FASTER, LIGHTER, BETTER: COMPRESSED EARTH OBSERVATION FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth observation foundation models (EOFMs) support diverse geospatial tasks, but adapting and deploying them could be computationally expensive. We introduce NewOrder, a framework for structured pruning of EOFMs before full downstream fine-tuning. Structural importance can change as the pretrained encoder and a newly initialized decoder learn together, while heterogeneous EO data make importance estimates sensitive to the calibration inputs. NewOrder addresses these challenges by combining output-sensitive scoring, progressive pruning, and representative calibration. Across four encoders and four tasks, NewOrder with 50% compression rate achieves the best performance among the compared pruning methods in 14 of 16 settings and exceeds dense adaptation in 8 of them. Additionally, full efficiency profiling shows up to 1.73× inference speedup, 1.72× training-step speedup, and 33.4% lower peak training memory. These results demonstrate that selecting which pretrained structures to retain during adaptation can improve both downstream accuracy and computational efficiency, making the selection of retained pretrained capacity an effective part of task adaptation.

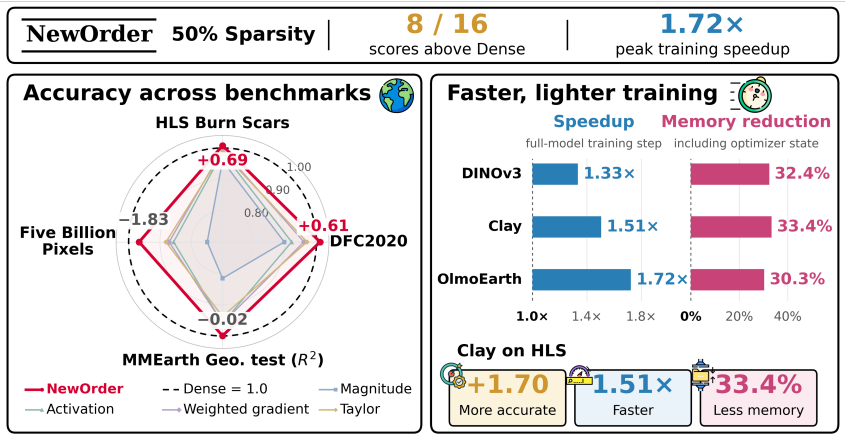


Figure 1: Accuracy and efficiency at the 50% sparsity. Left: task averages across four encoders, normalized by the dense average; annotations show absolute percentage-point differences. Right: full model training-step speedup and peak training-memory reduction on HLS Burn Scars.

1 INTRODUCTION

Understanding our rapidly changing planet increasingly depends on our ability to observe it at scale. Earth observation (EO) provides reusable, spatially extensive measurements of wildfires, land-use change, ecosystem dynamics, and terrestrial carbon, supporting applications in climate science, environmental monitoring, and disaster response (Clay Foundation, 2025; Jakubik et al., 2023). The breadth and heterogeneity of these observations have made foundation models (FMs) a natural paradigm for EO: by pretraining on large geospatial data, EOFMs learn representations that transfer across sensors, regions, and downstream tasks (Jakubik et al., 2025; Herzog et al., 2025; Siméoni et al., 2025). Yet the large encoders that support this breadth also make downstream adaptation and deployment computationally demanding (Marti Escofet et al., 2026; Snyder et al., 2026). A pretrained

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

EOFM is expected to accommodate diverse applications, whereas an individual downstream task may require only part of its capacity. Compressing the model only after full fine-tuning reduces deployment cost but still incurs the cost of adapting the entire encoder (Barreiro et al., 2026). We argue that adaptation should determine not only *how* the pretrained model is updated, but also *which pretrained structures should be retained* for the downstream task.

Structured pruning provides a natural solution because removing complete components produces physically smaller dense operators that remain compatible with standard accelerators (Molchanov et al., 2017; Xia et al., 2022; Cheng et al., 2024). Existing EO pruning methods, however, operate at different granularities or stages: prior works select modality-specific experts (Bi et al., 2026), remove entire Transformer layers before fine-tuning (Barreiro et al., 2026), or compress models after task-specific adaptation (Said et al., 2026). To the best of our knowledge, we are the *first* to study fine-grained head- and neuron-level structured pruning of EOFMs *before* full downstream adaptation.

Pruning early allows most downstream optimization to run directly on the compact architecture, reducing both adaptation and deployment cost. However, early pruning is substantially harder than pruning a task model whose representations have already stabilized. The pretrained encoder needs to adapt jointly with a newly initialized decoder, so both the encoder's representations and the decoder's use of those representations change during training. Structural importance is therefore *evolving*, meaning that structures that appear unimportant at initialization can become important later, thus making a fixed initial ranking a suboptimal solution. EO further complicates this estimation because downstream data span heterogeneous geographic and environmental conditions (Tong et al., 2023). Effective early pruning should therefore account for both the evolution of structural importance during adaptation and its variation across downstream conditions.

We introduce NewOrder, a framework for fine-grained structured pruning during early adaptation of Earth observation foundation models. NewOrder measures the output sensitivity of dependency-coupled attention heads and MLP neurons on a representative calibration set, progressively removes low-scoring structures, and re-estimates their importance as the encoder and decoder adapt jointly. We evaluate NewOrder across four EOFMs: DINOv3, Clay, OlmoEarth, and TerraMind, and four EO tasks spanning burn-scar segmentation, land-cover mapping, and dense biomass regression. As shown in Figure 1, at a 50% compression rate, NewOrder reaches up to $1.72\times$ faster training, 33.4% lower peak training memory, and outperforms their corresponding dense model in 8 of 16 settings. On Clay–HLS, for example, pruning half of the encoder raises mIoU from 90.10 to 91.80 while providing a $1.51\times$ training-step speedup and reducing peak training memory by 33.4%. Together, these results challenge the assumption that retaining the full pretrained model is always preferable for downstream adaptation: for EOFMs, deciding which pretrained structures to retain can itself be part of learning the task. Our main contributions are threefold:

- **Pruning before full EO adaptation.** To the best of our knowledge, NewOrder is the *first* framework for fine-grained structured pruning of EOFMs before full downstream adaptation, reframing pruning from post-hoc compression into the selective retention of pretrained structures during fine-tuning.
- **Adaptive pruning for heterogeneous EO data.** NewOrder combines output-sensitive scoring, progressive re-estimation, and representative calibration so that structural decisions evolve with the downstream fine-tuning.
- **Smaller, faster, and often more accurate models.** Across four encoders and four tasks, NewOrder achieves the best pruned result in 14 of 16 settings and exceeds dense adaptation in 8 of them while removing approximately half of the encoder parameters and delivering measured training, inference, and memory savings.

2 Proposed Method

NewOrder progressively prunes a pretrained EO encoder during its joint adaptation with a newly initialized task decoder. We estimate the importance of grouped structures through their influence on the current model's outputs, using representative calibration images to cover diverse geographic conditions. As training changes how the decoder uses the encoder's features, NewOrder updates these estimates and removes low-scoring structures. Figure 2 and Algorithm 1 illustrate this process.

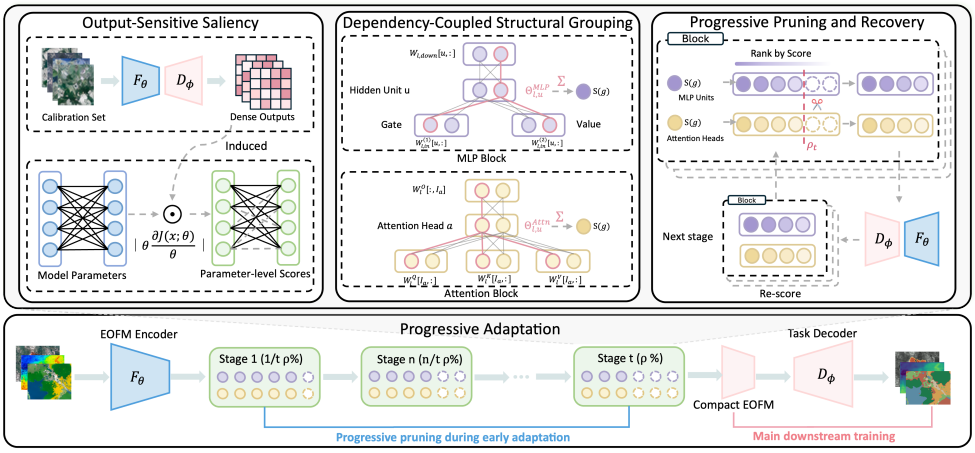


Figure 2: Overview of NewOrder. Output sensitivities are aggregated over the parameters belonging to each attention head or MLP neuron. Successive stages physically remove low-scoring structures and update their importance after encoder–decoder adaptation. Calibration supplies inputs for scoring; recovery uses the supervised downstream training objective.

2.1 Problem Setup

We attach a randomly initialized task decoder to a pretrained EO encoder and train both on the downstream task. We write $f(x;\theta)$ for the full encoder–decoder model, with θ collecting its parameters. Only the encoder is pruned: we entirely remove the selected attention heads and MLP neurons while preserving the encoder's output interface to the decoder.

The target pruning ratio specifies how many heads and neurons to remove from each Transformer block. We apply the same ratio to both structure types in every block, subject to integer widths and retaining at least one of each. Methods compared at a given target therefore have the same final architecture dimensions. NewOrder reaches these dimensions during early adaptation, without first training a dense task-specific model.

2.2 Output-Sensitive Structural Selection

For a calibration image $\mathbf{x}$, let $\Omega(\mathbf{x})$ denote the set of valid output locations. We use the model's outputs[1] as

$$J(x;\theta) = \frac{1}{|\Omega(x)|} \sum_{p\in\Omega(x)} f_p(x;\theta). \tag{1}$$

To interpret parameter importance, consider a multiplicative gate m_j on each encoder parameter θ_j. Let $\widetilde{\theta}(\mathbf{m})$ denote the parameters after replacing each encoder parameter by $m_j\theta_j$, with decoder parameters unchanged. At $\mathbf{m} = \mathbf{1}$, the model is unchanged, and the magnitude of its output sensitivity to gate m_j is

$$s_j(x) = \left| \frac{\partial J(x;\widetilde{\theta}(\mathbf{m}))}{\partial m_j}\bigg|_{\mathbf{m}=\mathbf{1}} \right| = \left| \theta_j \frac{\partial J(x;\theta)}{\partial \theta_j} \right|. \tag{2}$$

Thus, $s_j(x)$ measures local sensitivity to multiplicative changes in θ_j. The gates provide an interpretation of the score; computation uses the parameter values and their gradients directly.

Our intuition for applying this saliency score is that, during early adaptation stages, a supervised loss gradient reflects both how a structure influences the model output and how the current prediction differs from its target. The latter acts as an instantaneous correction signal. Consequently, a small loss-based saliency can indicate either that a structure has little effect on the model response, or simply that its effect is weakly weighted by the current prediction error.

[1]For segmentation, the outputs are the maximum of class logits. For regression, there is only one output channel, so it's the predicted scalar value.

Algorithm 1 NewOrder: Structured Pruning during Early Adaptation

Require: Pretrained encoder, training data $\mathcal{D}_{\text{train}}$, target pruning ratio ρ, total training steps K, pruning steps $k_1, \ldots, k_T$

Initialize the task decoder and form $f(\cdot;\theta)$
$\mathcal{C} \leftarrow \textsc{CoverageSelect}(\mathcal{D}_{\text{train}}, 32)$
for $t = 1, \ldots, T$ **do**
    Jointly train the current encoder and decoder until step k_t
    $\rho_t \leftarrow (t/T)\rho$
    Compute scores for all surviving structures:
    $$S(g) \leftarrow \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \sum_{j \in g} \left| \theta_j \frac{\partial J(x;\theta)}{\partial \theta_j} \right|$$
    for each Transformer block **do**
        Rank surviving heads and neurons separately by $S(g)$
        Select the lowest-scoring structures to reach cumulative pruning ratio ρ_t
        Physically remove their coupled parameter slices
    end for
end for
Continue joint training of the compact encoder and decoder until step K
return Adapted compact model

This distinction is particularly relevant in EO, where downstream data can be highly heterogeneous across geographic regions and land-cover conditions. A structural ranking dominated by the current label-conditioned errors could therefore become overly specific to the examples and errors represented in the calibration set. Instead, NewOrder measures the sensitivity of the model's currently expressed output response without using ground-truth labels. In this way, our score remains task-dependent through the downstream inputs and decoder, while avoiding the additional weighting induced by the supervised loss. It therefore identifies which encoder structures the current task response depends on, rather than which structures are most useful for reducing the current calibration loss.

For a structure g, we aggregate these sensitivities over its parameters and average over a calibration set $\mathcal{C}$ drawn from the downstream training data:

$$S(g) = \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \sum_{j \in g} s_j(x). \tag{3}$$

Here, g is the set of parameter indices belonging to one head or neuron, and $\mathcal{C}$ is a small calibration set drawn from the training data. Thus, $S(g)$ sums parameter importance within a structure and averages it across calibration images.

The groups follow the encoder's computational dependencies. An MLP unit contains the corresponding input-projection rows, biases, and output-projection column; a gated MLP includes both input branches. An attention head contains its query, key, and value projections and the matching output-projection columns. Removing all coupled slices produces smaller dense operators while leaving the dimensionality of retained heads and the block's output width unchanged. Appendix B gives the exact groups and physical slicing rules. The group score bounds the first-order change in the response when a structure is removed.

2.3 Progressive pruning and adaptation

An importance ranking computed before training reflects a pretrained encoder coupled to a decoder that has not yet learned the task. Pruning a large fraction of the encoder from this initial ranking risks removing useful capacity too early. To directly address this potential mismatch, NewOrder applies a progressive-style pruning (Zhu & Gupta, 2018) and removes structures gradually, allowing task training to inform subsequent decisions about which structures are redundant.

We divide the target pruning ratio ρ into T stages. At stage t, the cumulative fraction of heads and neurons removed from each block is

$$\rho_t = \frac{t}{T}\rho, \qquad t = 1, \ldots, T. \tag{4}$$

At each stage, we recompute Equation 3 on the current model and remove the lowest-scoring heads and neurons separately within each block until the target is reached. Between stages, both encoder

Table 1: Released EOFM encoders used in the controlled study. Parameter counts include the encoder and input adapter but exclude the downstream decoder.

Encoder	Input	Blocks	MLP width	Heads	Params
DINOv3 ViT-L	RGB	24	4096	16×64	303.1M
Clay v1.5	Multi.	24	4096	16×64	311.4M
OlmoEarth v1-Large	Multi.	24	4096	16×64	307.8M
TerraMind Base	Multi.	12	2048	12×64	86.1M

and decoder continue training on the supervised downstream objective. This training allows the decoder to learn to use the retained features while enabling the encoder to adapt to both the task and the preceding removals. The next pruning decision is therefore based on importance after these updates. This schedule allows adaptation to guide compression while retaining the computational benefit of training a smaller model.

2.4 Representative calibration data

Calibration data determines which inputs contribute to the importance estimates. EO datasets are highly heterogeneous, spanning different geographic regions, environmental conditions, and land-cover compositions. Even the same land-cover class can vary in appearance across regions. A randomly sampled calibration set might therefore capture only a narrow part of this diversity. Since structural importance is estimated from the model's responses to these images, structures needed for underrepresented scenes may receive low scores because the relevant conditions are missing from the calibration set.

We therefore select 32 training examples subject to coverage of the terrain categories represented in each task. For example, in Five-Billion-Pixels, which contains 24 land-cover classes, we select satellite images that collectively include all 24 classes, ensuring that less frequent categories also contribute to the importance estimates. This selection accounts for EO-specific variation in terrain, geographic environments, and land-cover composition, allowing importance estimation to draw on a representative range of downstream conditions while keeping the scoring budget fixed. Appendix A.4 gives the coverage constraint and provides more details on this process.

3 Experiments

In this section, we evaluate NewOrder through comparisons with dense adaptation and alternative pruning methods, followed by efficiency results, ablation study and parameter sensitivity analysis.

3.1 Experimental Setup

Pretrained Earth Foundation Models. We evaluate four released EOFMs with substantially different pretraining and input interfaces: DINOv3 (Siméoni et al., 2025), Clay (Clay Foundation, 2025), OlmoEarth (Herzog et al., 2025) and TerraMind (Jakubik et al., 2025). DINOv3 receives RGB input, while the remaining models use their released multispectral adapters and the compatible task bands. Table 1 summarizes the model details.

Downstream Tasks. Our downstream evaluation spans three segmentation tasks and one dense regression task. HLS Burn Scars uses six-band Harmonized Landsat–Sentinel imagery for burn-scar mapping (Phillips et al., 2023). DFC2020 evaluates eight-class land-cover segmentation from Sentinel-2 (Robinson et al., 2021). Five-Billion-Pixels contains Gaofen-2 imagery for 24-class high-resolution land-cover mapping (Tong et al., 2023). MMEarth biomass is a dense above-ground biomass regression task; we report its geographically disjoint test split (Gordon et al., 2026). We use mean intersection over union (mIoU) for segmentation and pixel-level R^2 for biomass regression.

Baselines and controlled protocol. We compare NewOrder with magnitude pruning (Han et al., 2015), activation-aware pruning (Sun et al., 2024), weighted gradient (Lee et al., 2019), and group Taylor (Molchanov et al., 2017). All methods target 50% of both MLP units and attention heads within every Transformer block and retain at least one unit of each type. We use 32 examples as calibration data. Following the standard approach (Pauls et al., 2025), every encoder is paired with the same newly initialized UPerNet decoder (Xiao et al., 2018). We optimize the complete encoder

Table 2: Main results across benchmarks and foundation models at a 50% joint pruning target. **Bold** and underlined values are the best and second-best pruned results. *Red italics* indicate a pruned result above its corresponding dense reference.

Method	DINOv3	Clay	OlmoEarth	TerraMind	Avg.
HLS Burn Scars (mIoU, %)					
Dense (unpruned)	86.76	90.10	86.95	90.42	88.56
Magnitude	75.62	90.03	83.59	90.37	84.90
Activation	83.19	*90.29*	85.80	89.82	87.28
Weighted gradient	84.62	*91.42*	**86.93**	89.61	88.14
Taylor	81.40	*91.47*	86.70	89.80	87.34
NewOrder	**86.75**	***91.80***	86.62	***91.82***	***89.25***
DFC2020 (mIoU, %)					
Dense (unpruned)	49.66	51.70	54.69	53.05	52.28
Magnitude	39.40	*51.73*	54.36	42.15	46.91
Activation	41.84	51.36	54.46	44.89	48.14
Weighted gradient	47.89	*52.27*	53.52	47.09	50.20
Taylor	49.31	48.91	51.09	*53.57*	50.72
NewOrder	***49.80***	***52.30***	***54.77***	***54.70***	***52.89***
MMEarth geographic-test (R^2, %)					
Dense (unpruned)	44.17	45.69	52.78	53.94	49.14
Magnitude	24.65	40.07	49.00	46.71	40.11
Activation	42.92	**45.62**	49.63	50.49	47.17
Weighted gradient	43.80	43.54	50.19	50.98	47.13
Taylor	41.58	41.78	50.55	50.49	46.10
NewOrder	***46.58***	44.70	**52.31**	**52.89**	**49.12**
Five Billion Pixels (mIoU, %)					
Dense (unpruned)	64.53	46.89	54.50	52.16	54.52
Magnitude	41.55	42.33	42.57	36.90	40.84
Activation	58.99	40.54	43.13	44.08	46.69
Weighted gradient	59.35	43.26	44.79	42.47	47.47
Taylor	56.28	45.85	44.70	45.26	48.02
NewOrder	**61.35**	*47.72*	**53.42**	**48.27**	**52.69**

and decoder for 20 epochs using AdamW, cosine learning-rate decay, 10% warmup, BF16 autocast on four NVIDIA H100 80GB GPUs. The encoder learning rate is one tenth of the decoder learning rate. Checkpoints are selected only by the validation metric and the held-out test split is evaluated once. Full split, preprocessing, optimization, and calibration details are provided in Appendix A.

3.2 Results and Analysis

Table 2 reports downstream performance at the 50% sparsity pruning target. NewOrder achieves the highest performance in fourteen of the sixteen model–task settings and the highest average across all four tasks. All pruning methods retain the same numbers of attention heads and MLP neurons within each Transformer block, so these comparisons evaluate the quality of compact models at matched architecture sizes. Compared with dense adaptation, NewOrder improves performance in eight of the sixteen settings while removing approximately half of the encoder parameters, specifically on HLS Burn Scars and DFC2020. The improvements show that retaining the entire pretrained encoder is not always necessary for the best downstream performance: selecting which structures to retain during adaptation can yield a smaller model that also performs better on the target task. This improvement is commendable since, by pruning half of the parameters, the model is able to achieve at most 1.7x speedup, as detailed in Section 3.3.

The gains over dense adaptation vary across tasks and are less consistent on the more challenging MMEarth and Five-Billion-Pixels benchmarks. Even in these settings, NewOrder achieves the highest average performance among the compared pruning methods, preserving downstream performance more effectively at the same structural budget, empirically demonstrating its effectiveness.

Table 3: Measured full-model efficiency on HLS Burn Scars using one H100 80GB GPU with BF16 and batch size 4. Inference latency measures the full forward pass; training latency includes the full training process. Peak memory is allocated GPU memory.

			Inference			Training		
Model	Variant	Params (M)	Lat. (ms)	Mem. (GiB)	Speedup	Step (ms)	Mem. (GiB)	Speedup
DINOv3	Dense	315.65	25.67	3.07	–	84.35	9.89	–
	NewOrder	164.58	19.82	2.23	**1.30**×	63.19	6.69	**1.33**×
Clay	Dense	323.97	60.01	2.93	–	216.12	19.49	–
	NewOrder	172.92	41.09	2.09	**1.46**×	143.59	12.99	**1.51**×
OlmoEarth	Dense	320.30	281.86	3.59	–	1020.90	51.24	–
	NewOrder	169.22	163.01	2.38	**1.73**×	592.11	35.72	**1.72**×
TerraMind	Dense	97.02	10.62	1.64	–	29.96	4.39	–
	NewOrder	54.55	9.56	1.40	**1.11**×	24.97	3.33	**1.20**×

Why NewOrder can improve over dense

We further examine the Clay–HLS checkpoints from Table 2 by partitioning the held-out test set according to whether a scene's source tile is represented in the training set. Our analysis shows that NewOrder improves over dense adaptation by **1.07 mIoU points** on scenes from training-observed tiles and by **3.00 points** on scenes from geographically unseen tiles. The larger gain on unseen tiles persists when the scenes are further stratified by burned-area fraction. This analysis suggests that NewOrder's improvement is not limited to geographic conditions represented during training, and is particularly pronounced under geographic shift, which provides evidence that NewOrder helps to improve the generalizability of the pretrained EO model.

3.3 Efficiency Analysis

We next measure how the parameter reductions translate into actual hardware acceleration. Table 3 reports the full profiles on HLS using one H100 80GB GPU, where encoders are pruned with NewOrder and the decoders are unchanged. Pruning reduces both latency and peak memory for every EO model. Training steps are at most $1.72\times$ faster, and memory requirement decreases by at most 33.4%. Still, the acceleration varies across encoder architectures despite the same structural pruning target. Larger encoders obtain greater speedups, while smaller encoders such as TerraMind achieve more modest gains of $1.11\times$ for inference and $1.20\times$ for training steps. Full-model speedup depends on the fraction of runtime affected by pruning: NewOrder reduces the internal attention and MLP projections, while the decoder and other unchanged operations continue to contribute to the measured cost. The same proportional reduction in encoder parameters therefore yields different runtime savings across models. NewOrder therefore offers the greatest practical acceleration when encoder computation constitutes a substantial portion of the end-to-end cost.

The efficiency results can be considered alongside the accuracy comparisons. On Clay–HLS, reducing the encoder to half of the parameters results in training-step and inference speedups of $1.51\times$ and $1.46\times$, and reduces the memory requirement from 20GB to 13GB. Impressively, without sacrificing performance, NewOrder is able to even improve the overall mIoU from 90.10 to 91.80. This setting therefore provides an empirical example in which the improvement over dense adaptation is obtained with lower latency and memory requirements.

3.4 Ablation & Parameter Study

Here we provide detailed ablation and parameter study on NewOrder. Each ablation changes the component or parameter under study while keeping the remaining experimental conditions fixed within the comparison. Further details are provided in Appendix A.6.

Effect of the saliency objective. To isolate the role of output-sensitive scoring, we replace NewOrder's output response with the supervised task loss while keeping the pruning schedule, calibration data, structural groups, and training protocol unchanged. As shown in Figure 3(a), output-sensitive NewOrder improves over the loss-gradient variant for three of the four encoders, while TerraMind differs by only 0.15 points in the opposite direction. This controlled comparison supports

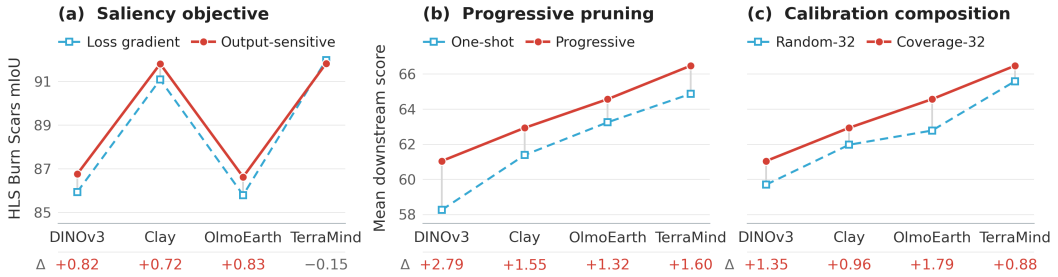


Figure 3: Ablations of the three components of NewOrder: (a) output-sensitive saliency, (b) progressive pruning, and (c) coverage-based calibration.

Table 4: Pruning ablations on Clay with HLS Burn Scars. (a) Performance across different sparsity levels. (b) Validation and test mIoU for different pruning-phase durations. All scores are mIoU (%). W. grad. denotes weighted gradient.

(a) Sparsity levels.

Target	Magnitude	Activation	W. grad.	Taylor	NewOrder
25%	91.01	91.43	**91.50**	90.74	91.39
50%	90.03	90.29	91.42	91.47	**91.80**
75%	89.33	89.50	89.41	88.97	**90.49**

(b) Pruning-phase duration

End epoch	Val	Test	Δ
2	92.89	91.24	−0.56
5	92.54	**91.80**	0.00
10	**93.15**	91.65	−0.15

our use of the model's current output response for structural selection rather than weighting saliency by the instantaneous supervised correction signal.

Effect of progressive pruning. To assess the role of the pruning-and-recovery schedule, we compare one-shot and progressive NewOrder using the same output-sensitive scoring rule. One-shot selects and removes all target structures before optimization, while progressive ends at epoch 5. Figure 3(b) shows that progressive pruning improves performance across all settings. With the scoring rule fixed, these results support distributing structural removal across adaptation and allowing recovery between successive pruning decisions. The evolution of structural rankings provides a complementary motivation for this schedule. We track attention-head rankings during dense HLS adaptation without applying pruning. Figure 4 illustrates how initially high-ranked heads can fall below the top-50% selection threshold, while initially excluded heads become highly ranked. Because structural importance can change during training, fixing all pruning decisions at initialization risks removing structures that would become important later in adaptation. Progressive pruning recomputes saliency and removes structures in stages, allowing later pruning decisions to follow the evolving importance rankings.

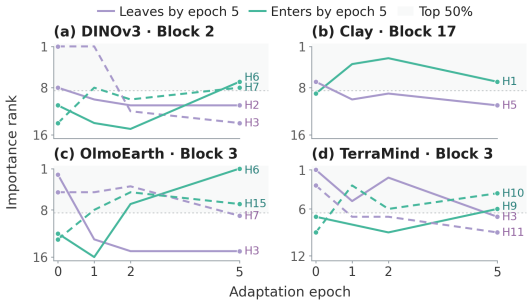


Figure 4: Examples of attention-head importance evolving during task adaptation. During downstream adaptation, initially high-ranked heads can fall below the top-50% selection threshold, while initially excluded heads rise into the selected set.

Effect of calibration composition. Figure 3(c) compares random and coverage-based calibration subsets, where Coverage-32 consistently improves performance across all settings. The largest gain is on TerraMind–Five-Billion-Pixels, because of this task's high heterogeneity. These results show that calibration data composition substantially affects pruning quality, supporting our observation that structural importance should be estimated on inputs that represent the range and coverage of downstream task conditions.

Sensitivity to the pruning ratio. We evaluate Clay–HLS at pruning sparsity levels of 25%, 50%, and 75%, using the same process budget for each method (Table 4(a)). At 25%, all methods exceed the dense reference, and NewOrder is competitive with the best baseline. At 50% and 75%, it achieves the highest mIoU. The difference is most consequential at 75%: all four baselines fall below dense, whereas NewOrder still reaches better performance compared with the full model (90.10). NewOrder therefore retains an accuracy advantage over dense adaptation even at this substantially reduced capacity, extending the main result beyond the 50% pruning setting.

Duration of the pruning phase. Finally, we vary the epoch at which Clay–HLS reaches the final pruning target to examine the trade-off between gradual compression towards the final compact model. Table 4(b) reports the results when pruning ends at epoch 2, 5, and 10, respectively. All three schedules exceed the dense reference. The epoch-10 schedule achieves the highest validation score, while epoch 5 achieves the highest test score, so extending the pruning phase does not consistently improve performance across the two splits.

4 Related Work

Earth observation foundation models. Earth observation foundation models (EOFMs) learn transferable representations from satellite imagery through contrastive learning (Stojnic & Risojevic, 2021; Mañas et al., 2021; Ayush et al., 2021) and masked autoencoders (He et al., 2022; Cong et al., 2022; Reed et al., 2023; Jakubik et al., 2023). Diffusion- (Khanna et al., 2024) and JEPA-based (Li et al., 2024; Astruc et al., 2025) approaches have also been studied. Recent EOFMs have emphasized heterogeneous observations: Clay (Clay Foundation, 2025) and DOFA (Xiong et al., 2025) utilize wavelength-aware dynamic components to generate patch embeddings, while RingMoE (Bi et al., 2026), OlmoEarth (Herzog et al., 2025) and TerraMind (Jakubik et al., 2025) learn representations across modalities. General-purpose ViTs have also shown strong performance when generalizing to EO tasks. DINOv3 introduces Gram anchoring to improve dense feature maps and achieves competitive performance with EOFMs (Siméoni et al., 2025).

Model pruning. Pruning reduces the computational cost of neural networks by removing weights or structures (LeCun et al., 1989). Unstructured pruning methods remove individual weights based on a saliency score but result in irregular sparsity patterns that require specialized software or hardware for speedups (Cheng et al., 2024). Pruning methods discard weights based on magnitude (Han et al., 2015), weight-activation norm products (Sun et al., 2024), or weight-gradient products (e.g., SNIP (Lee et al., 2019), SynFlow (Tanaka et al., 2020)). Structured pruning, on the other hand, reduces the size of neural networks by removing entire channels, attention heads, MLP neurons, or layers, achieving practical speedups without special software or hardware (Cheng et al., 2024). Molchanov et al. (2017) prunes CNNs using a first-order Taylor expansion, Xia et al. (2022) uses learnable masks and a layerwise distillation objective, while Zhu et al. (2021) learns sparse importance scores to prune the linear projection dimensions in ViTs. Yu et al. (2022) prunes heads and dimensions via learnable saliency and drops blocks using plug-in classifiers, NViT applies a Hessian-based criterion (Yang et al., 2023), and Isomorphic Pruning identifies and ranks isomorphic substructures using graph representations (Fang et al., 2024).

Efficient EOFMs. Several efforts have been made toward developing efficient EOFMs. Du et al. (2025) applies knowledge distillation to Prithvi-EO-2.0 for on-orbit satellite deployment. Other methods focus on the fine-tuning stage: Marti Escofet et al. (2026) evaluates LoRA, visual prompt tuning, and ViT adapters on several EOFMs, and Snyder et al. (2026) uses dynamical low-rank training during transfer learning to compress ViT-based geospatial models. Recent works have also explored pruning for EOFMs. SIMPLER (Barreiro et al., 2026) uses Centered Kernel Alignment to measure layer-wise representation similarity and truncates redundant layers, while Said et al. (2026) compares structured magnitude-, Taylor-, and Hessian-based pruning methods on fully fine-tuned model checkpoints. In contrast, NewOrder performs structured width pruning prior to full fine-tuning, removing attention heads and MLP neurons rather than full Transformer blocks.

5 Conclusion

We present NewOrder, a framework for structured pruning during the early stages of EOFM adaptation. It combines output-sensitive scoring with progressive removal and calibration data coverage. Across four encoders and four tasks, NewOrder removes approximately half of the encoder parameters while achieving the state-of-the-art pruned performance in most settings and exceeding dense adaptation in eight. Full-model profiling confirms reductions in training-step latency, inference latency, and memory consumption, and a complete training profile demonstrates a net reduction in adaptation time. These findings show that substantial compression can accompany improved downstream accuracy, supporting an approach to adaptation that learns both how to update pretrained parameters and which structures the downstream model should retain.

AI USE STATEMENT

In this work, we have used LLMs to help refine the paper writing and check grammar errors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

As a model compression method, NewOrder primarily targets reducing the resource demands of Earth foundation models without altering their behavior or capabilities. We do not foresee direct societal risks or misuse arising uniquely from this work. However, as with any efficiency technique, care should be taken to ensure that improved accessibility to large models does not inadvertently facilitate harmful or unintended applications.

REPRODUCIBILITY STATEMENT

To ensure our work is fully reproducible, we've provided extensive details on our experimental setup, hardware specifications, and algorithms, complete with pseudocode and illustrative figures in both the main body and in the appendix. For direct validation and easy adoption, the code implementation is available at https://anonymous.4open.science/r/Efficient_EOFM-F526/.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19530–19540, 2025. doi: 10.1109/CVPR52734.2025.01819.

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10181–10190, October 2021.

Víctor Barreiro, Johannes Jakubik, Francisco Argüello, and Dora B. Heras. Simpler: Efficient foundation model adaptation via similarity-guided layer pruning for earth observation, 2026. URL https://arxiv.org/abs/2603.19873.

Hanbo Bi, Yingchao Feng, Boyuan Tong, Mengyu Wang, Haichen Yu, Yongqiang Mao, Hao Chang, Wenhui Diao, Peijin Wang, Yue Yu, Hanyang Peng, Yehong Zhang, Kun Fu, and Xian Sun. Ringmoe: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (4):4388–4405, 2026. doi: 10.1109/TPAMI.2025.3643453.

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10558–10578, 2024. doi: 10.1109/TPAMI.2024.3447085.

Clay Foundation. Clay foundation model, version 1.5. https://clay-foundation.github.io/model/, 2025.

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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211. Curran Associates, Inc., 2022. doi: 10.52202/068431-0015. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/01c561df365429f33fcd7a7faa44c985-Paper-Conference.pdf.

Andrew Du, Roberto Del Prete, Alejandro Mousist, Nick Manser, Fabrice Marre, Andrew Barton, Carl Seubert, Gabriele Meoni, and Tat-Jun Chin. First on-orbit demonstration of a geospatial foundation model, 2025. URL https://arxiv.org/abs/2512.01181.

Gongfan Fang, Xinyin Ma, Michael Bi Mi, and Xinchao Wang. Isomorphic pruning for vision models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXX*, pp. 232–250, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73403-8. doi: 10.1007/978-3-031-73404-5_14. URL https://doi.org/10.1007/978-3-031-73404-5_14.

Lucia Gordon, Serge Belongie, Christian Igel, and Nico Lang. Mmearth-bench: Global model adaptation via multimodal test-time training, 2026. URL https://arxiv.org/abs/2602.06285.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01553.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. *arXiv preprint arXiv:2511.13655*, 2025.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence, 2023. URL https://arxiv.org/abs/2310.18660.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, et al. Terramind: Large-scale generative multimodality for earth observation. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=I5webNFDgQ.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/6c9882bbac1c7093bd25041881277658-Paper.pdf.

Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H.S. Torr. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2019.

Weijie Li, Wei Yang, Tianpeng Liu, Yuenan Hou, Yuxuan Li, Zhen Liu, Yongxiang Liu, and Li Liu. Predicting gradient is better: Exploring self-supervised learning for sar atr with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:326–338, 2024. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2024.09.013. URL https://www.sciencedirect.com/science/article/pii/S0924271624003514.

Francesc Marti Escofet, Benedikt Blumenstiel, Linus Scheibenreif, Paolo Fraccaro, and Konrad Schindler. Fine-tune smarter, not harder: Parameter-efficient fine-tuning for geospatial foundation models. In Rita P. Ribeiro, Bernhard Pfahringer, Nathalie Japkowicz, Pedro Larrañaga, Alípio M. Jorge, Carlos Soares, Pedro H. Abreu, and João Gama (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 516–532, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-06106-5.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9394–9403, 2021. doi: 10.1109/ICCV48922.2021.00928.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.

Jan Pauls, Max Zimmer, Berkant Turan, Sassan Saatchi, Philippe Ciais, Sebastian Pokutta, and Fabian Gieseke. Capturing temporal dynamics in large-scale canopy tree height estimation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48422–48438. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pauls25a.html.

Christopher Phillips, Sujit Roy, Ankur Kumar, and Rahul Ramachandran. HLS Foundation Burnscars Dataset, 2023. URL https://huggingface.co/ibm-nasa-geospatial/hls_burn_scars.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4088–4099, October 2023.

Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3185–3199, 2021. doi: 10.1109/JSTARS.2021.3063849.

Amina Said, Julia Dietlmeier, Margaret McCaul, and Noel E. O'Connor. When less is more: Evaluating structural pruning in geospatial foundation models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1403–1413, 2026. doi: 10.1109/WACVW68408.2026.00152.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Thomas Snyder, H. Lexie Yang, Stefan Schnake, and Steffen Schotthöfer. Compressing vision transformers in geospatial transfer learning with manifold-constrained optimization, 2026. URL https://arxiv.org/abs/2601.08882.

Vladan Stojnic and Vladimir Risojevic. Self-Supervised Learning of Remote Sensing Scene Representations Using Contrastive Multiview Coding . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1182–1191, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPRW53098.2021.00129. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW53098.2021.00129.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024.

Hidenori Tanaka, Daniel Kunin, Daniel Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6377–6389. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/46a4378f835dc8040c8057beb6a2da52-Paper.pdf.

Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:178–196, 2023. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2022.12.011. URL https://www.sciencedirect.com/science/article/pii/S0924271622003264.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1513–1528, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.107. URL https://aclanthology.org/2022.acl-long.107/.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*, pp. 418–434, 2018.

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 multimodal foundation model for earth observation, 2025. URL https://arxiv.org/abs/2403.15356.

Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18547–18557, 2023. doi: 10.1109/CVPR52729.2023.01779.

Fang Yu, Kun Huang, Meng Wang, Yuan Cheng, Wei Chu, and Li Cui. Width & depth pruning for vision transformers. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, pp. 3143–3151, 2022.

Michael H. Zhu and Suyog Gupta. To prune, or not to prune: Exploring the efficacy of pruning for model compression, 2018. URL https://openreview.net/forum?id=S1lN69AT-.

Mingjian Zhu, Yehui Tang, and Kai Han. Vision transformer pruning, 2021. URL https://arxiv.org/abs/2104.08500.

A Additional experimental details

A.1 Foundation models and shared decoder

We use the released checkpoints of DINOv3 ViT-L SAT-493M (Siméoni et al., 2025), Clay v1.5 (Clay Foundation, 2025), OlmoEarth v1-Large (Herzog et al., 2025), and TerraMind 1.0 Base (Jakubik et al., 2025). Table 5 reports the structures discovered by our pruning implementation. Parameter counts include the encoder and its input adapter but exclude the downstream decoder. DINOv3 uses RGB; the other encoders use their released multispectral input adapters and the available task bands.

Table 5: Released encoders used in the controlled study. "Heads" reports the number of heads times the per-head width.

Encoder	Input	Patch	Blocks	Width	MLP width	Heads	Params	Checkpoint license
DINOv3 ViT-L (SAT-493M)	RGB	16	24	1024	4096	16×64	303.1M	DINOv3 License
Clay v1.5	Multispectral	8	24	1024	4096	16×64	311.4M	OpenRAIL-M
OlmoEarth v1-Large	Multispectral	8	24	1024	4096	16×64	307.8M	OlmoEarth Artifact License
TerraMind 1.0 Base	Multispectral	16	12	768	2048	12×64	86.1M	Apache-2.0

Every encoder is paired with the same newly initialized UPerNet-style decoder (Xiao et al., 2018). A projection of the final ViT feature map is converted into a four-level feature pyramid at strides 4, 8, 16, and 32. The decoder uses 256 channels, pyramid-pooling bins $\{1, 2, 3, 6\}$, GroupNorm, GELU, and dropout 0.1, followed by bilinear resizing to the target resolution. We optimize every encoder and decoder parameter; no linear probing or frozen-encoder setting is used.

A.2 Tasks, splits, and preprocessing

Table 6 gives the exact splits consumed by the training pipeline. HLS Burn Scars (Phillips et al., 2023) contains six-band Harmonized Landsat–Sentinel imagery for binary burn-scar mapping. We deterministically split its 540 released training scenes into 486 training and 54 validation scenes and use the 264 scenes in the released validation directory only as the held-out test set. DFC2020 (Robinson et al., 2021) uses 13-band Sentinel-2 imagery; its labels are mapped to eight land-cover classes, with invalid pixels ignored. Five-Billion-Pixels (Tong et al., 2023) uses four-band Gaofen-2 imagery for 24-class land-cover mapping; unlabeled pixels are ignored. MMEarth (Gordon et al., 2026) is a dense above-ground biomass regression task. We select checkpoints on its validation split and report only the geographically disjoint test split in the main table.

Table 6: Downstream task inputs and split sizes. "Source input" describes the stored task sample before model-specific band selection.

Task	Source input	Targets	Train	Validation	Test	Metric
HLS Burn Scars	6×512×512	2 classes	486	54	264	mIoU
DFC2020	13×128×128	8 classes	41,535	4,617	8,874	mIoU
MMEarth biomass	12×128×128	Regression	10,665	2,286	3,156 (geo.)	Pixel R^2
Five-Billion-Pixels	4×256×256	24 classes	19,651	2,183	5,459	mIoU

Inputs are converted to reflectance and normalized with each checkpoint's released preprocessing statistics. DINOv3 receives the task's RGB bands. Clay, OlmoEarth, and TerraMind receive the compatible multispectral bands from their pretrained input adapters; atmospheric Sentinel-2 bands unsupported by an adapter are omitted. For Five-Billion-Pixels, each band is first clipped to its training-set mean plus or minus two standard deviations and scaled to $[0, 1]$, before model-specific normalization. Segmentation training uses random crops where required, horizontal and vertical flips, and rotations by multiples of 90 degrees; evaluation uses deterministic center crops. We do not augment MMEarth biomass samples.

A.3 Optimization and model selection

The main experiments use 20 training epochs, distributed data parallelism on four NVIDIA H100 80GB GPUs, and BF16 autocast. We optimize with AdamW, weight decay 0.01, cosine learning-rate

decay, linear warmup over the first 10% of optimizer steps, and gradient clipping at norm 1.0. The encoder learning rate is one tenth of the decoder learning rate. Segmentation uses cross-entropy over valid pixels; biomass regression uses Huber loss with $\delta = 10$.

For HLS, DFC2020, and MMEarth, each dense or pruned setting receives the same 20-configuration search budget. Decoder learning rates $\{3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$ are crossed with global batch sizes $\{8, 16, 32, 64\}$ for HLS, $\{32, 64, 128, 256\}$ for DFC2020, and $\{16, 32, 64, 128\}$ for MMEarth. Five-Billion-Pixels uses fixed decoder and encoder learning rates of 3×10^{-4} and 3×10^{-5}, respectively, with global batch 64 and a per-GPU microbatch of 4. Compatible per-GPU microbatches and gradient accumulation realize the requested global batch sizes.

Checkpoints and hyperparameters are selected by validation mIoU or R^2, and the selected checkpoint is evaluated on the held-out test set.

A.4 Pruning criteria and calibration

All pruning methods operate on the dependency-coupled attention-head and MLP-neuron groups described in Appendix B and use the same physical deletion rules. Magnitude sums absolute parameter values within each group (Han et al., 2015). Activation weights parameter magnitudes by the ℓ_2 norms of their input activations (Sun et al., 2024). Weighted gradient sums the magnitudes of the parameter–gradient products for the supervised task loss (Lee et al., 2019). Group Taylor takes the absolute value after summing the signed first-order terms within each group (Molchanov et al., 2017). NewOrder uses the output response in Equation 1 and aggregates its parameter sensitivities according to Equation 3.

Data-dependent scoring uses 32 training examples, calibration batch size 1, and sampling seed 42. Magnitude requires no calibration data, and dense adaptation performs no structural scoring. Activation uses input activations; weighted gradient and Taylor also use training targets. NewOrder computes its score from the encoder–decoder output without a supervised loss.

Terrain coverage. For NewOrder, we sample the calibration images at random subject to coverage of the terrain groups represented in the training set. Let $\mathcal{T}(x)$ denote the terrain groups associated with an image and $\mathcal{T}_{\mathrm{tr}}$ the set of groups in the training data. The calibration subset satisfies

$$|\mathcal{C}| = 32, \qquad \bigcup_{x\in\mathcal{C}} \mathcal{T}(x) = \mathcal{T}_{\mathrm{tr}}. \tag{5}$$

This rule applies to all four downstream tasks, including biomass regression. It uses terrain-group coverage to represent geographic heterogeneity within a fixed scoring budget. Table 8 compares this Coverage-32 composition with Random-32, which samples the same number of training examples without the coverage constraint. Calibration examples are selected from the training split.

A.5 Pruning schedule and structural budget

In the main comparison, each method removes 50% of the attention heads and 50% of the MLP neurons in every Transformer block, subject to integer counts and retaining at least one structure of each type. This target specifies the removable internal widths. Embeddings, the block output widths, and the decoder remain intact, so the percentage reduction in encoder parameters is approximately, rather than exactly, 50%.

Magnitude, Activation, weighted gradient, and Taylor compute their rankings once before optimization and train the resulting compact model for 20 epochs. NewOrder interleaves ten pruning stages with joint encoder–decoder training during the first five epochs. The cumulative targets are $\{5, 10, \ldots, 50\}\%$, with stage boundaries spaced uniformly in optimizer steps. At each stage, importance is recomputed on the current model, and the lowest-scoring surviving heads and neurons are removed separately within each block. Epochs 6–20 train the final compact architecture. The pruning phase is part of the same 20-epoch adaptation budget. All methods begin with released pretrained encoders and newly initialized decoders.

Table 7: Controlled NewOrder ablations on HLS Burn Scars. (a) Local versus global structural allocation. (b) Supervised loss-gradient versus output-sensitive saliency. Bold marks the better value in each pair; Δ is relative to the first variant.

(a) Structural allocation

Model	Global	Local	Δ
DINOv3	84.86	**86.75**	+1.89
Clay	91.60	**91.80**	+0.20
OlmoEarth	86.60	**86.62**	+0.02
TerraMind	90.38	**91.82**	+1.44

(b) Saliency objective

Model	Loss grad.	NewOrder	Δ
DINOv3	85.93	**86.75**	+0.83
Clay	91.08	**91.80**	+0.72
OlmoEarth	85.79	**86.62**	+0.83
TerraMind	**91.97**	91.82	-0.15

Table 8: Calibration composition for NewOrder.

Task	Model	Random-32	Coverage-32	Δ
HLS	DINOv3	85.03	**86.75**	+1.72
	Clay	91.52	**91.80**	+0.28
	OlmoEarth	83.94	**86.62**	+2.68
	TerraMind	91.03	**91.82**	+0.79
DFC2020	DINOv3	49.13	**49.80**	+0.67
	Clay	49.69	**52.30**	+2.61
	OlmoEarth	53.30	**54.77**	+1.47
	TerraMind	53.92	**54.70**	+0.78
MMEarth	DINOv3	44.93	**46.58**	+1.65
	Clay	**44.70**	**44.70**	+0.00
	OlmoEarth	51.10	**52.31**	+1.22
	TerraMind	51.82	**52.89**	+1.07
FBP	DINOv3	59.37	**61.35**	+1.98
	Clay	47.01	**47.72**	+0.72
	OlmoEarth	51.27	**53.42**	+2.15
	TerraMind	42.81	**48.27**	+5.46

A.6 Further Details on Ablation & Parameter Study

The main text examines the components of NewOrder through controlled ablations and parameter studies. Here, we provide the detailed numerical results for the saliency, progressive pruning, and calibration ablations summarized in the main-text figures. We additionally compare local and global structural allocation to examine the choice of pruning budget across Transformer blocks.

Local versus global allocation. NewOrder adopts a local-style pruning where each attention and MLP layer applies the same sparsity. We compare this local-style with a global-style pruning where structured across all layers are ranked together in Table 7(a). At the same pruning target, local allocation achieves higher performance than global allocation for all four models, supporting NewOrder's per-block local allocation in this setting.

B Structural Groups and Physical Deletion

Linear weights are stored with output channels as rows and input channels as columns. Each group g in Equation 3 contains the parameter coordinates removed together for one neuron or attention head. Groups are formed within each encoder block; decoder parameters are excluded.

MLP neurons. For neuron j, we remove row j of the input projection, its bias entry when present, and column j of the down projection. A gated MLP requires the same row and bias entry to be

Table 9: NewOrder with one-shot versus dynamic progressive pruning. Progressive uses ten stages ending at epoch 5. Bold marks the larger value in each pair.

Task	Model	One-shot	Progressive	Δ
HLS	DINOv3	83.51	**86.75**	+3.24
	Clay	91.44	**91.80**	+0.36
	OlmoEarth	85.90	**86.62**	+0.72
	TerraMind	91.01	**91.82**	+0.81
DFC2020	DINOv3	47.27	**49.80**	+2.53
	Clay	50.34	**52.30**	+1.96
	OlmoEarth	53.54	**54.77**	+1.23
	TerraMind	52.80	**54.70**	+1.90
MMEarth	DINOv3	43.99	**46.58**	+2.59
	Clay	42.38	**44.70**	+2.32
	OlmoEarth	50.31	**52.31**	+2.00
	TerraMind	50.79	**52.89**	+2.10

removed from both the gate and value branches. The down-projection bias is retained because its output dimension is unchanged.

Attention heads. For each removed head, we delete the rows and bias entries corresponding to its channels in the query, key, and value projections, together with the matching columns of the attention output projection. In a fused QKV matrix, these rows are taken from each of the three output segments. Each retained head keeps its original dimension and attention scaling. The attention output-projection bias is retained.

Physical deletion and interface. Retained indices are sorted to preserve channel order. Slicing the coupled rows, columns, and bias entries reduces the internal projection sizes while preserving the residual-stream width and the encoder outputs used by the decoder. The resulting projections execute as dense operators. These grouping rules cover the attention and MLP structures evaluated in the four encoders; additional operations that couple intermediate channels require their own dependency handling.

For the local response analysis, we represent deletion by zeroing the group's parameter coordinates in the unsliced model. Physical slicing has the same forward effect under the stated dependency rules when no normalization or other operation couples removed and retained intermediate channels.

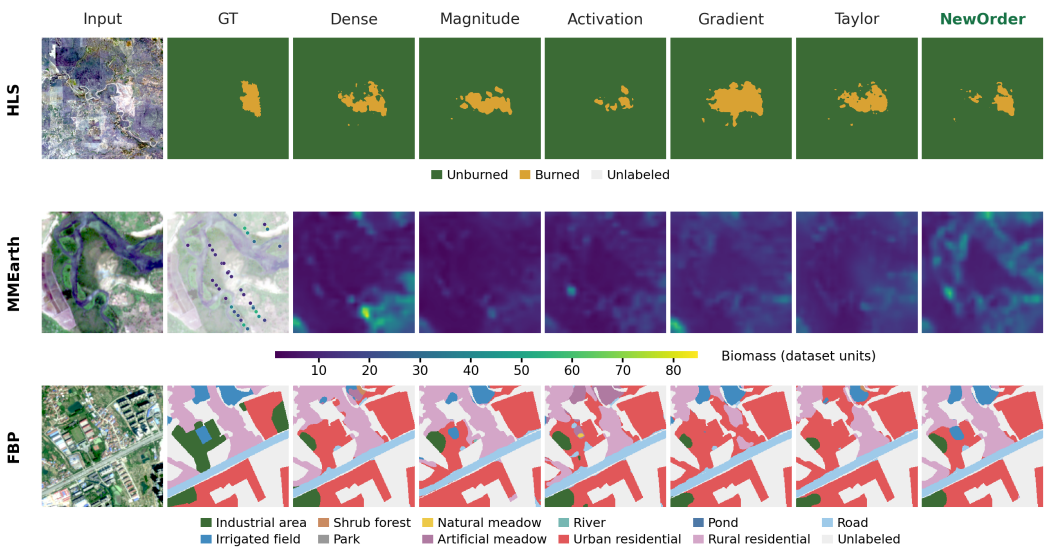


Figure 5: Representative qualitative predictions. We compare the dense reference and the 50%-pruned NewOrder model against the ground truth.

C Qualitative Case Study

In addition to the quantitative comparisons in Table 2, we visualize representative held-out predictions in Figure 5. Despite removing approximately half of the encoder parameters, NewOrder preserves the main spatial structures and prediction patterns of the dense model while remaining closely aligned with the ground truth. Across the displayed examples, the compact model produces qualitatively comparable predictions without an apparent loss of spatial detail. These cases complement the quantitative results and illustrate that the efficiency gains of NewOrder translate to meaningful dense predictions after structured compression.

D STRUCTURAL SALIENCY ANALYSIS

D.1 DO DIFFERENT SCORES PRESERVE THE SAME STRUCTURES?

Protocol. We analyze five saliency criteria on four released pretrained encoders (DINOv3, Clay, OlmoEarth, and TerraMind) and four downstream tasks (HLS Burn Scars, DFC2020, MMEarth biomass, and Five-Billion-Pixels).

Measurement. For two criteria with retained structure sets K_a and K_b, we report $J(K_a, K_b) = |K_a \cap K_b|/|K_a \cup K_b|$. An identity comprises the block, structure family, and original unit index. Figure 6 averages the 16 model–task contexts, separately for heads and neurons. Repeated copies of the full-mask statistic attached to different top-k diagnostics are counted only once.

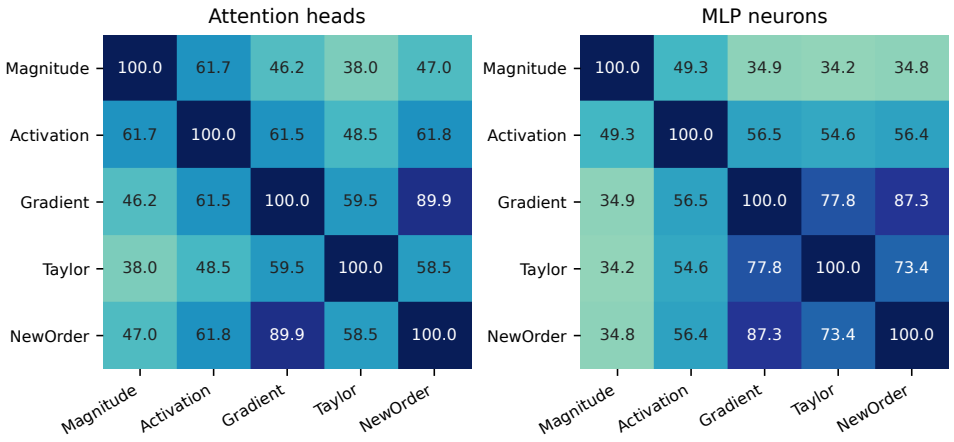


Figure 6: Mean retained-set Jaccard (%) for static saliency criteria.

Table 10: Mean retained-set Jaccard (%) against the static NewOrder score.

Comparator	Attention	MLP
Magnitude	46.98	34.80
Activation	61.76	56.40
Gradient	89.93	87.29
Taylor	58.53	73.42

Result and interpretation. The closest agreement is between NewOrder and weighted gradient: 89.93% for heads and 87.29% for neurons. Agreement with magnitude is much lower, at 46.98% and 34.80%, respectively. Taylor agrees more closely with NewOrder on neurons (73.42%) than on heads (58.53%). These observations show substantial shared preferences between gradient-based scores, together with nonidentical retained sets. They are consistent with NewOrder and weighted gradient sharing an absolute weight–gradient form while using different differentiation objectives.

D.2 ARE STRUCTURAL CHOICES TASK- AND MODEL-DEPENDENT?

Protocol. We next hold the pretrained encoder and scoring criterion fixed and compare retained sets across downstream tasks. Each encoder contributes six unordered task pairs. The full archive contains 240 unique comparisons: four encoders, five scores, six task pairs, and two structure families. Figure 7 shows the 48 comparisons for the NewOrder output score.

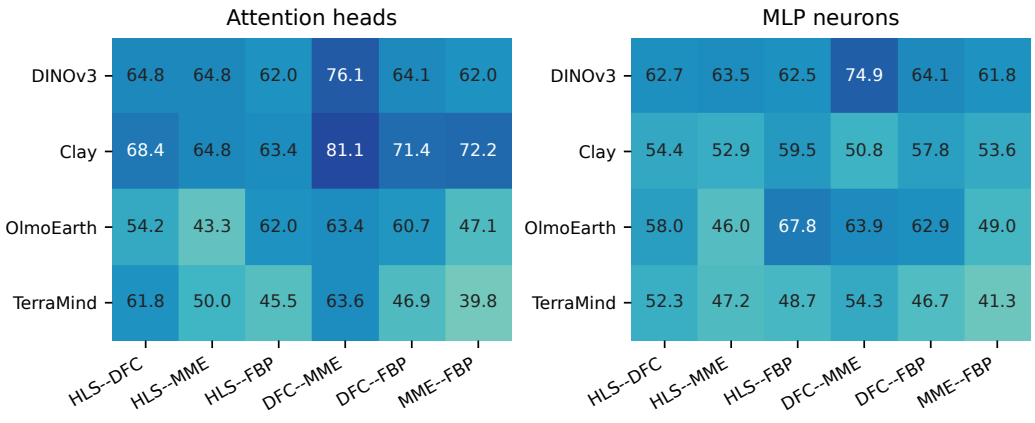


Figure 7: Cross-task retained-set Jaccard (%) for NewOrder, shown separately by encoder and structure family. MME denotes MMEarth.

Table 11: Cross-task agreement of the static NewOrder score. Values are mean Jaccard (%) over six task pairs within each encoder.

Encoder	Attention	MLP
DINOv3	65.65	64.91
Clay	70.23	54.83
OlmoEarth	55.12	57.92
TerraMind	51.27	48.43

Result. Across encoders and task pairs, mean NewOrder agreement is 60.57% for heads and 56.52% for neurons. Magnitude is nearly invariant across tasks (both family means round to 100.00%), consistent with scoring pretrained weights independently of data. The raw archive contains three Clay MLP comparisons at 99.9837%, so exact identity is not assumed. Activation gives 71.10% and 67.78%; weighted gradient gives 59.00% and 56.49%; Taylor gives 48.80% and 51.02%. Thus, data-dependent criteria select different subsets under different task contexts, whereas the data-independent control is nearly unchanged.

Model dependence. The effect is not uniform across representations. Clay has the highest mean cross-task head agreement (70.23%), whereas DINOv3 has the highest neuron agreement (64.91%). TerraMind has lower agreement in both families (51.27% and 48.43%). Even within Clay, heads and neurons differ substantially (70.23% versus 54.83%). These results support reporting the interaction between encoder, task, and structure family.

D.3 How concentrated are the saliency distributions?

Measurement. Raw score magnitudes have different scales across objectives and layers. We therefore summarize nonnegative scores with the scale-invariant Gini coefficient within each block and structure family. For sorted scores $s_{(1)} \le \cdots \le s_{(n)}$, we compute

$$G = \frac{2\sum_{i=1}^{n} i s_{(i)}}{n\sum_{i=1}^{n} s_{(i)}} - \frac{n+1}{n}, \tag{6}$$

with $G = 0$ when all scores are zero. We first take the median across blocks within each model–task context, then summarize the 16 context values. This avoids conflating between-layer scale differences with within-layer concentration, which is the relevant quantity for local pruning.

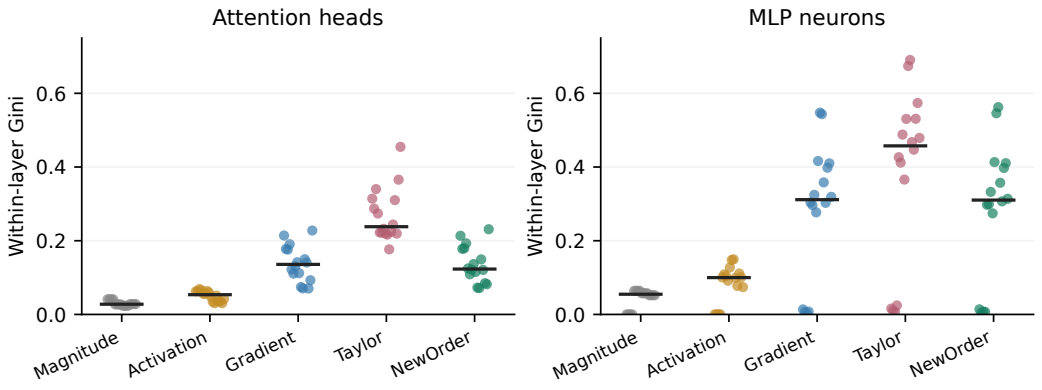


Figure 8: Within-layer score concentration. Each point represents one encoder–task context after taking its median across blocks; black bars mark the median of the 16 contexts. Higher Gini indicates more concentrated scores.

Table 12: Median within-layer Gini across 16 contexts.

Score	Attention	MLP
Magnitude	0.028	0.055
Activation	0.054	0.100
Gradient	0.136	0.311
Taylor	0.238	0.457
NewOrder	0.123	0.310

Result and interpretation. Magnitude has comparatively flat within-layer scores (median Gini 0.028 for heads and 0.055 for neurons). Taylor is more concentrated (0.238 and 0.457). The NewOrder distribution resembles weighted gradient, especially for neurons (0.310 versus 0.311).