Frozen Truths, Melting Lies: Hallucination and Overconfidence in Nordic Satellite VLMs

Abstract

Automated satellite image interpretation is becoming important for climate monitoring. However, the reliability of zero-shot vision-language models (VLMs) is under-researched, especially in the context of boreal and subarctic environments. Hence, we employ CLIP, RemoteCLIP, and BLIP-ITM on three diagnostic datasets, ranging from clean baseline conditions to Nordic seasonal shift and geographic OOD shift (n = 500–2000), to investigate the hallucination rate, calibration error, and the reliability of a training-free trust proxy (VTCS). Strikingly, we find that hallucination rates reach 49–89% and in the context of Nordic seasonal shift confidence scores become *anti-correlated with correctness in summer* (AUROC 0.48, below chance). The same signal becomes slightly more reliable in winter (AUROC 0.70). Conversely, VTCS is more stable across seasons (0.59 in summer, 0.57 in winter) and provides better overall discrimination, but it erodes under geographic shift. Subsequently, a cross-model comparison confirms that the failure is consistent across all tested architectures. Overall, these findings imply that per-image confidence thresholds are unreliable for operational Nordic deployment and season-specific recalibration is necessary. In Nordic summer conditions, filtering predictions by confidence systematically *increases* error.

CCS Concepts

• **Computing methodologies** → **Computer vision**; *Machine learning*.

Keywords

vision-language models, CLIP, hallucination, calibration, remote sensing, Nordic climate, seasonal shift, trustworthy AI

ACM Reference Format:

. 2026. Frozen Truths, Melting Lies: Hallucination and Overconfidence in Nordic Satellite VLMs. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Nordic and subarctic regions are disproportionately affected by climate change [15], and thus, reliable automated satellite image interpretation becomes critical for land-cover monitoring, flood detection, and cryosphere assessment [13, 14]. AI-based approaches demonstrated strong performance on curated remote sensing benchmarks [5, 17], but the core deployment challenge in Nordic contexts is slightly unique - the same landscape looks entirely different in summer (mixed boreal forest, wetlands, coastline) than in winter (snow-covered fields, frozen lakes, sea ice). Thus, in Nordic and subarctic monitoring settings, seasonal shift is a routine deployment condition [15].

Large vision-language models (VLMs) trained by contrastive learning [11] offer zero-shot scene classification without task-specific fine-tuning. Domain-adapted RS variants (RemoteCLIP [7], EarthGPT [17], SkyEyeGPT [16]) have extended this to remote sensing benchmarks, but the puzzlinng question is - whether any of these models remain *trustworthy* under the seasonal shifts of Nordic deployment - especially, when VLMs are known for Hallucination [8, 12] and miscalibration [3, 10]. This is crucial omission in the literature. Hence, we investigate:

RQ1 *What failure modes do zero-shot VLMs exhibit on Nordic satellite imagery, and how do they change under seasonal and geographic shift?*

RQ2 *Do training-free trust proxies reliably detect errors under Nordic shift, and does this hold across model architectures?*

RQ3 *What is the most actionable mitigation for practitioners deploying VLMs on Nordic satellite data?*

Contributions: Broadly, we enrich the extant literature by investigating a seasonally stratified Nordic evaluation protocol: three datasets designed to expose distinct failure modes, including the first controlled seasonal split (summer vs. winter, n=1,000 each) for zero-shot VLM evaluation on Nordic Sentinel-2 imagery, addressing RQ1 (§2). Second, we find that the confidence becomes *anti-correlated with correctness in summer* (AUROC 0.481, below chance) while recovering partial usefulness in winter (0.702) — a result explained by CLIP's texture-dominant embeddings and with direct practical implications for practitioners, addressing RQ2 across three architectures (§3). Lastly, our season-aware mitigation analysis shows that per-image recalibration is insufficient and seasonal metadata is the actionable lever, addressing RQ3 (§4).

2 Datasets and Method

Our *diagnostic system* is as following: each dataset targets a specific failure mode of zero-shot VLMs under Nordic conditions (Table 1). All three datasets use the same shared label vocabulary of 20 geospatial concepts (water, river, lake, ocean, snow, ice, glacier, forest, vegetation, wetland, farmland, cropland, urban, residential, road, bare land, mountain, cloud, smoke, coastline). Hence, our results are directly comparable and interpretable.

EuroSAT consists of 64×64 px RGB Sentinel-2 tiles covering 10 land-use/land-cover classes: AnnualCrop, Forest, HerbaceousVegetation, Highway, Industrial, Pasture, PermanentCrop, Residential, River, and SeaLake [4]. Tiles are centred on European locations spanning diverse climatic zones. We conisder 50 images per class (n=500) from the standard test split. EuroSAT is the upper bound: it represents the best case for zero-shot RS classification — a clean, balanced, single-label dataset with no distribution shift.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Sentinel-2 Nordic (S2 Nordic) is our core contribution to dataset design. This dataset is acquired from the Copernicus Sentinel-2 Level-2A archive within a Nordic bounding box (55°N–71°N, 4°E–32°E), covering Scandinavia, Finland, and the Baltic coast [1]. Images are 256×256 px RGB composites (bands B4–B3–B2) at 10 m ground resolution. Ground-truth labels are multi-label strings derived from land-cover metadata and manual verification, with 3–6 concepts per tile (e.g., *boreal, forest, vegetation, wetland*).

The seasonal split is the defining feature of this dataset. Summer tiles are acquired during June–August (peak vegetation, open water), and winter tiles during December–February (snow cover, frozen lakes, ice on coastal margins). We use 1 000 tiles per season (n=2,000 total), balanced across the geographic bounding box. Prompt templates include a Nordic-aware variant (*"a Nordic landscape with {}"*), in addition to generic RS templates, to investigate whether domain-specific prompting impacts confidence.

Floods OOD consists of 300 Sentinel-2 tiles capturing active flood events across four geographic regions: South Asia (Bangladesh, 2020 monsoon), sub-Saharan Africa (Mozambique, Cyclone Idai), South America (Paraguay river floods), and Southeast Asia (Mekong Delta). Ground truth is multi-label, reflecting the complex spectral mixing of floodwater with existing land cover — a typical tile carries labels such as *cropland, flood, inundation, river, vegetation, water* simultaneously. This dataset is interesting for two reasons: it is geographically outside the Nordic domain, and its dominant visual feature (turbid floodwater, sediment plumes) is absent from CLIP's standard RS training vocabulary. The label vocabulary is extended with flood-specific concepts (*flood*, *inundation*, *sediment*) that do not appear in EuroSAT or the Sentinel-2 Nordic label set. This setting intentionally combines label-space shift and geographic distribution shift to stress-test hallucination under the most demanding OOD conditions.

Dataset	N	Failure mode
EuroSAT [4]	500	Clean baseline (upper bound)
Sentinel-2 Nordic	2,000	Seasonal shift, confidence collapse
Floods OOD (4 regions)	300	Geographic OOD, hallucination

Table 1: Diagnostic datasets and targeted failure modes.

2.1 Zero-Shot Protocol and Models

Iniitally, all models run zero-shot with no fine-tuning or adaptation (except RemoteCLIP's RS pre-training). We evaluate three architectures: **CLIP** (`openai/clip-vit-base-patch32` [11]), **RemoteCLIP** [7] (same ViT-B/32 backbone fine-tuned on 105 k RS image-text pairs from RSICD, UCM, and NWPU-RESISC45), and **BLIP-ITM** [6] (`Salesforce/blip-itm-base-coco`, scored via cross attention image-text matching probability rather than cosine similarity).

We compute Label embeddings by averaging text features over three prompt templates (*"a satellite image of {}"*, *"an aerial image of {}"*, *"a remote sensing image of {"*) [9]. Following prior studies, predictions are argmax over the 20-concept vocabulary, and confidence is the maximum label-softmax probability. A sample is marked *correct* if the predicted concept appears in the ground-truth label set (any-match over the multi-label ground truth).

2.2 Metrics and Trust Signal

Hallucination Rate (HR): For single-label datasets, HR = 1 − Acc. For multi-label datasets (S2 Nordic and Floods OOD), a prediction *hallucinates* if the predicted concept is entirely absent from the ground-truth label set — in other words, the model invents a concept with no correspondence to the scene.

Expected Calibration Error (ECE): We group confidence scores into 10 equal-width bins. ECE is the weighted average of |accuracy−confidence| per bin [3, 10].

Visual-Textual Consistency Score (VTCS): We define VTCS as a training-free reliability estimate requiring no ground-truth labels at inference time as follows:

$$\mathrm{VTCS}(\mathbf{x}) = \alpha \cdot s(\mathbf{x}) - \beta \cdot H\big(p(\cdot|\mathbf{x})\big), \tag{1}$$

where $s(\mathbf{x})$ is the image-text cosine similarity to the top-predicted label, H is Shannon entropy of the label probability distribution, and α=1.0, β=0.25 are fixed (not tuned on labelled data). A trustworthy prediction should be both embedding-consistent (high s) and decisive (low H). We consider −VTCS, as an error risk score and compute it by *Error AUROC*, i.e., the probability that a randomly chosen incorrect prediction receives a higher risk score than a randomly chosen correct one. Bootstrap 95% CIs use B=1,000 stratified resamples.

3 Results

We find that Zero-shot CLIP exhibits hallucination, confidence collapse, and trust-signal erosion across all three datasets, with each failure mode intensifying as distribution shift increases (Table 2).

Dataset	Acc.	HR	ECE	VTCS [CI]	Conf.
EuroSAT	**0.486**	0.514	0.401	**0.825** [.787,.859]	**0.085**
S2 Nordic	0.507	0.493	0.456	0.567 [.542,.593]	0.051
Floods OOD	0.351	**0.892**	**0.300**	0.514 [.477,.550]	0.051

Table 2: CLIP (ViT-B/32) zero-shot results. VTCS = error-detection AUROC, bootstrap 95% CI (B=1000); bold = best per column.

Hallucination (RQ1): Alarmingly, hallucination is pervasive and worsens with shift. On EuroSAT, 51% of predictions introduce absent concepts. This rises to 89% on Floods OOD, where CLIP predicts land-use categories (*cropland*, *urban*) on scenes dominated by floodwater. Similarly, on S2 Nordic, 49% hallucinate: summer boreal scenes (mixed forest, wetland) are frequently predicted as *ice* or *snow*, while winter frozen lakes are predicted as *water* or *vegetation* — precisely, opposite-season hallucinations.

Confidence collapse (RQ2): Surprisingly, the mean confidence drops 40% from EuroSAT (0.085) to S2 Nordic and Floods (0.051), suggesting the model recognises distributional uncertainty at a coarse level. However, ECE remains high across all conditions (0.300–0.456), indicating uniform low-confidence across images rather than genuine calibration. The range of confidence scores

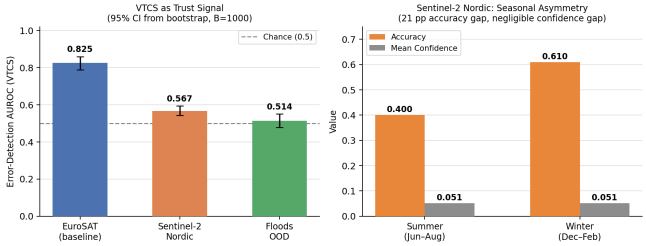

Figure 1: Left: Error-detection AUROC (VTCS) per dataset for CLIP. The signal is strong on clean data (0.825) and erodes to near-random under gradual shift (0.567, 0.514). Right: Seasonal accuracy vs. confidence on Sentinel-2 Nordic. Winter accuracy (61%) is 1.5× summer (40%) despite near-identical confidence — the model cannot signal its own seasonal performance gap.

collapses to < 0.003 on S2 Nordic because the model assigns near-identical probabilities to all images, making per-image confidence effectively useless for filtering.

VTCS erosion (RQ2): VTCS AUROC reaches 0.825 on EuroSAT (well above the MSP baseline of 0.726), but erodes monotonically: 0.567 on S2 Nordic, 0.514 on Floods OOD. The erosion follows the shift severity, confirming that training-free trust signals degrade precisely where deployment risk is highest.

VTCS Ablation: We consider VTCS, as a unifying formulation whose decomposition reveals *which component actually matters*: image-text similarity drives all discrimination, similarity-only equals full VTCS to three significant figures in every condition, and the entropy penalty contributes nothing. On S2 Nordic, every component falls near chance. MSP (0.527) is marginally better off, while the similarity-based signals invert (0.385), meaning the model is *more* cosine-similar to wrong labels under seasonal shift than to correct ones. This points to a fundamental embedding misalignment, not a calibration parameter problem that can be fixed post-hoc.

Dataset	MSP	Ent	Sim	VTCS
EuroSAT	0.726	0.680	0.809	0.809
S2 Nordic	0.527	0.427	0.385	0.385
Floods OOD	0.519	0.483	0.514	0.514

Table 3: Error-detection AUROC by signal component. MSP = max softmax prob.; Ent = $-H$; Sim = $s(\mathbf{x})$; VTCS = full Eq. (1).

Cross-Model Comparison (RQ3): To test whether trust-signal erosion is model-specific, we compare CLIP with domain-adapted RemoteCLIP and cross-attention BLIP-ITM across all three datasets (Table 4, Figure 2).

Domain adaptation is distribution-specific: RemoteCLIP's RS fine-tuning yields VTCS 0.827 on S2 Nordic (vs. CLIP's 0.567), confirming that domain adaptation improves trust on the matched distribution. On Floods OOD (unseen shift), the advantage reverses: RemoteCLIP falls to 0.488 (below chance), lower than CLIP's 0.514. Adaptation sharpens confidence for known conditions while making it more unreliable for unexpected ones.

Model		EuroSAT	S2 Nordic	Floods
CLIP	Acc	0.486	0.507	0.351
	VTCS	0.825	0.567	0.514
RC	Acc	0.187	0.387	0.343
	VTCS	0.543	0.827	**0.488**
BLIP	Acc	0.310	0.530	0.610
	VTCS	0.739	0.735	0.553

Table 4: Cross-model VTCS AUROC and accuracy. Below-chance AUROC in bold. RC = RemoteCLIP; BLIP = BLIP-ITM.

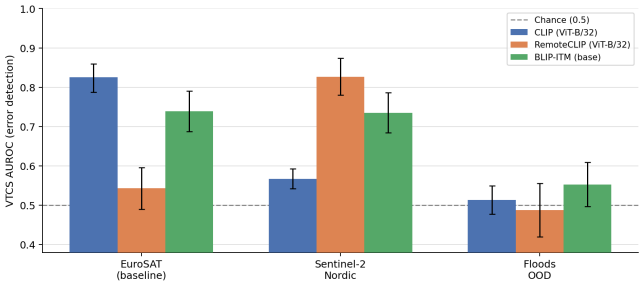

Figure 2: VTCS AUROC per model and dataset (95% CI error bars). Dashed line = chance (0.5). RemoteCLIP peaks on S2 Nordic (its adapted domain) but falls below chance on Floods OOD. BLIP-ITM is most stable but never reaches CLIP's peak.

Cross-attention provides stability, not peak performance: BLIP-ITM's VTCS range (0.553–0.739) is the narrowest across models, suggesting cross-attention scoring reduces the embedding-alignment aspect driving contrastive-model confidence under shift, at the cost of lower ceiling on clean data.

Erosion is model-agnostic: All three models show VTCS decline from EuroSAT to Floods OOD. The failure is a property of the distribution shift, not of any particular architecture.

4 Diagnosing the Confidence Signal

Analysing per-image scores from the full S2 Nordic run (n=2,000) reveals a striking asymmetry (Table 5). Raw confidence is a *misleading* error predictor in summer (AUROC 0.481, below chance), i.e., images where CLIP assigns higher confidence are not more likely to be correct, the opposite holds. Conversely, in winter, the same signal becomes genuinely useful (AUROC 0.702). VTCS shows a more stable pattern: it is slightly more informative in summer (0.594) than in winter (0.573), and remains above chance in both seasons, making it the safer choice when the season is unknown. Adjusting thresholds does not change these AUROCs. Season-specific VTCS thresholds improve F1 by only 0.003 pp, confirming that the problem lies in the signal quality, not the operating point.

Why Per-Image Recalibration Fails: The confidence collapse (range < 0.003) means all images receive near-identical softmax probabilities. Temperature scaling, Platt calibration, or threshold tuning cannot recover discrimination from a near-constant signal. The root cause is embedding misalignment (Table 3), i.r., the model is more cosine-similar to wrong labels under seasonal shift,

Signal	Overall	Summer	Winter
Raw confidence	0.620	0.481†	0.702
VTCS (Eq. 1)	0.567	0.594	0.573

Table 5: Trust signal reliability on S2 Nordic by season. AUROC < 0.5 (†) = below-chance, misleading predictor.

hence, similarity-based signals (VTCS, MSP) do not correlate with correctness.

The Way Forward: Season-Aware Deployment? Since neither per-image confidence nor VTCS provides reliable individual predictions under Nordic shift, the most actionable signal is *season as deployment metadata*:

- Summer accuracy is 40% and winter is 61% — a 20 pp gap predictable from the calendar, not from the model output.
- In winter, raw confidence is a valid error indicator (AUROC 0.702) and can be used for filtering.
- In summer, raw confidence is worse than random. We should use VTCS (AUROC 0.594), or fall back to season-level uncertainty by flagging all summer predictions at a higher review rate.

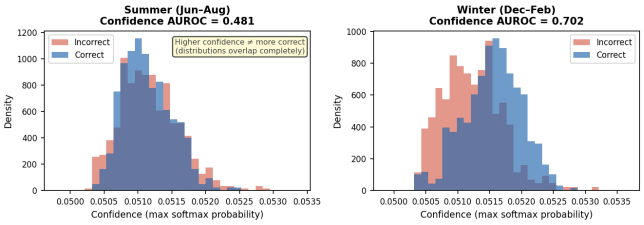

Figure 3: Confidence score distribution for correct vs. incorrect predictions on Sentinel-2 Nordic, split by season. In summer (left), incorrect predictions receive marginally *higher* confidence than correct ones — filtering on high-confidence predictions systematically retains errors. In winter (right), the relationship reverses and confidence becomes a useful signal. Both distributions are compressed into a < 0.003 range, confirming the overall confidence collapse.

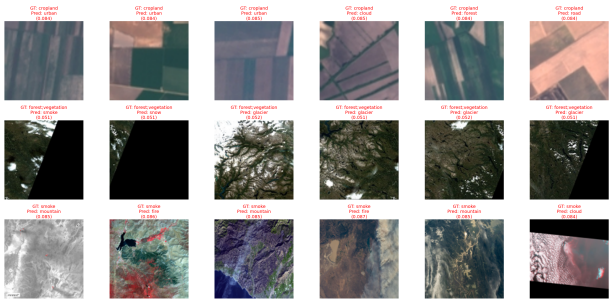

Figure 4: Representative failure cases. Ground-truth → CLIP prediction with confidence. EuroSAT (top): ambiguous land-use misclassifications. S2 Nordic (middle): seasonal hallucinations — summer vegetation predicted as *ice*, winter frozen lake as *ocean*. Floods OOD (bottom): floodwater predicted as *cropland* or *urban*.

5 Discussion

The winter accuracy paradox: Winter accuracy (61%) exceeds summer (40%) despite nominally harder conditions. The underlying mechanism is a spectral-semantic mismatch consistent with texture-dominant visual representations [2]: contrastive training on internet-scale image-text pairs produces embeddings that respond strongly to low-level texture features, i.e., uniform white surfaces score high cosine similarity with *snow* and *ice*, giving winter an accuracy boost that is not grounded in semantic understanding. Conversely, summer boreal landscapes (mixed forest, wetland, coastline) are spectrally heterogeneous and do not cluster around any single high-frequency concept in CLIP's vocabulary, so the model fails more often. The *anti-correlation* of confidence with correctness in summer (Figure 3) is intuitive: the model assigns slightly higher confidence to predictions that match a dominant surface texture, but in summer that texture heuristic fires on wrong labels - conversely, in winter, texture and label are aligned. Broadly, this spurious accuracy is driven by texture dominance, not semantic understanding. Figure 4 illustrates representative failure cases.

Practical Implications: Overall, our findings suggest that any zero-shot VLM deployment on Nordic satellite imagery should avoid using raw confidence as a sole quality filter, especially in summer. Logging the acquisition season and applying season-specific trust thresholds can be a low-cost improvement. in short, VTCS should be treated as a more stable but still imperfect proxy, with the expectation that its reliability degrades under geographic OOD shift beyond the Nordic domain.

Limitations: Our study elucidates some of the shortcomings of VLMs which may open up avenues for future research. *First,* we note all three models share ViT-B/32-scale encoders. Larger models (ViT-L, BLIP-2) may behave differently. *Second,*HR uses binary token matching, missing partial semantic overlap. *Third,* BLIP-ITM and RemoteCLIP were evaluated on smaller subsets (n=100–300) than CLIP (n=500–2000) due to inference cost, which limits head-to-head statistical comparison. *Lastly,* the seasonal split uses UTC acquisition dates as season labels. Actual phenological transitions vary by latitude within the bounding box.

6 Conclusion

We show that zero-shot VLMs deployed on Nordic satellite imagery exhibit hallucination rates of 49–89%, confidence scores that collapse to a near-uniform distribution under seasonal shift, and trust-signal erosion that is monotone with distribution shift severity and consistent across three architectures. The most alarming finding is the season-dependent reliability of confidence: raw softmax probabilities are below-chance error predictors in summer but useful in winter, while VTCS is more stable across seasons. Per-image recalibration cannot fix the underlying embedding misalignment. Hence, our study indicates that a potential mitigation strategy can be season-aware deployment with season-specific trust thresholds. Accordingly, future work should explore season-conditioned prompt adaptation and few-shot Nordic fine-tuning as more fundamental fixes.

References

[1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* 120 (2012), 25–36.

[2] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. ImageNet-Trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

[3] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 1321–1330.

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 7 (2019), 2217–2226.

[5] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Lodovici, Bianca Lambholt, Jonas Bernasconi, et al. 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. *arXiv preprint arXiv:2310.18660* (2023).

[6] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 12888–12900.

[7] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16.

[8] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

[9] Sachit Menon and Carl Vondrick. 2022. Visual Classification via Description from Large Language Models. *arXiv preprint arXiv:2210.07183* (2022).

[10] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the Calibration of Modern Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.

[12] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 4035–4045.

[13] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. 2021. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications* 12, 1 (2021), 4392.

[14] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling Climate Change with Machine Learning. *Comput. Surveys* 55, 2 (2022), 1–96.

[15] Julienne Stroeve and Dirk Notz. 2018. Changing state of Arctic sea ice across all seasons. *Environmental Research Letters* 13, 10 (2018), 103001.

[16] Yang Zhan, Zhitong Xiong, and Yuan Yuan. 2024. SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model. *arXiv preprint arXiv:2401.09712* (2024).

[17] Wei Zhang, Miaojing Shi, Chenguang Liu, Tao Xiang, and Timothy M. Hospedales. 2024. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–15.