	CLIP (Radford et al., 2021)†	ViT-L	Zero-shot	5.03	19.03	30.25	18.10
	Rahhal et al. (Rahhal et al., 2022)	ViT-B	RSICD	9.14	28.96	44.59	27.56
	CLIP-RSICD (Pal et al., 2021)†	ViT-B	RSICD	11.16	33.25	48.91	31.11
	CLIP-Cap-4†	ViT-L	Cap-4	13.83	39.07	56.05	36.32
	RemoteCLIP (Liu et al., 2024)	ViT-L	RemoteCLIP dataset	**14.73**	**39.93**	**56.58**	**37.08**
	RS-CapRet†	ViT-L	Cap-4	9.83	30.17	47.43	29.14
	RS-CapRet$_{finetuned}$†	ViT-L	Cap-4 + RSICD	10.25	31.62	48.53	30.13
UCM	KCR (Mi et al., 2022)	ResNet101	RSICD	17.43	57.52	80.38	51.78
	CLIP (Radford et al., 2021)†	ViT-B	Zero-shot	8.67	36.48	60.57	35.24
	CLIP (Radford et al., 2021)†	ViT-L	Zero-shot	10.76	46.00	73.33	43.37
	CLIP-RSICD (Pal et al., 2021)†	ViT-B	RSICD	13.81	57.05	91.24	54.03
	CLIP-Cap-4†	ViT-L	Cap-4	16.29	60.57	94.76	57.21
	RemoteCLIP (Liu et al., 2024)	ViT-L	RemoteCLIP dataset	**17.71**	62.19	**93.90**	57.93
	Rahhal et al. (Rahhal et al., 2022)	ViT-B	UCM	19.33	**64.00**	91.42	**58.25**
	RS-CapRet†	ViT-L	Cap-4	15.52	57.24	88.76	53.84
	RS-CapRet$_{finetuned}$†	ViT-L	Cap-4 + UCM	16.10	56.29	90.76	54.38

Emergent dialogue beyond our training



What can you see here ?

Five bunkers and a lake .

Where is this ?

At the golf course .

Show me one image like this with buildings beside.

Here is a lawn with a building and a lake . [RET]

Now one near the beach with sand .

Trees and a lake . [RET]

What can be seen here ?

Two planes parked in the open space next to the grass .

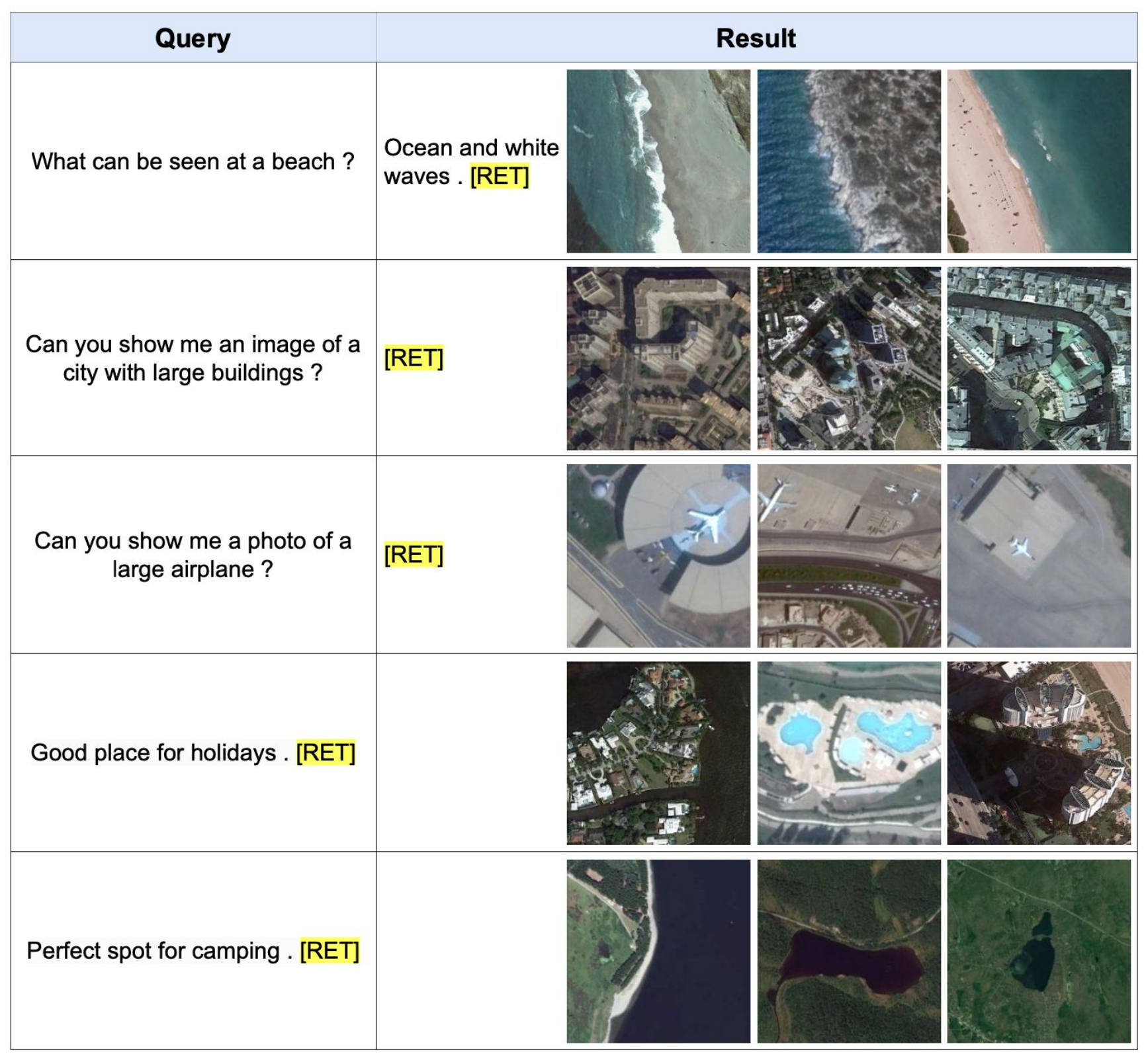
A photo with the same theme with lots of green grass .

[RET]

Now a picture with many airplanes .

Plentiful . [RET]

"OOD" Commonsense knowledge



Query	Result
What can be seen at a beach ?	Ocean and white waves . [RET]
Can you show me an image of a city with large buildings ?	[RET]
Can you show me a photo of a large airplane ?	[RET]
Good place for holidays . [RET]	
Perfect spot for camping . [RET]	

Conclusion

- Confirming potential of Vision Large Language Models for remote sensing, both for image captioning and text-image retrieval;
- Flexible and lightweight approach;
- Model shows characteristics learned beyond our training setup

Limitations and Future work

- Image-text retrieval is not SOTA.
- Robust from specific user instructions;
- Multilingual and high-resolution input images.

Acknowledgements

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by the Fundação para a Ciência e Tecnologia (FCT), specifically through the project with reference UIDB/50021/2020 (DOI: 10.54499/UIDB/50021/2020), and the project with reference UIDP/04516/2020 (DOI: 10.54499/UIDB/04516/2020).