remote sensing

MDPI

Article

LAPFuse: Language-Anchored and Physics-Informed Fusion of Sentinel and GEDI Data for Forest Aboveground Biomass Estimation

Yakun Li [1,2], Xiaolong Qu [1,2], Luhan Xu [1,2], Fangyu Yang [1,2] and Shiruo Hu [3,*]

[1] School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; lykun@bjfu.edu.cn (Y.L.); quxiaolong@bjfu.edu.cn (X.Q.); xlh5494@bjfu.edu.cn (L.X.); alexababab@bjfu.edu.cn (F.Y.)
[2] Hebei Key Laboratory of Smart National Park, Beijing 100083, China
[3] State Key Laboratory of Hydro-Science and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing 100084, China
* Correspondence: husr@mail.tsinghua.edu.cn

Highlights

What are the main findings?

- LAPFuse combines dense Sentinel observations and sparse GEDI L4A footprint records through language-anchored alignment and support-aware attention.
- The model achieves $R^2 = 0.882$ and 0.949 coverage for nominal 95% prediction intervals on 369 held-out plots.

What are the implications of the main findings?

- Cross-region tests show zero-shot R^2 values of 0.762–0.798 and small residual gaps after adaptation with 100 target plots.
- The framework provides plot-level AGB estimates and empirical uncertainty information for heterogeneous forest regions with complete sensor coverage.

Abstract

Accurate forest aboveground biomass (AGB) estimates are needed for forest inventories and carbon assessment, yet optical, synthetic aperture radar (SAR), and spaceborne lidar products describe forest properties at different spatial supports. We present LAPFuse, a multimodal framework that combines dense Sentinel-2 and Sentinel-1 observations with sparse GEDI Level 4A footprint-level aboveground biomass density records. Modality-specific encoders preserve the sensing characteristics of each input. Fixed language anchors regularize their projected representations, while distance-weighted sparse–dense attention relates GEDI footprint tokens to spatially distributed Sentinel features. A physics-informed neural process then produces a predictive mean and interval using an allometry-based soft constraint. We evaluate LAPFuse with 1847 permanent plots from the 2022–2024 growing seasons in three forest regions of China. On the held-out set of 369 plots, LAPFuse achieves $R^2 = 0.882$, RMSE $= 15.34$ Mg ha^{-1}, and MAE $= 9.87$ Mg ha^{-1}. Its 95% prediction intervals attain a coverage probability of 0.949 with a mean width of 67.2 Mg ha^{-1}. Sequential ablation shows progressive improvements as the alignment, attention, probabilistic, and allometric components are introduced. Zero-shot cross-region R^2 ranges from 0.762 to 0.798, and adaptation with 100 target-region plots reduces the performance gap to 0.1–1.5% relative to in-domain performance. These results demonstrate the value of support-aware multimodal fusion for plot-level AGB estimation under this sensor configuration and these regional conditions.

Check for updates

Academic Editor: Nikolay S. Strigul

Received: 14 September 2026
Revised: 20 October 2026
Accepted: 22 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: forest aboveground biomass; Sentinel-1; Sentinel-2; GEDI L4A; multimodal fusion; neural process; uncertainty quantification

1. Introduction

Forest aboveground biomass (AGB) is a fundamental variable for forest inventories, carbon-cycle assessment, and estimates of terrestrial carbon sinks [1,2]. Field inventories provide detailed measurements at the plot scale, but their cost and limited revisit frequency constrain spatially continuous monitoring. Satellite observations extend the spatial and temporal coverage of forest assessment, although converting remotely sensed signals to AGB remains difficult in structurally diverse forests [3–5]. The principal difficulties in converting remotely sensed signals to AGB include: (1) spectral saturation at high biomass in optical data, (2) variability in radar scattering influenced by viewing geometry and moisture, (3) sparse structural observations from spaceborne lidar, and (4) uncertainty in field-derived biomass references from plot allometry and sampling.

Optical, synthetic aperture radar (SAR), and spaceborne lidar observations offer complementary information. Sentinel-2 records multispectral reflectance that is sensitive to canopy composition and condition, but cloud contamination and saturation can limit its response in dense stands [6,7]. Sentinel-1 provides cloud-independent VV and VH backscatter, which responds to canopy structure and moisture but is also influenced by viewing geometry and terrain [8,9]. The Global Ecosystem Dynamics Investigation (GEDI) adds footprint-level observations that support biomass mapping, yet GEDI Level 4A (L4A) records remain spatially sparse relative to Sentinel imagery [10,11]. Their joint use therefore requires more than feature concatenation: the model must preserve differences between dense image support, sparse footprint support, and plot-level reference measurements.

Current multisource biomass mapping approaches range from stacked predictors and tree ensembles to convolutional networks, foundation models, and multimodal Transformers [12–16]. Tree ensembles remain effective when optical indices, radar statistics, terrain variables, and lidar-derived attributes can be summarized in a common table. Their performance, however, depends on feature engineering and does not preserve the native support of each observation. Convolutional and Transformer models reduce this dependence by learning spatial features from imagery [17–19]. Remote-sensing foundation models such as Prithvi further transfer representations learned from large Earth-observation archives [20].

Multimodal networks usually encode each sensor separately and combine the resulting features by concatenation, summation, gating, or attention [14,16,21]. These designs show that optical, radar, and lidar-derived variables can improve forest-attribute estimation when used together [22–24]. Yet, a pooled GEDI vector no longer records which footprint contributes to a particular part of an image window. Ordinary attention can learn feature compatibility, but it does not guarantee sensitivity to the distance between dense image tokens and sparse footprint observations. A support-aware fusion rule is therefore needed when geometry is part of the observation process.

Most biomass regressors also emphasize point accuracy. An R^2 or RMSE value summarizes average predictive performance but cannot show whether uncertainty changes across the biomass range. This omission matters because optical and radar responses can become less informative in dense stands, while plot allometry and footprint sampling add further error [3,5,9]. It also complicates model transfer: similar pooled error can conceal distinct regional biases or poorly calibrated high-biomass predictions. Point metrics therefore require joint examination with regional residuals, prediction-interval coverage, and interval width.

Representation alignment offers one route to managing sensor heterogeneity. Contrastive image–text pretraining maps visual and textual information into a shared embedding space [25,26]. Remote-sensing adaptations extend this principle to image retrieval, scene recognition, visual question answering, and geospatial dialogue [27–31]. Their success indicates that short descriptions can organize sensor representations even when images differ from ordinary ground-level photographs. Geospatial benchmarks nevertheless show sensitivity to resolution, viewing geometry, location, and task definition [32,33].

Remote sensing image captioning provides a related perspective on vision–language alignment. Early work introduced benchmark datasets (UCM-Captions, RSICD, Sydney-Captions) and encoder–decoder architectures that map satellite images to natural-language descriptions [34,35]. Subsequent methods added attention mechanisms, variational encoders, and reinforcement learning to improve description quality and semantic relevance [36,37]. More recent vision–language models extend CLIP-style contrastive pretraining to remote-sensing domains, enabling zero-shot scene classification and visual question answering [27,38]. These studies demonstrate that learned vision–language embeddings can organize sensor representations even when ground truth is specified numerically rather than textually. LAPFuse builds on this principle: fixed sensor-role descriptions anchor each modality encoder in a shared semantic space, but the regression target remains the field-derived AGB value, not generated text.

Unlike image captioning, where descriptions are generated from visual inputs, continuous biophysical regression uses language solely to align modality encoders while the numerical target comes from field measurements. The relevant opportunity is therefore representation regularization, not language-based prediction. A sentence describing an optical, SAR, or footprint role can provide a stable reference direction for a modality projection. Location-matched sensor pairs can then be aligned without discarding their separate encoders. In LAPFuse, these two constraints are used together: fixed descriptions distinguish sensor roles, while pairwise alignment encourages agreement among observations of the same plot.

Uncertainty and ecological consistency form a second methodological requirement. Monte Carlo dropout and deep ensembles approximate predictive uncertainty by repeated stochastic inference or repeated model fitting [39,40]. Neural processes instead model a distribution over target values conditioned on a set of observed input–response pairs [41,42]. Their context representation can adapt a prediction to the examples available for a region, while the decoder returns both a mean and a predictive scale. Spatial variants use this formulation for environmental prediction and interpolation [43–46].

A probabilistic head alone does not ensure ecologically credible predictions. Forest AGB is linked to tree diameter, height, and wood density through allometric relationships, and these relationships can provide a useful training prior [47]. Enforcing a single equation as a hard rule is inappropriate across forest types because its parameters have a calibration domain. A soft penalty offers a compromise: it discourages large departures from the allometric scale but permits the remote-sensing evidence to adjust the estimate. The combination of conditional prediction and soft physical regularization motivates the physics-informed neural process used here.

We propose LAPFuse, a language-anchored physics-informed multimodal fusion framework for plot-level AGB estimation. LAPFuse uses modality-specific encoders for Sentinel-2, Sentinel-1, and GEDI L4A records. It regularizes the encoded features with fixed sensor-role descriptions, relates sparse GEDI footprints to dense Sentinel tokens through distance-weighted attention, and predicts AGB with a physics-informed neural process. LAPFuse addresses the four challenges outlined above through three innovations: language-anchored alignment manages sensor heterogeneity (challenges

1–2), sparse–dense attention integrates irregular footprints (challenge 3), and a physics-informed neural process quantifies uncertainty (challenge 4). The framework is evaluated on 1847 permanent plots from Northeast, Southeast, and Southwest China using held-out prediction, sequential cumulative ablation, cross-region transfer, uncertainty diagnostics, modality-subset analysis, and computational measurements. The code is available at https://github.com/QuXiaolong0812/LAPFuse (accessed on 21 October 2026).

These three innovations are tightly coupled by design. Language-anchored alignment regularizes sensor representations but cannot relate sparse footprints to dense image tokens without the distance-weighted attention module. Attention without language anchors would learn token compatibility but could not guarantee that optical, SAR, and footprint representations remain distinguishable in the shared embedding space. The neural process provides context-conditioned uncertainty estimates but relies on aligned multimodal features to produce calibrated intervals across regions and biomass ranges. The allometric penalty further constrains the physical plausibility of predictions when remote-sensing signals alone become ambiguous. Together, the components address complementary aspects of the support-mismatch, sensor-heterogeneity, and uncertainty-quantification problems, and their combination forms a coherent methodological framework for plot-level AGB estimation under the stated sensor configuration.

These innovations constitute three methodological contributions:

- A language-anchored alignment objective provides a common representational reference for heterogeneous sensor features while preserving the field inventory as the sole regression target.
- A sparse–dense collaborative attention module retains GEDI footprint coordinates and uses distance-aware interactions to integrate irregular footprint records with dense Sentinel features.
- A physics-informed neural process combines plot-conditioned probabilistic prediction with a soft allometric penalty, allowing point accuracy, interval coverage, cross-region adaptation, and computational cost to be assessed within one framework.

2. Materials and Methods

2.1. Study Areas and Reference Data

As shown in Figure 1, the study uses 1847 permanent forest plots distributed across three regions of China. The Northeast region contains 680 plots in temperate larch and mixed broadleaf forests. The Southeast region contains 620 plots in subtropical evergreen broadleaf and bamboo forests, and the Southwest region contains 547 plots in mountain oak, pine, and fir forests. As summarized in Table 1, the three study regions collectively span a broad elevation gradient from 40 to 3620 m, with observed AGB values ranging from 51.5 to 262.1 Mg ha^{-1}.

The dataset contains field measurements from the 2022–2024 growing seasons. Each 0.06 ha plot records tree species, diameter at breast height (DBH), tree height, and wood density. Species-specific equations from the Chinese Tree Biomass Database are applied at the tree level, and the estimates are summed and normalized by plot area. For plot area A_{plot} and tree-level biomass estimate m_j, the target is:

$$y_i = \frac{1}{A_{\text{plot}}} \sum_{j=1}^{n_i} m_j, \qquad A_{\text{plot}} = 0.06 \text{ ha}. \tag{1}$$

The resulting y_i is plot-level AGB density in Mg ha^{-1}. DBH, height, species, and wood density are used to construct the reference response and the allometric regularizer described in Section 2.8. They are not remote-sensing predictor variables at inference.

The response and predictors describe related forest properties at different supports. Plot AGB integrates the trees within a fixed boundary, while a Sentinel value represents a regular image cell and GEDI L4A represents a laser footprint. A plot may therefore contain spatial variation that is not resolved by one sensor record. LAPFuse preserves these support labels during encoding and uses plot identifiers as the units for splitting and evaluation. This choice prevents the denser image representation from being mistaken for a larger number of independent field observations.

The three-region design provides variation in climate, forest composition, terrain, and biomass. Northeast China covers the lowest upper AGB bound, whereas Southeast China contains the highest observed values. Southwest China contributes the largest elevation range. These contrasts allow regional residual analysis and source–target transfer tests. They do not constitute a nationally representative probability sample, so inference remains tied to the sampled regional envelopes shown in Figure 1.

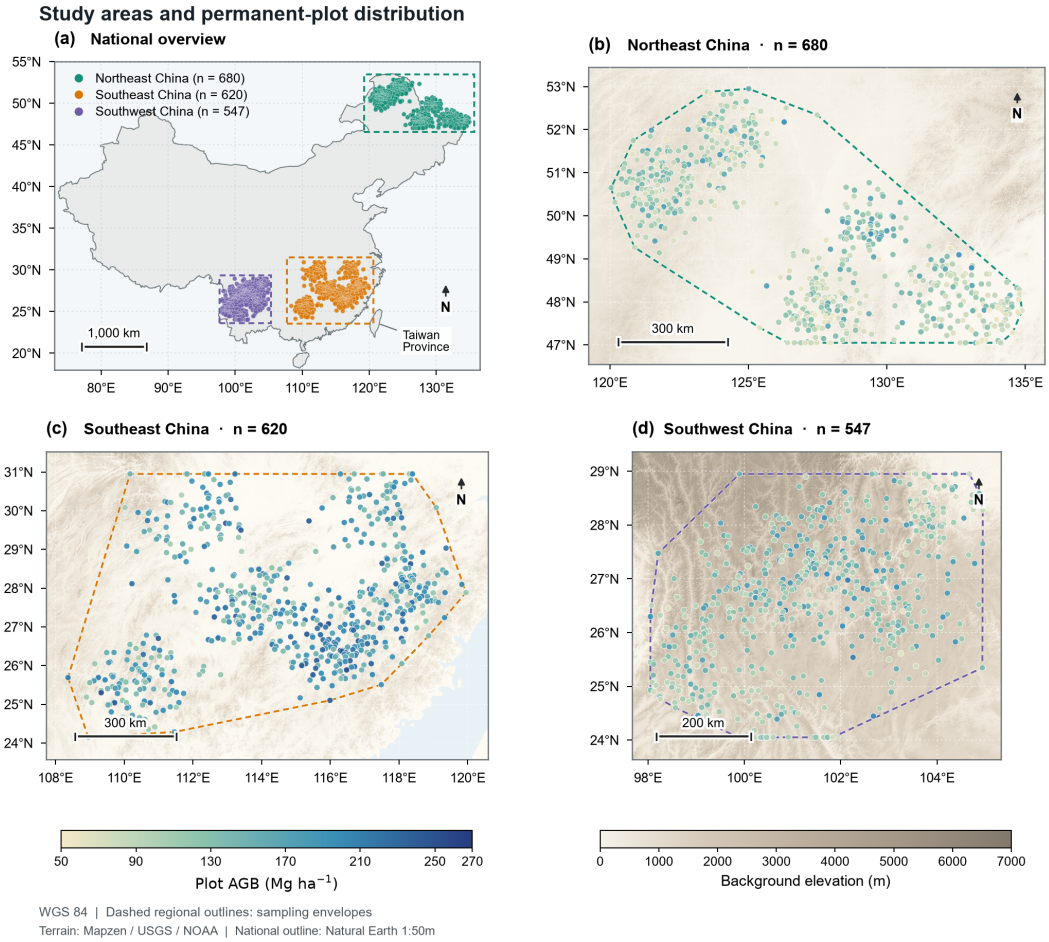

Figure 1. Study areas and permanent-plot distribution. (**a**) Locations of the three study regions in China; dashed rectangles show their sampled coordinate extents. (**b**–**d**) Plot distributions in Northeast, Southeast, and Southwest China, respectively. Each regional subplot displays an open-access digital elevation model (DEM) as the background layer, rendered with a unified elevation scale (0–7000 m) and consistent hillshade effect to reveal local topography. Point color represents field-derived AGB (Mg ha^{-1}), and dashed outlines are convex sampling envelopes. Coordinates are shown in WGS 84 (EPSG:4326), and the national outline uses the Natural Earth 1:50 m cartographic layer. DEM background is sourced from Mapzen Terrain Tiles (https://registry.opendata.aws/terrain-tiles/) (accessed on 21 October 2026).

Table 1. Study-region and reference-data summary for the 1847 permanent plots. Coordinate entries describe the sampled extents.

Region	Plots	Train/Val./Test	Coordinate Extent (°)	AGB Range (Mg ha^{-1})	Elevation (m)	Forest Types
Northeast China	680	402/140/138	47.05–52.95N; 120.05–134.88E	51.5–196.7	120–1476	Temperate larch and mixed broadleaf
Southeast China	620	379/120/121	24.15–30.95N; 108.37–119.95E	95.9–262.1	40–1099	Subtropical evergreen broadleaf and bamboo
Southwest China	547	327/110/110	24.05–28.95N; 98.05–104.95E	63.0–209.0	760–3620	Mountain oak, pine, and fir
All regions	1847	1108/370/369	–	51.5–262.1	40–3620	Surveyed during 2022–2024; 0.06 ha plots

2.2. Remote-Sensing Data and Preprocessing

As shown in Table 2, four Earth-observation products are associated with each field plot. Sentinel-2 Level-2A supplies ten optical bands: B2, B3, B4, and B8 at a native spatial resolution of 10 m, and B5, B6, B7, B8A, B11, and B12 at 20 m. Before band stacking and feature extraction, the six 20 m bands were resampled using bilinear interpolation onto the reference grid of the native 10 m bands, with consistent coordinate reference, grid origin, and pixel alignment. This procedure aligned all bands onto a common 10 m grid for constructing the multispectral inputs. The resulting pixel spacing does not imply that the original 20 m observations contain independently resolved 10 m spatial detail. Sentinel-1 ground-range-detected imagery supplies VV and VH SAR channels on the co-registered 10 m reference grid. Both Sentinel products were selected within ±30 days of the corresponding field survey. GEDI L4A v2.1 supplies irregular footprint coordinates and footprint-level AGBD values [11]. SRTM elevation at 30 m provides terrain context.

Table 2. Remote-sensing products and their roles in LAPFuse.

Product	Input	Spatial and Temporal Information	Model Role
Sentinel-2 L2A	B2–B8, B8A, B11, B12	10/20 m native, 10 m grid; ±30 d	Dense optical features
Sentinel-1 GRD	VV and VH	10 m grid; ±30 d	Dense SAR features
GEDI L4A v2.1	Footprint coordinates and AGBD	Sparse footprint records	Sparse biomass-related tokens
SRTM DEM	Elevation	30 m; static terrain layer	Terrain context

The products are co-registered before feature extraction. Sentinel-2 and Sentinel-1 windows are normalized with branch-specific statistics, while GEDI coordinates are expressed relative to the plot window. This procedure preserves the regular support of the image data and the irregular support of the footprint data. A sample is represented as:

$$\mathcal{Z}_i = \left(\mathbf{X}_i^{\text{opt}}, \mathbf{X}_i^{\text{sar}}, \{(\mathbf{p}_{in}, a_{in})\}_{n=1}^{N_i}, \mathbf{e}_i, y_i\right), \tag{2}$$

where $\mathbf{X}_i^{\text{opt}}$ and $\mathbf{X}_i^{\text{sar}}$ are dense image tensors, $\mathbf{p}_{in} \in \mathbb{R}^2$ is a projected GEDI footprint coordinate, a_{in} is its L4A AGBD value, $\mathbf{e}_i$ is the terrain feature, and N_i varies among plots. The GEDI branch processes the footprint-level product records as an irregular set.

Temporal matching uses the field-survey date as the reference. The ±30-day window limits large phenological offsets while allowing cloud-free optical observations to be selected. Across the 1847 plots, the mean absolute time difference between Sentinel-2 acquisition and field survey was 15.2 days (median: 14.5 days, range: 1–29 days). For Sentinel-1, the corresponding values were 15.1 days (median: 14.8 days, range: 4–25 days). Only 4.0% of plots had Sentinel-2 observations within ±7 days, 56.0% within ±15 days, and 100% within ±30 days. The ±30-day window was selected to balance temporal proximity with the practical need to obtain cloud-free optical observations in these subtropical and temperate forest regions [12,23]. Sentinel-1 and Sentinel-2 are normalized independently

because reflectance and backscatter have different distributions. The relative GEDI coordinates are retained because footprint rasterization assigns one product value to an arbitrary regular cell. SRTM is processed as an auxiliary terrain layer and does not replace the sensor-specific branches.

GEDI L4A AGBD is a model-derived biomass product calibrated using field inventory and airborne lidar data [11]. We could not establish whether the plots used in this study, historical measurements from the same permanent plots, or associated airborne lidar observations contributed to the training, calibration, or validation of the GEDI L4A product. Potential dependence between this external predictor and the field-derived reference data therefore cannot be excluded.

The preprocessing sequence also separates data quality from feature learning. Optical cloud screening determines which reflectance observations enter the model, while SAR calibration and terrain handling affect the comparability of backscatter. GEDI quality screening determines the membership of the footprint set. These operations precede the encoders, so the learned attention weights operate only on the observations retained for each plot. The resulting experiments evaluate plots for which the stated modality combination is available.

For each plot, remote-sensing features were extracted using plot-center coordinates as spatial references. Sentinel-2 and Sentinel-1 observations were extracted from a fixed 64×64 pixel window (4096 pixels) centered on each plot, corresponding to a 640 m $\times$ 640 m spatial footprint at the 10 m resolution. This window size was chosen to capture the forest context surrounding the 0.06 ha plot (circular radius $\approx$13.8 m) while providing sufficient spatial information for the convolutional and self-attention encoders.

GEDI L4A footprints within 250 m of each plot center were retained and processed as an irregular point set. Footprints were selected when their center coordinates fell within the circular search radius, and their complete AGBD values and coordinates were retained without clipping or area-weighting relative to the plot boundary. GEDI quality filtering required `algorithm_run_flag=1`, `l2_quality_flag=1`, `l4_quality_flag=1`, `degrade_flag=0`, and the absence of missing values in coordinates or AGBD. The number of quality-filtered footprints per plot varied across the dataset, reflecting GEDI's sparse and irregular sampling pattern.

SRTM elevation was extracted as the single 30 m pixel value at the plot-center coordinates. This single-pixel extraction provides a terrain reference for each plot location but does not resolve intra-plot elevation variation at the scale of the 0.06 ha plot area. The three remote-sensing inputs therefore operate at different native spatial supports: dense regular grids for Sentinel-2 and Sentinel-1, an irregular sparse point set for GEDI L4A, and a coarser single-pixel terrain reference for SRTM.

2.3. Sample Construction and Evaluation Design

The plot data are divided into training, validation, and test sets in a 60:20:20 ratio, yielding 1108, 370, and 369 plots, respectively. Stratification retains the regional and biomass distributions shown in Table 1. All sensor windows, GEDI records, and labels associated with a plot remain in the same partition. The validation set supports model selection, and the test set is reserved for final evaluation.

The random assignment does not enforce spatial separation between partitions. Based on plot coordinates, the minimum pairwise distances between training and test plots, training and validation plots, and validation and test plots are 1.103, 0.654, and 0.705 km, respectively. The median distance from each test plot to its nearest training-set plot is 13.376 km. These distances describe plot-center separation and do not account for potential overlap of Sentinel extraction windows or reuse of GEDI footprints across partitions. No

spatially blocked holdout evaluation was performed, so the reported metrics describe prediction performance under random plot-level partitioning within the sampled regions rather than spatially independent extrapolation.

Cross-region experiments assess geographic transfer. A model fitted in one region is applied to a different target region without target labels (zero-shot) or adapted with 10, 50, or 100 labelled target plots. The three directions are Northeast-to-Southeast (NE→SE), Northeast-to-Southwest (NE→SW), and Southeast-to-Southwest (SE→SW). Target responses used for evaluation are excluded from their own neural-process context sets. All performance values below use plots, rather than trees, footprints, or pixels, as the evaluation units.

The context set has a distinct role from the target set. Each context element contains a fused plot representation and its observed AGB, whereas a target element contains the representation whose response is to be predicted. During in-domain training, context and target subsets are drawn from labelled training data. During transfer, the target-region label budget determines how many target plots can enter the adaptation context. This construction allows the probabilistic head to change its regional context without redefining the sensor encoders.

The zero-shot and few-shot settings answer different questions. Zero-shot evaluation measures how the source-trained representation behaves under a regional shift without target labels. The 10-, 50-, and 100-shot settings measure adaptation as labelled target examples become available. Because the same three directions are evaluated at each budget, the successive values describe the marginal benefit of added regional context. The analysis does not extrapolate beyond the largest tested budget.

2.4. LAPFuse Architecture

As shown in Figure 2, LAPFuse contains modality-specific encoders, language-anchored alignment, sparse–dense collaborative attention, and a physics-informed neural-process predictor. The modality encoders first transform optical, SAR, and GEDI inputs into tokens with a common latent dimension d. Alignment losses regularize the projected token spaces. Dense Sentinel tokens then query GEDI footprint tokens through a distance-weighted attention module. Pooled optical, SAR, GEDI, and terrain features form a plot representation $\mathbf{x}_i$, which is combined with a labelled context set to predict the mean and scale of AGB.

For a batch of B plots, the encoders return $\mathbf{H}_{\text{opt}} \in \mathbb{R}^{B\times M_o\times d}$, $\mathbf{H}_{\text{sar}} \in \mathbb{R}^{B\times M_s\times d}$, and footprint sets $\mathbf{H}_{\text{gedi},i} \in \mathbb{R}^{N_i\times d}$. The token counts M_o, M_s, and N_i need not be equal. Pooling is performed only after the support-aware interaction so that the footprint coordinates remain available to the attention module.

The forward pass follows the same order during training and inference. First, each sensor is encoded without mixing its native values with another modality. Second, the projected features are aligned during the representation-learning stage. Third, the dense tokens interact with the coordinate-bearing footprint tokens. Finally, the fused representation and terrain context are pooled to the plot level. The neural-process decoder then combines this representation with the context summary. This order ensures that spatial information is available when the sparse and dense supports first interact.

2.5. Modality-Specific Encoders

The optical branch uses the Prithvi Earth-observation foundation model with a frozen backbone, a trainable adapter, and a 512-dimensional projected representation [20]. Sentinel-2 provides multispectral reflectance at 10-meter resolution in regular grids, capturing both spectral signatures and spatial context. Prithvi-EO-1.0 was pretrained on harmonized

Landsat–Sentinel (HLS) data through masked autoencoding, offering general-purpose multispectral representations. By freezing the pretrained backbone and training only the adapter, the model limits the number of parameters updated from the 1108 training plots, reducing the risk of overfitting to local patterns while preserving transfer knowledge from global multispectral pretraining.

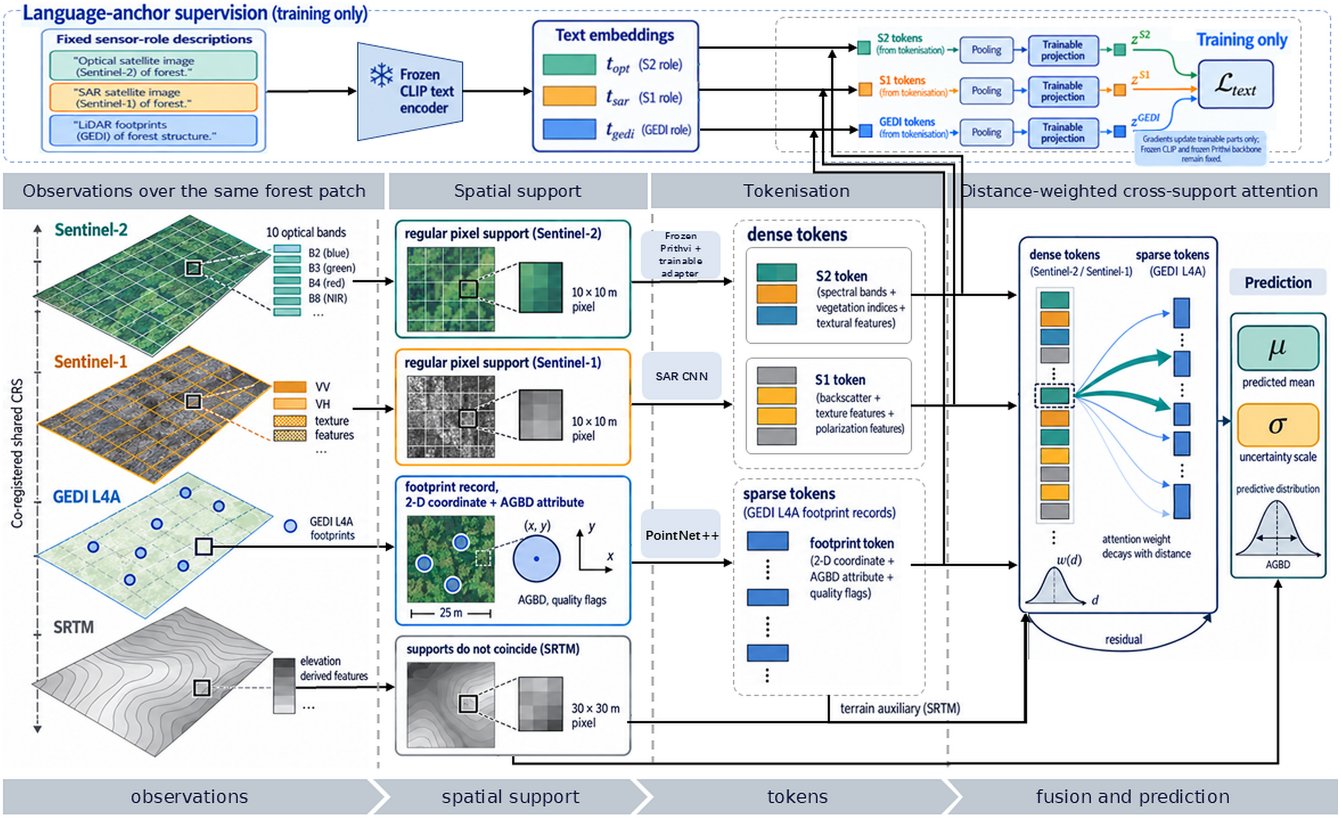

Figure 2. Architecture of LAPFuse. Co-registered Sentinel-2, Sentinel-1, GEDI L4A, and SRTM observations retain their spatial supports during tokenization. Dense Sentinel tokens interact with sparse GEDI footprint tokens through distance-weighted attention, while SRTM supplies terrain context. The neural-process predictor returns an AGB mean μ and predictive standard deviation σ. Fixed language anchors (orange box) are encoded by a frozen CLIP text encoder to produce modality-specific text embeddings ($\mathbf{t}_{\text{opt}}$, $\mathbf{t}_{\text{sar}}$, $\mathbf{t}_{\text{gedi}}$). These embeddings serve as semantic references during representation learning: each modality encoder's output is projected (via g_{opt}, g_{sar}, g_{gedi}) and aligned with its corresponding text embedding through contrastive learning ($\mathcal{L}_{\text{text}}$, blue arrows), establishing a shared semantic space before multimodal fusion. Loss components are color-coded: blue for language alignment ($\mathcal{L}_{\text{text}}$), green for pairwise alignment ($\mathcal{L}_{\text{pair}}$), orange for allometric regularization ($\mathcal{L}_{\text{allo}}$), and red for prediction loss ($\mathcal{L}_{\text{pred}}$).

The SAR branch contains four Conv2D–batch-normalization–ReLU blocks. It processes the VV and VH polarization channels together with the derived texture features used in the experiment. The convolutional layers progressively extract local spatial patterns from the backscatter intensity and texture, capturing structural information at neighbourhood scales. An independent branch allows the model to learn SAR-specific feature distributions without forcing its radiometric characteristics through the optical encoder. After four convolutional stages, global average pooling aggregates the spatial representation, and a linear projection maps it to the shared 512-dimensional latent space.

The GEDI branch applies PointNet++ to the irregular set of footprint coordinates and AGBD attributes [48]. GEDI L4A footprints form a sparse, irregularly distributed point set, with each observation containing a two-dimensional coordinate and a biomass density

estimate. PointNet++ aggregates features through hierarchical spatial neighbourhoods, learning representations that respect the geometric relationships among nearby footprints. This design avoids the need to rasterize the irregular footprint locations onto fixed grid cells, preserving the native structure of the sparse GEDI observations. Its output retains one latent token per sampled footprint before set aggregation. The three branches can be defined as:

$$\begin{aligned} \mathbf{H}_{\text{opt}} &= E_{\text{opt}}(\mathbf{X}_{\text{opt}}), \\ \mathbf{H}_{\text{sar}} &= E_{\text{sar}}(\mathbf{X}_{\text{sar}}), \\ \mathbf{H}_{\text{gedi}} &= E_{\text{PointNet++}}\left(\{(\mathbf{p}_n, a_n)\}_{n=1}^{N}\right). \end{aligned} \tag{3}$$

These encoders differ in architecture but share the same interface: each returns a sequence or set of d-dimensional tokens. Despite their architectural differences, all three encoders project their outputs to the same 512-dimensional latent space. This shared dimensionality provides a uniform interface for language-anchored alignment and cross-modal attention, enabling the framework to compare and fuse representations without assuming identical sensor statistics. The modality-specific encoders first preserve the intrinsic characteristics of each data source, and then map them to a common space where cross-modal reasoning becomes tractable. The interface permits cross-modal alignment without assuming identical sensor statistics. It also isolates the function of each branch in the modality-subset experiment. When one remote-sensing source is omitted, the corresponding branch is removed while the remaining branches retain their input definitions.

2.6. Language-Anchored Cross-Modal Alignment

Fixed descriptions of the three sensor roles are encoded with a frozen CLIP text encoder [25]. The descriptions in Table 3 provide reference directions for the modality projections. They are used during representation learning and do not supply the numerical AGB target.

Table 3. Fixed language anchors used for representation alignment.

Modality	Fixed Language Anchor	Represented Sensor Role
Sentinel-2	*optical reflectance reveals vegetation greenness*	Spectral vegetation cue
Sentinel-1	*SAR backscatter indicates canopy structure*	Scattering and structure cue
GEDI L4A	*footprint biomass records summarize local forest structure*	Sparse biomass-related cue

SRTM-derived terrain features ($\mathbf{e}_i$) are incorporated directly into the plot-level representation alongside the fused sensor features but are not assigned a separate language anchor. This design reflects SRTM's role as an auxiliary terrain predictor rather than a primary remote-sensing modality subject to language-guided alignment.

Let $\mathcal{M} = \{\text{opt}, \text{sar}, \text{gedi}\}$, $\mathbf{t}_m$ denote the normalized text embedding for modality m, and g_m denote a trainable projection of the pooled modality feature $\mathbf{h}_{m,i}$. With cosine similarity $\text{sim}(\cdot,\cdot)$ and temperature τ, the language-anchor loss is:

$$\mathcal{L}_{\text{text}} = -\frac{1}{B|\mathcal{M}|}\sum_{i=1}^{B}\sum_{m\in\mathcal{M}} \log \frac{\exp[\text{sim}(g_m(\mathbf{h}_{m,i}), \mathbf{t}_m)/\tau]}{\sum_{k\in\mathcal{M}} \exp[\text{sim}(g_m(\mathbf{h}_{m,i}), \mathbf{t}_k)/\tau]}. \tag{4}$$

A second term aligns location-matched sensor observations. For the modality-pair set $\mathcal{P} = \{(\text{opt}, \text{sar}), (\text{opt}, \text{gedi}), (\text{sar}, \text{gedi})\}$,

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{P}|}\sum_{(a,b)\in\mathcal{P}} [1 - \cos(q_a(\mathbf{h}_a), q_b(\mathbf{h}_b))]. \tag{5}$$

The two losses have complementary roles. Equation (4) draws each branch toward its declared sensor role, whereas Equation (5) reduces disagreement between co-located observations. The text embeddings remain fixed, so language acts as a representation regularizer rather than an inference-time biomass predictor.

The anchors are intentionally short because they identify broad sensing roles rather than individual land-cover classes. The optical sentence emphasizes spectral vegetation response, the SAR sentence emphasizes scattering and structure, and the GEDI sentence identifies a sparse biomass-related product. Their asymmetry keeps the three reference directions distinct. During optimization, the denominator in Equation (4) contrasts each modality feature with all three anchor embeddings, while the pairwise term preserves agreement at the plot level.

2.7. Sparse–Dense Collaborative Attention

The sparse–dense module relates the irregular GEDI footprints to the spatially distributed Sentinel representation. Let $\mathbf{D} \in \mathbb{R}^{M \times d}$ denote the combined dense tokens and $\mathbf{S} \in \mathbb{R}^{N \times d}$ the GEDI tokens. Sinusoidal positional encoding is added to the sparse tokens using the footprint coordinates [49]:

$$\mathbf{S}^{+} = \mathbf{S} + \mathrm{PE}(\mathbf{p}), \qquad \mathbf{Q} = \mathbf{D}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{S}^{+}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{S}\mathbf{W}_V. \tag{6}$$

For dense-token coordinate $\mathbf{r}_m$ and footprint coordinate $\mathbf{p}_n$, the attention logit contains both feature similarity and a Gaussian distance penalty:

$$\alpha_{mn} = \mathrm{softmax}_n\left(\frac{\mathbf{Q}_m \mathbf{K}_n^{\top}}{\sqrt{d_k}} - \frac{\|\mathbf{r}_m - \mathbf{p}_n\|^2}{2\sigma_d^2}\right), \qquad \widetilde{\mathbf{D}}_m = \mathbf{D}_m + \sum_{n=1}^{N} \alpha_{mn}\mathbf{V}_n. \tag{7}$$

The distance scale σ_d controls the spatial reach of the footprint contribution. The residual connection preserves the dense representation when attention to the sparse branch is weak. This design retains the distinction between a GEDI footprint and a Sentinel token while allowing their contributions to be learned jointly.

The two terms in the attention logit address different evidence. The scaled dot product measures compatibility in the learned feature space, whereas the squared-distance term favors geographically nearby footprints. When σ_d is small, the module approaches a local-neighbour aggregation. As σ_d increases, feature similarity contributes more strongly relative to distance. The softmax normalizes the footprint weights separately for each dense query, allowing different image locations to draw on different sparse observations.

Adding the weighted value to $\mathbf{D}_m$ rather than replacing it is important when footprint support is uneven. The dense token retains its Sentinel representation even when all footprint similarities are small. Conversely, a compatible nearby footprint can adjust that representation through the learned value projection. The module therefore treats GEDI as spatially localized auxiliary evidence instead of a substitute for wall-to-wall imagery.

2.8. Physics-Informed Neural Process

The fused plot representation is passed to a context-conditioned probabilistic predictor. For a context set $\mathcal{C} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_c}$, an encoder maps each feature–response pair to $\mathbf{r}_i = \mathrm{MLP}([\mathbf{x}_i; y_i])$. Attention pooling yields $\mathbf{r}_c = \sum_i \omega_i \mathbf{r}_i$, where the learned weights ω_i sum to one. Given target feature $\mathbf{x}_*$, the decoder returns:

$$(\mu_*, \log\sigma_*) = d_\theta([\mathbf{x}_*; \mathbf{r}_c]), \qquad p(y_* \mid \mathbf{x}_*, \mathcal{C}) = \mathcal{N}(\mu_*, \sigma_*^2), \tag{8}$$

with $\sigma_* > 0$. Model fitting uses the heteroscedastic Gaussian negative log-likelihood:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left[\frac{(y_i - \mu_i)^2}{2\sigma_i^2} + \log \sigma_i \right]. \tag{9}$$

The corresponding central 95% prediction interval is $[\mu_i - 1.96\sigma_i, \mu_i + 1.96\sigma_i]$.

Context aggregation lets the predictor respond to the labelled examples available for a region. The encoder summarizes each context feature together with its response, and attention pooling determines its contribution to $\mathbf{r}_c$. The target response is absent from this calculation. The decoder therefore estimates a conditional distribution from the target's remotely sensed features and the independent context summary.

The negative log-likelihood couples point accuracy and predictive scale. A large residual is penalized more strongly when the predicted scale is small, while the logarithmic term prevents uniformly wide intervals from minimizing the objective. This balance makes interval width interpretable only alongside empirical coverage. The reported PICP and MPIW consequently evaluate complementary aspects of the predictive distribution.

Ecological knowledge enters through a soft allometric penalty. The reference relationship follows Chave et al. [47]:

$$\widehat{A}_{\mathrm{allo}} = 0.0673(\rho\, \mathrm{DBH}^2 H)^{0.976}, \tag{10}$$

where ρ is wood density in g cm^{-3}, DBH is in cm, H is tree height in m, and tree biomass is in kg. After aggregation and conversion to the plot-level density used by the response, the auxiliary estimate enters the loss as:

$$\mathcal{L}_{\mathrm{allo}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left| \mu_i - \widehat{A}_{\mathrm{allo},i} \right|. \tag{11}$$

The full training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \lambda_1 \mathcal{L}_{\mathrm{text}} + \lambda_2 \mathcal{L}_{\mathrm{pair}} + \lambda_3 \mathcal{L}_{\mathrm{allo}}. \tag{12}$$

The allometric term regularizes the predictive mean without imposing equality to the reference relationship. This allows the remotely sensed evidence and plot context to modify the prediction when the generic allometric form does not match local conditions.

The regularizer is used only during model fitting. At inference, LAPFuse receives the remote-sensing inputs and any labelled context allowed by the evaluation setting. Field measurements used to calculate $\widehat{A}_{\mathrm{allo}}$ are not appended to the target feature vector. This training–inference separation does not establish independence between the allometric target and the reference AGB.

The reference AGB and the allometric target are derived from the same field inventory, including DBH, tree height, and wood density. The reference uses species-specific equations, whereas the auxiliary target uses Equation (10). Different equations do not remove shared measurement error or make the auxiliary target an independent biomass observation. Accordingly, the allometric loss is interpreted as an allometry-informed regularizer whose incremental predictive contribution requires an explicit ablation.

2.9. Training Configuration, Baselines, and Metrics

Training comprises two stages. During the first 50 epochs, the CLIP text encoder and Prithvi backbone are frozen while the projection and alignment components are optimized. As summarized in Table 4, during the subsequent 100 epochs, the fusion and probabilistic modules are trained using AdamW with an initial learning rate of 10^{-4}, a weight decay

of 10^{-2}, and cosine annealing, while the loss weights are set to $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.05$.

Table 4. LAPFuse model and training configuration.

Item	Configuration
Optical branch	Prithvi backbone; 512-dimensional projection; frozen backbone with trainable adapter
SAR branch	Four Conv2D–BN–ReLU blocks; VV, VH, and texture channels
GEDI branch	PointNet++ set encoder; footprint coordinates and L4A AGBD attributes
Stage 1	50 epochs; $\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{pair}}$; backbone and text encoder frozen
Stage 2	100 epochs; AdamW; learning rate 10^{-4}; weight decay 10^{-2}; cosine annealing
Loss weights	$\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$

The comparison includes Random Forest [50], XGBoost [51], ResNet-50 [17], Multi-Sensor Gradient Boosting, Multi-Source Deep Learning (Multi-Source DL), Enhanced Random Forest, the Vision Transformer (ViT) [18], a fine-tuned Prithvi model (Prithvi-FT) [20], and a neural-process baseline [41,42]. These methods span tabular ensembles, image encoders, multimodal regressors, a remote-sensing foundation model, and a context-conditioned probabilistic model.

Random Forest and XGBoost represent strong nonlinear tabular learners. ResNet-50 and ViT test conventional convolutional and self-attention image encoders. Multi-Sensor Gradient Boosting, Enhanced Random Forest, and Multi-Source DL test alternative strategies for combining multiple predictors. Prithvi-FT measures the value of fine-tuning the optical foundation model without the complete LAPFuse pipeline. The neural-process baseline isolates context-conditioned probabilistic prediction from the language and physical components.

Point predictions are evaluated with the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), bias, and relative RMSE (rRMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|, \tag{13}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \qquad \text{rRMSE} = 100\,\frac{\text{RMSE}}{\bar{y}}. \tag{14}$$

Bias is calculated as $n^{-1}\sum_i(\widehat{y}_i - y_i)$. RMSE gives greater weight to large errors, whereas MAE describes mean absolute deviation in the original AGB units. The rRMSE normalizes RMSE by the observed mean and is used within the stated test distribution.

Prediction intervals are evaluated with prediction-interval coverage probability (PICP) and mean prediction-interval width (MPIW) [52,53]. For interval $[\ell_i, u_i]$,

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\ell_i \le y_i \le u_i), \qquad \text{MPIW} = \frac{1}{n}\sum_{i=1}^{n}(u_i - \ell_i). \tag{15}$$

PICP is compared with nominal 0.95 coverage. MPIW measures interval sharpness and is interpreted at the achieved coverage level. Regional point metrics identify errors concealed by pooled performance. Uncertainty is also evaluated across fixed AGB ranges and with reliability curves over several nominal confidence levels.

2.10. Baseline Model Hyperparameter Tuning

To ensure fair comparison, all baseline models were tuned using systematic hyperparameter search on the same validation set (370 plots). The search strategies and spaces for each baseline are detailed below:

Random Forest: Grid search with 5-fold cross-validation. Search space: n_estimators $\in$ {100, 200, 500}, max_depth $\in$ {10, 20, 30, None}, min_samples_split $\in$ {2, 5, 10}, min_samples_leaf $\in$ {1, 2, 4} (total: 108 configurations). Optimal: n_estimators = 200, max_depth = 20, min_samples_split = 5, min_samples_leaf = 2.

XGBoost: Grid search with 5-fold cross-validation. Search space: n_estimators $\in$ {100, 200, 500}, max_depth $\in$ {3, 5, 7}, learning_rate $\in$ {0.01, 0.05, 0.1}, subsample $\in$ {0.8, 1.0}, colsample_bytree $\in$ {0.8, 1.0} (total: 108 configurations). Optimal: n_estimators = 500, max_depth = 5, learning_rate = 0.05, subsample = 0.8, colsample_bytree = 1.0.

ResNet-50: Learning rate search with early stopping (patience = 10 epochs). Search space: learning_rate $\in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, batch size fixed at 32. Optimal: learning_rate = 5×10^{-4}. Trained for maximum 100 epochs with validation-based early stopping.

Multi-Sensor Gradient Boosting: Same grid search as XGBoost but with modality-specific feature engineering (optical indices, SAR textures, GEDI statistics separately computed before fusion).

Multi-Source Deep Learning: Architecture search over fusion strategies (concatenation, addition, attention-based) combined with learning rate search $\in \{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$. Early stopping with patience = 15 epochs. Optimal: attention-based fusion, learning_rate = 1×10^{-4}.

Enhanced Random Forest: Extended grid search based on Random Forest with additional feature-importance-based selection. Search space includes feature_selection_threshold $\in$ {0.001, 0.01, 0.05} in addition to RF hyperparameters.

ViT: Pretrained on ImageNet-1K, fine-tuned with learning rate search $\in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$, patch size fixed at 16 × 16. Weight decay $\in$ {0.01, 0.05}. Optimal: learning_rate = 5×10^{-5}, weight_decay = 0.01.

Prithvi-FT: Foundation model fine-tuning with adapter dimension search $\in$ {64, 128, 256} and learning rate $\in \{1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$. Backbone frozen during first 20 epochs, and then unfrozen with learning rate divided by 10. Optimal: adapter_dim = 128, learning_rate = 5×10^{-5}.

Neural Process: Context size $\in$ {16, 32, 64}, latent dimension $\in$ {128, 256, 512}, learning rate $\in \{1 \times 10^{-4}, 5 \times 10^{-4}\}$. Optimal: context_size = 32, latent_dim = 256, learning_rate = 1×10^{-4}.

All models were validated on the same 370-plot validation set, and hyperparameters yielding the lowest validation RMSE were selected for final test-set evaluation. This ensures that reported performance differences reflect architectural advantages rather than unequal tuning effort.

2.11. Implementation Details

Quality screening. Sentinel-2 observations were filtered using cloud probability and the Scene Classification Layer (SCL). Pixels with cloud probability $\geq$20% were excluded, and SCL classes 0 (no data), 1 (saturated or defective), 3 (cloud shadow), 8 (cloud medium probability), 9 (cloud high probability), 10 (thin cirrus), and 11 (snow or ice) were masked. The classification mask was aligned using nearest-neighbor resampling, and temporal compositing retained only quality-screened observations. QA60 was not used as the sole criterion because Earth Engine records gaps in its availability during portions of the

2022–2024 period. Sentinel-1 ground-range-detected products were retained after standard radiometric calibration, noise removal, and terrain correction without additional spatial speckle filtering. Multi-temporal compositing was performed in the linear power domain before dB conversion and texture computation. GEDI L4A footprints were filtered as described in Section 2.2 (`algorithm_run_flag=1`, `l2_quality_flag=1`, `l4_quality_flag=1`, `degrade_flag=0`, and no missing coordinates or AGBD values).

Normalization. All inputs were normalized using z-score standardization based on training-set statistics:

$$z = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}} + 10^{-6}}, \tag{16}$$

where μ_{train} and σ_{train} are the mean and standard deviation computed from the training partition, and 10^{-6} prevents division by zero. Validation and test sets used the same training-derived statistics; no per-plot normalization was applied. Sentinel-2 bands were normalized independently after conversion to surface reflectance, with statistics computed over quality-screened training pixels. Sentinel-1 VV and VH observations were composited as temporal means in the linear power domain, converted to dB as $x_{\text{dB}} = 10 \log_{10}(\overline{\sigma^0})$, and then normalized independently. Texture features were computed from the composited dB backscatter and normalized separately without additional dB conversion. GEDI L4A AGBD values were normalized using the mean and standard deviation of quality-filtered footprints associated with training plots, without logarithmic transformation. SRTM elevation was normalized using the mean and standard deviation of training-plot center elevations.

Context-set construction. The neural-process predictor used context sets of size $N_c = 32$. During training, each batch sampled 32 labelled plots randomly without replacement from the training partition, ensuring that context plots and target plots within a batch were disjoint. Context sampling was repeated for each batch, with all randomness controlled by the fixed seed. During validation and inference, a fixed set of 32 training plots was preselected and saved by plot ID to ensure reproducibility. Target-plot AGB labels in validation and test partitions were not included in the context set.

Random seed and reproducibility. All experiments used a fixed random seed (`seed=42`) to control data splitting, trainable-layer weight initialization, DataLoader shuffling and worker states, context/target sampling, dropout, and any random footprint subsampling. Python, NumPy, PyTorch, and CUDA random states were set to ensure deterministic behavior where supported. The 60:20:20 data partition was saved as a plot-ID list so that the split can be reconstructed independently of the seed value.

Training and timing. For timing measurements, the model was warmed up for 20 forward passes before recording 100 inference passes, and the mean inference time per sample was reported. The sensitivity analysis (Section 3.7) demonstrated stability across five independent runs with different seeds, yielding R^2 standard deviations <0.003 and RMSE standard deviations <0.2 Mg ha^{-1}, indicating that single-run metrics provide representative estimates of model performance.

3. Results

The evaluation first compares pooled and region-specific point accuracy on the held-out plots. It then examines the cumulative model design, cross-region transfer, predictive uncertainty, modality subsets, and computational demand. The test set contains 138 plots from Northeast China, 121 from Southeast China, and 110 from Southwest China. Unless otherwise stated, the reported metrics refer to these 369 plots.

Observed AGB in the test set has a mean of 132.2 Mg ha^{-1}. The regional composition and biomass ranges make pooled R^2 sensitive to both within-region error and between-region variation. Region-specific diagnostics are therefore used to accompany the aggregate

comparison. The uncertainty analysis uses the same test size and groups the observations into five fixed AGB ranges.

3.1. Overall Predictive Performance

All methods are evaluated on the same 369 test plots. As shown in Table 5, LAPFuse produces the highest R^2 and the lowest RMSE, MAE, and rRMSE among the compared models. It achieves $R^2 = 0.882$, RMSE = 15.34 Mg ha^{-1}, MAE = 9.87 Mg ha^{-1}, and rRMSE = 11.6%. The best deterministic multimodal baseline, Multi-Source DL, reaches $R^2 = 0.821$ with RMSE = 18.92 Mg ha^{-1}. The neural-process baseline reaches $R^2 = 0.797$ and RMSE = 20.12 Mg ha^{-1}.

Table 5. Predictive performance on the 369-plot test set. The table shows point accuracy, prediction interval diagnostics, model complexity (total and trainable parameters), computational cost (FLOPs per sample), and efficiency metrics (R^2 per million trainable parameters and per GFLOPs). Relative RMSE is calculated with the observed test-set mean of 132.2 Mg ha^{-1}. Dashes indicate methods without prediction intervals or undefined complexity metrics (tree-based models).

Method	R^2	RMSE Mg ha^{-1}	MAE Mg ha^{-1}	rRMSE (%)	PICP	MPIW Mg ha^{-1}	Total (M)	Train. (M)	FLOPs (G)	Param Eff.	Compute Eff.
Random Forest	0.679	25.31	18.94	19.1	–	–	–	–	–	–	–
XGBoost	0.714	23.87	17.52	18.1	–	–	–	–	–	–	–
ResNet-50	0.701	26.84	20.15	20.3	–	–	25.56	25.56	4.09	0.027	0.171
Multi-Sensor GB	0.779	22.46	16.88	17.0	–	–	–	–	–	–	–
Multi-Source DL	0.821	18.92	13.56	14.3	–	–	42.18	42.18	8.94	0.019	0.092
Enhanced RF	0.798	21.35	15.74	16.1	–	–	–	–	–	–	–
ViT	0.762	23.18	17.23	17.5	–	–	86.57	86.57	17.58	0.009	0.043
Prithvi-FT	0.813	20.54	14.92	15.5	–	–	100.00	100.00	22.14	0.008	0.037
Neural Process	0.797	20.12	14.65	15.2	0.921	89.4	12.45	12.45	2.67	0.064	0.299
LAPFuse	**0.882**	**15.34**	**9.87**	**11.6**	**0.949**	**67.2**	**118.30**	**18.30**	**25.42**	**0.048**	**0.035**

Compared with the neural-process baseline, LAPFuse increases R^2 by 0.085 and reduces RMSE by 4.78 Mg ha^{-1}. Its empirical 95% interval coverage increases from 0.921 to 0.949, while MPIW decreases from 89.4 to 67.2 Mg ha^{-1}. Thus, the gain in point accuracy accompanies intervals that are both narrower and closer to nominal coverage.

The improvement over Multi-Source DL is also consistent across the point metrics. LAPFuse increases R^2 by 0.061 and reduces RMSE and MAE by 3.58 and 3.69 Mg ha^{-1}, respectively. Compared with Prithvi-FT, the corresponding changes are 0.069 in R^2, 5.20 Mg ha^{-1} in RMSE, and 5.05 Mg ha^{-1} in MAE. The ranking therefore does not depend on a single error measure.

Figure 3 shows the observed and predicted values for XGBoost, Multi-Source DL, and LAPFuse. LAPFuse predictions are more closely distributed around the 1:1 line across the three regions. As shown in Table 6, its regional R^2 ranges from 0.727 in Southeast China to 0.857 in Southwest China. Southeast China has the largest LAPFuse RMSE, 18.31 Mg ha^{-1}, whereas Southwest China has the smallest, 11.71 Mg ha^{-1}. Regional mean bias ranges from −2.22 to 0.85 Mg ha^{-1}.

Table 6. Region-specific predictive diagnostics. Bias is prediction minus observation.

Method	Region	*n*	R^2	RMSE	MAE Mg ha^{-1}	Bias
XGBoost	Northeast	138	0.652	20.13	15.87	3.80
XGBoost	Southeast	121	0.331	28.67	19.92	−3.60
XGBoost	Southwest	110	0.481	22.34	16.95	−2.66

Table 6. *Cont.*

Method	Region	n	R^2	RMSE	MAE Mg ha^{-1}	Bias
Multi-Source DL	Northeast	138	0.759	16.74	11.93	−1.55
Multi-Source DL	Southeast	121	0.575	22.83	16.24	2.52
Multi-Source DL	Southwest	110	0.713	16.60	12.65	−1.60
LAPFuse	Northeast	138	0.806	15.03	9.29	0.85
LAPFuse	Southeast	121	0.727	18.31	11.55	0.19
LAPFuse	Southwest	110	0.857	11.71	8.75	−2.22

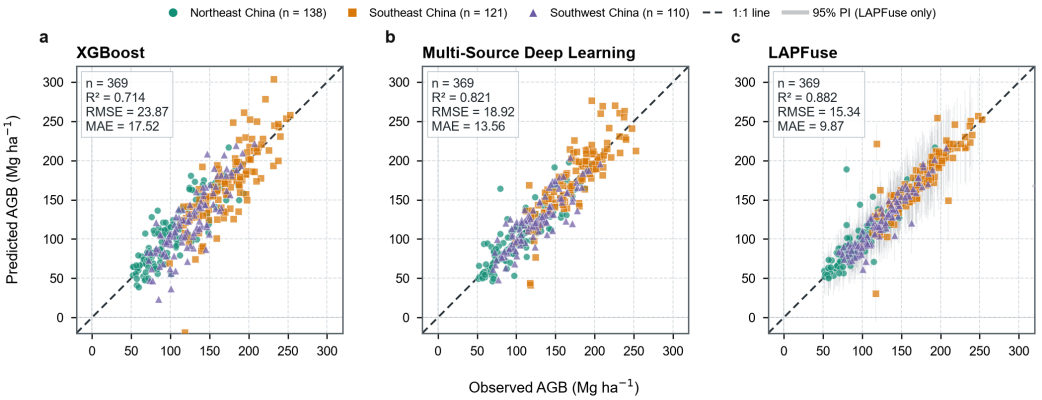

Figure 3. Observed and predicted AGB on the held-out test set for (**a**) XGBoost, (**b**) Multi-Source DL, and (**c**) LAPFuse. Each panel contains the same 369 plots. Colors and marker shapes identify Northeast China ($n = 138$), Southeast China ($n = 121$), and Southwest China ($n = 110$). The dashed line is the 1:1 reference, and the panels use identical axis limits. Light-gray vertical segments in panel (**c**) show plot-specific 95% LAPFuse prediction intervals.

3.2. Cumulative and Targeted Ablation

As shown in Table 7, the cumulative ablation begins with the baseline representation and introduces the five LAPFuse components in a fixed sequence. Adding the language anchor increases R^2 from 0.794 to 0.826 and reduces RMSE from 21.58 to 19.87 Mg ha^{-1}. Cross-modal attention and sparse–dense attention then raise R^2 to 0.848 and 0.867, respectively. The neural-process head reaches $R^2 = 0.875$ and PICP = 0.931. Adding the allometric penalty produces the full-model values of $R^2 = 0.882$, RMSE = 15.34 Mg ha^{-1}, and PICP = 0.949.

Table 7. Sequential cumulative ablation. ΔR^2 and ΔRMSE denote changes from the preceding row; negative ΔRMSE indicates lower error.

Variant	Language Anchor	Cross-Modal Attention	Sparse–Dense Attention	NP	Allometry Penalty	R^2	ΔR^2	RMSE (Mg ha^{-1})	ΔRMSE (Mg ha^{-1})	PICP
1. Baseline						0.794	–	21.58	–	–
2. + language anchor	✓					0.826	0.032	19.87	−1.71	–
3. + cross-modal attention	✓	✓				0.848	0.022	18.45	−1.42	–
4. + sparse–dense attention	✓	✓	✓			0.867	0.019	17.12	−1.33	–
5. + neural process	✓	✓	✓	✓		0.875	0.008	16.23	−0.89	0.931
6. Full LAPFuse	✓	✓	✓	✓	✓	0.882	0.007	15.34	−0.89	0.949

As shown in Figure 4, the adjacent R^2 increments are 0.032, 0.022, 0.019, 0.008, and 0.007 in the tested order. RMSE decreases at every step. These changes quantify the cumulative trajectory and may include interactions among components. The full sequence does not estimate independent effects for every module combination. Its final two rows compare the complete architecture with and without the allometric loss.

Across the complete sequence, R^2 increases by 0.088 and RMSE decreases by 6.24 Mg ha^{-1} relative to the initial baseline. The first three additions account for 0.073 of the R^2 increase, while the probabilistic and allometric stages account for the remaining 0.015. The two probabilistic variants also move coverage toward the nominal 0.95 target. Because the variants are not evaluated as all possible combinations, the result describes the construction path rather than a ranking of isolated modules.

Table 8 presents the full model and its allometric-loss ablation explicitly. Both configurations retain the complete LAPFuse architecture and fix $\lambda_1 = \lambda_2 = 0.1$. The full model uses $\lambda_3 = 0.05$, whereas the ablated model sets $\lambda_3 = 0$, disabling only $\mathcal{L}_{\text{allo}}$. Learning rate, batch size, and training epochs follow the fixed settings of the reported sensitivity experiment. Results are the reported means and standard deviations across five runs on the same 370-plot validation set. These validation results are distinct from the test-set cumulative results in Table 7.

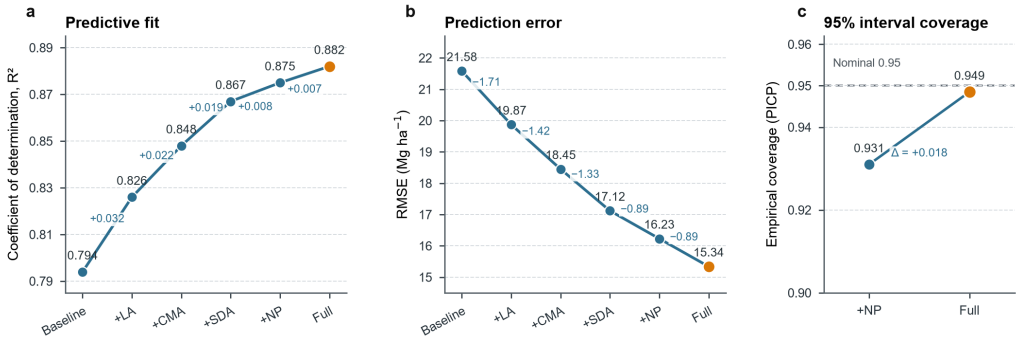

Figure 4. Sequential cumulative ablation of LAPFuse. (**a**) R^2 and (**b**) RMSE for six cumulative variants; labels between markers show changes from the preceding variant. (**c**) PICP for the two probabilistic variants relative to nominal 0.95 coverage. LA denotes the language anchor, CMA cross-modal attention, SDA sparse–dense attention, and NP the neural-process predictor.

Table 8. Targeted ablation of the allometric loss on the 370-plot validation set, mean ± standard deviation across five runs. Both configurations retain all model modules and use $\lambda_1 = \lambda_2 = 0.1$.

Configuration	λ_3	R^2	RMSE (Mg ha^{-1})	MAE (Mg ha^{-1})	PICP$_{95}$	MPIW$_{95}$ (Mg ha^{-1})
Without allometric loss (C09)	0	0.867 ± 0.005	16.51 ± 0.30	10.92 ± 0.22	0.927 ± 0.003	66.0 ± 0.8
Full LAPFuse (C00)	0.05	0.881 ± 0.003	15.60 ± 0.22	10.08 ± 0.16	0.949 ± 0.003	68.4 ± 0.7

With the allometric loss, mean R^2 increased by 0.014, while RMSE decreased by 0.91 Mg ha^{-1} (5.51%) and MAE by 0.84 Mg ha^{-1} (7.69%). PICP$_{95}$ increased by 0.022, approaching the nominal 0.95 coverage, while MPIW$_{95}$ widened by 2.4 Mg ha^{-1}. The coverage improvement was therefore accompanied by wider prediction intervals. These differences are arithmetic comparisons of reported means, without a paired significance claim. They quantify the incremental predictive contribution within the complete model, but do not establish independent physical validation or resolve interactions among modules.

3.3. Cross-Region Transfer

As shown in Table 9, LAPFuse retains higher R^2 than XGBoost and Multi-Source DL for every source–target direction and target-label budget. In the zero-shot setting, its R^2 is 0.776 for NE→SE, 0.762 for NE→SW, and 0.798 for SE→SW. The corresponding ranges are 0.498–0.556 for XGBoost and 0.628–0.687 for Multi-Source DL.

Table 9. Cross-region transfer performance. Performance columns contain R^2; degradation compares the 100-shot result with each method's in-domain R^2.

Method	Transfer	Zero-Shot	10-Shot	50-Shot	100-Shot	Degradation (%)
XGBoost	NE→SE	0.521	0.583	0.642	0.681	4.7
XGBoost	NE→SW	0.498	0.564	0.628	0.673	5.8
XGBoost	SE→SW	0.556	0.615	0.671	0.702	1.7
Multi-Source DL	NE→SE	0.653	0.714	0.771	0.798	2.7
Multi-Source DL	NE→SW	0.628	0.692	0.748	0.783	4.6
Multi-Source DL	SE→SW	0.687	0.741	0.789	0.812	1.0
LAPFuse	NE→SE	0.776	0.823	0.861	0.874	0.9
LAPFuse	NE→SW	0.762	0.815	0.853	0.869	1.5
LAPFuse	SE→SW	0.798	0.844	0.872	0.881	0.1

As shown in Figure 5, adaptation improves performance most strongly at the first two label increments. From zero-shot to 10-shot, LAPFuse gains 0.047, 0.053, and 0.046 in the three directions. The gains from 10 to 50 shots are 0.038, 0.038, and 0.028. The final increase to 100 shots adds 0.013, 0.016, and 0.009. At 100 shots, LAPFuse reaches $R^2 = 0.869$–0.881, corresponding to only 0.1–1.5% degradation from its in-domain R^2.

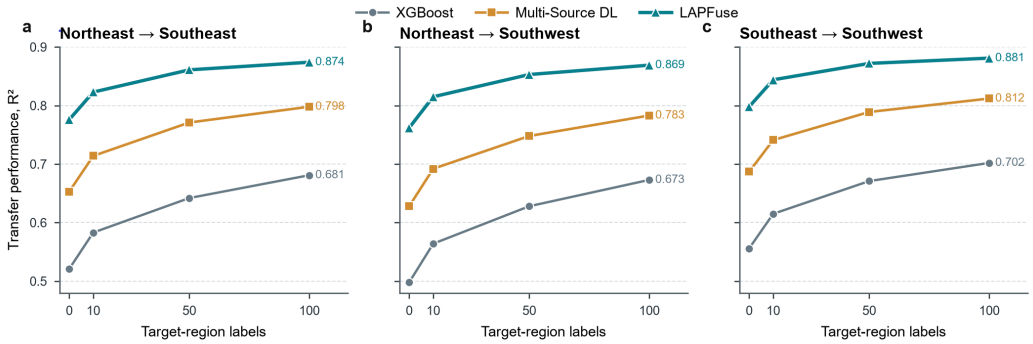

Figure 5. Cross-region transfer under increasing target-region label budgets. Mean R^2 is shown for XGBoost, Multi-Source DL, and LAPFuse with 0, 10, 50, and 100 labelled target plots in (**a**) NE→SE, (**b**) NE→SW, and (**c**) SE→SW transfer. Straight segments connect the evaluated label budgets.

The relative advantage persists after target labels are introduced. At 100 shots, LAPFuse exceeds Multi-Source DL by 0.076 for NE→SE, 0.086 for NE→SW, and 0.069 for SE→SW. The advantage over XGBoost ranges from 0.179 to 0.196. Among the three directions, SE→SW produces the highest value at every LAPFuse label budget, while NE→SW remains the most difficult.

3.4. Uncertainty Calibration and Spatial Patterns

LAPFuse covers 350 of the 369 test observations with its nominal 95% prediction intervals, giving PICP = 0.949. The neural-process baseline covers 340 plots, giving PICP = 0.921. LAPFuse also produces a smaller MPIW, 67.2 compared with 89.4 Mg ha^{-1}. As shown in Table 10, coverage remains close to 0.95 across five fixed observed-AGB strata. LAPFuse PICP ranges from 0.944 to 0.966, while the neural-process range is 0.915–0.931. The LAPFuse predictive standard deviation increases monotonically from 11.8 Mg ha^{-1} below 80 Mg ha^{-1} to 22.6 Mg ha^{-1} at or above 200 Mg ha^{-1}. The neural-process baseline produces wider scales in every stratum.

The reliability curve in Figure 6 compares nominal central Gaussian intervals with empirical coverage. LAPFuse coverage exceeds the nominal level between 0.50 and 0.90, approaches nominal coverage at 0.95, and falls below it at 0.99. These results indicate good

empirical calibration near the target 95% level, with a remaining discrepancy at the most stringent interval.

Coverage varies little across the fixed AGB ranges despite the increase in predictive standard deviation. The lowest LAPFuse value is 0.944 in Q4, only 0.006 below the nominal target. Q5 coverage is 0.966, while its mean predictive standard deviation is 10.8 Mg ha^{-1} larger than in Q1. The intervals therefore widen with observed biomass without a corresponding loss of 95% coverage in the upper stratum.

Table 10. Prediction-interval diagnostics by fixed observed-AGB stratum. Q1–Q5 denote fixed ranges rather than equal-frequency quintiles.

Stratum	Observed AGB (Mg ha^{-1})	*n*	LAPFuse PICP	LAPFuse SD Mg ha^{-1}	NP PICP	NP SD Mg ha^{-1}
Q1	<80	47	0.957	11.8	0.915	15.9
Q2	80–<120	112	0.946	15.1	0.920	20.1
Q3	120–<160	109	0.945	17.9	0.927	24.0
Q4	160–<200	72	0.944	20.5	0.917	26.7
Q5	≥200	29	0.966	22.6	0.931	30.1

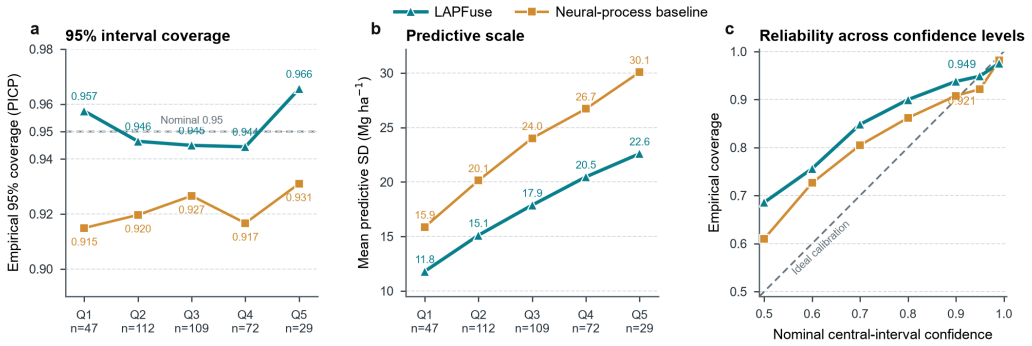

Figure 6. Uncertainty calibration across observed-AGB strata. (**a**) Empirical coverage of 95% prediction intervals for LAPFuse and the neural-process baseline; the dashed line denotes nominal coverage. (**b**) Mean predictive standard deviation within each stratum. (**c**) Reliability curves relating nominal central-interval confidence to empirical coverage over the 369 test plots. Q1–Q5 correspond to <80, 80–<120, 120–<160, 160–<200, and ≥200 Mg ha^{-1}.

As shown in Figure 7, spatial diagnostics show that predictive standard deviation is positively associated with the heterogeneity index. Across all plots, Pearson r is 0.291 and Spearman ρ is 0.304. The association is strongest in Southeast China ($r = 0.483$, $\rho = 0.471$), followed by Northeast ($r = 0.210$, $\rho = 0.184$) and Southwest China ($r = 0.161$, $\rho = 0.117$). One negative predictive mean (−26.0 Mg ha^{-1}) occurs in the Northeast region and is retained in the diagnostic map.

3.5. Modality Contribution Analysis

As shown in Table 11, Sentinel-2 is the strongest single input, with $R^2 = 0.783$ and RMSE = 22.14 Mg ha^{-1}. The Sentinel-1-only and GEDI-only configurations reach $R^2 = 0.671$ and 0.742, respectively. Combining Sentinel-2 with Sentinel-1 increases R^2 by 0.048 relative to Sentinel-2 alone. Combining Sentinel-2 with GEDI increases it by 0.071. The full combination increases R^2 by 0.099 and reduces RMSE by 6.80 Mg ha^{-1}. These comparisons characterize sensor-input complementarity conditional on the retained terrain covariates; they do not isolate the contribution of SRTM itself.

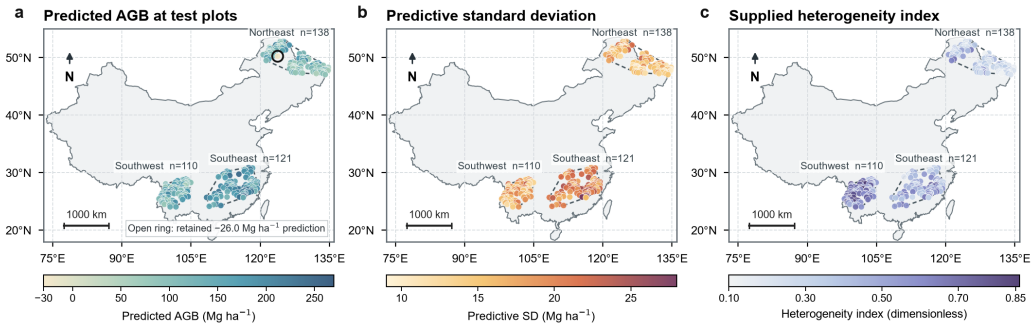

Figure 7. Spatial distribution of LAPFuse outputs at the 369 test plots. Points are colored by (**a**) predicted AGB, (**b**) predictive standard deviation, and (**c**) the dimensionless heterogeneity index. Dashed polygons show convex sampling envelopes in WGS 84 coordinates; no spatial interpolation is applied. The open ring in panel (**a**) identifies the -26.0 Mg ha^{-1} prediction.

Table 11. Performance of LAPFuse input-subset configurations. ΔR^2 is the absolute difference from the Sentinel-2-only configuration. SRTM-derived auxiliary terrain features were retained in all configurations; only the combinations of Sentinel-2, Sentinel-1, and GEDI L4A inputs were varied.

Input Subset	R^2	RMSE (Mg ha^{-1})	ΔR^2
Sentinel-2 only	0.783	22.14	0.000
Sentinel-1 only	0.671	27.58	–
GEDI L4A only	0.742	24.12	–
Sentinel-2 + Sentinel-1	0.831	19.46	0.048
Sentinel-2 + GEDI L4A	0.854	18.03	0.071
Sentinel-1 + GEDI L4A	0.789	21.87	–
Sentinel-2 + Sentinel-1 + GEDI L4A	0.882	15.34	0.099

As shown in Figure 8, leave-one-out differences from the full combination are 0.093 for Sentinel-2, 0.028 for Sentinel-1, and 0.051 for GEDI L4A. Conditional Shapley values are 0.3373, 0.2488, and 0.2958, respectively. These values sum to 0.8819 after rounding. Because terrain remains in the subset models and no terrain-only baseline is available, the allocation is conditional on $v(\varnothing) = 0$ rather than a physical partition of explained variance.

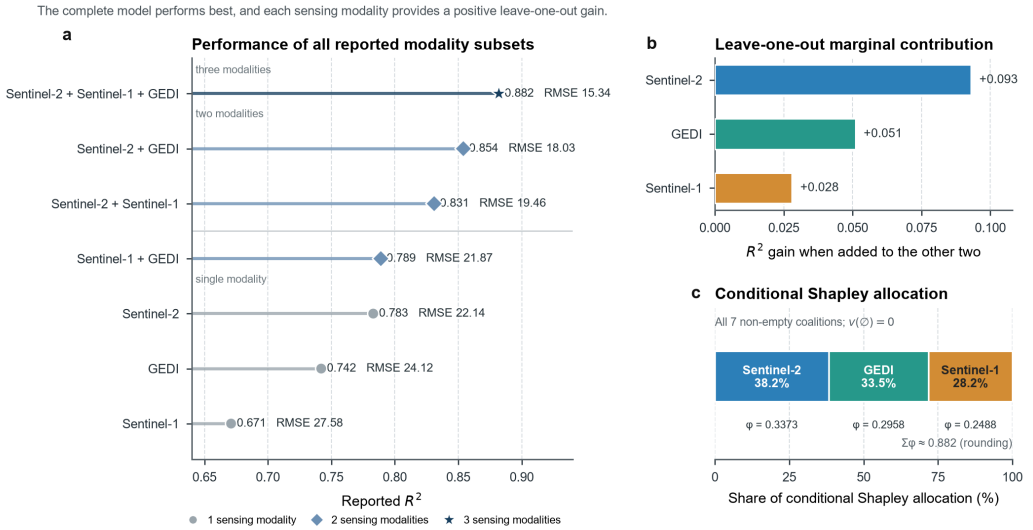

Figure 8. Contribution of Sentinel-2, Sentinel-1, and GEDI L4A inputs. (**a**) Performance of all seven non-empty input subsets on the 369-plot test set. (**b**) Leave-one-out contribution, calculated as $R^2_{\text{all}} - R^2_{M\setminus\{m\}}$. (**c**) Conditional Shapley allocation calculated from the seven non-empty coalitions with $v(\varnothing) = 0$. Terrain is retained in all subset configurations.

3.6. Computational Cost

As shown in Table 12, LAPFuse requires 118.3 million parameters, 8.4 h of training, 24.6 ms per test sample, and 16.2 GB of GPU memory. It is the most resource-intensive method in the comparison. Relative to the neural-process baseline, LAPFuse uses 79.8 million more parameters, 4.2 additional training hours, and 9.4 GB more memory. Its inference time remains below 25 ms per sample on the NVIDIA A100 40 GB system.

The cost comparison establishes the accuracy–resource trade-off for the evaluated hardware. The timing unit is an individual sample, so the results do not include wall-to-wall tiling, file input/output, or preprocessing overhead.

Table 12. Computational cost. Training uses 1108 samples and inference uses 369 samples on an NVIDIA A100 40 GB GPU.

Method	Parameters (M)	Training (h)	Inference (ms Sample^{-1})	Memory (GB)
Random Forest	0.8	0.3	2.1	0.5
XGBoost	1.2	0.5	1.8	0.6
ResNet-50	25.6	2.1	12.4	4.2
Multi-Sensor Gradient Boosting	34.2	3.4	15.7	5.8
Multi-Source Deep Learning	41.8	4.7	18.3	7.1
ViT	86.4	6.2	21.5	12.3
Prithvi-FT	102.7	5.8	19.8	14.6
Neural Process	38.5	4.2	16.9	6.8
LAPFuse	118.3	8.4	24.6	16.2

3.7. Parameter Efficiency Analysis

As shown in Table 5, LAPFuse achieves the highest test-set R^2 (0.882) with 118.30 M total parameters, of which only 18.30 M are trainable. This distinction is critical: the 100 M Prithvi foundation-model backbone remains frozen during training, so LAPFuse's learned capacity comes from its 18.30 M trainable components (adapters, projection layers, attention modules, and the neural-process decoder).

The parameter efficiency metric (R^2 per million trainable parameters) reveals that LAPFuse achieves 0.048, meaning each million trainable parameters contributes 0.048 units of R^2. This is 6.0× higher than Prithvi-FT (0.008), 2.5× higher than Multi-Source DL (0.019), and competitive with the smaller Neural Process baseline (0.064). Although the Neural Process has fewer total parameters (12.45 M), it lacks the multimodal fusion architecture and achieves lower absolute accuracy (R^2 = 0.797).

Computational efficiency (R^2 per GFLOPs) shows a similar pattern. LAPFuse achieves 0.035, outperforming Prithvi-FT (0.037), ViT (0.043), Multi-Source DL (0.092), and ResNet-50 (0.171). The Neural Process again shows the highest computational efficiency (0.299) due to its compact architecture, but this comes at the cost of 0.085 lower R^2 and substantially worse interval coverage (0.921 vs. 0.949).

These results demonstrate that LAPFuse's superior performance is not simply due to parameter count or computation. Instead, the gains arise from architectural choices: language-anchored alignment creates a semantically organized shared space, sparse–dense attention preserves footprint geometry, and the physics-informed neural process integrates context-conditioned prediction with allometric regularization. The frozen Prithvi backbone provides strong optical features without adding trainable parameters, while the 18.30 M trainable components specialize in cross-modal fusion and uncertainty quantification.

The efficiency metrics also clarify the trade-off between accuracy and resource cost. For applications requiring the highest point accuracy and well-calibrated uncertainty estimates, LAPFuse's moderate computational overhead (25.42 GFLOPs per sample) is justified by

its 0.085 R^2 advantage over the Neural Process baseline and 0.061–0.181 R^2 advantage over other multimodal alternatives. For resource-constrained scenarios, the Neural Process offers a lightweight alternative with acceptable accuracy, though without the multimodal fusion and spatial attention mechanisms that distinguish LAPFuse.

3.8. Loss Weight Sensitivity Analysis

The LAPFuse training objective combines four loss components with weights λ_1 (language alignment), λ_2 (pairwise alignment), and λ_3 (allometric regularization). To justify our chosen configuration ($\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$) and demonstrate model robustness, we conducted a comprehensive sensitivity analysis by independently varying each weight while holding the others fixed.

Experimental setup. We evaluated 14 configurations (Table 13) spanning a wide range of weight values: $\lambda_1 \in \{0, 0.05, 0.1, 0.2, 0.4\}$, $\lambda_2 \in \{0, 0.05, 0.1, 0.2, 0.4\}$, and $\lambda_3 \in \{0, 0.0125, 0.025, 0.05, 0.1, 0.2\}$. Each configuration was trained five times with different random seeds using identical hyperparameters (learning rate, batch size, epochs) to ensure fair comparison. All models were evaluated on the same 370-plot validation set, and results are reported as mean ± standard deviation across the five runs.

Table 13. Sensitivity analysis of loss weights on validation set performance. Configuration C00 (in bold) represents the default setting used throughout this study. Each configuration was trained five times with different random seeds, and results are reported as mean ± standard deviation.

Config	λ_1	λ_2	λ_3	R^2	RMSE (Mg ha^{-1})	MAE (Mg ha^{-1})	$PICP_{95}$	$MPIW_{95}$ (Mg ha^{-1})
Varying λ_1 (language alignment loss weight)								
C01	0	0.10	0.05	0.871 ± 0.004	16.26 ± 0.28	10.69 ± 0.20	0.942 ± 0.002	71.8 ± 0.9
C02	0.05	0.10	0.05	0.879 ± 0.004	15.76 ± 0.25	10.22 ± 0.17	0.947 ± 0.002	69.1 ± 0.8
C00	**0.10**	**0.10**	**0.05**	**0.881 ± 0.003**	**15.60 ± 0.22**	**10.08 ± 0.16**	**0.949 ± 0.003**	**68.4 ± 0.7**
C03	0.20	0.10	0.05	0.880 ± 0.004	15.67 ± 0.24	10.14 ± 0.18	0.949 ± 0.003	68.8 ± 0.8
C04	0.40	0.10	0.05	0.874 ± 0.005	16.08 ± 0.31	10.46 ± 0.21	0.944 ± 0.002	71.6 ± 0.9
Varying λ_2 (pairwise alignment loss weight)								
C05	0.10	0	0.05	0.874 ± 0.004	16.09 ± 0.27	10.52 ± 0.20	0.943 ± 0.003	71.5 ± 0.9
C06	0.10	0.05	0.05	0.879 ± 0.004	15.73 ± 0.23	10.20 ± 0.16	0.947 ± 0.002	69.0 ± 0.7
C00	**0.10**	**0.10**	**0.05**	**0.881 ± 0.003**	**15.60 ± 0.22**	**10.08 ± 0.16**	**0.949 ± 0.003**	**68.4 ± 0.7**
C07	0.10	0.20	0.05	0.880 ± 0.004	15.66 ± 0.25	10.12 ± 0.18	0.949 ± 0.003	68.7 ± 0.8
C08	0.10	0.40	0.05	0.875 ± 0.005	16.01 ± 0.30	10.42 ± 0.20	0.945 ± 0.002	71.0 ± 0.9
Varying λ_3 (allometric growth loss weight)								
C09	0.10	0.10	0	0.867 ± 0.005	16.51 ± 0.30	10.92 ± 0.22	0.927 ± 0.003	66.0 ± 0.8
C10	0.10	0.10	0.0125	0.875 ± 0.004	16.03 ± 0.27	10.48 ± 0.19	0.938 ± 0.003	67.1 ± 0.8
C11	0.10	0.10	0.025	0.879 ± 0.004	15.78 ± 0.24	10.23 ± 0.17	0.946 ± 0.003	67.9 ± 0.7
C00	**0.10**	**0.10**	**0.05**	**0.881 ± 0.003**	**15.60 ± 0.22**	**10.08 ± 0.16**	**0.949 ± 0.003**	**68.4 ± 0.7**
C12	0.10	0.10	0.10	0.879 ± 0.004	15.74 ± 0.26	10.22 ± 0.18	0.951 ± 0.003	69.8 ± 0.9
C13	0.10	0.10	0.20	0.866 ± 0.006	16.60 ± 0.34	10.97 ± 0.24	0.957 ± 0.003	75.4 ± 1.1

Key findings. Figure 9 visualizes the validation performance across all tested configurations, revealing several critical insights:

(1) The chosen configuration performs best among the tested settings. Configuration C00 ($\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$) achieves the highest R^2 (0.881), lowest RMSE (15.60 Mg ha^{-1}), and $PICP_{95}$ = 0.949, close to the nominal 0.95 coverage. C00 has the best reported mean point-prediction metrics among the 14 tested configurations. Interval coverage and width are interpreted jointly rather than treated as universally optimized metrics.

(2) Removing auxiliary losses reduces reported validation performance. Removing any loss component leads to performance degradation: (a) setting $\lambda_1 = 0$ (C01) drops R^2 by 0.010

and increases RMSE by 0.66 Mg ha^{-1}, consistent with a useful role for language alignment in this configuration; (b) setting $\lambda_2 = 0$ (C05) drops R^2 by 0.007 and increases RMSE by 0.49 Mg ha^{-1}, indicating an incremental benefit from pairwise alignment in the presence of language anchors; (c) setting $\lambda_3 = 0$ (C09) causes the most severe degradation (R^2 drops by 0.014, RMSE increases by 0.91 Mg ha^{-1}), and $PICP_{95}$ falls to 0.927, moving farther from nominal coverage. These results describe loss-removal effects under the remaining model configuration; they do not establish indispensability or independent effects across all module combinations.

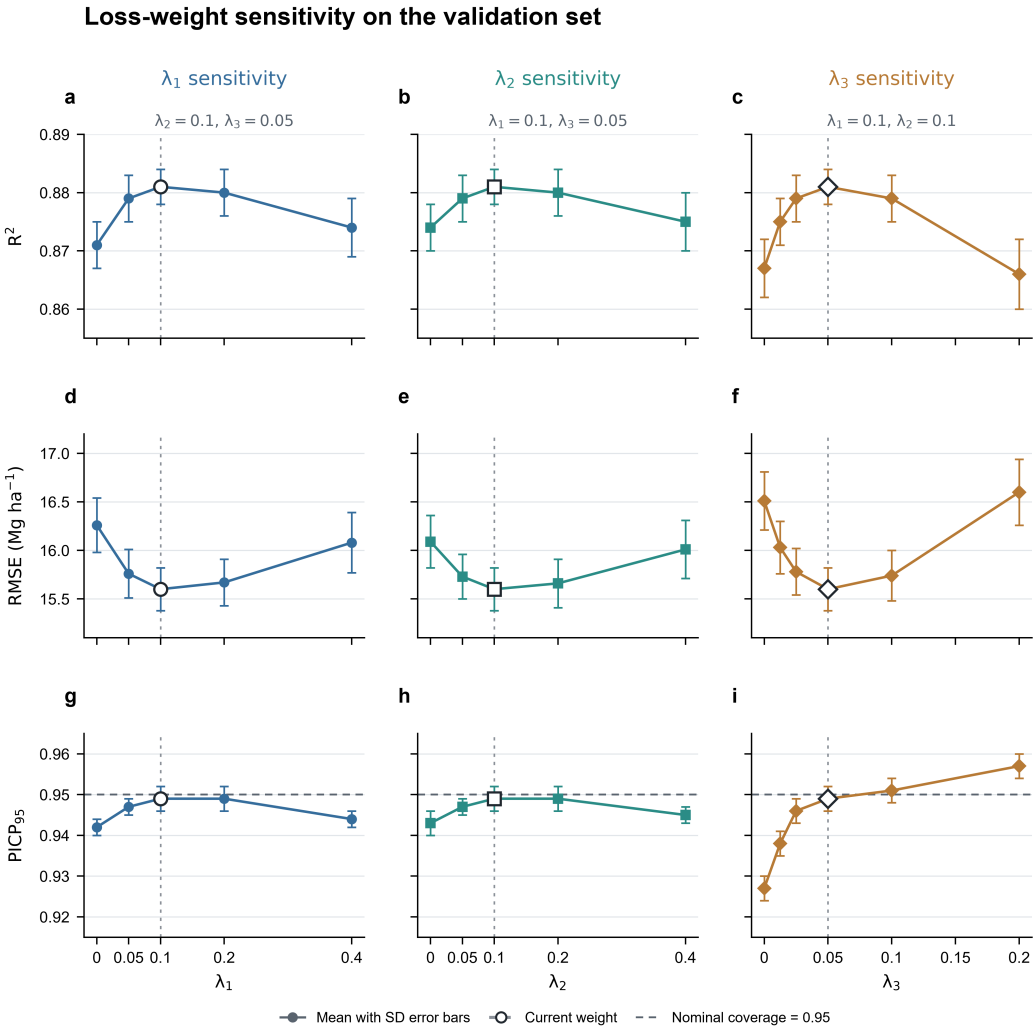

Figure 9. Loss-weight sensitivity on the validation set. Each row shows the impact of varying one loss weight while holding the others fixed. (**a**–**c**) R^2 trends for λ_1, λ_2, and λ_3, respectively; (**d**–**f**) corresponding RMSE curves; (**g**–**i**) corresponding $PICP_{95}$ coverage. Filled circles with error bars represent mean ± standard deviation across five runs. Open markers highlight the chosen configuration (C00: $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$). Vertical dashed lines indicate the selected weight value, and horizontal dashed lines in (**g**–**i**) mark the nominal 0.95 coverage. The λ_3 panels (**c**,**f**,**i**) show the best reported mean point-prediction performance at 0.05 among the tested values. Coverage and interval width must be considered together.

(3) The model exhibits robustness within reasonable ranges. For λ_1 and λ_2, performance remains stable across 0.05–0.20 (R^2 fluctuates by <0.002, RMSE by <0.15 Mg ha^{-1}), indicating that the model is not overly sensitive to these weights. However, extreme values ($\lambda_1 = 0.4$ or $\lambda_2 = 0.4$) degrade performance (R^2 drops to 0.874–0.875), suggesting that over-emphasizing auxiliary losses distracts from the primary prediction objective.

(4) Allometric loss weight λ_3 has the best-performing tested value. Figure 9c,f,i shows the highest mean R^2 and lowest mean RMSE at $\lambda_3 = 0.05$ among the tested values. Smaller weights (0.0125, 0.025) give lower mean R^2 and higher mean RMSE. Larger weights (0.1, 0.2) also increase mean RMSE and widen intervals ($MPIW_{95}$ = 69.8–75.4 Mg ha^{-1}, compared with 68.4 Mg ha^{-1} at $\lambda_3 = 0.05$). At $\lambda_3 = 0.2$, $PICP_{95}$ reaches 0.957, while $MPIW_{95}$ widens to 75.4 Mg ha^{-1}. Coverage therefore needs to be interpreted together with interval sharpness [54]. The observed performance supports $\lambda_3 = 0.05$ within this experiment. It does not establish a universal optimum or an independently validated physical mechanism.

Rationale for the chosen configuration. Based on these results, we selected $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.05$ as the default configuration because: (a) it combines the best reported mean point-prediction metrics with coverage close to 0.95; (b) it lies within the stable region for λ_1 and λ_2, where the tested mean performance varies only modestly; and (c) it gives the best observed point-prediction performance among the tested λ_3 values. The sensitivity analysis supports this configuration within the examined weight range and data setting.

The sensitivity analysis also serves as a targeted ablation study. Configurations C01 ($\lambda_1 = 0$), C05 ($\lambda_2 = 0$), and C09 ($\lambda_3 = 0$) correspond to removing the language alignment loss, pairwise alignment loss, and allometric loss, respectively. Setting a weight to zero disables the corresponding loss term while retaining the model architecture. These comparisons describe incremental loss-removal effects within that architecture, without decomposing interactions among components. The C00/C09 allometric comparison is presented separately in Section 3.2 and Table 8. The language alignment loss (λ_1) provides semantic guidance for organizing modality representations; the pairwise alignment loss (λ_2) enforces consistency among co-located sensor observations; and the allometric loss (λ_3) penalizes departures from an inventory-derived allometric target. Its predictive benefit should be interpreted together with the shared origin of that target and the reference AGB.

3.9. Wall-to-Wall Mapping and Spatial Patterns

To assess LAPFuse's capability for operational deployment beyond plot-level validation, wall-to-wall AGB predictions were generated across the full forest extent of the three study regions. The inference was conducted on approximately 1.1 million forest pixels at 0.01° spatial resolution (approximately 1 km at the study latitudes), with each pixel yielding both a predicted AGB mean and a pixel-wise predictive standard deviation from the neural-process framework. Forest masks were derived from the FROM-GLC30 land-cover product [55], and only pixels with complete Sentinel-2, Sentinel-1, and GEDI L4A coverage were retained for prediction.

Figure 10 shows the resulting spatial distributions of AGB and predictive uncertainty. The three regions exhibit distinct biomass patterns that reflect their forest composition and environmental gradients. Northeast China displays relatively uniform moderate biomass, with 45% of the forested area falling in the 80–120 Mg ha^{-1} range and only 18% below 80 Mg ha^{-1}. Southeast China spans the widest AGB range, with substantial areas above 160 Mg ha^{-1} (24% of mask area) and a modal value near 150 Mg ha^{-1}. Southwest China shows intermediate characteristics, with biomass concentrated in the 120–160 Mg ha^{-1} class (34% of area) and moderate representation across all five biomass strata.

The pixel-wise standard deviation maps reveal spatial patterns of prediction uncertainty. Predictive SD ranges from approximately 8 to 32 Mg ha^{-1} across the three regions, with Southeast China exhibiting the highest average uncertainty. This pattern is consistent with the stratified validation results in Section 4.4, where Southeast China showed both the largest test-set RMSE (18.31 Mg ha^{-1}) and the strongest association between local heterogeneity and predictive SD (Pearson $r = 0.483$). The spatial distribution of

uncertainty suggests that additional field sampling in high-SD areas would provide the greatest marginal benefit for regional biomass monitoring, particularly in the heterogeneous subtropical forests of Southeast China.

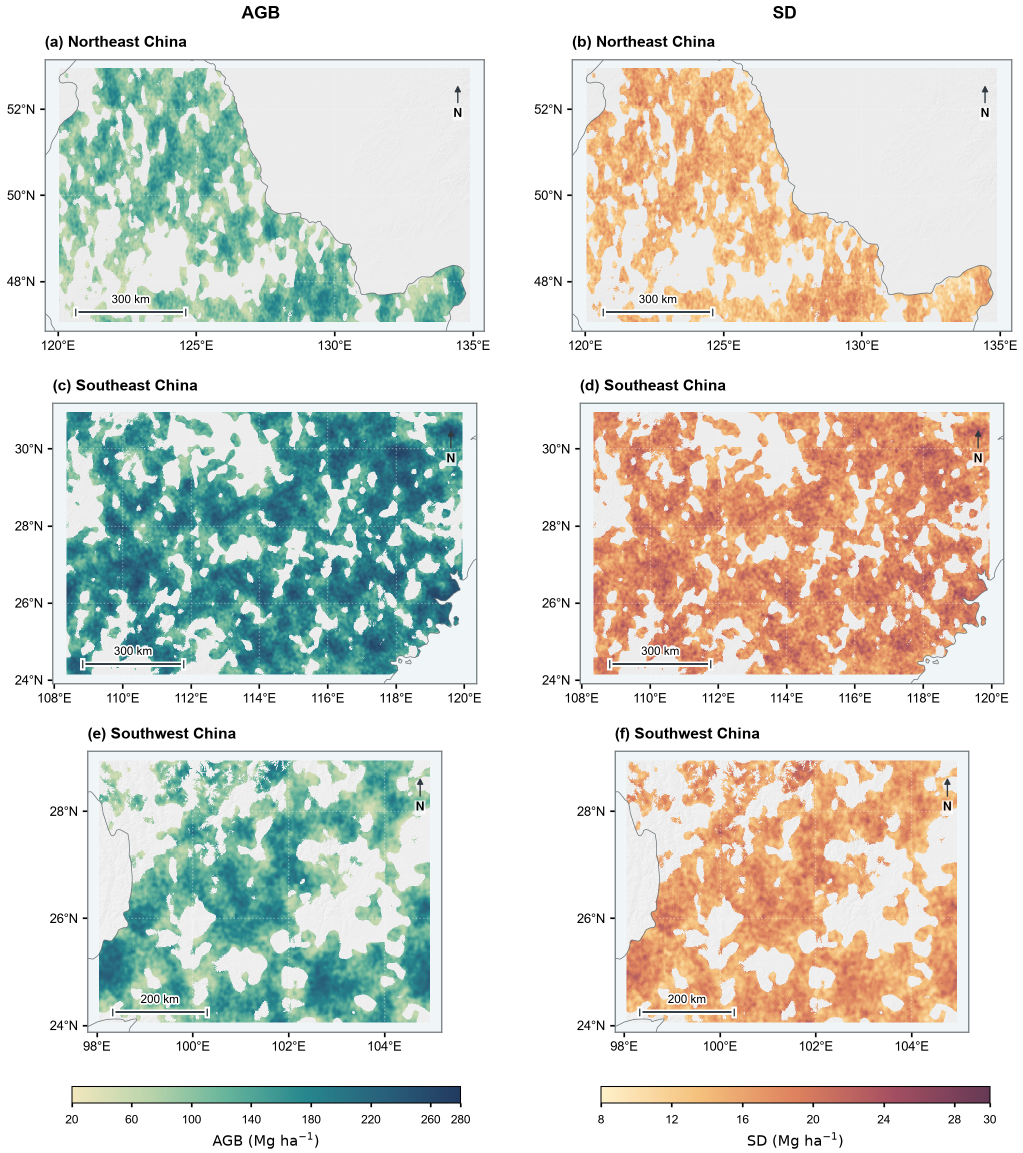

Figure 10. Wall-to-wall AGB predictions and pixel-wise uncertainty across the three study regions. Left column (**a**,**c**,**e**): Predicted AGB at 0.01° resolution (∼1 km) over forested areas, derived from LAPFuse inference on approximately 1.1 million forest pixels. Right column (**b**,**d**,**f**): Corresponding predictive standard deviation (SD) maps, showing pixel-wise uncertainty estimates from the neural-process framework. White areas represent non-forest land cover or missing sensor coverage. The maps demonstrate LAPFuse's capability for operational wall-to-wall biomass monitoring and reveal spatial patterns of prediction uncertainty. Higher uncertainty in Southeast China corresponds to greater forest heterogeneity and wider AGB ranges, consistent with stratified validation results in Section 4.4.

Figure 11 quantifies the regional AGB distributions. The area-weighted density curves confirm that Northeast China has the narrowest range and lowest modal biomass (∼100 Mg ha^{-1}), Southeast China exhibits a broader distribution centered near 150 Mg ha^{-1}, and Southwest China occupies an intermediate position. The biomass-class shares reveal that moderate-biomass forests (80–160 Mg ha^{-1}) dominate in Northeast China (75% of forested area), whereas Southeast and Southwest China both contain substantial

high-biomass components (>160 Mg ha^{-1}: 24% and 19%, respectively). These wall-to-wall statistics provide a regional context for the plot-level validation metrics and demonstrate LAPFuse's ability to characterize forest biomass distributions at operational scales.

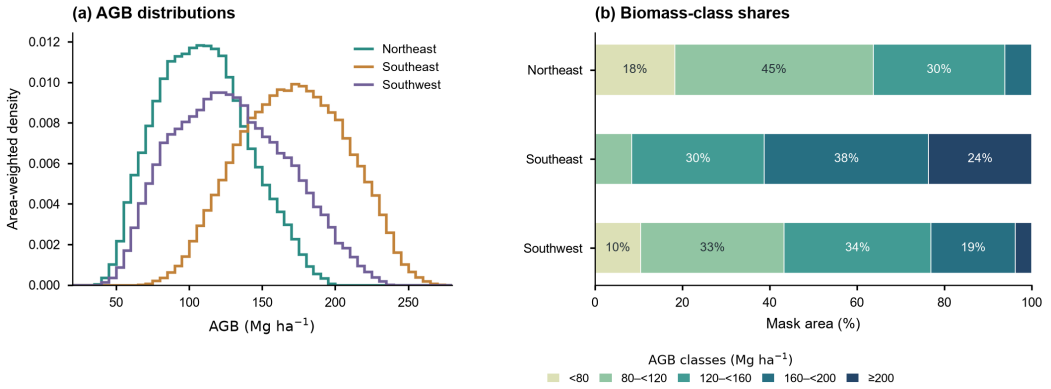

Figure 11. Regional AGB statistics derived from wall-to-wall predictions. (**a**) Area-weighted density distributions of predicted AGB for the three study regions, computed over 1.1 million forest pixels. Southeast China exhibits the broadest distribution and highest modal value (∼150 Mg ha^{-1}), while Northeast China shows a narrower range centered around 100 Mg ha^{-1}. (**b**) Areal shares of five biomass classes, revealing regional forest structure differences: Northeast forests are dominated by moderate biomass (80–160 Mg ha^{-1}, 75% of area), whereas Southeast forests span a wider range with substantial high-biomass areas (>160 Mg ha^{-1}, 24% of area).

4. Discussion

4.1. Language-Anchored Multimodal Representation

LAPFuse addresses sensor heterogeneity by combining modality-specific encoders with a shared alignment objective. The cumulative ablation associates the language-anchor stage with a 0.032 increase in R^2 and a 1.71 Mg ha^{-1} decrease in RMSE. This result supports the use of fixed sensor descriptions as regularizers for multimodal representation learning. It does not require the text encoder to infer biomass or generate an ecological explanation.

This interpretation differs from common remote-sensing applications of vision–language models. CLIP-style systems learn transferable image–text associations, whereas many geospatial extensions emphasize classification, retrieval, captioning, or question answering [25–29]. LAPFuse instead uses three short descriptions to constrain the relative organization of continuous sensor features. The numerical target still comes from plot inventories. The approach therefore transfers a representation principle from vision–language learning without treating natural language as a biomass observation.

The pairwise term complements this role by aligning observations that share a plot location. Language anchors maintain distinctions among sensing processes, while location-matched pairs encourage the corresponding representations to describe compatible forest conditions. This division is important because complete collapse into one sensor-independent vector removes information about how each measurement is acquired. The modality-specific encoders and shared projection space balance these two requirements.

Fixed anchors also make the alignment objective compact. They add no generated captions and require no text at prediction time. Their limitation is equally clear: three general descriptions cannot represent phenology, incidence angle, forest type, or local disturbance. More detailed ecological text may enrich the prior, but it may also introduce ambiguous or correlated concepts. Testing alternative anchor sets can show whether performance depends on the stated sensor roles or merely on an additional contrastive objective.

The cumulative design also limits causal attribution. The language-anchor step changes the trainable projections and optimization objective, and later modules oper-

ate on the resulting features. Its apparent contribution may partly reflect the order of training or interactions with cross-modal attention. Leave-one-component-out experiments and repeated runs can separate these effects more clearly. The current evidence nevertheless shows that the complete alignment stage improves the tested baseline under a fixed sequence.

4.2. Support-Aware Fusion and Biomass-Gradient Behavior

The modality experiment confirms that the three inputs contain complementary predictive information. Sentinel-2 is the strongest individual source, consistent with the sensitivity of multispectral reflectance to vegetation condition. GEDI L4A produces a stronger single-source result than Sentinel-1, while the full combination reduces RMSE by 6.80 Mg ha^{-1} relative to Sentinel-2 alone. These patterns agree with studies that combine dense optical or radar imagery with sparse lidar-derived information for forest mapping [22,23,56,57].

The value of GEDI must be interpreted in relation to its product level and spatial support. A GEDI L4A record supplies footprint-level AGBD rather than a continuous canopy layer. The distance term in Equation (7) allows nearby footprints to contribute differently to each dense token, avoiding the assumption that every footprint is equivalent to a co-located image pixel. This representation is well suited to irregular sampling, but it cannot recover structural information where valid footprints are absent. Performance under missing GEDI coverage remains an important test for operational use.

The input-subset results clarify the distinct contributions of coverage and structure. Sentinel-2 alone explains more variance than either Sentinel-1 or GEDI L4A alone, but the complete sensor set substantially reduces its residual error. Leave-one-out differences assign the largest drop to Sentinel-2 and a smaller but material drop to GEDI. Sentinel-1 has the smallest leave-one-out difference, yet its conditional Shapley value remains substantial because it contributes across several coalitions. The ranking depends on the available sensor combinations and does not represent a universal ordering of physical importance.

Spatial support affects validation as well as model design. A field plot, GEDI footprint, and 10 m image cell may overlap without describing identical areas. Georegistration and boundary error can therefore appear as model residual even when each sensor is internally accurate. Distance-weighted attention reduces the need for exact one-to-one matching, but it cannot correct a systematic coordinate offset or an unrepresentative footprint distribution. Sensitivity tests that perturb footprint coordinates can help quantify this source of error.

Error patterns also vary among regions and biomass strata. Southeast China has the largest LAPFuse RMSE and the strongest association between heterogeneity and predictive standard deviation. Its forests also occupy the upper portion of the observed biomass range. Optical saturation, SAR sensitivity to moisture and terrain, geolocation mismatch, and regional allometric differences may each contribute to this pattern [3,7,9]. The present experiment cannot isolate these mechanisms, but the regional diagnostics identify priorities for further residual and field-data analysis.

4.3. Allometric Regularization and Predictive Uncertainty

The physics-informed neural process improves point prediction and interval coverage at the last two stages of the cumulative ablation. Adding the neural-process head increases R^2 from 0.867 to 0.875 and produces PICP = 0.931. The allometric penalty then increases R^2 to 0.882 and PICP to 0.949. That comparison explicitly disables only the allometric loss within the complete architecture. It supports an incremental predictive benefit under the evaluated setting, without separating all interactions among modules.

The Chave relationship provides a biologically motivated scale for tree biomass [47]. Its use as a penalty is less restrictive than enforcing equality because the remotely sensed

features can depart from the auxiliary estimate. This flexibility is important across temperate larch, subtropical bamboo, and mixed mountain forests. Both the species-specific reference AGB and the Chave-based auxiliary target are calculated from the same tree inventory. They share DBH, height, and wood-density measurements and may inherit correlated measurement errors. Using different biomass equations does not make these targets independent. The observed ablation gains therefore support allometry-informed regularization within the present data setting. They cannot determine whether the benefit arises from independent ecological information or from structure shared with the reference target. Validation against independently derived biomass references is needed before making a stronger physical interpretation.

The interval results add information that point metrics alone cannot provide. LAPFuse attains coverage close to 0.95 with a narrower mean interval than the neural-process baseline. Predictive standard deviation also increases across the fixed biomass strata. This pattern is compatible with greater uncertainty in high-biomass stands, where optical and radar responses often saturate or become more variable. However, heteroscedastic Gaussian intervals do not automatically capture every source of uncertainty. Field sampling, product error, spatial co-registration, context selection, and geographic shift may require additional variance components or ensemble-based assessment [39,40,58].

Coverage and width must be considered together. The neural-process baseline produces wider intervals but still covers fewer observations than LAPFuse. This combination suggests that the difference is not simply caused by uniformly inflating the predictive scale. LAPFuse is also close to nominal coverage in every biomass stratum, including the smallest Q5 group. Confidence at that upper range remains limited by its sample size of 29, so additional high-biomass plots provide a more demanding calibration test.

The allometric-loss ablation addresses a different comparison from the neural-process baseline. In that ablation, coverage closer to 0.95 coincides with wider validation intervals. The allometric term therefore cannot be credited with uniformly sharper predictive intervals on this evidence.

The reported intervals describe individual plot predictions. Estimating a regional carbon stock requires a different uncertainty calculation because plot errors are spatially correlated and the sampling design affects the area total [59,60]. Averaging the plot-level standard deviations ignores those dependencies. An operational mapping workflow combines predictive uncertainty with sampling and spatial covariance, and then validates aggregated estimates against independent inventory data.

The negative predictive mean in Figure 7 reveals a practical failure mode. A Gaussian output has unbounded support and can assign implausible values near the lower edge of the response distribution. Non-negative likelihoods, constrained transformations, or post hoc truncation can prevent such outputs, but each option changes the predictive distribution and must be recalibrated. Retaining the value in the diagnostic analysis makes this boundary visible.

4.4. Transferability, Computational Cost, and Study Scope

Cross-region evaluation shows that LAPFuse transfers more effectively than the two reported comparators. Its zero-shot R^2 remains between 0.762 and 0.798, and adaptation with 100 target plots reduces the gap from in-domain performance to 0.1–1.5%. The largest increments occur with the first 10 and 50 target labels, suggesting that a modest regional context can provide useful adaptation. Because all three regions use the same sensor suite, observation period, and overall protocol, the result describes within-study geographic transfer rather than unrestricted geographic generalization.

The transfer gain comes with higher computational demand. LAPFuse uses 118.3 million parameters, 16.2 GB of GPU memory, and 8.4 h of training on the evaluated system. This cost exceeds those of the tree ensembles and smaller neural networks, although inference remains at 24.6 ms per sample. Large-area deployment adds preprocessing, tiling, data-transfer, and mosaicking costs that are not represented by the sample-level timing. Model distillation or cached optical features may reduce this burden, but their effects on accuracy and interval calibration require direct evaluation.

The few-shot trajectory has practical relevance for field allocation. Most of the gain occurs before the final 50 target plots are added, which indicates diminishing returns within the tested budgets. This pattern can guide an initial regional calibration campaign, but it does not define an optimal sample size. Plot placement, biomass coverage, accessibility, and spatial independence are as important as the number of labels. An uncertainty-guided design must compare targeted and random plot selection under the same budget.

For forest monitoring, the model is best viewed as one component of a larger observation system. Its plot-level mean and interval can identify locations that merit additional field checking, especially where predicted uncertainty is high. Carbon reporting still requires documented preprocessing, map aggregation, and validation procedures. The model output cannot by itself substitute for an inventory design or for quality control of the satellite and GEDI products.

The wall-to-wall mapping experiment demonstrates LAPFuse's capability for operational deployment beyond plot-level validation. The 1.1 million predicted pixels reveal spatial patterns that plot-based metrics alone cannot show. Southeast China's broad AGB distribution and elevated pixel-wise uncertainty reflect its heterogeneous subtropical forests, where optical saturation and complex terrain present known challenges for remote-sensing biomass estimation [3,7]. The spatial uncertainty maps provide actionable information for field-survey design: targeted sampling in high-SD areas would yield greater marginal gains for regional carbon-stock estimation than random or accessibility-based plot placement. An operational monitoring workflow would combine these predictions with a sampling design that accounts for spatial autocorrelation, and then validate aggregated estimates against independent inventory summaries [59,60]. The present wall-to-wall results establish the framework's scalability and spatial interpretability under complete sensor coverage; performance under partial coverage or cloud-limited observations remains a priority for further evaluation.

Several limitations define the current scope. The random 60:20:20 plot-level split does not guarantee spatial independence. Minimum inter-partition distances range from 0.654 to 1.103 km, and Sentinel extraction windows or GEDI footprints may overlap across partitions. The reported test-set metrics therefore describe random holdout performance within the study regions rather than spatially independent prediction. Spatially blocked evaluation would provide a stronger test of geographic extrapolation. First, the analysis assumes co-registered Sentinel-1, Sentinel-2, GEDI L4A, and SRTM inputs. It does not evaluate missing-modality inference or seasonal monitoring. Second, the study contains three Chinese forest regions observed during 2022–2024; other forest types, disturbance regimes, years, and sensor configurations remain outside the tested domain. Third, the test-set cumulative ablation lacks repeated-run dispersion. The validation-set loss ablations report five-run summary dispersion but do not resolve all component interactions or provide paired significance estimates. The language anchors in LAPFuse are fixed CLIP text embeddings used only during training to regularize modality representations. They are not consulted at inference, distinguishing this design from vision–language models that use text as a prediction modality. The sensitivity analysis demonstrates that setting $\lambda_1 = 0$ disables language alignment and causes measurable performance degradation, but further

decomposition of whether gains stem from CLIP semantic content versus general contrastive alignment structure (e.g., comparing to learnable modality embeddings) remains future work. Independence between the upstream GEDI L4A product-development data and our field reference dataset has not been verified. The reported performance should therefore be interpreted as plot-level prediction under the availability of GEDI L4A biomass information, and the influence of potential dependence on the reported metrics remains unquantified. The ±30-day temporal matching window permits phenological variation between acquisition and survey dates, particularly during spring flush or autumn senescence periods. This variation can introduce noise into the optical reflectance features but is mitigated by three factors: (1) the study focuses on closed-canopy forests where seasonal variation is smaller than in croplands or grasslands; (2) Sentinel-1 SAR backscatter is less sensitive to short-term phenological changes; (3) the multimodal fusion design allows the model to learn robust representations across modest temporal offsets. Prior studies using similar ±30-day windows report acceptable performance in forest monitoring [12,23]. Sensitivity analysis with narrower temporal windows (±7 or ±15 days) would provide a stronger test but would also reduce the number of cloud-free observations available for many plots, particularly in Southeast China. Fourth, the interval analysis evaluates empirical plot-level coverage and does not propagate error to an area-wide carbon total. Finally, the allometric target and reference AGB share field measurements, so correlated errors may contribute to the observed gains. Region-specific relationships and independently derived biomass references are needed to assess the physical interpretation of this regularizer.

These limitations point to focused extensions. A missing-modality version requires explicit training and calibration rather than replacement of absent inputs with zeros. Repeated spatial splits and external sites test sensitivity to sampling configuration and geographic shift. Multi-temporal inputs examine biomass change, while region-specific allometry clarifies the role of physical regularization. Such tests determine whether the advantages observed here persist beyond the current complete-modality, plot-level setting.

5. Conclusions

LAPFuse integrates dense Sentinel-2 and Sentinel-1 observations with sparse GEDI L4A footprint records for plot-level forest AGB estimation. Its main components are fixed language anchors, distance-weighted sparse–dense attention, and a physics-informed neural-process predictor. Together, they address representation heterogeneity, spatial-support mismatch, and predictive uncertainty within a single framework. On 369 held-out plots from three Chinese forest regions, LAPFuse achieves $R^2 = 0.882$, RMSE $= 15.34$ Mg ha^{-1}, and MAE $= 9.87$ Mg ha^{-1}. The nominal 95% intervals attain PICP $= 0.949$ with MPIW $= 67.2$ Mg ha^{-1}. Zero-shot cross-region R^2 ranges from 0.762 to 0.798, and the 100-shot results are within 0.1–1.5% of in-domain performance. Wall-to-wall inference over 1.1 million forest pixels demonstrates the framework's operational scalability and reveals interpretable spatial patterns of biomass and uncertainty. These findings support language-regularized, support-aware fusion as a practical approach for uncertainty-aware AGB estimation under the evaluated data regime.

The conclusions are bounded by the three study regions, the 2022–2024 observation period, and the availability of all input products. The allometric-loss ablation supports an incremental predictive contribution under the evaluated data setting. Because the constraint and reference AGB share inventory measurements, this contribution does not constitute independent physical validation. External validation, explicit missing-modality training, repeated ablations, and independent evaluation of the allometric constraint remain necessary for broader forest-monitoring applications.

Author Contributions: Conceptualization, Y.L. and X.Q.; methodology, Y.L.; software, Y.L. and L.X.; validation, Y.L.; formal analysis, X.Q., L.X., and F.Y.; investigation, F.Y.; data curation, Y.L.; writing—original draft preparation, Y.L. and X.Q.; writing—review and editing, S.H.; visualization, Y.L. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (Grant No. 52539001) and Shenzhen Science and Technology Program (AI20260020).

Data Availability Statement: Sentinel-1 and Sentinel-2 data are publicly available through the Copernicus Data Space Ecosystem and the Alaska Satellite Facility. GEDI L4A Version 2.1 AGBD is available from NASA ORNL DAAC (DOI: 10.3334/ORNLDAAC/2056), and SRTM elevation is available through USGS EarthExplorer. Plot-level field measurements and derived model outputs are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pan, Y.; Birdsey, R.A.; Fang, J.; Houghton, R.; Kauppi, P.E.; Kurz, W.A.; Phillips, O.L.; Shvidenko, A.; Lewis, S.L.; Canadell, J.G.; et al. A large and persistent carbon sink in the world's forests. *Science* **2011**, *333*, 988–993. [CrossRef] [PubMed]
2. Harris, N.L.; Gibbs, D.A.; Baccini, A.; Birdsey, R.A.; de Bruin, S.; Farina, M.; Fatoyinbo, L.; Hansen, M.C.; Herold, M.; Houghton, R.A.; et al. Global maps of twenty-first century forest carbon fluxes. *Nat. Clim. Change* **2021**, *11*, 234–240. [CrossRef]
3. Lu, D.; Chen, Q.; Wang, G.; Liu, L.; Li, G.; Moran, E. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. *Int. J. Digit. Earth* **2016**, *9*, 63–105. [CrossRef]
4. Fassnacht, F.E.; White, J.C.; Wulder, M.A.; Næsset, E. Remote sensing in forestry: Current challenges, considerations and directions. *For. Int. J. For. Res.* **2024**, *97*, 11–37. [CrossRef]
5. Zolkos, S.G.; Goetz, S.J.; Dubayah, R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sens. Environ.* **2013**, *128*, 289–298. [CrossRef]
6. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
7. Mutanga, O.; Adam, E.; Cho, M.A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 399–406. [CrossRef]
8. Torres, R.; Snoeij, P.; Geudtner, D.; Bibby, D.; Davidson, M.; Attema, E.; Potin, P.; Rommen, B.; Floury, N.; Brown, M.; et al. GMES Sentinel-1 mission. *Remote Sens. Environ.* **2012**, *120*, 9–24. [CrossRef]
9. Santoro, M.; Cartus, O.; Carvalhais, N.; Rozendaal, D.M.A.; Avitabile, V.; Araza, A.; de Bruin, S.; Herold, M.; Quegan, S.; Rodríguez-Veiga, P.; et al. The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. *Earth Syst. Sci. Data* **2021**, *13*, 3927–3950. [CrossRef]
10. Dubayah, R.; Blair, J.B.; Goetz, S.; Fatoyinbo, L.; Hansen, M.; Healey, S.; Hofton, M.; Hurtt, G.; Kellner, J.; Luthcke, S.; et al. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. *Sci. Remote Sens.* **2020**, *1*, 100002. [CrossRef]
11. Dubayah, R.O.; Armston, J.; Kellner, J.R.; Duncanson, L.; Healey, S.P.; Patterson, P.L.; Hancock, S.; Tang, H.; Bruening, J.M.; Hofton, M.A.; et al. GEDI L4A Footprint Level Aboveground Biomass Density, Version 2.1. 2022. Available online: https://www.earthdata.nasa.gov/data/catalog/ornl-cloud-gedi-l4a-agb-density-v2-1-2056-2.1 (accessed on 9 September 2026).
12. Dalponte, M.; Bruzzone, L.; Gianelle, D. Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1416–1427. [CrossRef]
13. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
14. Schwartz, M.; Ciais, P.; Sean, E.; de Truchis, A.; Vega, C.; Besic, N.; Fayad, I.; Wigneron, J.P.; Brood, S.; Pelissier-Tanon, A.; et al. Retrieving yearly forest growth from satellite data: A deep learning based approach. *Remote Sens. Environ.* **2025**, *330*, 114959. [CrossRef]
15. Roy, S.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515620 . [CrossRef]
16. Ma, X.; Zhang, X.; Pun, M.O.; Liu, M. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403215. [CrossRef]
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA , 2016; pp. 770–778.

18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
19. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 9992–10002. [CrossRef]
20. Jakubik, J.; Roy, S.; Phillips, C.E.; Fraccaro, P.; Zadrozny, B.; Olofsson, P.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. Foundation Models for Generalist Geospatial Artificial Intelligence. *SSRN Electron. J.* **2024**. [CrossRef]
21. Hong, D.; Gao, L.; Yokoya, N.; Yao, J.; Chanussot, J.; Du, Q.; Zhang, B. More diverse means better: Multimodal deep learning meets remote sensing imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 4340–4354. [CrossRef]
22. Guo, Q.; Du, S.; Jiang, J.; Guo, W.; Zhao, H.; Yan, X.; Zhao, Y.; Xiao, W. Combining GEDI and Sentinel data to estimate forest canopy mean height and aboveground biomass. *Ecol. Inform.* **2023**, *78*, 102348. [CrossRef]
23. Schwartz, M.; Ciais, P.; de Truchis, A.; Chave, J.; Ottlé, C.; Vega, C.; Wigneron, J.P.; Nicolas, M.; Jouaber, S.; Liu, S.; et al. FORMS: Forest Multiple Source height, wood volume, and biomass maps in France at 10 to 30 m resolution based on Sentinel-1, Sentinel-2, and Global Ecosystem Dynamics Investigation (GEDI) data with a deep learning approach. *Earth Syst. Sci. Data* **2023**, *15*, 4927–4945. [CrossRef]
24. Emick, E.; Babcock, C.; White, G.W.; Hudak, A.T.; Domke, G.M.; Finley, A.O. An approach to estimating forest biomass while quantifying estimate uncertainty and correcting bias in machine learning maps. *Remote Sens. Environ.* **2023**, *295*, 113678. [CrossRef]
25. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020. Published in Proceedings 1154 of the 38th International Conference on Machine Learning (ICML 2021).
26. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2021; pp. 4904–4916.
27. Huo, C.; Chen, K.; Zhang, S.; Wang, Z.; Yan, H.; Shen, J.; Hong, Y.; Qi, G.; Fang, H.; Wang, Z. When Remote Sensing Meets Foundation Model: A Survey and Beyond. *Remote Sens.* **2025**, *17*, 179. [CrossRef]
28. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [CrossRef]
29. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; Li, X. RSGPT: A remote sensing vision language model and benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 272–286. [CrossRef]
30. Zhang, Z.; Shen, H.; Zhao, T.; Chen, B.; Guan, Z.; Wang, Y.; Jia, X.; Cai, Y.; Shang, Y.; Yin, J. GeoRSMLLM: A Multimodal Large Language Model for Vision-Language Tasks in Geoscience and Remote Sensing. *arXiv* **2025**, arXiv:2503.12490. [CrossRef]
31. Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M.A.; Danish, M.S.; Fraccaro, P.; Watson, C.D.; Klein, L.J.; Khan, F.S.; et al. EarthDial: Turning Multi-sensory Earth Observations to Interactive Dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2025; pp. 14303–14313. [CrossRef]
32. Danish, M.S.; Munir, M.A.; Shah, S.R.A.; Kuckreja, K.; Khan, F.S.; Fraccaro, P.; Lacoste, A.; Khan, S. GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 7132–7142. [CrossRef]
33. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; Khan, F.S. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 27831–27840. [CrossRef]
34. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In *Proceedings of the 2016 International Conference on Computer and Information Technology (CIT)*; IEEE: New York, NY, USA, 2016; pp. 1–7. . [CrossRef]
35. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
36. Wang, B.; Lu, X.; Zheng, X.; Li, X. Remote sensing image captioning via variational autoencoder and reinforcement learning. *Knowl.-Based Syst.* **2020**, *203*, 105920. [CrossRef]
37. Li, Y.; Du, L.; Wei, D. Multiscale CNN based on component analysis for SAR ATR. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5211212. [CrossRef]
38. Zhang, Y.; Shi, H.; Huang, W.; Li, X.; Sun, X.; Yang, F.; Sun, K.; Tao, D. RSGPT: A remote sensing vision language model and benchmark. *arXiv* **2023**, arXiv:2307.15266.
39. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2016; pp. 1050–1059.
40. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6405–6416.

41. Garnelo, M.; Rosenbaum, D.; Maddison, C.; Ramalho, T.; Saxton, D.; Shanahan, M.; Teh, Y.W.; Rezende, D.; Eslami, S.M.A. Conditional neural processes. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR: New York, NY, USA, 2018; pp. 1704–1713.
42. Kim, H.; Mnih, A.; Schwarz, J.; Garnelo, M.; Eslami, S.M.A.; Rosenbaum, D.; Vinyals, O.; Teh, Y.W. Attentive neural processes. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
43. Vaughan, A.; Tebbutt, W.; Hosking, J.S.; Turner, R.E. Convolutional conditional neural processes for local climate downscaling. *Geosci. Model Dev.* **2022**, *15*, 251–268. [CrossRef]
44. Bao, L.L.; Zhang, J.S.; Zhang, C.X. Spatial multi-attention conditional neural processes. *Neural Netw.* **2024**, *173*, 106201. [CrossRef] [PubMed]
45. Bao, L.L.; Zhang, C.X.; Zhang, J.S.; Guo, R. A two-stage spatial prediction modeling approach based on graph neural networks and neural processes. *Expert Syst. Appl.* **2025**, *258*, 125173. [CrossRef]
46. Li, G.; Cao, G. Neural process for uncertainty-aware geospatial modeling. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 106–109. [CrossRef]
47. Chave, J.; Réjou-Méchain, M.; Búrquez, A.; Chidumayo, E.; Colgan, M.S.; Delitti, W.B.C.; Duque, A.; Eid, T.; Fearnside, P.M.; Goodman, R.C.; et al. Improved allometric models to estimate the aboveground biomass of tropical trees. *Glob. Change Biol.* **2014**, *20*, 3177–3190. [CrossRef] [PubMed]
48. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114.
49. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
51. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
52. Khosravi, A.; Nahavandi, S.; Creighton, D.; Atiya, A.F. Lower Upper Bound Estimation Method for Construction of Neural Network-Based Prediction Intervals. *IEEE Trans. Neural Netw.* **2011**, *22*, 337–346. [CrossRef] [PubMed]
53. Kuleshov, V.; Fenner, N.; Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR: New York, NY, USA, 2018; pp. 2796–2804.
54. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
55. Gong, P.; Wang, J.; Yu, L.; Zhao, Y.; Zhao, Y.; Liang, L.; Niu, Z.; Huang, X.; Fu, H.; Liu, S.; et al. Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. *Int. J. Remote Sens.* **2013**, *34*, 2607–2654. [CrossRef]
56. Potapov, P.; Li, X.; Hernandez-Serna, A.; Tyukavina, A.; Hansen, M.C.; Kommareddy, A.; Pickens, A.; Turubanova, S.; Tang, H.; Silva, C.E.; et al. Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sens. Environ.* **2021**, *253*, 112165. [CrossRef]
57. Liang, M.; Duncanson, L.; Silva, J.A.; Sedano, F. Quantifying aboveground biomass dynamics from charcoal degradation in Mozambique using GEDI Lidar and Landsat. *Remote Sens. Environ.* **2023**, *284*, 113367. [CrossRef]
58. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5574–5584.
59. McRoberts, R.E.; Næsset, E.; Saatchi, S.; Quegan, S. Statistically rigorous, model-based inferences from maps. *Remote Sens. Environ.* **2022**, *279*, 113028. [CrossRef]
60. Hunka, N.; May, P.; Babcock, C.; de la Rosa, J.A.A.; de Los Ángeles Soriano-Luna, M.; Mayorga Saucedo, R.; Armston, J.; Santoro, M.; Requena Suarez, D.; Herold, M.; et al. A geostatistical approach to enhancing national forest biomass assessments with Earth Observation to aid climate policy needs. *Remote Sens. Environ.* **2025**, *318*, 114557. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.