EVALUATING SELF-SUPERVISED OBJECTIVES FOR SPECTROSCOPIC SATELLITE EMBEDDINGS

Manuel Pérez-Carrasco[*,1] **Core Francisco Park**[2,3] **Qindan Zhu**[1] **Rocco di Tella**[1]
Zolal Ayazpour[1] Gonzalo Gonzalez Abad[1] Cecilia Garraffo[1]
manuel.perez_carrasco@cfa.harvard.edu*,
corefranciscopark@g.harvard.edu, {qindan.zhu, rocco.di_tella,
zolal.ayazpour, ggonzalezabad, cgarraffo}@cfa.harvard.edu
[1]Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA 02138, USA
[2]Prior Computers, Cambridge, MA, 02139, USA
[3]Center for Brain Science, Harvard University, Cambridge, MA 02138, USA
* corresponding author

ABSTRACT

Deep learning enables efficient compression of hyperspectral satellite observations, but the choice of self-supervised objective significantly impacts what information is preserved. We compare three paradigms on NASA's Tropospheric Emissions: Monitoring of Pollution (TEMPO) data: variational autoencoders (VAE), autoregressive generation (GIVT), and masked autoencoders (MAE). Our experiments reveal a trade-off between reconstruction fidelity and atmospheric information preservation. VAE and GIVT achieve near-perfect reconstruction (MSE $\sim$ 0.0002) but encode atmospheric products less effectively. MAE produces substantially worse reconstructions (MSE $\sim$ 0.02–0.33) yet consistently outperforms when retrieving NO_2, O_3, HCHO, and cloud fraction, with improvements up to 69% for NO_2 retrieval over VAE at 64$\times$ compression (R^2 = 0.49 vs. 0.29). Aggregated across products, MAE improves over VAE by 17% (MLP probes) to 32% (linear probes). This trade-off diminishes at aggressive compression but grows as representational capacity increases. Our experiments provide empirical evidence that reconstruction quality is a poor proxy for representation utility in spectroscopic retrieval tasks, with direct implications for pretraining objective selection in climate foundation model design.

1 INTRODUCTION

Geostationary hyperspectral satellites generate unprecedented data volumes that challenge storage infrastructure and limit data distribution. NASA's TEMPO mission (Zoogman et al., 2017) exemplifies this challenge, measuring 1028 spectral channels across the 290-490nm and the 540-740nm wavelength ranges hourly over North America. Deep learning offers a compelling solution through learned embeddings that reduce data volumes while preserving task-relevant information (Brown et al., 2025; Feng et al., 2025; Klemmer et al., 2025). Recent work demonstrates that variational autoencoders can achieve 514$\times$ compression of TEMPO data while maintaining spectral reconstruction fidelity (Park et al., 2025). However, a critical question remains: does optimizing for reconstruction fidelity yield representations that best preserve atmospheric information?

This question has direct implications for Earth foundation models. While recent approaches span a range of self-supervised objectives (Cong et al., 2022; Xiong et al., 2024; Hong et al., 2024; Szwarcman et al., 2024; Wang et al., 2025b), it remains an open empirical question whether reconstruction-oriented pretraining preserves the spectral-spatial structure most relevant to atmospheric retrievals. Atmospheric science fundamentally concerns spatially coherent phenomena such as pollution plumes, cloud structures, and tropospheric gas distributions, and alternative self-supervised objectives may better align learned representations with these geophysical quantities. Understanding how objective choice shapes downstream retrieval performance is therefore important for designing foundation models that serve climate science applications (Bodnar et al., 2025; Zhu et al., 2026).

We systematically compare three self-supervised paradigms on TEMPO observations: variational autoencoders (Kingma & Welling, 2014, VAE) optimizing reconstruction, autoregressive generation in latent space (Tschannen et al., 2024, GIVT), and masked autoencoders (He et al., 2022, MAE) predicting spatial context. Our evaluation spans four atmospheric products representing diverse retrieval challenges: tropospheric NO_2, total column O_3, formaldehyde, and cloud fraction.

Our central finding reveals a trade-off between reconstruction quality and atmospheric information preservation. VAE and GIVT achieve near-perfect spectral reconstruction (MSE $\sim$ 0.0002) but

encode atmospheric products less effectively. MAE produces substantially worse reconstructions (MSE = 0.02–0.33 depending on compression rate) yet consistently outperforms alternatives for atmospheric retrieval, with NO_2 extraction improving by 69% over VAE at 64× compression. This trade-off diminishes at aggressive compression where all methods converge in performance, but becomes pronounced as representational capacity increases.

These findings indicate that reconstruction quality is not a reliable indicator of representation utility for spectroscopic retrieval tasks. While our comparison involves different architectures alongside different objectives, the consistent ordering across compression rates and products suggests that the pretext task plays a significant role. For climate foundation models targeting scientific applications, our results indicate that patch-level prediction may better align learned representations with geophysical quantities of interest, even at the cost of reconstruction fidelity.

2 METHODS

2.1 PROBLEM SETUP AND DATA

We evaluate self-supervised pretext tasks on TEMPO Zoogman et al. (2017) satellite observations. TEMPO provides hourly hyperspectral measurements across North America from geostationary orbit. We focus on the UV-visible 1028 spectral channels spanning 290–490nm with a spatial resolution of approximately 2.1 km × 4.75 km.

Similar to Park et al. (2025), our dataset comprises 50 TEMPO granules from January 2025. We extract 64×64 pixel tiles, resulting in hyperspectral cubes of shape $[1028 \times 64 \times 64]$. Raw radiance undergoes log transformation $\log(\max(I, 1.0))$, per-channel z-score normalization (from 2.4M pixels), and clipping to $[-10, 10]$.

We evaluate four atmospheric products from TEMPO Level-2 retrievals (Park et al., 2025): tropospheric NO_2 (Nowlan et al., 2025), total column O_3, HCHO vertical columns (Gonzalez Abad et al., 2025), and cloud fraction (Wang et al., 2025a). Each product is derived from the same UV/visible radiance spectra used as model input, via spectroscopic fitting of specific absorption features across the 1028-channel range. The retrieval task is a spatially-resolved, per-pixel prediction: given a latent representation $\mathbf{z} \in \mathbb{R}^{d \times 16 \times 16}$, probes predict a 16×16 map of atmospheric product values, where each spatial location corresponds to an approximately $4\times$ downsampled region of the original 64×64 input tile. To match this latent spatial resolution, we apply 4×4 average pooling to reduce Level-2 products from 64×64 to 16×16 pixels prior to training. Each product uses specialized normalization (asinh for NO_2/HCHO, z-score for O_3, logit for cloud fraction; see Park et al. (2025) for details). The reported R^2 scores are computed pixel-wise across the 16×16 prediction grid.

2.2 SELF-SUPERVISED PRETEXT TASKS

We compare three representative paradigms for representation learning.

Variational Autoencoder (VAE): Reconstruction. The VAE (Kingma & Welling, 2014) learns probabilistic compressed representations via the evidence lower bound. The encoder E_ϕ maps input $\mathbf{x} \in \mathbb{R}^{1028 \times 64 \times 64}$ to Gaussian parameters $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi, \text{diag}(\boldsymbol{\sigma}^2_\phi))$ where $\mathbf{z} \in \mathbb{R}^{d \times 16 \times 16}$, and d is the latent dimension. The decoder D_θ reconstructs from sampled $\mathbf{z}$ using the reparameterization trick. The objective combines reconstruction and KL regularization: $\mathcal{L}_{\text{VAE}} = \|\mathbf{x} - D_\theta(\mathbf{z})\|_1 / \exp(\log \sigma^2) + \log \sigma^2 + D_{\text{KL}}(q_\phi \| p(\mathbf{z}))$ where $p(\mathbf{z}) = \mathcal{N}(0, I)$, explicitly optimizing for pixel-level fidelity.

GIVT: Autoregressive Generation in Latent Space. GIVT (Tschannen et al., 2024) extends VAE by modeling latent dependencies autoregressively. After encoding to real-valued tokens $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ where $N = 16 \times 16$, a causal transformer predicts $p(\mathbf{z}_{1:N}) = \prod_{i=1}^{N} p(\mathbf{z}_i | \mathbf{z}_{<i})$ via raster-scan ordering. Unlike discrete token models, GIVT's output head predicts parameters of a k-component Gaussian mixture model (GMM) with diagonal covariance: for each token $\mathbf{z}_i \in \mathbb{R}^d$, it outputs $2kd + k$ parameters comprising means $\boldsymbol{\mu}^{(1:k)}$, standard deviations $\boldsymbol{\sigma}^{(1:k)}$, and mixture weights $\boldsymbol{\pi}^{(1:k)}$. The model is trained via negative log-likelihood enabling rich modeling of continuous latent distributions.

Masked Autoencoder (MAE): Spatial Context Prediction. MAE (He et al., 2022) learns by predicting masked regions from visible context. The method randomly masks input patches. An encoder processes only visible patches: $\mathbf{h}_{\text{vis}} = E_\phi(\mathbf{x}_{\text{vis}})$. A decoder reconstructs masked patches from encoded context and learned mask tokens: $\hat{\mathbf{x}}_{\text{mask}} = D_\theta([\mathbf{h}_{\text{vis}}, \mathbf{m}_{\text{mask}}])$. Loss applies only to masked regions: $\mathcal{L}_{\text{MAE}} = \|\mathbf{x}_{\text{mask}} - \hat{\mathbf{x}}_{\text{mask}}\|_2^2$. This forces spatial understanding rather than local pattern memorization. For atmospheric data, predicting concentration at location A may require un-

derstanding transport from location B. We use Vision Transformer architecture (Dosovitskiy, 2020) with patches of size 4×4 for all our experiments.

We note that contrastive learning approaches (e.g., Chen et al. (2020); He et al. (2020)) and joint-embedding methods (Assran et al., 2023, JEPA) were excluded due to the fundamental challenge of defining physics-preserving augmentations for atmospheric hyperspectral data. Standard transformations (rotations, crops, color jittering) can invalidate spectral absorption features or disrupt spatial pollution patterns. Developing domain-appropriate augmentation strategies remains an important direction for future work.

2.3 Evaluation Framework

We assess compression quality via pixel-level reconstruction error (MSE) and atmospheric information preservation via probing using the coefficient of determination (R^2). All models were trained for 200K steps (batch 32) using AdamW (lr=10^{-4}, weight decay=0.05) with gradient clipping (norm 1.0) on identical hardware for fair comparison. We evaluate all models over 5-fold cross-validation at the scene level. Each TEMPO granule represents a distinct temporal snapshot separated by 30-60 minutes, reducing but not eliminating temporal and spatial autocorrelation between folds. This autocorrelation could lead to optimistic performance estimates, and future work should evaluate generalization across longer temporal gaps and distinct meteorological regimes.

For atmospheric information preservation, after training each model we freeze encoders and train supervised probes mapping latent $\mathbf{z} \in \mathbb{R}^{d\times16\times16}$ to products $\mathbf{y} \in \mathbb{R}^{16\times16}$. **Linear probes** ($\hat{\mathbf{y}} = W\mathbf{z} + b$) assess linear accessibility. **MLP probes** (with hidden dimensions $[d \to 512 \to 256 \to 1]$) use ReLU, and dropout 0.1 reveal non-linear encoded information. For VAE, we probe the encoder's latent representation; for GIVT and MAE, we probe the final-layer output of the transformer.

3 Results

We evaluate the three self-supervised pretext tasks, on their ability to compress TEMPO hyperspectral data while preserving atmospheric information.

3.1 Reconstruction Quality

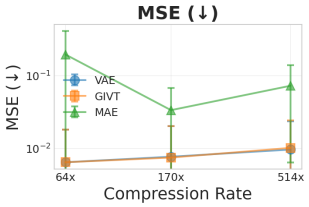
VAE and GIVT achieve excellent reconstruction across all compression rates, with MSE approximately two orders of magnitude lower than MAE (Figure 1a). At 170× compression, both methods reconstruct individual spectral channels with MSE $\sim$ 0.0002, preserving fine spatial details and spectral absorption features across all 1028 channels (Figure 1b).

MAE produces substantially degraded reconstructions, exhibiting visible patch artifacts and systematic spectral offsets. This is expected by design: MAE optimizes exclusively for masked patch prediction and lacks an explicit full-reconstruction objective. Moreover, each 4×4 patch must encode 1028 spectral channels from a single spatial context, making faithful pixel-level reconstruction particularly challenging without a dedicated loss. Unlike VAE and GIVT, which are trained end-to-end for complete input reconstruction, MAE learns representations through spatial context prediction. The reconstruction gap thus reflects a difference in training objective rather than a failure of representation quality.

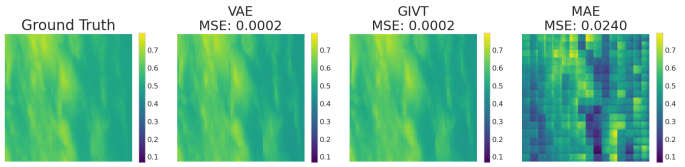
This effect is also visible in the spectral continuum (channels 200–1000), where MAE shows pronounced deviations from ground truth (Appendix Figure 2).

3.2 Atmospheric Product Extraction

Despite inferior reconstruction, MAE consistently outperforms alternatives for atmospheric product extraction. Table 1 presents aggregated R^2 scores across the four atmospheric products and compression rates. At 64× compression, MAE achieves mean R^2 = 0.75 with MLP probes versus 0.64 for VAE, representing a 17% improvement. The gap widens for linear probes: MAE achieves R^2 = 0.62 versus 0.47 for VAE (32% improvement). This superior linear probe performance suggests that MAE's latent dimensions are more directly aligned with atmospheric properties, requiring less nonlinear transformation to extract geophysical information.
The trade-off exhibits strong compression dependence. At aggressive compression (514×), all methods converge in downstream performance despite persistent reconstruction differences. MAE and GIVT achieve comparable extraction ($R^2 \sim 0.70$ for MLP probes), both outperforming VAE (R^2 = 0.64). As compression relaxes to 64×, MAE's advantage grows substantially: R^2 = 0.49 for NO_2 extraction compared to 0.29 for VAE (69% improvement) and R^2 = 0.95 for O_3 compared to 0.77 (23% improvement). GIVT occupies an intermediate position, consistently outperforming VAE but falling short of MAE across all compression rates (See Tables 2–4 for details).



(a) Reconstruction metrics vs. compression rate. MSE (lower is better). Error bars indicate standard deviation across 5 cross-validation samples.



(b) Spatial reconstruction comparison at 170× compression. Ground truth (left) compared with reconstructions from VAE, GIVT, and MAE. VAE and GIVT preserve fine spatial details with MSE ≈ 0.0002, while MAE exhibits visible patch artifacts and significantly higher error (MSE $= 0.0240$).

Figure 1: Reconstruction performance evaluation. (a) Quantitative comparison of reconstruction error across compression rates. (b) Qualitative visualization of a single spectral channel reconstruction, highlighting the spatial fidelity differences between methods.

	MLP Probes			**Linear Probes**		
Method	64×	170×	514×	64×	170×	514×
VAE	0.64 ± 0.24	0.66 ± 0.24	0.64 ± 0.24	0.47 ± 0.22	0.48 ± 0.21	0.48 ± 0.22
GIVT	0.72 ± 0.23	$\mathbf{0.72 \pm 0.21}$	0.70 ± 0.22	0.58 ± 0.23	0.54 ± 0.22	0.51 ± 0.22
MAE	$\mathbf{0.75 \pm 0.21}$	$\mathbf{0.72 \pm 0.22}$	$\mathbf{0.70 \pm 0.23}$	$\mathbf{0.62 \pm 0.23}$	$\mathbf{0.56 \pm 0.23}$	$\mathbf{0.52 \pm 0.22}$

Table 1: Aggregated atmospheric product extraction performance (R^2 scores averaged across NO_2, O_3, HCHO, and Cloud). Results shown as mean ± std the four products. Per-product results with cross-validation uncertainty are reported in Tables 2–4.

4 Discussion

Our results reveal a clear trade-off between reconstruction fidelity and atmospheric information preservation. MAE achieves the worst reconstruction quality yet produces representations most useful for downstream retrievals. This finding adds empirical grounding to an important design consideration for Earth foundation models: reconstruction fidelity alone is not a reliable indicator of representation utility for atmospheric retrieval tasks.

We attribute MAE's advantage primarily to its spatial context prediction objective, which we hypothesize encourages learning of spatially coherent structures that may relate more directly to retrieval targets than pixel-level spectral fidelity. VAE and GIVT optimize for spectral fidelity at each pixel independently, capturing local absorption features but not the spatial relationships governing atmospheric phenomena. This interpretation is further supported by scaling behavior: MAE performance improves consistently with model capacity, while GIVT plateaus despite comparable parameter increases (See Appendix Figure 4).

This tension echoes a broader open problem: current vision systems rely on separate experts for semantic understanding, fine-grained segmentation, and pixel-level generation (Radford et al., 2021; Kirillov et al., 2023; Tong et al., 2024), with no unified representation spanning all levels. Our results ground this challenge concretely in atmospheric remote sensing.

5 Conclusion

We systematically compared self-supervised pretext tasks for hyperspectral atmospheric data compression, revealing a clear trade-off between reconstruction quality and downstream task performance. MAE produces the poorest reconstructions yet achieves the best atmospheric product extraction, outperforming VAE by up to 69% for trace gas retrieval. This advantage grows with representational capacity, suggesting that spatial context prediction captures atmospheric structure that reconstruction objectives miss. For Earth foundation models targeting climate science applications, our results indicate that reconstruction fidelity is not a reliable indicator for atmospheric retrieval accuracy. Future work should explore hybrid objectives that balance spectral fidelity with geophysical information preservation, and investigate whether architectural differences contribute to the performance gaps.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*, 2025.

G. Gonzalez Abad, C. Nowlan, K. Chance, X. Liu, J. Carr, H. Chong, J. E. Davis, J. Fitzmaurice, D. E. Flittner, J. Geddes, B. Henderson, W. Hou, J. Houck, L. Judd, H.-A. Kwon, K. E. Knowland, C. Chan Miller, E. O'Sullivan, J. Park, B. Pierce, and the TEMPO team. Tropospheric emissions: Monitoring of pollution (tempo) nitrogen dioxide and formaldehyde retrievals. In *EGU General Assembly 2025*, pp. EGU25–14296, Vienna, Austria, Apr–May 2025. doi: 10.5194/egusphere-egu25-14296. URL https://doi.org/10.5194/egusphere-egu25-14296.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Konstantin Klemmer, Esther Rolf, Marc Russwurm, Gustau Camps-Valls, Mikolaj Czerkawski, Stefano Ermon, Alistair Francis, Nathan Jacobs, Hannah Rae Kerner, Lester Mackey, et al. Earth embeddings: Towards ai-centric representations of our planet. 2025.

C. R. Nowlan, G. González Abad, X. Liu, H. Wang, and K. Chance. Tempo nitrogen dioxide retrieval algorithm theoretical basis document. Technical report, NASA Algorithm Publication Tool, February 2025.

Core Francisco Park, Manuel Pérez-Carrasco, Caroline Nowlan, and Cecilia Garraffo. Hyperspectral variational autoencoders for joint data compression and component extraction. *ArXiv*, abs/2511.18521, 2025. URL https://api.semanticscholar.org/CorpusID:283244087.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arévolo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. *arXiv preprint arXiv:2412.02732*, 2024.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pp. 292–309. Springer, 2024.

H. Wang, C. R. Nowlan, G. Gonzalez Abad, H. Chong, W. Hou, J. C. Houck, W. Qin, A. P. Vasilkov, J. S. Joiner, R. Spurr, N. A. Krotkov, Ap Code 614, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA, T. S. Team, and J. A. Fitzmaurice. Algorithm theoretical basis for version 3 tempo o2-o2 cloud product. *Earth and Space Science*, 12:e2024EA004165, 2025a. doi: 10.1029/2024EA004165. URL https://agupubs.onlinelibrary.wiley.com.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision, 2025b. URL https://arxiv.org/abs/2503.11849.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024.

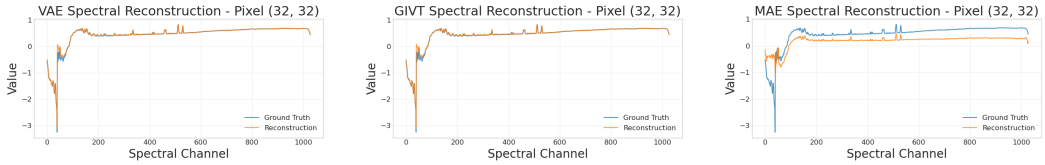
Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J. Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. *Nature Communications Earth & Environment*, 7(1):103, 2026. ISSN 2662-4435. doi: 10.1038/s43247-025-03127-x. URL https://doi.org/10.1038/s43247-025-03127-x.

P. Zoogman, X. Liu, R.M. Suleiman, W.F. Pennington, D.E. Flittner, J.A. Al-Saadi, B.B. Hilton, D.K. Nicks, M.J. Newchurch, J.L. Carr, S.J. Janz, M.R. Andraschko, A. Arola, B.D. Baker, B.P. Canova, C. Chan Miller, R.C. Cohen, J.E. Davis, M.E. Dussault, D.P. Edwards, J. Fishman, A. Ghulam, G. González Abad, M. Grutter, J.R. Herman, J. Houck, D.J. Jacob, J. Joiner, B.J. Kerridge, J. Kim, N.A. Krotkov, L. Lamsal, C. Li, A. Lindfors, R.V. Martin, C.T. McElroy, C. McLinden, V. Natraj, D.O. Neil, C.R. Nowlan, E.J. O'Sullivan, P.I. Palmer, R.B. Pierce, M.R. Pippin, A. Saiz-Lopez, R.J.D. Spurr, J.J. Szykman, O. Torres, J.P. Veefkind, B. Veihelmann, H. Wang, J. Wang, and K. Chance. Tropospheric emissions: Monitoring of pollution (tempo). *Journal of Quantitative Spectroscopy and Radiative Transfer*, 186:17–39, 2017. ISSN 0022-4073. doi: https://doi.org/10.1016/j.jqsrt.2016.05.008. URL https://www.sciencedirect.com/science/article/pii/S0022407316300863. Satellite Remote Sensing and Spectroscopy: Joint ACE-Odin Meeting, October 2015.

A APPENDIX

A.1 Spectral Reconstruction Quality

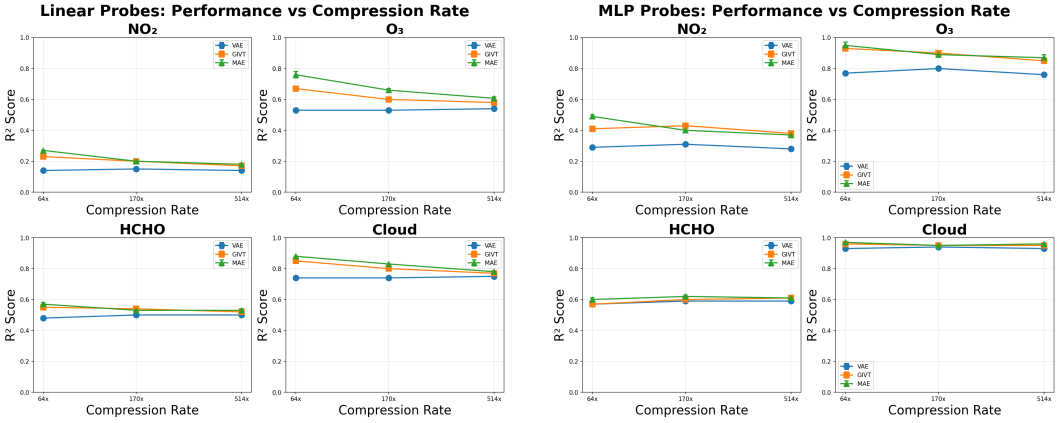
Figure 2 shows the spectral reconstruction fidelity for a representative pixel at location (32, 32). The full spectrum spans 1028 channels covering the UV-visible range measured by TEMPO. VAE and GIVT produce near-identical reconstructions that closely follow the ground truth across all spectral channels, including the complex absorption features below channel 200. In contrast, MAE reconstructions exhibit a systematic offset in the continuum level (channels 200–1000) and fail to capture fine spectral structure, consistent with the higher MSE observed in spatial reconstructions. Despite this spectral degradation, MAE achieves superior downstream retrieval performance (Tables 2–4), suggesting that spatial context encoded in the representations compensates for reduced spectral fidelity at individual pixels.



(a) VAE (b) GIVT (c) MAE

Figure 2: Spectral reconstruction comparison at pixel (32, 32) for 170× compression. Ground truth (blue) and reconstruction (orange) are shown across all 1028 spectral channels. VAE and GIVT achieve high-fidelity reconstruction with the curves nearly overlapping, while MAE shows significant deviation, particularly in the continuum region (channels >200).

A.2 Detailed Downstream Task Results The main trends across compression rates are more clearly visualized in Fig. 3, which summarizes linear and MLP probe performance for all atmospheric products.



(a) Linear probes (b) MLP probes

Figure 3: **Probe performance vs. compression rate.** R^2 scores for NO_2, O_3, HCHO, and cloud fraction across compression rates for VAE, GIVT, and MAE representations using linear (left) and MLP (right) probes.

Tables 2–4 present the complete per-product R^2 scores for atmospheric retrieval at each compression rate, complementing the aggregated results in the main text. Performance is reported for four products: tropospheric NO_2, total column O_3, formaldehyde (HCHO), and cloud fraction.

Across all compression rates, cloud fraction is the easiest to retrieve ($R^2 > 0.74$), while NO_2 proves most challenging due to its lower signal-to-noise ratio and finer spatial variability. MAE achieves the best linear probe performance in most settings, indicating that its representations capture atmospheric structure in a more linearly separable manner. With MLP probes, GIVT and MAE perform comparably, suggesting that the additional nonlinearity compensates for differences in representation geometry. The performance gap between methods narrows at higher compression, where all approaches face increased information loss.

	Linear Probes				MLP Probes			
Method	NO_2	O_3	HCHO	Cloud	NO_2	O_3	HCHO	Cloud
VAE	0.14 ± 0.00	0.54 ± 0.00	0.50 ± 0.01	0.75 ± 0.00	0.28 ± 0.01	0.76 ± 0.01	0.59 ± 0.01	0.93 ± 0.00
GIVT	0.17 ± 0.00	0.58 ± 0.00	0.52 ± 0.01	0.77 ± 0.00	**0.38 ± 0.00**	0.85 ± 0.01	**0.61 ± 0.01**	0.94 ± 0.00
MAE	**0.18 ± 0.00**	**0.61 ± 0.01**	**0.53 ± 0.01**	**0.78 ± 0.00**	0.37 ± 0.00	**0.87 ± 0.02**	**0.61 ± 0.01**	**0.96 ± 0.00**

Table 2: **Atmospheric product extraction** (R^2 scores at 514x compression rate). Linear and MLP probes trained on frozen representations.

	Linear Probes				MLP Probes			
Method	NO_2	O_3	HCHO	Cloud	NO_2	O_3	HCHO	Cloud
VAE	0.15 ± 0.00	0.53 ± 0.01	0.50 ± 0.01	0.74 ± 0.00	0.31 ± 0.00	0.80 ± 0.00	0.59 ± 0.01	0.94 ± 0.00
GIVT	**0.20 ± 0.00**	0.60 ± 0.01	**0.54 ± 0.01**	0.80 ± 0.01	**0.43 ± 0.01**	**0.90 ± 0.01**	0.60 ± 0.01	0.95 ± 0.00
MAE	**0.20 ± 0.00**	**0.66 ± 0.01**	0.53 ± 0.00	**0.83 ± 0.00**	0.40 ± 0.01	0.89 ± 0.01	**0.62 ± 0.01**	**0.95 ± 0.00**

Table 3: **Atmospheric product extraction** (R^2 scores at 170x compression rate). Linear and MLP probes trained on frozen representations.

A.3 Model Architectures

The Variational Autoencoder (VAE) encoder processes input $x \in \mathbb{R}^{1028\times64\times64}$ through a series of convolutional blocks with progressive downsampling. Table 5 presents the complete architecture specifications for each compression rate.

The encoder uses two downsampling layers (stride-2 convolutions) for $64\times$ compression and three for higher compression rates. Each channel configuration represents a residual block with the specified number of filters.

The GIVT (Generative Infinite-Vocabulary Transformers) extends the VAE by adding an autoregressive transformer over the latent space. After VAE encoding, the latent tokens are processed in raster-scan order. Table 6 provides transformer specifications.

All configurations use dropout of 0.1 and model the conditional distribution using 16 Gaussian mixtures per token.

The MAE model employs a Vision Transformer architecture with an asymmetric encoder–decoder design. Table 7 details the specifications.

All MAE models use 4×4 patches, resulting in 256 total patches (16×16 grid) from the 64×64 input. The masking ratio is 75% for all experiments. Both encoder and decoder use standard transformer blocks with LayerNorm and GELU activations.

A.4 Parameter Analysis and Scaling Behavior

Figure 4 shows downstream retrieval performance as a function of encoder model parameters. MAE performance improves consistently with model capacity across all products, while GIVT plateaus despite comparable parameter increases.

The VAE uses a fixed encoder-decoder architecture across all compression rates, varying only the latent dimensionality. This deliberate design choice preserves high expressivity regardless of compression level, yet downstream performance remains lower, suggesting that additional capacity in reconstruction-oriented architectures does not translate to better atmospheric information encoding.

The total parameter counts between MAE and GIVT differ due to input and output interfaces, not core modeling capacity. At $64\times$ compression, both architectures share identical transformer blocks. MAE's larger total arises from its patch embedding layer, which projects raw spectral patches into lower-dimensional embeddings. GIVT's total includes a GMM output head for mixture model prediction but receives already-compressed VAE latents, requiring a smaller input projection.

	Linear Probes				MLP Probes			
Method	NO_2	O_3	HCHO	Cloud	NO_2	O_3	HCHO	Cloud
VAE	0.14 ± 0.00	0.53 ± 0.01	0.48 ± 0.01	0.74 ± 0.01	0.29 ± 0.01	0.77 ± 0.01	0.57 ± 0.01	0.93 ± 0.00
GIVT	0.21 ± 0.01	0.64 ± 0.01	0.55 ± 0.01	0.84 ± 0.01	0.37 ± 0.00	0.88 ± 0.01	0.57 ± 0.01	0.95 ± 0.00
MAE	**0.27 ± 0.00**	**0.76 ± 0.02**	**0.57 ± 0.01**	**0.88 ± 0.00**	**0.49 ± 0.01**	**0.95 ± 0.02**	**0.60 ± 0.01**	**0.97 ± 0.00**

Table 4: **Atmospheric product extraction** (R^2 scores at 64x compression rate). Linear and MLP probes trained on frozen representations.

Table 5: VAE architecture specifications across compression rates.

Compression	Embed Dim	Channels	Z-Channels	Latent Shape
64×	256	[512, 256, 128]	256	$256 \times 16 \times 16$
170×	96	[512, 256, 128]	96	$96 \times 8 \times 8$
514×	32	[512, 256, 128]	32	$32 \times 8 \times 8$

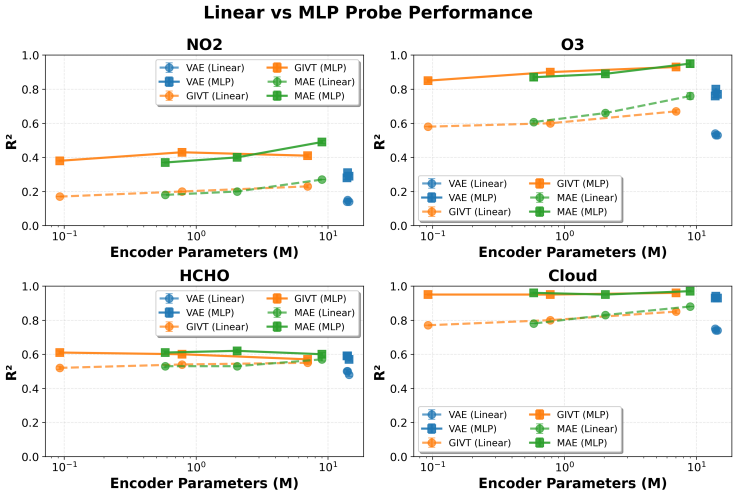


Figure 4: Downstream retrieval performance vs. total encoder parameters. Dashed lines indicate linear probes; solid lines indicate MLP probes. MAE (orange) scales more favorably with model capacity than VAE (blue) and GIVT (purple) across all atmospheric products.

Table 6: GIVT transformer specifications across compression rates.

Compression	d_{model}	Heads	Layers	MLP Ratio	Mixtures	Block Size
64×	256	8	6	4	16	256
170×	96	8	4	4	16	256
514×	32	8	4	4	16	256

Table 7: MAE architecture specifications across compression rates.

Compression	Patch	Encoder Dim	Enc. Depth	Enc. Heads	Decoder Dim	Dec. Depth	Dec. Heads
64×	4×4	256	6	8	256	6	8
170×	4×4	96	4	8	96	4	8
514×	4×4	32	4	8	32	4	8