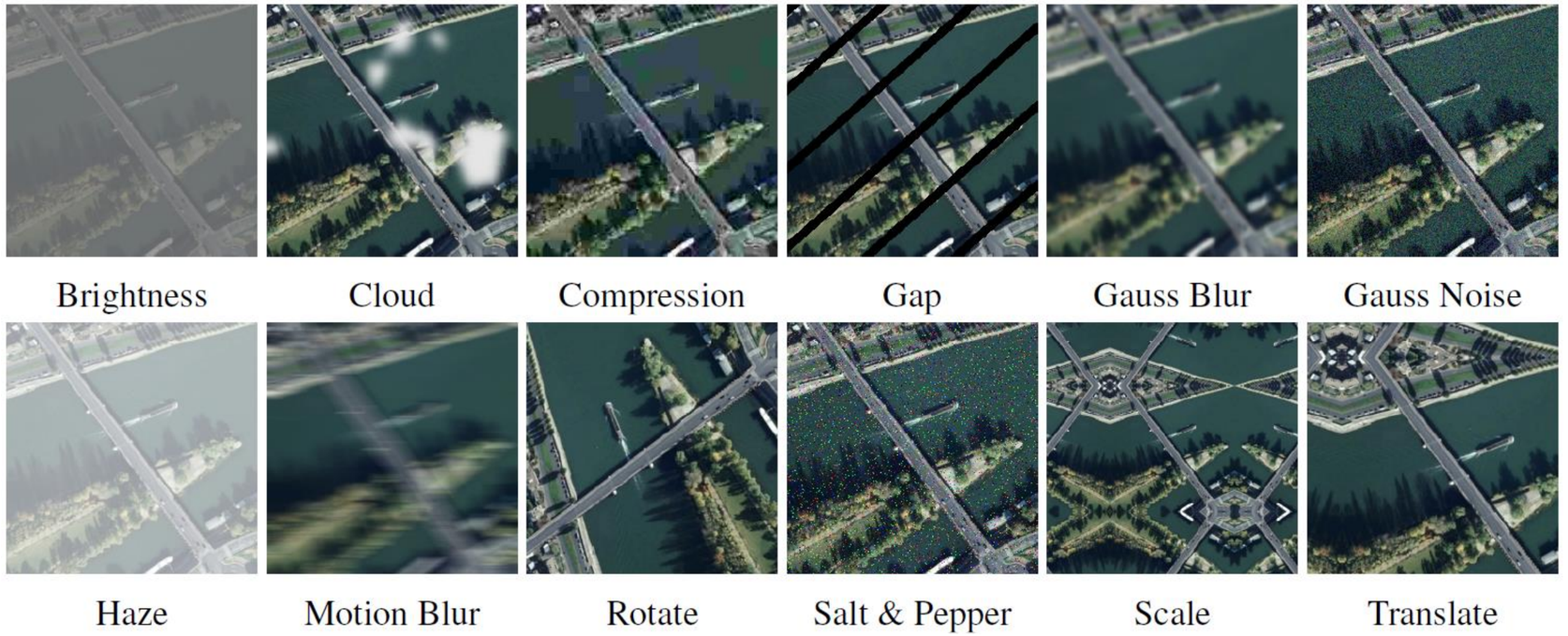
REOBench:Benchmarking Robustness of Earth Observation Foundation Models



Xiang Li[1*], Yong Tao[2*], Siyuan Zhang[3*], Siwei Liu[4], Zhitong Xiong[5], Chunbo Luo[2],
Lu Liu[2], Mykola Pechenizkiy[6], Xiao Xiang Zhu[5], Tianjin Huang[2,6†]

NEURAL INFORMATION PROCESSING SYSTEMS
University of BRISTOL
University of Exeter
1933
1495 UNIVERSITY OF ABERDEEN
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Robustness

Background:

Earth observation **foundation models** (EOFMs) have shown strong generalization across multiple tasks.

Motivation:

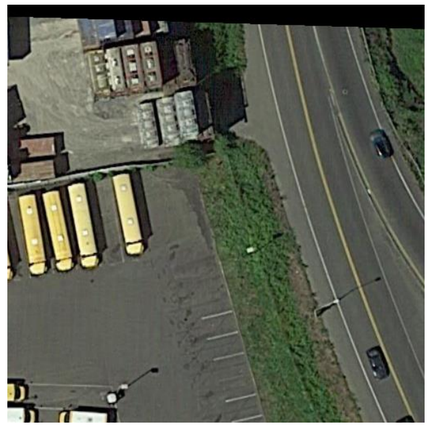
Real-world images are often affected by corruptions such as noise, blur, haze, and geometric distortions.

Question:

How robust are EOFMs under these real-world corruptions?

REOBench

A comprehensive benchmark to systematically evaluate robustness across tasks and corruptions



Clean Image

Corrupted Image

What color are the large vehicles seen in the image?
GT: Yellow, **GeoChat**: Teal, **SkySenseGPT**: teal, **VHM**: yellow.

How many small vehicles are visible in the image?
GT: 2, **GeoChat**: 11, **SkySenseGPT**: 2, **VHM**: 3.

Is there a vehicle located at the top-most position in the provided image?
GT: Yes, **GeoChat**: Yes, **SkySenseGPT**: Yes, **VHM**: no.

What is the orientation of the road in the image?
GT: North-South, **GeoChat**: Right, **SkySenseGPT**: horizontal, **VHM**: north-south.

REOBench Dataset

6 Task

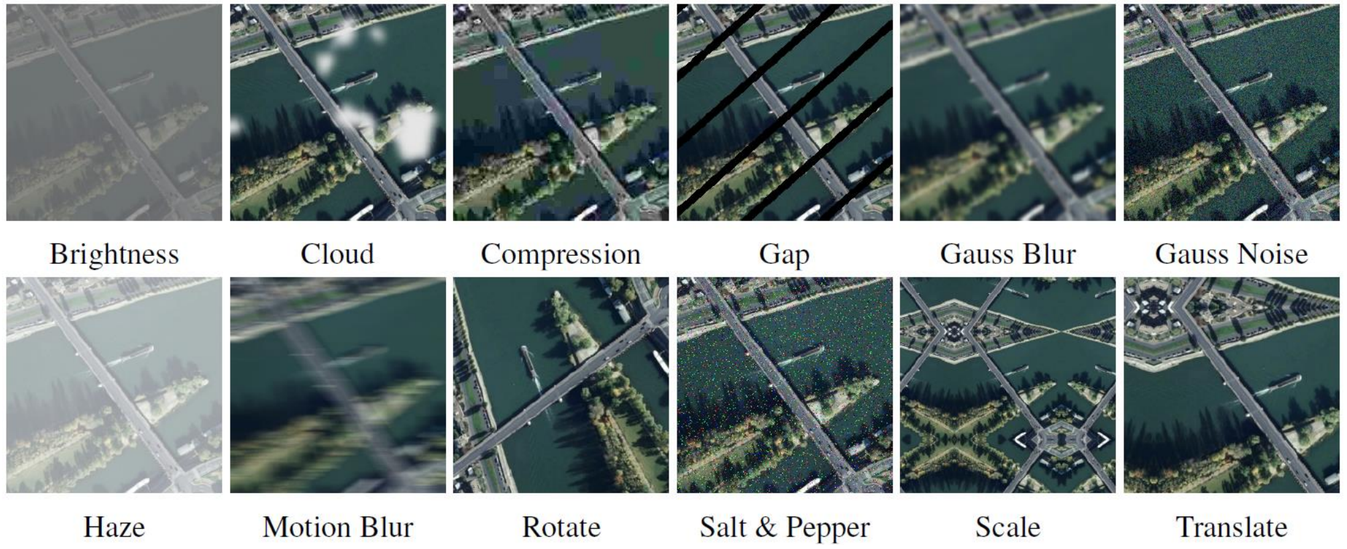
- Classification
- Segmentation
- Object Detection
- Caption
- VQA
- Grounding

12 synthetic corruptions

Environmental Corruptions
Cloud, Brightness, Haze distortions that caused by atmospheric and lighting changes

Sensor-induced Corruptions
Blur, Noise, Compression and Sensor Gaps during sensor capture or data transmission

Geometric Corruptions
Rotation, Scale, and Translation distortions caused by platform motion misalignment



5 Severity

Definition of Corruption Robustness

- We define robustness as a model's ability to maintain task performance under realistic geophysical and sensor-induced corruptions.
- Robustness is quantified by **Relative Task Performance Drop ($\mathcal{R}_{TP}$)** — the relative decrease in accuracy under corrupted vs. clean data:

Performance on clean images Performance on corrupted images

$$\mathcal{R}_{\mathrm{TP}} = \frac{\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x)=y] - \mathbb{E}_{c\sim\mathcal{C}}\left[\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(c(x))=y]\right]}{\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x)=y]}$$

Smaller $\mathcal{R}_{TP}$ ↓ indicates **higher robustness ↑**

$C = \{C_1, \ldots, C_K\}$ is a set of physically plausible corruption operator, such as haze, cloud occlusion, or sensor noise.

Benchmark Robustness on EOFMs

Evaluate **robustness** across **different tasks, model paradigms, and backbone scales**.

- 6 Tasks: classification, segmentation, detection, captioning, VQA, and grounding.
- 3 Model Types: **MIM-based** (SatMAE, RVSA…), **CL-based** (RemoteCLIP, GeoRSCLIP), and **LLM-based** (GeoChat, RS-LLaVA…)

Table 1: Scene classification performance on AID dataset across different image perturbations. *zs* denotes zero-shot evaluation.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{TP}$
MIM-based																
SATLAS [57]	Swin-B	90.85	82.54	84.32	73.36	67.23	78.10	79.16	80.46	32.44	72.54	77.56	72.54	88.54	74.07	18.47
SatMAE [5]	ViT-L	72.05	44.82	59.58	67.26	46.49	71.33	71.25	28.31	63.85	69.15	70.45	59.74	66.12	59.86	16.92
Scale-MAE [7]	ViT-L	75.75	51.80	72.65	39.60	43.69	31.65	46.31	55.24	17.49	66.15	47.27	61.58	69.84	50.27	33.64
RVSA [46]	ViT-B	84.60	56.84	77.33	56.07	53.14	53.53	32.51	49.19	23.45	76.88	35.12	71.78	77.22	55.26	34.69
SatMAE++ [47]	ViT-L	91.35	64.62	82.64	62.69	60.70	48.23	76.98	62.56	29.43	85.49	73.22	75.79	87.61	67.50	26.11
CL-based																
RemoteCLIP$_{zs}$ [9]	ViT-L	81.10	78.32	80.64	73.91	79.43	76.83	76.72	80.10	57.02	82.80	70.90	68.39	80.72	75.48	6.93
RemoteCLIP [9]	ViT-B	96.85	90.80	95.36	91.13	88.96	89.18	94.25	87.46	63.75	96.22	91.43	83.62	**95.42**	88.97	8.15
RemoteCLIP [9]	ViT-L	95.45	93.11	93.80	88.77	94.21	92.47	94.20	93.37	**74.45**	95.01	86.99	83.37	94.06	90.32	5.38
GeoRSCLIP$_{zs}$ [10]	ViT-L	66.05	62.41	65.47	60.45	64.41	62.03	62.32	62.24	44.20	65.52	58.88	52.59	64.25	60.40	8.55
GeoRSCLIP [10]	ViT-B	96.90	93.59	96.04	91.01	93.34	92.60	92.99	92.91	57.78	95.70	88.22	75.70	93.87	88.65	8.51
GeoRSCLIP [10]	ViT-L	**97.40**	**96.27**	**96.45**	**92.28**	**95.62**	**96.09**	**95.68**	**96.00**	71.03	**97.20**	**92.75**	**77.16**	95.15	**91.80**	5.74
LLM-based																
GeoChat [12]	ViT-L	65.85	64.67	65.26	60.71	64.61	63.32	62.34	64.54	48.68	65.05	62.32	56.21	62.91	61.72	6.27
LHRS-Bot [50]	ViT-L	87.75	87.38	86.82	78.53	85.95	84.94	82.62	87.62	67.76	87.31	76.73	79.07	86.71	82.79	5.65
RS-LLaVA [51]	ViT-L	67.55	65.36	68.95	63.05	67.69	65.73	63.13	65.97	43.86	66.55	68.24	54.89	63.40	63.07	6.63
SkySenseGPT [53]	ViT-L	87.35	87.72	87.91	79.66	87.67	84.78	83.08	87.71	63.11	86.86	83.61	75.49	85.25	82.74	5.28
VHM [52]	ViT-L	80.60	79.81	80.67	76.10	80.82	78.81	76.92	79.55	59.74	80.47	75.43	72.57	79.73	76.72	**4.81**

Overall degradation of Scene classification

- All models drop largely under corruptions
- MIM-based methods are most affected

Robustness comparison

- CL & LLM models are more robust than MIM
- High-level semantics less sensitive to noise

Performance ranking

- CL-based models perform best overall
- GeoRSCLIP → highest accuracy
- VHM → smallest performance drop

Benchmark Robustness on EOFMs

Table 2: Semantic segmentation performance (mIoU) on the ISPRS Potsdam dataset under different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{TP}$
MIM-based																
SatMAE [5]	ViT-L	59.51	50.18	37.39	48.23	51.15	57.89	41.83	44.95	57.52	56.02	36.07	54.81	59.09	49.59	16.67
ScaleMAE [7]	ViT-L	68.92	64.37	65.43	**49.96**	41.84	64.86	**54.89**	**63.89**	64.49	64.51	**52.12**	65.45	68.38	**60.02**	12.91
RVSA [46]	ViT-B	**69.82**	**64.71**	**65.67**	47.99	45.34	**66.89**	48.89	61.52	**65.25**	**65.58**	45.26	**66.87**	**69.23**	59.43	14.88
SatMAE++ [47]	ViT-L	62.68	53.91	59.06	49.74	**53.94**	60.38	44.32	48.80	60.44	58.34	39.45	58.13	61.94	54.04	13.78
CL-based																
RemoteCLIP [9]	ViT-B	50.28	42.32	45.27	39.33	37.78	50.26	48.46	36.61	50.06	46.48	46.91	48.39	49.63	45.12	10.26
RemoteCLIP [9]	ViT-L	56.69	54.51	52.73	43.24	51.19	50.82	45.12	50.47	51.82	53.68	38.98	54.59	56.53	50.31	11.25
GeoRSCLIP [10]	ViT-B	51.44	42.89	46.41	40.37	38.64	51.35	49.79	38.56	51.15	47.89	48.24	49.28	50.87	46.29	**10.01**
GeoRSCLIP [10]	ViT-L	56.81	54.97	52.53	43.28	41.37	50.49	42.7	49.41	51.36	53.54	36.98	54.66	56.64	48.99	13.77

Table 3: Object detection performance (mAP) on the DIOR dataset across different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{TP}$
MIM-based																
SatMAE [5]	ViT-L	62.30	56.84	57.86	55.80	58.36	55.38	58.44	59.34	56.92	56.60	53.76	51.58	60.90	56.82	8.81
ScaleMAE [7]	ViT-L	70.20	64.80	65.98	62.50	64.46	62.58	**63.82**	66.10	**63.08**	63.44	**60.50**	53.08	68.26	63.22	9.94
RVSA [46]	ViT-B	**70.96**	60.59	65.02	61.58	64.60	62.35	62.87	63.98	62.88	**64.04**	56.61	55.97	**69.69**	62.51	11.91
SatMAE++ [47]	ViT-L	65.20	59.44	61.02	60.30	59.88	59.66	61.06	61.72	59.56	59.14	58.64	48.48	64.70	59.47	**8.79**
CL-based																
RemoteCLIP [9]	ViT-B	60.40	56.72	56.28	56.78	54.56	53.68	57.36	55.90	53.42	54.54	54.40	44.92	59.72	54.86	9.17
RemoteCLIP [9]	ViT-L	70.20	**66.52**	**66.62**	63.84	**65.40**	**63.62**	63.68	**66.76**	62.66	63.52	59.16	**57.42**	68.64	63.99	8.85
GeoRSCLIP [10]	ViT-B	60.20	56.28	56.04	56.08	55.46	53.38	56.92	55.50	53.38	53.98	53.48	46.98	59.32	54.73	9.09
GeoRSCLIP [10]	ViT-L	69.80	66.12	65.34	**65.34**	64.96	**63.62**	62.90	66.04	62.02	62.68	56.04	57.40	68.10	**63.38**	9.20

Semantic Segmentation

- Both MIM & CL models drop >10% mIoU under corruptions
- MIM-based models perform better on both clean & noisy data,but decline more under corruptions

Object Detection

- Both MIM & CL models drop >8% mAP under corruptions
- Comparable results on clean and corrupted data

Benchmark Robustness on EOFMs

Table 4: Image captioning performance (CLAIR) on the VRSBench-Cap dataset across different image perturbations. *ft* denotes models trained on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{TP}$
GeoChat [12]	ViT-L	41.39	40.06	40.45	37.65	40.20	39.76	38.48	40.38	39.92	37.61	39.59	4.35
SkySenseGPT [53]	ViT-L	48.29	47.21	46.64	44.22	46.25	45.52	44.97	46.14	45.13	44.36	45.60	5.57
VHM [52]	ViT-L	52.02	50.19	50.82	50.26	50.57	51.22	50.46	50.39	50.72	49.48	50.46	3.00
RS-LLaVA [51]	ViT-L	51.30	51.15	50.43	51.78	50.54	52.01	47.84	50.57	49.88	48.12	50.26	**2.03**
Falcon [54]	DaViT-B	**61.90**	**59.98**	**60.09**	**57.13**	**59.48**	**57.43**	**56.31**	**59.85**	**59.94**	**51.83**	**58.01**	6.28
GeoChat$_{ft}$ [12]	ViT-L	71.26	69.00	68.93	66.60	69.45	68.63	67.83	69.98	69.02	63.87	68.15	4.36

Table 5: VQA performance (Accuracy) on the VRSBench-VQA dataset across different image perturbations. *ft* indicates models fine-tuned on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{TP}$
GeoChat [12]	ViT-L	56.63	53.89	54.82	55.14	55.99	55.88	55.44	56.22	54.08	54.04	55.06	2.77
LHRS-Bot [50]	ViT-L	35.72	35.72	35.69	35.72	35.72	35.72	35.72	35.72	35.34	35.72	35.56	**0.45**
SkySenseGPT [53]	ViT-L	60.21	59.26	59.73	57.93	59.64	59.21	58.27	59.63	59.17	57.27	58.90	2.18
VHM [52]	ViT-L	**61.72**	**60.91**	**61.07**	**60.40**	**61.49**	**60.91**	**60.91**	**61.12**	**59.97**	**60.39**	**60.90**	1.33
RS-LLaVA [51]	ViT-L	57.25	57.04	57.14	55.45	57.25	57.14	55.97	57.21	55.25	55.82	56.47	1.36
Falcon [54]	DaViT-B	33.27	32.83	32.70	32.19	33.30	33.43	32.85	32.76	32.97	31.55	32.73	1.59
GeoChat$_{ft}$ [12]	ViT-L	75.79	75.13	74.97	73.84	75.63	74.89	74.46	75.43	74.76	72.77	74.65	1.50

Table 6: Visual grounding performance on the VRSBench-Ref dataset across different image perturbations. We report grounding accuracy at an IoU threshold of 0.5. * indicates the GeoGround model includes VRSBench in its training data. *ft* indicates models fine-tuned on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{TP}$
GeoChat [12]	ViT-L	18.96	17.09	16.54	16.52	16.19	16.61	16.93	17.09	16.91	16.57	16.72	11.81
VHM [52]	ViT-L	37.20	34.66	35.29	34.18	35.48	35.01	35.78	35.54	32.21	34.58	34.74	6.61
GeoGround* [59]	ViT-L	**75.93**	**73.57**	**71.57**	**71.30**	**72.23**	**73.23**	**72.92**	**74.06**	**72.11**	**71.77**	**72.53**	**4.48**
Falcon [54]	DaViT-B	73.30	71.31	69.92	65.83	68.61	70.79	64.28	71.04	68.17	59.53	67.72	7.61
GeoChat$_{ft}$ [12]	ViT-L	55.50	53.79	52.20	50.51	53.06	53.11	51.57	54.23	52.99	49.82	52.36	5.66

Image Captioning

- All models degrade under noise.
- Fine-tuned GeoChat improves performance with similar robustness.

Visual Question Answering (VQA)

- All LLM-based models show moderate decline under perturbations.
- Fine-tuned GeoChat outperforms zero-shot models and improves robustness.

Visual Grounding

- All models drop under perturbations.
- Fine-tuned GeoChat improves accuracy and robustness.

Robustness Analysis

- **Vision-language** models more robust than vision-centric models
- **Classification** tasks most sensitive to corruptions
- **Larger backbones** are more robust but can hurt fine-grained tasks.
- **Motion blur** causes largest performance drops
- Compound corruptions amplify performance degradation
- Multispectral models still brittle under perturbations

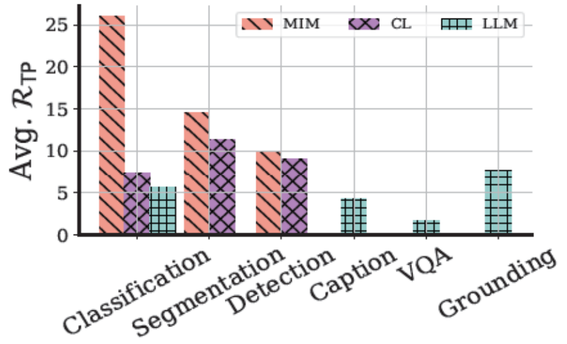


Figure 2: Robustness across different tasks and model architectures. We report the average $\mathcal{R}_{TP}$ across models.

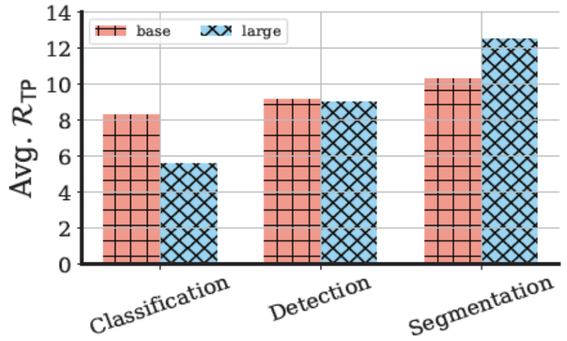


Figure 3: Robustness across different backbone sizes. We report the average $\mathcal{R}_{TP}$ for RemoteCLIP and GeoRSCLIP.

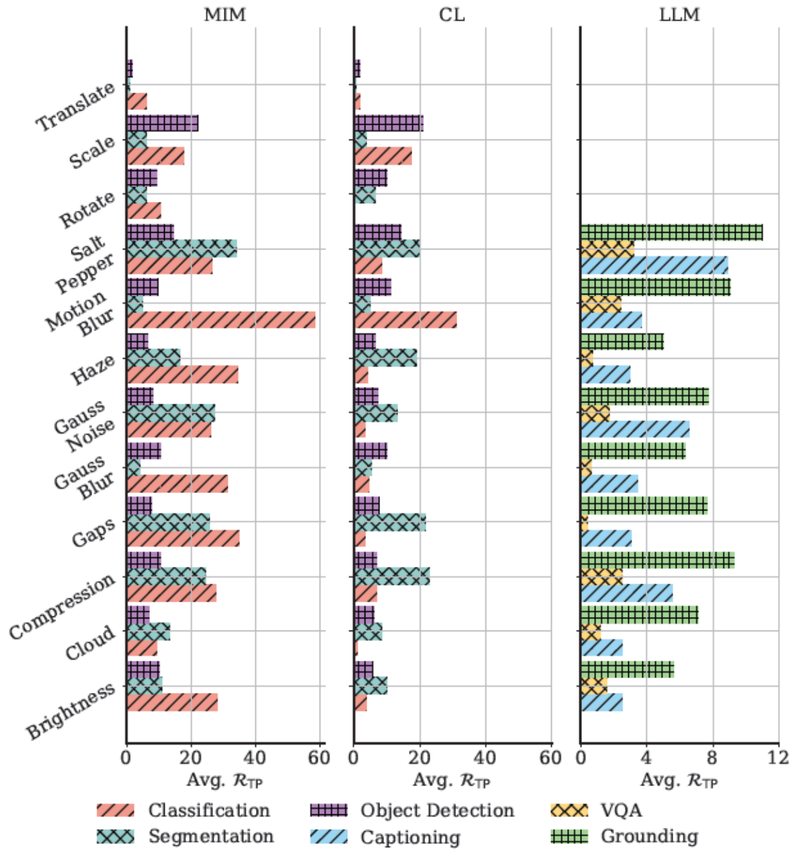


Figure 4: Robustness across different types of corruptions. We report the $\mathcal{R}_{TP}$ across models.

Robustness Analysis

- Vision-language models more robust than vision-centric models
- Classification tasks most sensitive to corruptions
- Larger backbones are more robust but can hurt fine-grained tasks.
- Motion blur causes largest performance drops
- **Compound corruptions** amplify performance degradation
- **Multispectral** models still brittle under corruptions

Table 7: Object Detection performance (mAP) on DIOR-R under compound corruptions

Model	Clean	Brightness	Clouds	Compression	Brightness + Clouds	Brightness + Compression	Clouds + Compression	All Three
Brightness	✗	✓	✗	✗	✓	✓	✗	✓
Clouds	✗	✗	✓	✗	✓	✗	✓	✓
Compression	✗	✗	✗	✓	✗	✓	✓	✓
RemoteCLIP	60.40	56.72	56.28	56.78	49.98	45.36	52.57	40.58
GeoRSCLIP	60.20	56.28	56.04	56.08	50.27	44.90	52.39	40.94

Table 8: Scene classification performance on the multispectral dataset across different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{\text{TP}}$
SatMAE [5]	fMoW-S2 [60]	59.75	37.46	58.69	33.58	50.01	29.35	36.64	41.57	40.12	38.93	51.79	43.35	59.68	43.43	27.31
EarthDial [62]	BigEarthNet [61]	46.521	31.47	46.48	45.40	34.97	36.37	38.69	33.25	39.84	29.51	22.71	39.18	45.30	36.93	20.62

Thank you for your Attention

Github available here: