ANCHORED LATENT PREDICTIONS FOR EARTH FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised Earth-observation encoders learn scene context, but mapping clouds and floods also requires local detail. In two runs of a latent-prediction recipe with weak pixel-target weights, extending training from 10B to 40B tokens cost 5.6 and 5.7 points of frozen-probe cloud mIoU; thin-cloud IoU fell by 28 to 34%, and crop-type and flood segmentation also degraded. Meanwhile, land-cover and GEO-Bench accuracy peaked at the worst cloud checkpoints, and feature rank rose. Better scene recognition can therefore hide a loss of spatial detail, motivating fixed pixel-level targets that keep evolving representations tied to the observations. We propose *anchored latent prediction* (ALP): complement evolving latent targets with strongly weighted, frozen random projections of the pixels, predicted at twice the token resolution with no extra encoder cost. In single-run tests, switching to ALP's loss weights, including a higher cloud-loss weight, halted the decline; an ALP run kept improving until about 20B tokens. EarthFM, a 91M-parameter encoder trained with ALP plus map and cloud-model targets, **is on par with or better than OlmoEarth v1.2 Base on six of eight public test sets** from GEO-Bench, GEO-Bench-2, and Sen1Floods11, with **1.8 to 3 times less hardware-adjusted pretraining compute**: ahead on clouds, within noise on five tasks, and behind on EuroSAT and So2Sat. Its shorter 18.75B-token branch leads on clouds with a third of OlmoEarth's compute (+1.15 mIoU, 95% CI 0.53 to 1.79); only TerraMind-L scores higher, with 7.2 times as much compute.

1 INTRODUCTION

Satellite imagery supports agricultural mapping, flood response, and environmental monitoring. These uses need both scene context and local detail: recognizing a landscape is different from locating a cloud shadow or a field boundary. Many Earth-observation foundation models learn by latent prediction, forecasting a teacher network's features for hidden parts of an image (Assran et al., 2023; Tseng et al., 2025), and they keep growing in size and pretraining compute (Herzog et al., 2026; Jakubik et al., 2025). Their pretraining is monitored with scene-level probes and feature statistics, and they are compared by aggregate rankings over many tasks (Marsocci et al., 2026; Simumba et al., 2025). How local detail changes as training continues is rarely tracked.

The problem. Local detail can quietly degrade. We trained a latent-prediction recipe that puts little weight on its pixel-level targets, which we call *latent-heavy*, for 40B tokens in two independent runs (Section 3). The last 30B tokens cost about as much compute as training EarthFM end to end and made the cloud maps worse: under a frozen linear probe, cloud mIoU fell by 5.6 and 5.7 points, from level with OlmoEarth v1.2 Base to 4.9 and 5.9 points behind it, and crop-type and flood segmentation fell as well. Meanwhile land-cover and GEO-Bench accuracy peaked at the final checkpoint, each run's worst for clouds, and by 25B tokens feature rank was the highest of the 37 encoders we measured. A model chosen by those checks would have been one of each run's worst for cloud maps. Similar dense degradation has been reported for natural-image models (Siméoni et al., 2026; Dai et al., 2025), so the problem is not specific to our data. We hypothesize that it is a general risk of latent prediction: latent targets are a moving goal. The teacher is an average of the model itself, and when pixel-level targets carry little weight, little forces its features to keep describing the pixels. If that is the cause, the remedy is targets that cannot drift, weighted enough to matter.

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.

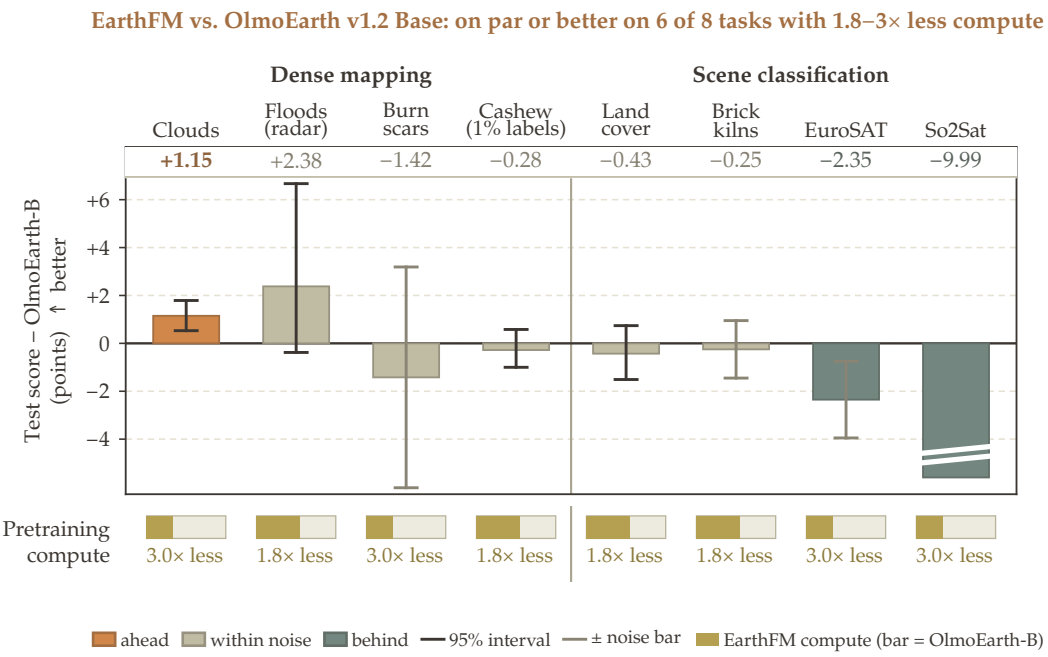

Figure 1: EarthFM is on par with or better than OlmoEarth v1.2 Base on six of eight official test sets, with 1.8 to 3 times less pretraining compute. Bars: EarthFM minus OlmoEarth v1.2 Base in each task's own metric (points). Whiskers: the paired 95% interval or, where none exists, ± the noise bar from the task's pre-registered rule, so a whisker that crosses zero means within noise. So2Sat's bar is cut (break marks); its value is printed above. Strip: pretraining compute of the EarthFM checkpoint used, where the full bar is OlmoEarth's. Clouds use EarthFM-short; burn scars, EuroSAT, and So2Sat an 18.75B-token checkpoint; the rest EarthFM. The clouds lead is a tie on validation, and pretraining uses a cloud model's labels.

Our solution. We propose *anchored latent prediction* (ALP; Section 4, Figure 3). ALP brings together complementary target families: I-JEPA and Galileo predict evolving latent targets (Assran et al., 2023; Tseng et al., 2025), and OlmoEarth uses frozen random projections of the pixels, which it reports stabilizes training (Herzog et al., 2026). ALP keeps both. The latent targets carry context; the frozen projections, which we call anchors, cannot drift, and ALP weights them so that they send the encoder about as much gradient as the latent term or more (0.9 to 2.1 times). ALP also predicts the anchors below the token grid, for each 80 m patch and each of its four 40 m sub-patches, through queries that carry only position, time, and sensor, so the finer targets add no encoder tokens and no inference cost. As a by-product, how well the features still recover the anchors gives a label-free warning sign for lost detail.

Evidence that it works. In single-run tests (Section 5, Figure 2b), switching the declining run to ALP's loss weights mid-training halts its decline, and the reverse switch costs an ALP run 1.9 points; the switch also changes the cloud weight. Removing the anchors early in training costs 4 to 7 cloud mIoU points even where they carry little weight. ALP extends the useful training window, with learning-rate scheduling and checkpoint selection guiding later training.

What it buys. We build EarthFM, a 91M-parameter multisensor encoder trained with ALP plus targets from existing maps and a cloud model (Figure 1); its comparisons test this whole recipe. We compare it with OlmoEarth v1.2 Base, a recent open encoder of the same size, on eight public benchmarks with official test splits from GEO-Bench, GEO-Bench-2, and Sen1Floods11 (Lacoste et al., 2023; Simumba et al., 2025; Bonafilia et al., 2020), four dense and four scene-level; prior papers or the GEO-Bench-2 leaderboard report each of them, and we re-measure every baseline under one protocol, except Galileo-B on land cover (Section 6). EarthFM is on par with or better than OlmoEarth on six of the eight, with 1.8 to 3 times less pretraining compute: it is ahead on clouds, within noise on floods, burn scars, cashew plantations, land cover, and brick kilns, and behind on EuroSAT and So2Sat. We will release code, configurations, checkpoints, and per-seed results.

2 Related work

Targets for Earth-observation pretraining. Galileo combines shallow and deep targets with different masking strategies; its shallow targets are the EMA encoder's own input projection, so they evolve with the model, and its predictor queries carry only position, time, and channel embeddings (Tseng et al., 2025). OlmoEarth replaces the evolving target encoder with a frozen random projection, which it reports stabilizes training, and adds map targets (Herzog et al., 2026); v1.2 changes its tokenization, masking, loss, and positional encoding and cuts pretraining compute (Tseng et al., 2026). TerraMind learns from pixels and discrete tokens across many modalities (Jakubik et al., 2025), CROMA aligns radar and optical views (Fuller et al., 2023), and DeCUR separates their common and modality-specific representations (Wang et al., 2024). ALP keeps both kinds of target, weights the anchors so that they send the encoder at least about as much gradient as the latent term, and predicts them at twice the linear resolution of the token grid through Galileo-style metadata-only queries.

Dense degradation. DINOv3 reports that dense features degrade during long training and repairs them with Gram anchoring, which pulls the student's patch-similarity matrix toward that of an earlier snapshot (Siméoni et al., 2026). Dai et al. (2025) show that image-level diagnostics can miss such degradation, and Cross-Scale Pretraining transfers spatial structure from higher-resolution imagery (Waithaka et al., 2026). Our results are consistent with these reports. We add the multisensor setting, two-seed replication with scene accuracy rising at the same time, and token-level RankMe (Garrido et al., 2023) ranking two strongly degraded checkpoints highest. ALP differs from Gram anchoring in that it predicts fixed pixel features from the first step and needs no snapshot; we do not compare the two.

3 The problem: longer training, worse maps

Setting. We pretrain a 12-block transformer on multisensor imagery by latent prediction (architecture in Section 4). The recipe studied here, which we call *latent-heavy*, predicts an EMA teacher's features, averaged over its top four blocks. It also predicts the same frozen pixel anchors that ALP uses, but puts most of the loss on the latent term: $(\lambda_p, \lambda_s, \lambda_l) = (0.3, 0.2, 6)$ against ALP's $(1, 0.5, 4)$, so the anchors send the encoder only 0.15 to 0.71 times as much gradient as the latent term (Section 4). Like our ALP runs, it has map and cloud-model heads, so the decline below happens with a cloud target in the objective. It is a later variant of our ALP recipe that also changed several other settings (Table 4); we present ALP second because it answers the problem, not because it came first. We follow cloud segmentation (dense) and land-cover classification (scene) with frozen linear heads on the official CloudSEN12 and BigEarthNet test splits, which we read repeatedly during development, so we treat these trajectories as exploratory and report validation alongside. Other tasks use development panels, fixed samples of each task's validation split, or GEO-Bench test splits. *Annealed* checkpoints are short learning-rate cooldown branches; the others come from the main run, whose learning rate in the latent-heavy runs is constant until 26B tokens and then decays to 40B.

Four times the training, worse cloud maps. Between 10B and 40B tokens, the latent-heavy recipe loses 5.6 and 5.7 cloud mIoU points in its two pretraining seeds (63.50 to 57.87 and 62.51 to 56.84; Figure 2a). For scale: at 10B both runs are level with OlmoEarth v1.2 Base under the same probe (62.76); at 40B they are 4.9 and 5.9 points behind it and at or below Galileo-B (58.78), and the drop is larger than the whole 4.0-point gap between those two published models. All four pretraining-seed and head-seed pairs lose 5.5 to 5.9 points, about twenty times the median head-seed spread (0.26) and five times the gap between pretraining seeds. Validation shows the same fall (-5.8 and -6.9), all four cloud classes fall, and thin clouds, the finest class, lose 28 and 34% of their IoU (35.3 to 25.5 and 33.1 to 21.9). The 30B extra tokens cost 233 to 243 B200 GPU-hours per run, about as much as EarthFM's entire pretraining.

Segmentation falls; classification does not. The loss is not specific to clouds (Figure 2c). Between the annealed 12.5B and 25B checkpoints, the latent-heavy runs lose 3.3 and 3.0 points of crop-type mIoU and 2.4 and 5.5 points of flood IoU on development panels, while land cover and So2Sat hold or rise, EuroSAT dips by at most 1.2 points, and canopy-height error does not worsen. Over the same branch points, one ALP run holds all three segmentation tasks within 0.6 points. The scene metric

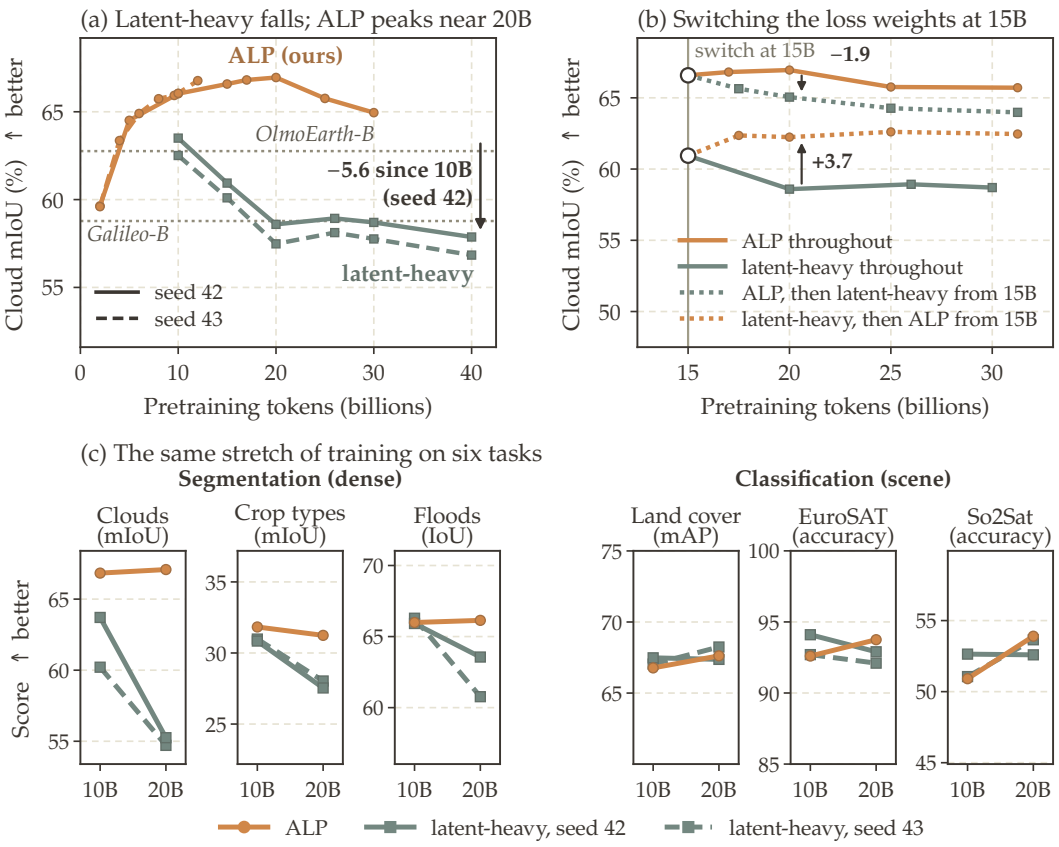

Figure 2: Longer training erodes segmentation under latent-heavy weights; ALP's weights keep it higher for longer (one ALP seed after 12B). Frozen linear probes; higher is better everywhere; latent-heavy puts most loss weight on the latent term. (a) Cloud mIoU, CloudSEN12 test (read repeatedly during development; validation agrees); dotted lines: OlmoEarth v1.2 Base and Galileo-B, same probe. (b) Loss weights switched at 15B, one run per arm; lines end at annealed 31.25B checkpoints, except latent-heavy (30B, mid-decay). (c) Scores at the start and end of the same stretch of training on six tasks, each panel spanning 15 points: ALP at 10B and 20B; latent-heavy at the cooldown checkpoints branched there (12.5B and 25B). Development panels, except EuroSAT and So2Sat (test). Canopy-height error, a dense regression task, did not worsen.

never warned: most of the cloud loss comes between 10B and 20B tokens while land-cover mAP is flat, and land-cover mAP on test and validation and mean GEO-Bench accuracy all peak at 40B, each run's worst checkpoint for clouds.

Feature rank rises; recoverable detail falls. Between the annealed 12.5B and 25B checkpoints of each seed, token-level RankMe rises from 539 to 673 and from 545 to 669; the two late checkpoints have the highest rank of the 37 encoders we measured, above OlmoEarth v1 Large. Over the same interval, the *anchor-recovery* R^2, how well a ridge probe on the features recovers the frozen 40 m anchors, falls from 0.72 to 0.56 and from 0.72 to 0.53, while under ALP it stays between 0.76 and 0.79 through 18.75B, the last ALP checkpoint measured (Figure 6a,b; development panel). Rank rises with tokens in every run, so it tracks training length rather than map quality. Probes on intermediate layers show where the detail goes: in the late checkpoint, layer 3 of 12 still scores 62.2 mIoU while the output layer scores 54.4 (two head seeds; Figure 6c). This fits our hypothesis, since this recipe's teacher targets average its top four blocks, which drift toward scene-level features.

4 Our solution: anchored latent prediction

ALP keeps the encoder, and through its moving average the latent targets, tied to the pixels. What it adds to the latent targets of I-JEPA and Galileo and the frozen pixel projections of OlmoEarth is

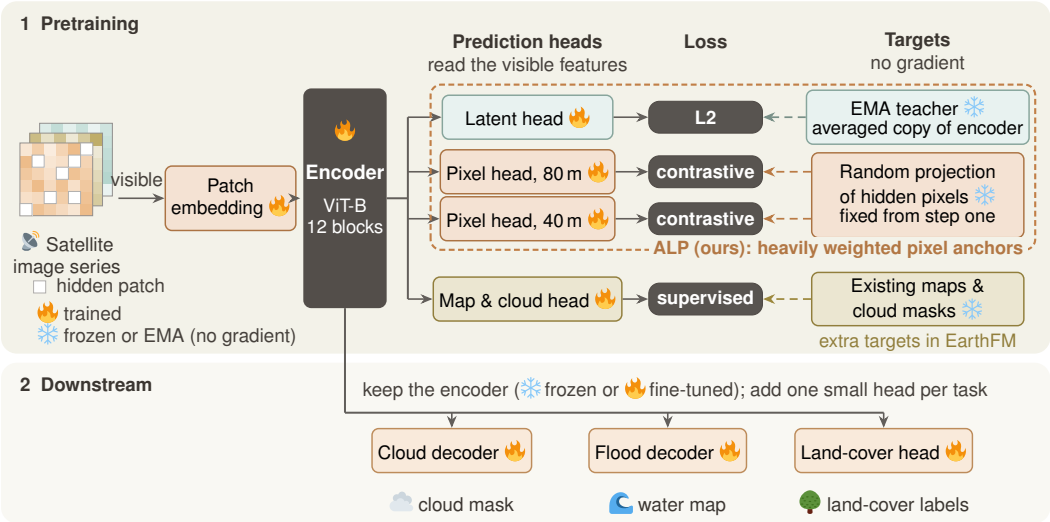

Figure 3: Anchored latent prediction (ALP), our pretraining objective, as used in EarthFM. (1) The encoder sees only the visible patches. Each loss compares a head's prediction (solid arrow) with its target (dashed arrow): an EMA teacher's features, which drift as the encoder learns, or fixed random projections of the hidden pixels at 80 m and 40 m, which never change. ALP (dashed box) predicts at hidden patches; the map and cloud head, part of EarthFM's recipe but not of ALP, also queries visible ones. Not shown: a small CNN inside each patch that feeds the 40 m head, and two pooled heads (class presence and view alignment). (2) After pretraining, only the encoder is kept, and each task trains its own small head.

how the anchors are used: they carry at least about as much of the encoder's gradient as the latent term, and they are predicted at twice the token resolution at no encoder cost. We call runs trained with ALP's weights from the start ALP runs; every run except one ablation also carries external map and cloud-model targets.

Encoder. A training example covers one location with several sensors (Sentinel-2, Sentinel-1, Landsat, elevation) and up to 12 dates, all on a 10 m grid. We draw two views of it. A per-sensor MLP embeds each 8×8-pixel patch (80 m at 10 m resolution) and adds an embedding of position, time, and sensor. A 12-block transformer (width 768, 12 heads, spatial and temporal rotary encodings; 85M parameters) processes the visible tokens of one view, $H = f_\theta(X_V)$. Hidden pixels never enter the encoder. With the patch embeddings, the transferred encoder has 91M parameters; pretraining also trains about 86M parameters of heads that are discarded afterwards, and runs a forward pass of the EMA teacher.

Prediction heads: detail without encoder cost. Each target family has its own head: two cross-attention blocks whose queries contain only the metadata of a target location and whose keys are the visible features H of the same view. Such metadata-only queries follow Galileo's predictor and I-JEPA's mask tokens (Tseng et al., 2025; Assran et al., 2023); ALP's change is to place some of them below the token grid. Because a query carries no pixels, it can ask for a target at any location or scale without adding encoder tokens, so ALP supervises at 40 m while the encoder, and its inference cost, stay at the 80 m grid of an 8×8-patch ViT-B.

Latent targets: context that drifts. The teacher is an exponential moving average (EMA) of the encoder, $\bar{\theta} \leftarrow m_t\bar{\theta} + (1 - m_t)\theta$, with m_t rising from 0.996 to 1 on a cosine schedule. It sees both visible and hidden patches, and its layer-normalized final-block features are the targets $y_j = \mathrm{sg}[\mathrm{LN}\{f_{\bar{\theta}}(X_V \cup X_M)_j\}]$. The latent head is trained with squared error,

$$\mathcal{L}_{\mathrm{lat}} = \frac{1}{|M|\,d} \sum_{j \in M} \left\| p_{\mathrm{lat}}(q_j, H) - y_j \right\|_2^2, \tag{1}$$

where M indexes hidden targets and $d = 768$; I-JEPA uses the same squared-error objective (Assran et al., 2023). These targets carry context from the whole view, but they change as the encoder changes.

Anchor targets: pixels that cannot drift. For each sensor, a randomly initialized extractor ϕ maps the calibrated pixels and their validity mask through a per-pixel linear layer and ReLU, then projects the patch to a normalized 256-dimensional vector. Its weights are frozen for all of pretraining, so its targets always describe the same pixels in the same way. A second extractor of the same form produces one target for each 4×4-pixel (40 m) quarter of a patch. The 80 m and 40 m pixel heads must pick out the correct target $\phi_s(x_j)$ among up to 4,096 targets from the same sensor, $\mathcal{N}_j$:

$$\mathcal{L}_s = -\frac{1}{|Q_s|} \sum_{j \in Q_s} \log \frac{\exp(u_j^\top \phi_s(x_j)/\tau)}{\exp(u_j^\top \phi_s(x_j)/\tau) + \sum_{k \in \mathcal{N}_j} \exp(u_j^\top \phi_s(x_k)/\tau)}, \qquad s \in \{\text{patch}, \text{sub}\}, \tag{2}$$

with $\tau = 0.1$; copies of the target patch itself (another processing level of the same acquisition, or the overlapping view) are excluded from the negatives. A small CNN (two 3×3 convolutions, 64 channels) runs inside each visible patch and gives the 40 m pixel head extra keys; it never crosses a patch border and is discarded after pretraining.

Anchor weight: a large share, not an auxiliary term. The objective is

$$\mathcal{L} = \lambda_p \mathcal{L}_{\text{patch}} + \lambda_s \mathcal{L}_{\text{sub}} + \lambda_l \mathcal{L}_{\text{lat}} + \mathcal{L}_{\text{ext}}. \tag{3}$$

ALP sets $(\lambda_p, \lambda_s, \lambda_l) = (1, 0.5, 4)$. Because loss terms have different scales, the weights alone do not say how much each term shapes the encoder. Measured on one fixed diagnostic batch per run at checkpoints from 1B to 40B tokens, the two anchor terms together send the encoder 0.9 to 2.1 times as much gradient as the latent term under ALP, and 0.15 to 0.71 times as much under the latent-heavy recipe (two seeds each). Levels across runs are indicative, but the ranges do not overlap.

External targets (EarthFM). A map head predicts WorldCover classes, WorldCereal crop layers, and canopy height, and a cloud head predicts the fractions of clear, thin-cloud, thick-cloud, and shadow pixels from masks produced by the SEnSeIv2 cloud model (Francis, 2024). These queries sit at visible or hidden locations. Two pooled heads predict which land-cover classes are present and align the two views. $\mathcal{L}_{\text{ext}}$ collects these weighted terms (Appendix A). We do not count them as part of ALP, but every run except the external-head ablation carries them, and Table 1 measures what they add.

5 Does ALP fix the problem?

ALP sustains stronger dense features through longer training: switching a declining run to ALP's loss weights halts its decline, and the anchors contribute local detail from early in training. Each condition below uses one pretraining run; the late-training results also show how optimization complements anchoring.

Switching a declining run to ALP's loss weights. Switching the latent-heavy run to ALP's loss weights at 15B halts its decline: it holds between 62.2 and 62.6 mIoU through 31.25B tokens, 3.65 points above the unswitched run at 20B, though below its own 10B peak of 63.5 (Figure 2b). The reverse switch costs the ALP run 1.90 points. The switch changes all five loss weights, including doubling or halving the cloud weight. An interim decomposition (one run per arm, read at 17.5B; Appendix E) separates them: restoring only the anchor weights recovers 43% of the switch's validation cloud gain (32% on a validation subset), restoring only the cloud weight 31% (none on the subset), and the remainder comes from the latent and view-alignment weights or from interactions.

Removing the anchors. Removing the anchor targets from an early latent-heavy variant costs 4.2 to 6.6 cloud mIoU points and 9.5 to 16.3 thin-cloud points at every checkpoint from 2B to 6.4B tokens in one run, while land cover moves by less than 2 points (Table 1). The cloud target stays in place, and the local CNN reaches the encoder only through the sub-patch loss, so the loss comes from the anchor pathway. The anchors here carried little weight, showing that even weakly weighted anchors contribute substantially to dense features; ALP's heavier weighting and its 40 m anchors were not ablated separately.

Table 1: Pixel anchors protect fine detail; the land-cover score barely registers the difference. Each block compares our recipe (shaded) with one change to it; scores are frozen linear probes with two head seeds, one pretraining run per row. Bold: our recipe scores higher on a cloud metric. Every bold gap exceeds the noise threshold except the ALP run against its switch (1.9 and 2.6 points) and the thin-cloud gap for the external targets (3.9); every land-cover gap is within noise. OlmoEarth v1.2 Base scores 62.8 cloud and 35.2 thin-cloud IoU under this probe on test.

	Cloud mIoU ↑	Thin-cloud IoU ↑	Land-cover mAP ↑
Remove the anchor targets (early latent-heavy variant; validation subset, 6.4B tokens)			
no anchors	54.9	20.5	61.6
with anchors	**61.4**	**36.7**	62.3
Switch the loss weights at 15B tokens (official test, read at 20B)			
latent-heavy run, kept its weights	58.6	25.4	65.7
latent-heavy run → ALP weights	**62.2**	**31.1**	66.2
ALP run, kept its weights	**67.0**	**40.3**	66.3
ALP run → latent-heavy weights	65.1	37.7	66.1
Remove the external map and cloud targets (official test, 8B tokens)			
no external targets	63.1	34.9	63.7
ALP with external targets	**65.7**	**38.8**	63.9

Where the detail lives. In runs trained with ALP's weights from the start, the best layer stays in the upper half: in EarthFM, layer 12 scores 66.5 against 61.2 at layer 3, and the ALP run at 30B peaks at layers 6 to 9, the opposite of the late latent-heavy checkpoint (Figure 6c), though the recipes differ in more than the anchors. The anchor-recovery R^2 is a free warning sign: it fell by 0.16 and 0.19 in the declining runs. It is not a stopping rule, since it can also fall slightly while accuracy improves.

Sustaining the gains through later training. ALP maintains a substantial advantage over the latent-heavy runs through 30B tokens, while its cloud mIoU moves from a peak of 66.95 at 20B to 64.95 at 30B. The decline coincides with saturated clipping: after 19B tokens at its constant peak learning rate, every update hit the clipping limit. From one shared 20B checkpoint, halving the learning rate improved validation cloud mIoU by 0.65 points, while raising the anchor weights further added 0.18 (single continuations; Appendix E). These results motivate combining ALP's preservation of local detail with learning-rate schedules and checkpoint selection suited to the downstream task.

6 Putting ALP to work: EarthFM against open encoders

EarthFM combines ALP's pixel anchoring with map and cloud-model supervision; these comparisons evaluate the resulting full recipe. Without any external heads, an 8B-token ALP run already matches OlmoEarth v1.2 Base under the frozen linear cloud probe (63.08 against 62.76, within noise; one run; official test, read repeatedly). Under the adapted protocols of Table 2, removing the external targets at 8B costs 1.0 cloud mIoU (UNet decoder), 0.6 land-cover mAP (fine-tuning), and 0.5 flood IoU, the last within seed noise (validation).

From the ALP run to EarthFM. EarthFM continues the ALP run from 20B to 31.25B tokens with even heavier anchors, $(\lambda_p, \lambda_s, \lambda_l) = (1.5, 1, 2)$, and a learning rate that decays from 25B; its cloud score held while the unbranched run fell and reached 67.31 at 31.25B (frozen linear head), with a 0.18-point validation gain from heavier anchors in the matched-learning-rate comparison. EarthFM-short branches at 15B, decays to 18.75B tokens, and adds a view-statistics head (not part of ALP) worth $+0.91$ cloud points in one test comparison (Appendix E).

Compute. We report pretraining compute in H100-equivalent GPU-hours: GPU-hours times the device's dense BF16 peak, divided by the H100 SXM peak. EarthFM took 243 B200 GPU-hours and EarthFM-short 148, counting every ancestor checkpoint, or 553 and 338 H100-equivalent hours. OlmoEarth v1.2 Base reports 1,012 H100 GPU-hours (Tseng et al., 2026), so EarthFM and EarthFM-short use 1.8 and 3.0 times less compute (55% and 33% of it; 65% and 40% if OlmoEarth used the slower H100 NVL). This proxy measures allocated hardware, not executed FLOPs, and excludes data preparation and downstream adaptation for every model (Appendix B).

Table 2: Official test results: EarthFM is on par with or better than OlmoEarth v1.2 Base on six of eight tasks, with 1.8 to 3 times less pretraining compute. Scores in percent, higher is better; best and second among the models evaluated on each row ("–": not evaluated). GPU-h: H100-equivalent pretraining compute of the EarthFM checkpoint used on that row. Bold: EarthFM above OlmoEarth v1.2 Base. Δ: ▲ ahead, ≈ within noise (paired 95% interval spans zero or gap inside the task's pre-declared noise bar), ▼ behind. On clouds, CROMA-B (25 GPU-h) scores 68.41 and DeCUR (38) 63.14. Against Galileo-B, trained with similar compute, EarthFM scores higher on three of four shared tasks. [§]Three seeds and a smaller search.

	EarthFM (ours)	GPU-h	Δ vs v1.2 Base	OlmoEarth v1.2 Base	TerraMind v1 Large	Galileo Base	OlmoEarth v1 Large
Pretraining GPU-h ↓	335–553			1,012	2,422	573	11,915
Dense mapping							
Clouds	**69.52**	338	+1.15 ▲	68.37	70.80	66.97	–
Floods (radar)	**68.22**	553	+2.38 ≈	65.84	–	–	–
Burn scars	81.66	335	−1.42 ≈	83.08	–	83.30	82.25
Cashew (1% labels)	29.48	553	−0.28 ≈	29.75	28.45	23.00	–
Scene classification							
Land cover	72.74	553	−0.43 ≈	73.17	74.39	70.37[§]	–
Brick kilns	97.70	553	−0.25 ≈	97.95	–	–	–
EuroSAT	95.20	335	−2.35 ▼	97.55	–	–	98.40
So2Sat	57.35	335	−9.99 ▼	67.34	–	–	66.99

Benchmarks. We compare on every standard benchmark with an official test split that we scored for both EarthFM and OlmoEarth v1.2 Base under a symmetric, validation-selected protocol: eight public tasks, four dense and four scene-level, with cashew read last. Clouds (CloudSEN12+, Aybar et al., 2024), land cover (BigEarthNet v2, Clasen et al., 2025), and burn scars (Jakubik et al., 2023; Phillips et al., 2023) use GEO-Bench-2 test partitions (Simumba et al., 2025); we fixed the first two before any transfer result and pre-registered burn scars. EuroSAT, So2Sat, brick kilns, and cashew come from GEO-Bench (Lacoste et al., 2023), whose tasks OlmoEarth and Galileo also report (Herzog et al., 2026; Tseng et al., 2025); of the other GEO-Bench tasks they report, we use BigEarthNet v2 in place of m-BigEarthNet, exclude SA-crop because of an input-unit error on our side (EarthFM 36.5 against 39.9), and did not run ForestNet. Floods use the hand-labelled Sen1Floods11 test split (Bonafilia et al., 2020), as in OlmoEarth, Galileo, and PANGAEA (Marsocci et al., 2026). We measured every OlmoEarth number ourselves.

Evaluation. We compare against OlmoEarth v1.2 Base and v1 Large (Herzog et al., 2026; Tseng et al., 2026), TerraMind-L (Jakubik et al., 2025), Galileo-B (Tseng et al., 2025), CROMA-B (Fuller et al., 2023), and DeCUR (Wang et al., 2024) at their native input resolutions, except 4×4-pixel patches for EarthFM and OlmoEarth on land cover and each encoder's own validation-selected readout on EuroSAT and So2Sat (for EarthFM on So2Sat, with Sentinel-1). Every model gets the same learning-rate search except Galileo-B on land cover; settings are chosen on validation, though earlier checkpoints were also scored on these test splits during development. Intervals are paired 95% bootstrap intervals over test chips, seeds, or flood events (Appendix C).

The count. Against OlmoEarth v1.2 Base, EarthFM is on par or better on six of the eight official test sets: ahead on clouds, within noise on floods, burn scars, cashew, land cover, and brick kilns, and behind on EuroSAT and So2Sat (Figure 1, Table 2). *On par* means no resolved difference: the paired 95% interval spans zero, or the gap is inside the task's pre-declared noise bar. It does not show equivalence, and on point estimates EarthFM is higher on two of the eight. By task type, EarthFM is on par or better on all four dense tasks and on two of the four scene tasks. The verdicts use three checkpoints of one run: EarthFM-short (clouds) and an 18.75B-token checkpoint (burn scars), both declared in advance; the same checkpoint on EuroSAT and So2Sat, replacing pre-registered ones; and EarthFM elsewhere, chosen for land cover partly on an earlier test sweep.

Clouds. With frozen encoders, cloud segmentation tests how readily pretrained features support dense mapping. Using a UNet decoder, EarthFM-short reaches 69.52 test mIoU and EarthFM 69.50, against 68.37 for OlmoEarth v1.2 Base: +1.15 points (95% CI 0.53 to 1.79) at one-third

of OlmoEarth's compute, and $+1.13$ (0.54 to 1.75) at 55%. Their similar scores favor choosing training duration for the downstream task. Both comparisons are ties on validation (-0.20 and $+0.23$; Appendix C). EarthFM-short also leads CROMA-B, Galileo-B, and DeCUR; only TerraMind-L scores higher, with 7.2 times its compute (Figure 5). EarthFM combines pixel anchoring with cloud supervision from SEnSeIv2, whose audited checkpoint scores 70.08 on this test set; removing external heads costs 2.5 to 3.1 points under the frozen linear probe (Table 1). Finer 4×4-pixel patches at evaluation increase the lead to $+1.49$ (CI 0.77 to 2.24), at four times the inference tokens, suggesting that inference resolution offers a further way to exploit spatial detail.

Within noise. Adaptation and evaluation diversity shape the gains these benchmarks can resolve. On radar floods, EarthFM leads by 2.38 water-IoU points (68.22 versus 65.84; CI -0.38 to $+6.67$) and in four of five seeds, but the test covers ten events and the gains are concentrated in two. An earlier validation comparison found little separation from training from scratch (Appendix C), suggesting a substantial role for downstream adaptation here. On burn scars and brick kilns, the gaps (-1.42 and -0.25) fall inside the pre-declared noise bars (4.6 and 1.2); for burn scars, seed variation exceeds the spread across models. Land cover is also within noise (-0.43, CI -1.51 to $+0.74$; behind on validation), with TerraMind-L leading by 1.65. These tasks thus support competitive transfer at lower pretraining compute (Figure 5).

Behind. EarthFM trails OlmoEarth on EuroSAT (-2.35) and So2Sat (-9.99), which classify images containing just 8×8 and 4×4 of our 80 m tokens, compared with 32×32 during pretraining. This gives the encoder substantially less spatial context at transfer, a plausible mismatch between pretraining and downstream inputs. OlmoEarth pretrains on varied input sizes, and random crop sizes improved So2Sat by 2.9 to 6.4 points in the latent-heavy recipe. Input-size diversity is therefore a promising direction for improving small-scene transfer alongside ALP's preservation of local detail.

Few labels. Cashew mapping shows why label efficiency should be considered together with the adaptation budget. With 13 labeled images and the pre-declared validation choice of 80 or 400 updates, EarthFM scores 29.48 test mIoU, level with OlmoEarth's 29.75 (-0.28, CI -1.00 to $+0.58$), and above TerraMind-L (28.45) and Galileo-B (23.00). At 80 updates it leads OlmoEarth by 1.47 (CI $+0.15$ to $+3.32$; a tie on validation), while longer matched fine-tuning favors OlmoEarth (-1.34 at 1,600 updates). This pattern is consistent with features that adapt quickly: early transfer performance and the gains from additional optimization capture different strengths. The test spans two sites, with likely pretraining imagery covering 51% for EarthFM and none for OlmoEarth, so geographic exposure also matters when interpreting this comparison (Appendix D.1).

7 Limitations

Each test in Table 1 was run once, no experiment isolates ALP's heavier weighting or its 40 m anchors, and we measured the decline only with frozen linear probes. That latent targets drifting from the pixels cause it is a hypothesis, not a tested cause. The transfer comparisons test ALP together with EarthFM's map and cloud-model targets, from one pretraining run; the one resolved lead, on clouds, is a tie on validation, uses a cloud model's labels, and comes from a test split scored during development. We study one architecture at one scale, do not compare ALP with Gram anchoring, and convert compute between devices by peak throughput.

8 Conclusion

Scene benchmarks and feature rank can improve while a model's maps get worse, plausibly because latent targets drift from the pixels when nothing heavy holds them there. Anchored latent prediction adds heavily weighted, frozen pixel anchors at no encoder cost; it halted such a decline when switched in and delayed it from the start, and EarthFM is on par with or better than OlmoEarth v1.2 Base on six of eight test sets with 1.8 to 3 times less compute. We hope ALP and the anchor-recovery R^2 make local detail a measured and protected target.

Reproducibility statement

Section 4 and Appendix A specify the architecture, objective weights, and checkpoint lineage; Appendix C gives every evaluation protocol, selection rule, and interval definition. We will release code, configurations, checkpoints, and per-seed results for every number in the paper.

Ethics statement

Errors in satellite maps can affect land-use, agricultural, and disaster-response decisions. Performance varies across regions and land-cover types, and the model can inherit errors and coverage gaps from its map and cloud-model targets. We report supervision sources, evaluation geography, and uncertainty to help users judge these limits.

AI use statement

Generative AI assistants, including Claude Opus 5.5, GPT Astra, and OpenAI Codex, supported literature discovery, experiment planning, implementation and debugging, analysis, and drafting and polishing the text and figures of this manuscript. The authors reviewed and edited all AI-generated content, checked the code, verified all numerical claims against recorded results, and take full responsibility for the final manuscript.

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 (CVPR)*, pp. 15619–15629, 2023.

Cesar Aybar, Lesly Bautista, David Montero, Julio Contreras, Daryl Ayala, Fernando Prudencio, Jhomira Loja, Luis Ysuhuaylas, Fernando Herrera, Karen Gonzales, et al. CloudSEN12+: The largest dataset of expert-labeled pixels for cloud and cloud shadow detection in Sentinel-2. *Data in Brief*, 56:110852, 2024. doi: 10.1016/j.dib.2024.110852.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 835–845, 2020. doi: 10.1109/CVPRW50498.2020.00113.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reBEN: Refined BigEarthNet dataset for remote sensing image analysis. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1264–1268, 2025.

Siran Dai, Qianqian Xu, Peisong Wen, Yang Liu, and Qingming Huang. Exploring Structural Degradation in Dense Representations for Self-Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alistair Francis. Sensor independent cloud and shadow masking with partial labels and multimodal inputs. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024. doi: 10.1109/TGRS.2024.3391625.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5506–5538, 2023.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974, 2023.

Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, Josh Hansen, Andrew Howe, Patrick Alan Johnson, Mark Otterlee, Ted Schmitt, Hunter Pitelka, Stephen Daspit, Rachel Ratner, Christopher Wilhelm, Sebastian Wood, Mike Jacobi, Hannah Kerner, Evan Shelhamer, Ali Farhadi, Ranjay Krishna, and Patrick Beukema. OlmoEarth: Stable Latent Image Modeling for Multimodal Earth Observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34806–34817, 2026.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence, 2023.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. TerraMind: Large-Scale Generative Multimodality for Earth Observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jakubik_TerraMind_Large-Scale_Generative_Multimodality_for_Earth_Observation_ICCV_2025_paper.html.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 51080–51093, 2023.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, 14(1):245–285, 2026. doi: 10.1109/MGRS.2025.3628194.

NVIDIA. NVIDIA A100 Tensor Core GPU specifications, 2026a. URL https://www.nvidia.com/en-us/data-center/a100/. Accessed September 26, 2026. BF16 dense throughput 312 TFLOPS.

NVIDIA. NVIDIA H100 Tensor Core GPU specifications, 2026b. URL https://www.nvidia.com/en-us/data-center/h100/. Accessed September 26, 2026. SXM and NVL specifications distinguished.

NVIDIA. NVIDIA HGX: B200 platform specifications, 2026c. URL https://www.nvidia.com/en-us/data-center/hgx/. Accessed September 26, 2026. BF16 dense throughput is half the listed sparse throughput.

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

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longepe, Xiao Xiang Zhu, Hannah Kerner, Juan Bernabe-Moreno, and Alexandre Lacoste. GEO-Bench-2: From Performance to Capability, Rethinking Evaluation in Geospatial AI, 2025. URL https://arxiv.org/abs/2511.15658.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning Global &

Local Features of Many Remote Sensing Modalities. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60280–60300, 2025.

Gabriel Tseng, Yawen Zhang, Favyen Bastani, Henry Herzog, Joseph Redmon, Hadrien Sablon, Piper Wolters, Ando Shah, Patrick Alan Johnson, Christopher Wilhelm, and Patrick Beukema. OlmoEarth v1.2: A more efficient family of OlmoEarth models, 2026. URL https://arxiv.org/abs/2605.20804.

John Waithaka, Gustave Bwirayesu, and Moise Busogi. Cross-Scale Pretraining: Enhancing Self-Supervised Learning for Low-Resolution Satellite Imagery for Semantic Segmentation, 2026. URL https://arxiv.org/abs/2601.12964.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pp. 286–303, 2024. doi: 10.1007/978-3-031-73397-0_17.

A RECIPES, CHECKPOINTS, AND NAMES

Names. The paper names models by role. Table 3 maps each name to its internal code, so every number can be traced to a registered checkpoint.

Table 3: Paper names, internal codes, and lineage. All EarthFM checkpoints descend from one ALP run (seed 42) unless marked. ALP names the objective; the ALP runs (V3-B) are the full recipe that uses it, including external heads, and differ from the latent-heavy recipe in several settings besides the loss weights (Table 4).

Paper name	Code	Tokens	Lineage
EarthFM	X7	31.25B	ALP to 20B, then anchor weights $(1.5, 1, 2)$
EarthFM-short	X8	18.75B	ALP to 15B, then learning-rate decay with the view-statistics head
EarthFM, 18.75B tokens, no view-statistics head	cd15b	18.75B	ALP to 15B, then learning-rate decay
EarthFM, 12B tokens, seed 43	S43	12B	independent ALP run, seed 43
ALP recipe (ALP)	V3-B	–	seeds 42 and 43
Latent-heavy recipe	V4-B	–	seeds 42 and 43, 40B schedule
Anchor ablation	V4-A pixel-free	2–6.4B	early latent-heavy variant without anchor targets
External-head ablation	X3	2–8B	ALP (seed 42, 8B schedule) without map, cloud, presence heads
Weight switches	X4, X5	15–31.25B	switch all five main weights at 15B
Late continuations	x9, X10, X11, X12	20–31.25B	from the shared ALP 20B state

Table 4: Objective coefficients and settings. Map and presence weights are 0.15 and 0.20 throughout. The latent-heavy recipe also averages the teacher's top four blocks for its targets (others use the final block), uses a four-block latent head, samples random crop extents, predicts 4,096 targets per example, samples the three sources at 0.30/0.20/0.50, and decays its learning rate over the last 35% of training (others 20%). Loss terms have different scales, so coefficients do not set gradient shares.

Recipe	Patch	Sub-patch	Latent	Cloud	View align	View stats	Peak LR
ALP (V3-B)	1.0	0.5	4.0	0.30	0.02	0	$2 \cdot 10^{-4}$
Latent-heavy (V4-B)	0.3	0.2	6.0	0.15	0.05	0	$2 \cdot 10^{-4}$
EarthFM (X7)	1.5	1.0	2.0	0.30	0.02	0	$2 \cdot 10^{-4}$
EarthFM-short (X8)	1.0	0.5	4.0	0.30	0.02	0.1	$2 \cdot 10^{-4}$
x9	1.0	0.5	4.0	0.30	0.02	0	10^{-4}
X10	1.5	1.0	4.0	0.30	0.02	0	$2 \cdot 10^{-4}$
X11	1.5	1.0	4.0	0.30	0.02	0	10^{-4}
X12	1.5	1.0	4.0	0.30	0.02	0.1	10^{-4}

Optimization. AdamW with $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, gradient clipping at 1, and 256 examples (512 views) per step. Schedules warm up over 2% of training and decay to 1% of the peak learning rate. The EMA schedule is tied to each run's declared token budget, so branches with different budgets have different teacher schedules at equal token counts.

Data. The ALP recipe draws from SSL4EO-S12, TerraMesh, and OlmoEarth tiles in the mixture 0.25/0.40/0.35; the latent-heavy recipe uses 0.30/0.20/0.50. The approved release holds 6.6M source keys (14.9 TB), but sampling is uneven: in a trace of a latent-heavy run, 94% of tokens came from about 500,000 keys. Each example yields two 256×256-pixel views with up to 2,048 visible tokens each. Of the 85% of views that are not single-modality, masks hide spatial blocks (40%), whole dates (30%), whole sensors (20%), or later dates (10%); the other 15% show only radar or only optical data.

External targets. Maps come from OlmoEarth tiles: WorldCover (11-way classification of each 80 m cell), WorldCereal (8 crop layers), and canopy height (smoothed L1 on the standardized cell mean). Cloud targets come from SEnSeIv2 masks shipped with SSL4EO-S12 v1.1 and TerraMesh, reduced to clear, thin cloud, thick cloud, and shadow. Half of the cloud queries sit at visible patches; examples with at least 30% cloud are shown three times; cloud-fraction metadata is removed from visible tokens so the head cannot read the answer. The view-statistics head (EarthFM-short, X12) predicts per-band means and standard deviations, WorldCover fractions, and cloud fractions over all of a view's pixels.

B Compute accounting

We convert GPU-hours H_g on device g to H100-equivalent hours as $\widetilde{H}_g = H_g\, P_g / P_{\mathrm{H100}}$, where P_g is the specified dense BF16 peak: 2.25 PFLOP/s for B200, 0.9895 for H100 SXM, and 0.312 for A100 (NVIDIA, 2026c;b;a). This measures the hardware capacity allocated to pretraining, not executed FLOPs or runtime. EarthFM costs include every ancestor checkpoint, reconstructed from logs, receipts, and checkpoint timestamps; ranges reflect two missing start times, not statistical uncertainty.

Table 5: Pretraining compute. OlmoEarth v1.2 Base and Galileo-B report H100 hours without the variant; the last column assumes both used the slower H100 NVL instead of SXM. *Derived from 32 A100 GPUs for 10 days.

Model	Reported GPU-h	H100-equiv.	% of OlmoEarth-B	% if NVL
EarthFM-short (X8, 18.75B tokens)	147–148 B200	335–338	33	39–40
EarthFM (X7, 31.25B tokens)	241–243 B200	548–553	54–55	64–65
EarthFM (cd15b, 18.75B tokens)	146–147 B200	332–335	33	39
EarthFM (S43, 12B tokens, seed 43)	96 B200	219	22	26
OlmoEarth v1.2 Base (Tseng et al., 2026)	1,012 H100	1,012	100	100
OlmoEarth v1 Large (Herzog et al., 2026)	5,240 B200	11,915	1,177	1,394
TerraMind-L (Jakubik et al., 2025)	7,680 A100*	2,422	239	283
Galileo-B (Tseng et al., 2025)	573 H100	573	57	57
CROMA-B (Fuller et al., 2023)	80 A100	25	2.5	3.0
DeCUR (Wang et al., 2024)	120 A100	38	3.7	4.4

What the comparison excludes. Every row excludes data preparation, teacher or tokenizer training, and downstream adaptation. Our full research program used about 3,841 B200 GPU-hours of pretraining experiments, 889 of downstream work, and 62 of early screens; these are research costs, not the cost of a checkpoint. EarthFM and OlmoEarth v1.2 Base have the same analytic encoder cost at inference, about 155.8 GFLOP per 224×224 tile at 8×8 patches.

Training exposure. EarthFM and EarthFM-short saw 11.6M and 7.0M training examples, counting repeats; OlmoEarth v1.2's batch size and length imply 341.6M. The examples differ in size (about 2,690 versus 420 encoder tokens), so we do not convert this into a data-efficiency ratio.

C Evaluation protocols

Clouds (CloudSEN12). A UNet decoder is trained on frozen features tapped at one-quarter, one-half, three-quarters, and full depth (blocks 3, 6, 9, and 12 for 12-block encoders), with the same six-learning-rate search and epoch selection for every model. The split has 4,000 training, 535 validation, and 975 test chips; decoder seeds are 205 and 913. Validation / test mIoU: TerraMind-L 73.00 / 70.80, EarthFM 71.41 / 69.50, EarthFM-short 70.98 / 69.52, OlmoEarth-B 71.18 / 68.37, CROMA-B 70.47 / 68.41, Galileo-B 69.04 / 66.97, DeCUR 65.09 / 63.14. Intervals resample test chips with decoders fixed. EarthFM-short's paired test differences are $+1.11$ [0.54, 1.68] against CROMA-B, $+2.55$ [1.84, 3.23] against Galileo-B, $+6.38$ [5.62, 7.15] against DeCUR, and -1.28 [-1.94, -0.68] against TerraMind-L. The gains over OlmoEarth remain positive when the 12 test chips imaged on the same date as, and within 10 km of, a pretraining view with cloud targets are excluded. In a separate follow-up, both EarthFM and OlmoEarth select 4×4-pixel patches on validation (72.30 and 71.62) and score 70.08 and 68.59 on test.

Cloud supervision and test history. The cloud targets come from SEnSeIv2 masks. The audited candidate checkpoint was trained on 8,490 CloudSEN12 v1 chips and other data and scores 71.22 validation and 70.08 test mIoU; we could not confirm that this exact checkpoint produced every stored mask. No scored chip lies within 2 km of a pretraining sample. During development, earlier checkpoints of this lineage, including EarthFM and EarthFM-short, were evaluated with linear heads on the CloudSEN12 test split. The decoder results in Section 6 fix every setting on validation, but the test split is not untouched.

Land cover (BigEarthNet). Full fine-tuning with 4×4-pixel patches for EarthFM and OlmoEarth (TerraMind-L at its native 16×16 patches on 224×224 inputs) for 2,000 updates, learning rate 3×10^{-4}, layer-wise decay 0.9, six seeds; each model's EMA weights and training step are chosen on validation. Validation / test mAP: EarthFM 78.45 / 72.74, OlmoEarth-B 79.11 / 73.17, TerraMind-L 80.20 / 74.39. Hierarchical intervals resample seeds and chips. On validation, EarthFM trails OlmoEarth by 0.65 [0.03, 1.29]. Two sensitivities also favor OlmoEarth: excluding the two rarest classes (-0.69 [-1.12, -0.23]) and six-member prediction ensembles (73.10 versus 73.92). Galileo-B had a three-seed, smaller search at 8×8 patches (70.37 test); EarthFM under that same protocol scores 71.64, a matched gap of $+1.27$ ($+2.27$ on validation, chip CI 1.52 to 3.08).

Radar floods (Sen1Floods11). Full fine-tuning of radar-only inputs with validation-selected learning rates (3×10^{-5} for EarthFM, 3×10^{-4} for OlmoEarth) and five fresh seeds; water IoU pools confusion matrices over seeds on 89 test chips from ten events (one published chip has no observation). The gain is concentrated in two events (Pakistan $+17.9$, Somalia $+10.9$). A stronger stored OlmoEarth frozen-encoder reference scores 67.09 ($+1.13$ [-2.11, $+5.58$] for EarthFM). On validation the difference is $+0.73$ [-0.72, $+2.27$]. Under an earlier fine-tuning recipe, a randomly initialized encoder reached 67.02 validation IoU against EarthFM's 67.77, so pretraining helps little on this task.

Burn scars, EuroSAT, So2Sat, brick kilns. Burn scars (HLS), EuroSAT (band-corrected), and So2Sat use frozen 1×1 linear heads with each model's best setting from a shared grid and two head seeds; burn scars has 120 test chips, and its five-model spread (1.64 points) is smaller than one model's seed spread (2.31). The pre-declared EarthFM checkpoint for burn scars is the 18.75B-token checkpoint (81.66); the independent 12B-token run scores 82.58. Brick kilns fine-tune with a shared learning-rate grid and two seeds; published fine-tuning results on this task range from 98.1 to 99.0, above both models here.

Geographic scope. We audited exclusions between the pretraining release and CloudSEN12, BigEarthNet, Sen1Floods11, and PASTIS. Geographic overlap is known for burn scars and brick kilns. For cashew, a replay of EarthFM's training stream suggests imagery covering about 56% of the average validation chip; OlmoEarth's audited pretraining index has no overlap with the benchmark.

D Additional results

Figure 4 shows the per-run scores behind the land-cover, flood, and cashew comparisons, Table 6 the cloud results with a frozen linear head, and Figure 5 every task against pretraining compute.

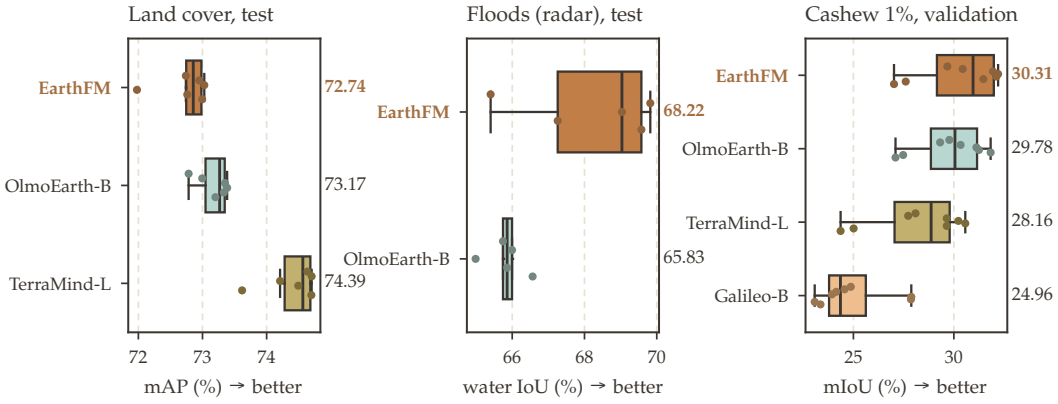

Figure 4: Per-run scores behind three comparisons: EarthFM is 0.4 points behind OlmoEarth-B on land cover and 2.4 ahead on floods (both within noise), and level on cashew at 80 updates (validation). Each dot is one fine-tuning seed or label draw; boxes show quartiles and the median, whiskers 1.5 times the interquartile range, and the number on the right is the mean. With five to eight runs, the boxes describe spread rather than estimate a distribution; the paired intervals in Section 6 account for the shared seeds.

Table 6: Cloud segmentation with a frozen linear head (CloudSEN12 test, two head seeds). This protocol was evaluated on many checkpoints during development, so we report it as supporting evidence. It is the only transfer result replicated across two independent pretraining runs.

Encoder	Test mIoU ↑	vs OlmoEarth-B
EarthFM-short	67.90	+5.14
EarthFM	67.31	+4.55
EarthFM, 18.75B tokens, no view-statistics head	66.99	+4.23
EarthFM, 12B tokens, seed 43	66.77	+4.02
OlmoEarth v1.2 Base	62.76	–
OlmoEarth v1 Large	61.84	−0.91
Galileo-B	58.78	−3.98

Against a model of similar compute. Against Galileo-B, which used a similar budget (573 GPU-hours), EarthFM is ahead on clouds (+2.52, CI 1.78 to 3.23), cashew (+6.48, CI 4.00 to 10.36), and land cover (+1.27 on test under Galileo-B's three-seed protocol, no interval; +2.27 on validation, chip-level CI 1.52 to 3.08); on burn scars, the 18.75B-token checkpoint is within noise (−1.64).

Where EarthFM trails. On marine debris (MADOS validation, two seeds), EarthFM scores 78.92 and 76.50 against 80.24 and 79.57. With 10% of burn-scar labels (validation), EarthFM scores 82.03 against 86.08. On multi-temporal crop types (PASTIS, development fold, temporal attention head), the independent 12B-token run scores 42.23 against 51.84. On a canopy-height development panel, EarthFM-short has a balanced MAE of 10.60 m against 10.42 m (lower is better). Only crop types is strongly temporal; the other three are dense tasks, so EarthFM's cloud advantage does not extend to every dense task.

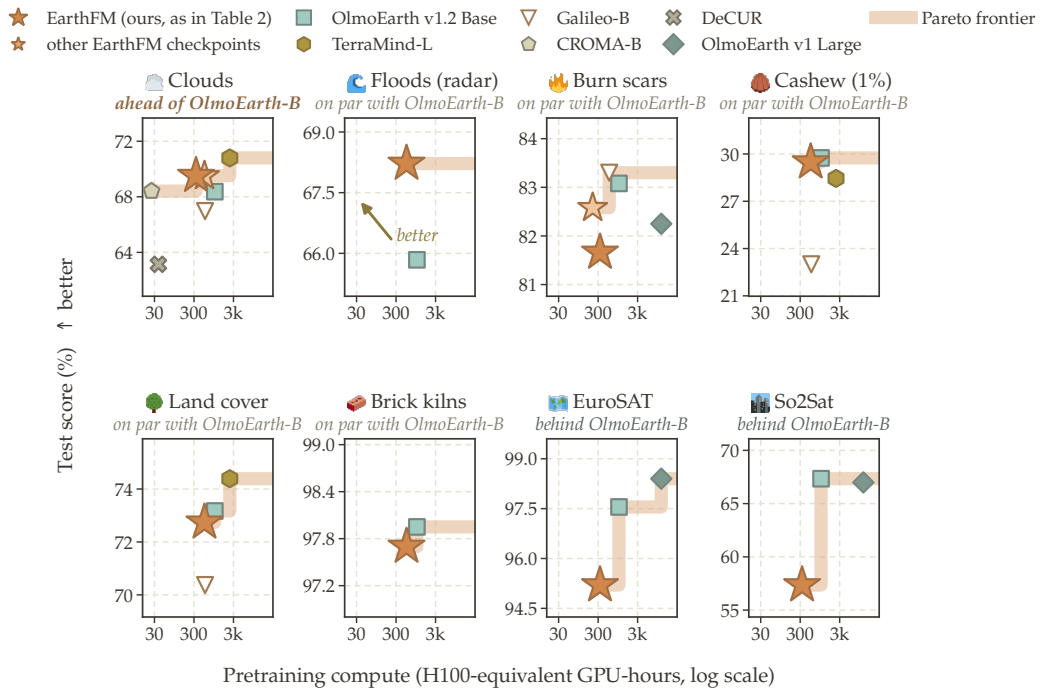

Figure 5: Accuracy against pretraining compute on the eight official test sets. Up and to the left is better; the shaded line is the Pareto frontier (no evaluated model is both cheaper and more accurate than a point on it). Large stars are the EarthFM checkpoints counted in Table 2; small stars are other EarthFM checkpoints (on burn scars, the independent 12B-token run). An EarthFM checkpoint is on every frontier, on several only because it is the cheapest model evaluated there; each panel has its own score axis, so compare gaps with Table 2, not across panels. Under the H100 NVL assumption of Table 5, Galileo-B would be slightly cheaper than EarthFM.

D.1 Cashew with 1% labels

The 1% setting samples 13 of 1,350 training images. Four fixed label draws and two adaptation seeds give eight paired runs. Every encoder chooses between a frozen encoder and full fine-tuning at three learning rates, with 80 updates per trial and no epoch selection. The 400 validation images come from 16 sites and are used both for selection and reporting, so intervals (hierarchical bootstrap over runs and sites) are conditional on that selection.

Table 7: Validation mIoU for each label draw and adaptation seed (80 updates). Highlight marks the best, second, and third model in each row.

Draw / seed	EarthFM	OlmoEarth-B	TerraMind-L	Galileo-B
B / 234	27.02	27.10	24.36	23.36
B / 235	27.60	27.48	25.01	23.07
C / 234	31.47	31.83	29.65	27.87
C / 235	32.12	31.25	30.57	27.88
D / 234	32.20	31.12	30.22	23.93
D / 235	31.95	30.34	29.65	24.15
E / 234	30.44	29.32	27.73	24.56
E / 235	29.68	29.78	28.09	24.87
Mean	30.31	29.78	28.16	24.96

Official test read. Every cashew cell was read once on the official test split (50 images from two sites), keeping the validation selection. The pre-declared claim uses the union selection, which picks 80 or 400 updates on validation and here picks 400: EarthFM 29.48, OlmoEarth 29.75 (-0.28 [-1.00, $+0.58$]), TerraMind-L 28.45, Galileo-B 23.00. At 80 updates alone, EarthFM scores 25.36 against 23.89 ($+1.47$ [$+0.15$, $+3.32$]; chip-level $+1.15$ to $+1.83$), 21.99, and 19.88, ahead of OlmoEarth in all eight runs, but the lead passes only when all eight runs are pooled. Under matched fine-tuning recipes, OlmoEarth leads on test at 800 updates (-0.79 [-1.65, -0.03]) and 1,600 updates (-1.34 [-2.24, -0.51]); with each model's locked recipe the two tie at 1,600 (-0.33 [-1.76, $+0.82$]). With only two test sites, the site-level resampling has few distinct outcomes; the chip-level interval treats correlated images of the same site as independent and is therefore optimistic. Test scores come from deterministic re-runs; their validation scores matched the originals within 0.06 points for the 80-update cells (TerraMind-L up to 0.63) and within 0.10 (EarthFM), 0.32 (OlmoEarth), 0.54 (Galileo-B), and 1.78 (TerraMind-L) for the 400-update cells. Imagery that EarthFM very likely consumed during pretraining (from a replay of its data stream) covers 51% of the average test image (88% of one test site, 14% of the other), though none of it from the same date or the benchmark's time window; none of OlmoEarth's pretraining imagery overlaps the benchmark (nearest 22.7 km).

Longer adaptation. On validation, with 400 updates, every model gains 3.2 to 4.5 points and the order of EarthFM and OlmoEarth flips (34.02 versus 34.12); TerraMind-L scores 32.70 and Galileo-B 28.18. With recipes fixed on draws B and C and extended schedules, OlmoEarth leads at 1,600 updates (37.12 versus 35.51; -1.61 [-2.48, -0.78]); on test at 1,600 updates the two tie (-0.33 [-1.76, $+0.82$]). On validation, EarthFM adapts quickly but plateaus earlier; on test, its locked recipe is still improving from 800 to 1,600 updates.

E Diagnostics and late-training details

Where the loss sits. On a development panel (two head seeds, final-step readout), linear probes on intermediate layers give the following cloud mIoU (Figure 6c). Under latent-heavy weights, the best layer moves down to layer 3 late in training (layer 3 beats the output by 7.2 and 6.7 points at 30B for the two head seeds, paired chip-bootstrap intervals above zero); in runs trained with ALP's weights from the start it stays in the upper half, and in EarthFM the output is within noise of the best layer; the latent-heavy trunk switched to ALP weights at 15B (X5) has a flat profile from layer 3 to layer 6.

What R^2 does and does not predict. Across the 36 encoders with cloud scores, the rank correlation between anchor-recovery R^2 and thin-cloud IoU is 0.33; across matched-token pairs, changes agree

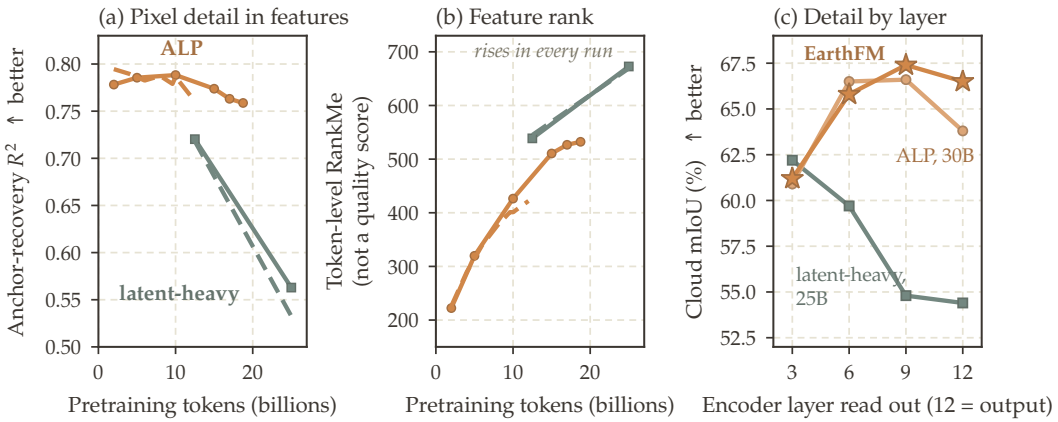

Figure 6: Diagnostics along training (CloudSEN12 development panel). In (a) and (b), solid lines are seed 42 and dashed lines seed 43. (a) Anchor-recovery R^2: how well a ridge probe on frozen tokens recovers the fixed 40 m pixel anchors; higher means more pixel detail is linearly recoverable. The latent-heavy runs were measured only at their 12.5B and 25B cooldown checkpoints. (b) Token-level RankMe rises with tokens in every run, whatever happens to the maps. (c) Linear cloud probes on intermediate encoder layers (12 is the output used everywhere else; two head seeds, seed-42 runs; values in Table 8).

Table 8: Cloud mIoU by encoder layer (development panel, mean of two head seeds). Highlight marks each row's best layer.

Checkpoint	Layer 3	Layer 6	Layer 9	Layer 12 (output)
Latent-heavy, 25B (cooldown)	62.2	59.7	54.8	54.4
Latent-heavy, 30B	61.1	59.6	55.2	54.2
ALP, 30B	60.9	66.5	66.6	63.8
Latent-heavy trunk → ALP weights (X5), 31.25B	61.9	62.3	59.6	59.2
ALP trunk → latent-heavy weights (X4), 31.25B	59.6	66.1	64.0	61.1
EarthFM, 31.25B	61.2	65.8	67.4	66.5

in direction in 39 of 47 pairs, but the pairs share checkpoints. R^2 can fall while accuracy rises (ALP run, 10B to 18.75B), and OlmoEarth v1 Large has higher R^2 than v1.2 Base but lower cloud accuracy. We therefore use R^2 to compare recipes, not to decide when to stop. Pooled-feature rank behaves differently from token rank; all ranks in the paper are token-level RankMe.

Ablations. The anchor ablation sets the patch and sub-patch weights to zero in an early latent-heavy variant (V4-A) and compares it with the same variant at every checkpoint from 2B to 6.4B tokens, on a 64-chip development panel with two head seeds and one pretraining seed. It removes the local CNN's only consumer, so it ablates targets and CNN together. The external-head ablation (X3) removes the map, cloud, and presence heads from the ALP recipe and is compared with the seed-43 ALP run (12B schedule) on the official test split, because the seed-42 run has no 4B–8B checkpoints; the 8B ablation point is annealed while the reference is not.

Weight switches. Each switch resumes a run at 15B with the other recipe's five main weights (patch, sub-patch, latent, cloud, view alignment) and a 31.25B budget, which also changes the EMA schedule relative to the original runs. The swapped latent-heavy run holds 62.2 to 62.6 test mIoU from 17.5B to 31.25B tokens: it halts the decline but does not return to its 10B peak (63.5).

Which weights matter (interim). To separate the anchor weights from the cloud weight, three further branches resume the latent-heavy run at 15B on the same schedule as the switch, changing only the anchor weights, only the cloud weight, or nothing (one run each; read at 17.5B; the registered comparison at 20B had not been read when this draft was written).

Table 9: Decomposing the switch to ALP's weights (latent-heavy run, switched at 15B tokens, read at 17.5B; one pretraining run per row; validation head-seed spread at most 0.5 points for cloud mIoU and up to 3.5 for thin cloud). The full switch changes five weights: the 80 m and 40 m anchor weights, the latent weight, the cloud weight, and the view-alignment weight. Cloud and thin-cloud IoU on the official validation split; land cover on the development panel.

Weights changed at 15B	Cloud mIoU ↑	Thin-cloud IoU ↑	Land-cover mAP ↑
none	59.6	32.7	66.1
anchor weights only	61.0	34.6	65.9
cloud weight only	60.6	34.8	66.3
all five ALP weights	**62.9**	**38.8**	67.2

External targets under the adapted protocols. Removing the map, cloud, and presence heads (8B tokens; validation; two seeds) lowers frozen-encoder UNet cloud mIoU from 68.6 to 67.6, fine-tuned land-cover mAP from 74.6 to 74.0, and fine-tuned flood IoU from 68.7 to 68.3 (seeds disagree in sign), much less than the 2.7-point cloud cost under the frozen linear probe in Table 1.

Table 10: Late-training changes (frozen linear heads, two head seeds, one pretraining run). All rows except the last branch from one shared 20B ALP state with a 31.25B budget and use official validation; the last row uses the repeatedly read official test split. For scale, a continuation that changed only the sensor mix moved test cloud mIoU by $+0.13$. ▭ gain of 0.5 points or more.

Change	Held fixed	Tokens	Δ Cloud	Δ BEN
Halve the learning rate	anchor weights raised	25B	+0.65	−0.04
Raise anchor weights	low learning rate	25B	+0.18	−0.04
Raise latent weight ($2 \to 4$)	anchor weights raised	25B	+0.01	+0.03
Add view-statistics targets	low learning rate	25B→31.25B	+0.70	−0.02
Add view-statistics targets (test split)	ALP, from 15B	18.75B	+0.91	+0.08

Why the learning rate matters late. In the ALP run at constant peak learning rate, the fraction of clipped updates rose from 0.13 at 15B tokens to 1.00 from 19B onward, while the pre-clip gradient norm rose from 0.96 to 1.36. **Caveat:** the branches in Table 10 share one declared budget and therefore one EMA schedule, so their differences are free of any schedule effect. Every branch scores above the unbranched run at 25B: by 0.8 points for X7 and X10, which share its learning rate, and by

1.3 to 1.5 points for x9 and X11 at half the learning rate. The unbranched run uses a longer budget and a different EMA schedule, so that comparison mixes the treatment with the schedule, and we do not credit EarthFM's gain over the unbranched run to its anchor weights alone. The run-to-run noise of a continuation was not measured.