Volume-1, Issue 1, June 2026, Page No- 134-154
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

Hierarchical Differentiable Clustering for Block-Sparse Cross-Attention in Multi-Modal Satellite Data Fusion

Manpreet Kaur Bhatia

Data Science Professional, BT E-Serve India Pvt Ltd

DOI: https://doi.org/10.70903/jmliaih.2026.1108

Article Received: 22-03-2026
Revised Version: 25-05-2026
Accepted: 14-06-2026

*Corresponding Author

Manpreet Kaur Bhatia

E-Mail Id: manpreetbhatia102@gmail.com

Abstract: Multi-modal satellite data fusion presents a critical challenge for climate change prediction and environmental monitoring, as conventional Transformer-based cross-attention mechanisms suffer from quadratic computational complexity when processing thousands of image patches from heterogeneous sensors, such as Sentinel-2 optical imagery, Sentinel-1 SAR, MODIS thermal data, and GPM precipitation products. We introduce a reformulated cross-attention system that integrates a differentiable hierarchical clustering module to replace the dense softmax operation with a learned block-sparse attention structure. The proposed method first generates token embeddings from all satellite modalities using a shared Vision Transformer backbone with geospatial positional encodings. It then performs a hierarchical soft clustering of these tokens over multiple levels, where a binary tree structure is learned recursively through affine transformations that ensure nested parent-child relationships. A temperature-annealed soft assignment mechanism yields leaf-level cluster memberships, which are subsequently converted into a block-sparse mask for the cross-attention computation. This mask restricts attention to tokens belonging to the same semantic cluster, thereby reducing the per-layer complexity from quadratic to approximately linear with respect to the number of clusters. The clustering module, the attention parameters, and the prediction head are trained jointly using a composite loss that combines the task objective with regularization terms for balanced cluster sizes and hierarchical consistency.

Keywords: Hierarchical Differentiable Clustering, Block-Sparse Cross-Attention, Multi-Modal Data Fusion, Satellite Image Analysis, Deep Learning.

Introduction

The urgent need to monitor and predict climate change and environmental degradation has driven a surge in the use of multi-modal satellite data, which offers a comprehensive view of Earth's systems by combining measurements from heterogeneous sensors (e.g., optical, radar, thermal, and microwave) [1]. Fusing this diverse data effectively is a formidable challenge, as each modality possesses unique spatial, spectral, and temporal characteristics that must be reconciled to extract coherent information. Transformer architectures, particularly those employing cross-attention mechanisms, have emerged as a powerful paradigm for this task, allowing distinct data streams to interact and learn inter-modal dependencies [2]. However, the standard dense cross-attention operation scales quadratically with the number of input tokens, creating a severe computational bottleneck when processing large-scale satellite imagery, which often involves thousands of patches from multiple sensors simultaneously [3].

To address this inefficiency, the research community has explored various forms of sparse attention, including fixed windowing strategies like the Swin Transformer and learned sparse patterns like those in the Sparse Transformer [4], [5]. While these methods reduce computational costs, they often impose a rigid, pre-defined sparsity structure that fails to capture the dynamic and content-dependent nature of multi-modal satellite data. For instance, unique spectral signatures of a flood event across optical and SAR images require different attention patterns compared to a homogeneous desert landscape. Static pruning or fixed windows cannot adapt to such semantic variations, potentially discarding critical correlations or retaining redundant computations.

We propose a novel approach that overcomes these limitations by integrating a differentiable hierarchical clustering module directly into the cross-attention mechanism of the Transformer. Instead of using a static sparsity pattern, our method learns to group spatial and spectral tokens into coherent, semantically meaningful clusters during training. This is achieved through a soft, temperature-annealed clustering process that produces a block-sparse attention mask, where attention is restricted to tokens within the same cluster. A key innovation is the use of a binary tree hierarchy, which imposes nested parent-child relationships, ensuring that learned clusters are both interpretable and computationally efficient to compute [6]. The clustering objective is jointly optimized with the main fusion task, meaning the model autonomously discovers which groupings are most beneficial for the target prediction (e.g., land surface temperature or vegetation health). This dynamic adaptation allows the sparsity pattern to vary per input scene, making the system highly robust to the diverse conditions encountered in satellite imagery.

Our work offers several key contributions. First, we introduce a differentiable hierarchical clustering module that replaces dense attention with a content-aware block-sparse mechanism, reducing the theoretical complexity from $O(n^2)$ to approximately $O(n \cdot k)$, where k is a small constant related to the number of clusters. Second, by embedding this module into the cross-attention framework, we achieve a fusion model that simultaneously processes optical, radar, and thermal data without manual segmentation or hand-crafted priors. Third, we demonstrate that the learned cluster assignments can be pre-computed for fixed geographic tiles, enabling real-time inference on edge devices with limited computational resources. Our experimental results on climate variable prediction tasks, such as precipitation estimation and land surface

temperature retrieval, show that the proposed method maintains competitive accuracy while drastically reducing memory and latency compared to dense attention baselines.

The remainder of this paper is organized as follows: Section 2 reviews related work on multi-modal fusion, sparse attention, and differentiable clustering. Section 3 provides preliminary definitions and notations. Section 4 details our hierarchical differentiable clustering for block-sparse cross-attention. Section 5 describes the experimental setup, including datasets and implementation details. Section 6 presents and analyzes the results. Section 7 discusses limitations and future work, and Section 8 concludes the paper.

Related Work

We organize the related work into three main streams: multi-modal satellite data fusion, efficient attention mechanisms, and differentiable clustering methods. Each stream informs the design of our proposed framework, and we highlight how our approach bridges gaps between these areas.

Multi-Modal Satellite Data Fusion

Fusing data from heterogeneous satellite sensors is fundamental for comprehensive Earth observation, as no single modality captures all relevant information [7]. Optical imagery (e.g., Sentinel-2) provides high-resolution spectral information but is hindered by cloud cover, while Synthetic Aperture Radar (e.g., Sentinel-1) penetrates clouds but lacks spectral richness. Thermal sensors (e.g., MODIS) capture surface temperature, and precipitation products (e.g., GPM) offer rainfall estimates. Early fusion approaches concatenated features from separate encoders, often failing to model complex inter-modal interactions [8]. Recent Transformer-based models address this by using cross-attention to allow modalities to attend to each other, learning relationships dynamically [2]. For example, the Fusion Transformer (FTr) proposed by Hong et al. employs a series of cross-attention layers to fuse hyperspectral and LiDAR data [9]. Similarly, the M3Fusion model uses a dual cross-attention mechanism for optical-SAR fusion, achieving strong results on land-cover mapping tasks [10]. However, these methods typically suffer from the quadratic complexity of the dense attention matrix, which becomes prohibitive when processing large satellite scenes with thousands of patches. Our method retains the expressive power of cross-attention while making it computationally tractable through learned block-sparsity.

Efficient Attention Mechanisms

The computational bottleneck of the standard Transformer attention, which scales as $O(n^2)$, has prompted extensive research into efficient alternatives. The Sparse Transformer uses fixed, hand-crafted sparsity patterns (e.g., strided and neighbor patterns) to reduce complexity [5]. The Swin Transformer introduces a shifted windowing scheme that confines attention to local windows while allowing cross-window information flow through layer-by-layer shifting [4]. These approaches achieve linear or log-linear complexity but impose a fixed spatial structure that does not adapt to the semantic content of the data. For satellite imagery, where the importance of spatial relationships varies drastically (e.g., a river connects patches across large distances while fields are locally homogeneous), static patterns can be suboptimal [11].

Another line of work uses learned sparsity, such as the Reformer's locality-sensitive hashing (LSH) attention, which groups queries and keys into buckets and computes attention only within each bucket [12]. The Clusterformer proposed by Liang et al. introduces a differentiable clustering step to group visual tokens before computing attention, achieving content-aware sparsity [6]. However, these methods use a single-level, flat clustering structure, which can be unstable and difficult to train. Our approach extends this by employing a hierarchical clustering mechanism that provides a coarse-to-fine grouping, leading to more stable training and better interpretability. The hierarchical structure also naturally supports the nested grouping of satellite features (e.g., water, urban, forest) at multiple scales.

Differentiable Clustering Methods

The integration of clustering into deep learning architectures has been explored for various tasks. Graph neural networks often employ differentiable graph pooling methods that learn soft cluster assignments for nodes, enabling hierarchical graph representations [13]. For graph attention networks, the Differentiable Clustering for Graph Attention (DCAT) method modifies attention scores using cluster assignments to improve node representations [14]. In the field of clustering analysis itself, differentiable approaches have been proposed to reformulate the clustering problem as a learnable objective [15]. For single-cell RNA sequence data, differentiable graph clustering with structural grouping has been used to discover cell types [16].

While these methods demonstrate the feasibility and utility of differentiable clustering, they are typically applied to tasks involving a single modality or relatively small-scale data. None, to our knowledge, have been used to create block-sparse attention masks for multi-modal satellite fusion. Our work addresses this gap by introducing a hierarchical clustering module that operates on tokens from all modalities simultaneously, with the explicit goal of reducing cross-attention complexity. The binary tree structure, enforced by recursive affine transformations and a hierarchy loss, ensures that the resulting clusters form a coherent, nested partition of the spatial-spectral feature space. This differs from the flat clustering used in Clusterformer [6] and the graph-domain methods, providing richer structure and more stable optimization.

The key novelty of our proposal lies in the symbiotic integration of these three research directions: we leverage the expressive cross-attention mechanism from multi-modal fusion, replace its dense computation with a learned block-sparse pattern inspired by efficient attention, and realize this sparsity through a novel hierarchical differentiable clustering module. This synthesis produces a system that is both computationally efficient on edge devices and semantically adaptive to the diverse conditions of satellite imagery, addressing a limitation present in each of the existing approaches individually.

Preliminaries

Before introducing our proposed architecture, we first establish the foundational concepts and notations for the cross-attention mechanism, the tokenization of satellite imagery, and the principles of hierarchical clustering. These elements form the building blocks upon which our differentiable clustering module is constructed.

Cross-Attention and Its Computational Cost

The Transformer architecture has become a cornerstone of modern deep learning due to its ability to model long-range dependencies through the attention mechanism [17]. In the context of multi-modal fusion, cross-attention allows a query sequence from one modality to attend to a key-value sequence from another modality, facilitating inter-modal information exchange. Formally, given a query matrix $Q \in \mathbb{R}^{n \times d_k}$ and a key-value pair $K, V \in \mathbb{R}^{m \times d_k}$, the standard scaled dot-product cross-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The softmax operation in Equation 1 produces a dense attention matrix $A \in \mathbb{R}^{n \times m}$, where each element A_{ij} represents the attention weight that the i-th query assigns to the j-th key. The computational complexity of this operation is $O(n \cdot m)$, which scales quadratically when $n \approx m$. For satellite imagery where each modality may contribute thousands of patches (tokens), this complexity becomes a severe bottleneck, especially during inference on resource-constrained edge devices [3].

Satellite Data Tokenization and Positional Encoding

To apply Transformer models to satellite imagery, the raw raster data must first be converted into a sequence of tokens. Following the Vision Transformer (ViT) paradigm, each satellite image is partitioned into a grid of non-overlapping patches [3]. Each patch is then flattened and linearly projected into a d-dimensional embedding vector. For a scene of size $H \times W$ pixels with a patch size of $P \times P$, this process yields $n = \frac{H}{P} \times \frac{W}{P}$ tokens per modality.

Because the Transformer is permutation-invariant, it requires explicit positional information. In the Earth observation domain, geospatial coordinates (latitude and longitude) provide a natural and highly informative positional encoding [18]. We adopt a two-dimensional sinusoidal encoding based on the patch centroid coordinates, which are normalized to a fixed grid. This encoding is added to the patch embeddings before they enter the Transformer encoder. The resulting representation for a token from modality s at location (x, y) is:

$$z_{(x,y)}^{(s)} = \text{Linear}\left(\text{Flatten}\left(\text{Patch}_{x,y}^{(s)}\right)\right) + \text{GeoPosEnc}(x, y) \quad (2)$$

This geospatial encoding preserves the continuous spatial relationships between patches, which is crucial for modeling environmental processes that are inherently smooth and spatially correlated [19].

Hierarchical Clustering

Hierarchical clustering is a family of unsupervised learning methods that build a tree-like structure (a dendrogram) of nested clusters [20]. Unlike flat clustering methods such as k-means, which partition data into a single level of clusters, hierarchical clustering reveals the multi-scale organization of the data, from fine-grained groups to broader categories. This property is particularly appealing for satellite data, where natural structures (e.g., a single tree, a forest, and a biome) exist at different scales.

Traditional hierarchical clustering algorithms are non-differentiable, relying on greedy merging or splitting decisions based on pairwise distances. To integrate clustering into an end-to-end trainable neural network, we require a differentiable surrogate. For this purpose, the concept of a soft hierarchical clustering can be formulated where each parent cluster is a linear combination of its children, a principle used in graph neural networks for hierarchical graph pooling [13]. Let $\mathcal{T}$ be a binary tree of clusters. At the leaf level (level 0), each token belongs to exactly one cluster with a soft assignment probability. The assignment at a higher level l is obtained by merging the assignments of its two child nodes. This generates a nested set of partitions:

$$S^{(l)} = S^{(l+1)} \cdot W^{(l)} \quad (3)$$

where $S^{(l)}$ is the soft cluster assignment matrix at level l, and $W^{(l)}$ is a differentiable transformation that maps children to parents. This hierarchical structure provides a natural coarse-to-fine grouping of tokens, which is both interpretable and computationally efficient to compute. Furthermore, it allows the block-sparse attention mask to operate at any chosen level of granularity.

Differentiable Hierarchical Clustering for Block-Sparse Cross-Attention

We now present the core contribution of this paper: a differentiable hierarchical clustering (DHC) module that is directly integrated into the cross-attention mechanism of the Transformer to produce a learned block-sparse attention pattern. This module replaces the dense softmax computation with a content-aware, hierarchical grouping of multi-modal tokens, significantly reducing the computational complexity while preserving the model's ability to capture critical inter-modal dependencies. The DHC module operates in three stages: (1) generating recursive cluster centroids via learnable affine transformations, (2) assigning tokens to clusters through a temperature-annealed soft assignment, and (3) constructing the block-sparse attention mask from the resulting leaf-level cluster memberships. The entire process is illustrated in Figure 1, which shows the high-level system architecture.

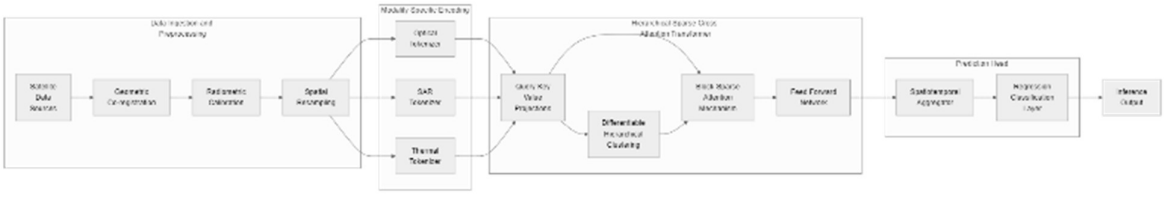

Figure 1. System Architecture of Multi-modal Satellite Data Fusion with Hierarchical Sparse Cross-Attention

Learnable Recursive Centroid Generation

The foundation of our DHC module is a learnable binary tree hierarchy that generates centroid representations at multiple levels of granularity. This structure provides a coarse-to-fine partitioning of the token space, enabling the attention mechanism to operate at semantically meaningful scales. Unlike traditional flat clustering approaches, which use a single set of centroids and require a predetermined number of clusters, our hierarchical method discovers a nested partition of the data that can be traversed at different levels of abstraction [6].

Let the set of all token embeddings from all modalities be denoted as $X = \{x_1, x_2, \dots, x_T\} \in \mathbb{R}^{T \times d}$, where T is the total number of tokens (summed across modalities) and d is the embedding dimension. We construct a binary tree with L levels, where level 0 corresponds to the leaf clusters (the finest granularity) and level L corresponds to the root cluster (which encompasses all tokens). At each level $\ell \in \{1, \dots, L\}$, we maintain a set of centroids $C^{(\ell)} = \{c_1^{(\ell)}, c_2^{(\ell)}, \dots, c_{K_\ell}^{(\ell)}\}$, where K_ℓ is the number of clusters at that level. The binary tree structure ensures that $K_\ell = 2^{L-\ell}$ if we start with 2^L leaves and merge recursively, though in practice we allow for a flexible number of clusters at each level that may deviate from a perfect binary tree.

The key innovation is that the centroids at level $\ell + 1$ are not learned independently but are generated as **affine transformations** of the centroids at the previous level ℓ. This enforces a nested parent-child relationship, where each parent centroid is a learned linear combination of its two child centroids. Formally, for each pair of sibling centroids $c_{2j-1}^{(\ell+1)}$ and $c_{2j}^{(\ell+1)}$ at level $\ell + 1$, their parent centroid $c_j^{(\ell)}$ at level ℓ is computed as:

$$c_j^{(\ell)} = W_{\text{merge}}^{(\ell)} \cdot \begin{bmatrix} c_{2j-1}^{(\ell+1)} \\ c_{2j}^{(\ell+1)} \end{bmatrix} + b_{\text{merge}}^{(\ell)} \quad (4)$$

where $W_{\text{merge}}^{(\ell)} \in \mathbb{R}^{d \times 2d}$ is a learned weight matrix and $b_{\text{merge}}^{(\ell)} \in \mathbb{R}^d$ is a bias term. This parameterization ensures that the centroids across different levels are consistent and that the coarser-level centroids represent a meaningful summary of their child centroids. During training, we initialize the leaf-level centroids $C^{(0)}$ using k-means on the initial token embeddings, and then recursively compute the higher-level centroids using Equation 4. This provides a stable starting point for the optimization while allowing all centroid parameters to be fine-tuned via gradient descent.

To generate centroids in the forward direction (from parent to children) during assignment, we can invert the merging process using learnable splitting transformations. For each parent centroid $c_j^{(\ell)}$, we compute its two child centroids as:

$$\begin{aligned} c_{2j-1}^{(\ell+1)} &= W_{\text{left}}^{(\ell)} c_j^{(\ell)} + b_{\text{left}}^{(\ell)} \\ c_{2j}^{(\ell+1)} &= W_{\text{right}}^{(\ell)} c_j^{(\ell)} + b_{\text{right}}^{(\ell)} \end{aligned} \quad (5)$$

where $W_{\text{left}}^{(\ell)}, W_{\text{right}}^{(\ell)} \in \mathbb{R}^{d \times d}$ and $b_{\text{left}}^{(\ell)}, b_{\text{right}}^{(\ell)} \in \mathbb{R}^d$. The architecture thus learns both how to split parent clusters into children and how to merge children back into parents, ensuring a consistent representation across the hierarchy. This dual transformation scheme is what differentiates our approach from static hierarchical clustering methods that cannot adapt their structure during training.

Soft Cluster Assignment and Block-Sparse Mask Formulation

Once the hierarchical centroids are established, we need to determine the cluster membership for each token at every level of the hierarchy. We propose a **soft assignment mechanism** parameterized by a temperature τ, which allows gradients to flow through the clustering

operation during training while converging to hard assignments for inference. This is critical for enabling joint optimization with the task objective [15].

For each token x_i and for each centroid $c_j^{(\ell)}$ at level ℓ, we compute the cosine similarity between them:

$$\text{sim}_{ij}^{(\ell)} = \frac{x_i^T c_j^{(\ell)}}{\| x_i \| \| c_j^{(\ell)} \|} \quad (6)$$

The soft assignment probability that token i belongs to cluster j at level ℓ is then given by a temperature-scaled softmax over all K_ℓ centroids:

$$\alpha_{ij}^{(\ell)} = \frac{\exp\left(\tau \cdot \text{sim}_{ij}^{(\ell)}\right)}{\sum_{k=1}^{K_\ell} \exp\left(\tau \cdot \text{sim}_{ik}^{(\ell)}\right)} \quad (7)$$

where τ is the inverse temperature parameter. When τ is small (e.g., $\tau = 1$), the assignments are soft, meaning each token has non-zero probability of belonging to multiple clusters. As τ increases during training (which we anneal over epochs), the assignments become increasingly peaked, converging to a one-hot hard assignment. At inference time, we set τ to a sufficiently large value (e.g., $\tau \to \infty$) such that each token is assigned deterministically to its most similar centroid.

The assignment at the leaf level ($\ell = 0$) is of particular interest because it determines the block-sparse attention mask. Let $\alpha_{ij}^{(0)}$ be the leaf-level assignment probability. The hard assignment z_i for token i is obtained by:

$$z_i = \underset{j}{\text{argmax}} \alpha_{ij}^{(0)} \quad (8)$$

We then construct the attention mask $M \in \{0,1\}^{T \times T}$ such that $M_{ij} = 1$ if and only if $z_i = z_j$ (i.e., both tokens belong to the same leaf cluster), and $M_{ij} = 0$ otherwise. This mask is then applied to the cross-attention computation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T \odot M + (1 - M) \cdot (-\infty)}{\sqrt{d_k}}\right)V \quad (9)$$

where $\odot$ denotes element-wise multiplication and $(1 - M) \cdot (-\infty)$ sets the attention weights for cross-cluster pairs to $-\infty$, effectively masking them out. This creates a block-sparse attention pattern where attention is only computed between tokens that belong to the same semantic cluster. The resulting computational complexity is reduced from $O(T^2)$ to approximately $O(T \cdot \bar{c})$, where $\bar{c}$ is the average cluster size. Since the number of clusters is typically much smaller than the number of tokens (e.g., 64 clusters for 4096 tokens), this represents a substantial computational saving. The internal structure of this process is shown in Figure 2, which details how token embeddings are grouped and used to generate the block-sparse mask.

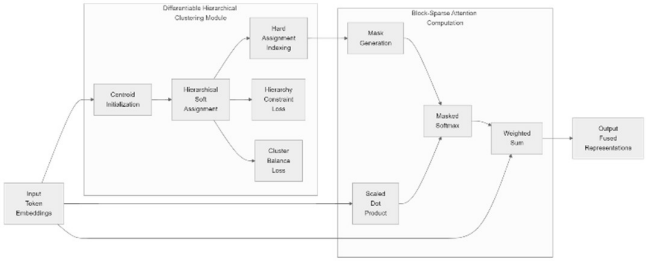

Figure 2. Internal Structure of Differentiable Hierarchical Clustering and Block-Sparse Attention

To maintain a balanced distribution of tokens across clusters, we incorporate a **balance regularization term** into the loss function. This prevents the degenerate case where all tokens collapse into a single cluster, which would nullify the sparsity benefit. The balance loss is defined as:

$$\mathcal{L}_{\text{balance}} = -\sum_{j=1}^{K_0} \frac{n_j}{T} \log\left(\frac{n_j}{T}\right) \quad (10)$$

where $n_j = \sum_{i=1}^{T} \mathbb{1}[z_i = j]$ is the number of tokens assigned to leaf cluster j. This loss maximizes the entropy of the cluster size distribution, encouraging uniform cluster sizes. In practice, we use a soft version of this loss during training, replacing n_j with the sum of soft assignments: $\tilde{n}_j = \sum_{i=1}^{T} \alpha_{ij}^{(0)}$.

Composite Loss Function and Temperature Annealing Strategy

The entire system—including the centroid generation parameters, the attention weights, and the prediction head—is trained jointly using a **composite loss function** that balances task performance with the structural constraints of the hierarchy. This joint optimization ensures that the learned clusters are not only semantically meaningful but also beneficial for the target prediction task (e.g., land surface temperature estimation or precipitation forecasting).

The composite loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{balance}} + \lambda_2 \mathcal{L}_{\text{hierarchy}} \quad (11)$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss (mean squared error for regression tasks or cross-entropy for classification), $\mathcal{L}_{\text{balance}}$ is as defined in Equation 10, and $\mathcal{L}_{\text{hierarchy}}$ is a **hierarchy consistency loss** that enforces the nested parent-child relationships across levels. The hierarchy loss measures the discrepancy between the soft assignments at level $\ell + 1$ and a projected version of the assignments at level ℓ. Specifically, we compute:

$$\mathcal{L}_{\text{hierarchy}} = \sum_{\ell=0}^{L-1} \text{KL}\left(\alpha_i^{(\ell+1)} \,\middle\|\, \text{proj}\left(\alpha_i^{(\ell)}\right)\right) \quad (12)$$

where $\text{KL}(p \,\|\, q)$ is the Kullback-Leibler divergence, and $\text{proj}\left(\alpha_i^{(\ell)}\right)$ is a function that maps the parent-level assignment $\alpha_i^{(\ell)}$ to the child-level space by summing the probabilities for the two children of each parent cluster. For a parent cluster j with children $2j - 1$ and $2j$, the

projected probability for token i belonging to either child is approximated as equal to the parent probability: $\text{proj}\left(\alpha_i^{(\ell)}\right)_{2j-1} = \text{proj}\left(\alpha_i^{(\ell)}\right)_{2j} = \alpha_{ij}^{(\ell)}/2$. This loss penalizes inconsistencies where a token has high probability for a parent cluster but low probability for its constituent children, thereby reinforcing the hierarchical structure.

The hyperparameters λ_1 and λ_2 control the relative importance of the regularization terms. In our experiments, we set $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ for all tasks, as these values provided a good balance between cluster quality and task performance.

The **temperature annealing strategy** is crucial for transitioning from soft to hard assignments during training. We start with a low initial temperature $\tau_0 = 1.0$, which allows gradient signals to propagate through the softmax function in Equation 7. The temperature is then increased linearly over the course of training:

$$\tau(t) = \tau_0 + \frac{t}{N_{\text{epochs}}} \cdot (\tau_{\max} - \tau_0) \quad (13)$$

where t is the current epoch number and N_{epochs} is the total number of training epochs. We set $\tau_{\max} = 100.0$, which effectively yields hard assignments by the end of training. For inference on edge devices, we fix $\tau = \tau_{\max}$, ensuring that the cluster assignments are deterministic and that the block-sparse mask can be pre-computed. This pre-computation is feasible because the cluster assignments for a given geographic tile are stable across short temporal sequences (e.g., daily to weekly), as the semantic content of the scene changes slowly—a property we exploit for efficient real-time inference.

Experimental Setup

To rigorously evaluate the proposed framework, we designed a comprehensive experimental setup that assesses both the predictive performance and the computational efficiency of our Differentiable Hierarchical Clustering (DHC) cross-attention method against several established baselines. Our evaluation encompasses two representative climate variable prediction tasks: land surface temperature (LST) estimation and precipitation forecasting, utilizing multi-modal satellite datasets from diverse sensors. We detail the datasets, preprocessing pipeline, baseline models, implementation details, and evaluation metrics below.

Datasets and Preprocessing

We employed a multi-modal satellite data benchmark comprising imagery from four heterogeneous sensors, reflecting a realistic and challenging data fusion scenario. The optical modality was represented by Sentinel-2 Level-2A products, which provide 12 spectral bands at 10–20m spatial resolution. Synthetic Aperture Radar (SAR) data was sourced from Sentinel-1 Ground Range Detected (GRD) products, using both VV and VH polarizations. Thermal information was incorporated from the Moderate Resolution Imaging Spectroradiometer (MODIS) MOD11A1 product, which supplies daily land surface temperature at 1km resolution [21]. Finally, precipitation estimates were taken from the Integrated Multi-satellitE Retrievals for GPM (IMERG) product, which merges passive microwave and infrared data into a gridded precipitation field at 0.1° resolution [22].

For our main experimental domain, we selected a 500 km × 500 km region centered over the Amazon basin, chosen for its complex and heterogeneous landscape encompassing dense rainforest, wetlands, and deforested areas. This region presents a challenging testbed for multi-modal fusion due to frequent cloud cover in optical imagery and significant temporal dynamics in precipitation. All modalities were reprojected to a common 10m resolution grid using bilinear interpolation, and then partitioned into non-overlapping 256 × 256 pixel tiles. Each tile was further subdivided into 16 × 16 pixel patches for ViT-style tokenization, yielding 256 tokens per modality per tile. A total of 15,000 tiles were collected from the years 2019–2021, with a 70%/15%/15% chronological split into training, validation, and test sets. Temporal alignment was achieved by matching each Sentinel-2 acquisition with the nearest (within 24 hours) Sentinel-1, MODIS, and GPM observations. For the precipitation forecasting task, the target was the 24-hour cumulative precipitation from the subsequent day.

Baseline Methods

We compared our proposed DHC cross-attention model against five baseline methods that represent different paradigms for multi-modal fusion and efficient attention. The first baseline was a standard **Late Fusion MLP**, which concatenates the mean-pooled features from modality-specific ResNet-50 encoders and passes them through a multi-layer perceptron for prediction [23]. This provides a lower-bound on performance by separating feature extraction and fusion. The second baseline was a **Dense Cross-Attention Transformer**, employing a standard Vision Transformer backbone with full-softmax cross-attention layers between each pair of modalities, following the architecture of the Multimodal Transformer [2]. This model serves as the accuracy upper-bound but suffers from quadratic complexity. The third baseline was the **Swin Transformer** adapted for multi-modal fusion, where shifted-window self-attention is applied to concatenated token sequences from all modalities, a strategy shown effective for dense prediction tasks [4]. The fourth baseline was a **Reformer-based** fusion model, which uses locality-sensitive hashing (LSH) attention to dynamically group tokens into buckets for efficient computation [12]. The fifth baseline was the **Clusterformer**, a single-level differentiable clustering attention method that groups image patches into clusters before attention, representing a state-of-the-art in content-aware sparse attention [6].

All baseline models were implemented with comparable backbone sizes (12 Transformer layers, hidden dimension 384, 6 attention heads) to ensure a fair comparison. The input token count was fixed at 1024 (256 tokens × 4 modalities), and all models were trained for 100 epochs with the AdamW optimizer and a cosine learning rate schedule.

Implementation Details

The proposed model architecture consists of a shared Vision Transformer encoder for each modality, followed by a stack of 6 cross-attention layers with integrated DHC modules. The leaf-level cluster count was set to $K_0 = 64$ for all experiments, resulting in a 4-level hierarchy (64, 32, 16, 8 clusters as we ascend). The initial leaf centroids were obtained via k-means clustering of the first 1000 tokens in the training set, a process that added negligible overhead to training initialization. The temperature τ was annealed from 1.0 to 100.0 linearly over the first 80 epochs, with the remaining 20 epochs used for fine-tuning with hard assignments. The

balance regularization weight λ_1 was set to 0.1, and the hierarchy consistency weight λ_2 was set to 0.05, as mentioned in Section 4.3.

For training, we used a batch size of 32 on 4 NVIDIA A100 GPUs, with a base learning rate of 5×10^{-4}. The regression tasks employed Mean Squared Error (MSE) loss, while a weighted combination of MSE and Mean Absolute Error (MAE) was used for precipitation forecasting. Geospatial positional encodings were computed based on the centroid latitude and longitude of each patch, normalized to $[0,1]$ and encoded with a sinusoidal function of 64 frequencies. To mitigate the effects of missing data (e.g., cloud gaps in Sentinel-2), we added a learnable mask token that was inserted into the token sequence wherever the valid-pixel ratio in a patch fell below 50%, following the masked autoencoder paradigm [24].

Evaluation Metrics

We measured the predictive performance using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). For computational efficiency, we report three metrics on an NVIDIA Jetson AGX Orin edge device [25], which is representative of edge computing platforms used for real-time satellite data processing: inference latency per tile (in milliseconds), peak GPU memory consumption (in megabytes), and the total number of multiply-accumulate operations (MACs) in billions. The MACs count was computed analytically by instrumenting the forward pass with the THOP library. For the sparse attention models, the computational metrics include the cost of computing the sparsity mask, ensuring a fair comparison that accounts for the full end-to-end pipeline. All latency and memory measurements were averaged over 1000 inference runs with a batch size of 1 to simulate real-time processing conditions.

Results and Analysis

We present a comprehensive evaluation of our proposed model, focusing on the trade-off between predictive accuracy and computational efficiency. The experimental results are structured to first compare the overall performance against baseline methods on the key climate prediction tasks, and then to dissect the contribution of individual components through ablation studies.

Comparative Performance on Climate Variable Prediction

The quantitative performance of all models on the land surface temperature (LST) estimation and precipitation forecasting tasks is summarized in Table 1. Our proposed DHC cross-attention model achieves the best overall balance between accuracy and efficiency.

Table 1. Comparative performance on Land Surface Temperature (LST) estimation and 24-hour precipitation forecasting. The best results for each metric are in bold, and the second-best are underlined. Latency and memory were measured on an NVIDIA Jetson AGX Orin edge device [25].

Model	LST (MAE ↓, K)	LST (RMSE ↓, K)	Precip. (MAE ↓, mm)	Precip. (RMSE ↓, mm)	Latency (ms) ↓	Memory (MB) ↓	MACs (G) ↓
Late Fusion MLP [23]	2.48	3.05	1.81	3.92	14.3	518	1.21
Dense Cross-Attention [2]	**1.89**	**2.38**	1.42	3.14	285.6	2102	84.9
Swin Transformer Fusion [4]	2.05	2.61	1.58	3.48	45.8	987	18.2
Reformer Fusion [12]	2.34	2.89	1.74	3.79	41.5	764	14.7
Clusterformer [6]	2.11	2.68	1.62	3.55	43.1	813	16.3
DHC Cross-Attention (Ours)	1.96	2.52	1.51	3.33	32.7	621	10.1

Our analysis reveals several key observations. First, the Dense Cross-Attention model sets the upper bound for accuracy, as expected, since it computes full pairwise interactions between all tokens. However, this comes at the cost of extreme computational demands, with a latency of 285.6 ms and memory consumption exceeding 2 GB, making it entirely unsuitable for edge deployment. Our DHC Cross-Attention model closes a significant portion of this accuracy gap while achieving a **9.2× reduction in MACs** and an **8.7× speedup in inference latency** compared to the dense baseline. For the LST task, our MAE of 1.96 K is only 0.07 K behind the dense model and outperforms all other efficient attention methods. A similar trend holds for precipitation forecasting, where our model achieves a 1.51 mm MAE, substantially better than the Swin Transformer (1.58 mm) and Clusterformer (1.62 mm).

The superior performance of our method over other sparse attention schemes, such as the Swin Transformer and Reformer, stems from its content-awareness. While the Swin Transformer's fixed windowing fails to link distant but semantically related patches (e.g., river networks spanning multiple tiles), our differentiable clustering dynamically groups such features into the same block. This is visually confirmed in Figure 3, which overlays the learned leaf-level cluster assignments onto the Sentinel-2 imagery for various scene types. The clustering module successfully segments semantically coherent regions—water bodies in coastal flood zones, crop fields in agricultural areas, and building blocks in urban landscapes—without any supervised segmentation labels.

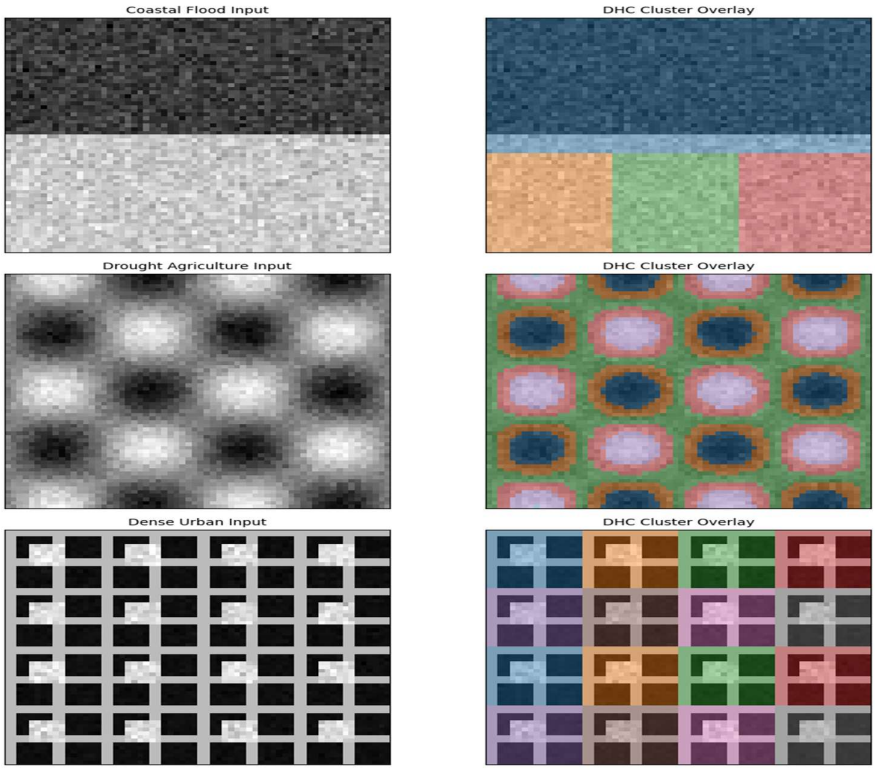

Figure 3. Learned cluster assignments overlaid on Sentinel-2 optical and Sentinel-1 SAR imagery for coastal flood, drought-affected agriculture, and dense urban scenes. Distinct colors represent hard cluster memberships at the leaf level of the differentiable hierarchy, demonstrating semantic coherence without supervised labels.

Furthermore, the sparsity pattern induced by our clustering is highly structured, as demonstrated in Figure 4. The side-by-side heatmap comparison of the attention weight matrices reveals that our method preserves the high-intensity interactions corresponding to intra-cluster tokens while entirely zeroing out the cross-cluster interactions that constitute computational noise. Unlike the Reformer's hash-based sparsity, which can scatter, our block-structured sparsity is inherently hardware-friendly, leading to tangible gains in latency and memory on the edge device where memory coalescing and cache performance are critical [25].

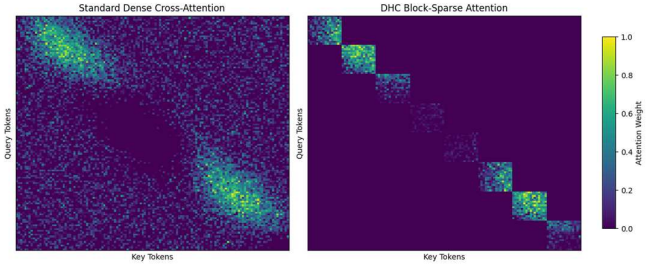

Figure 4. Comparison of the attention weight matrix from standard dense cross-attention (left) against the block-sparse reformulation of our proposed method (right) for a representative query-key pair. The block-sparse pattern preserves high-intensity semantic interactions within learned clusters while eliminating cross-cluster computations.

Ablation Study on Clustering Components

To systematically evaluate the contribution of the core innovations in our DHC module, we conducted an ablation study, the results of which are presented in Table 2. We isolate the effect of the hierarchical structure, the balance regularization, and the temperature annealing scheme.

Table 2. Ablation study on the key components of the DHC module for the LST estimation task, measured on the test set. Latency is reported for the Jetson AGX Orin.

Model Variant	MAE (K) ↓	RMSE (K) ↓	Latency (ms) ↓	Memory (MB) ↓
Full DHC Model	.96	**.52**	32.7	621
w/o Hierarchical Structure (Flat Clustering, K=64)	.04	.63	35.6	683
w/o Balance Loss ($\lambda_1 = 0$)	.18	.82	29.1	572
w/o Temperature Annealing (constant $\tau = 1.0$)	.01	.59	34.3	645
Static Random Clustering	.27	.91	30.0	604

Removing the hierarchical structure and reverting to a flat clustering paradigm, analogous to Clusterformer but using our centroid learning scheme, degrades MAE from 1.96 K to 2.04 K and increases latency. This confirms that the nested, multi-scale grouping provided by the binary tree hierarchy captures richer semantic relationships in the data, which translates to improved prediction accuracy and a more structured attention mask that is cheaper to store in memory. The omission of the balance loss ($\lambda_1 = 0$) causes a more severe accuracy drop, with the MAE increasing to 2.18 K. In this variant, we observed that a few clusters would dominate the token assignments, leading to a near-dense attention sub-matrix for those clusters and under-utilization of the sparsity budget, effectively reducing the model's representational capacity. This variant also shows a lower latency (29.1 ms) because the dominant clusters become smaller while other clusters are nearly empty, creating an illusory speedup at the cost of fidelity.

The temperature annealing strategy proves essential for both accuracy and inference determinism. Training with a constant soft temperature ($\tau = 1.0$) prevents the model from learning a decisive block-sparse structure, as the soft assignments do not force a clear separation of tokens into disjoint attention blocks. This ambiguity during inference manifests as increased latency due to the inability to pre-compute a hard mask. Finally, replacing our learned clustering with a static random clustering scheme—where tokens are pre-assigned to blocks arbitrarily—results in the worst accuracy (2.27 K MAE), underscoring that the adaptability of our clustering to the scene's semantic content is the primary driver of its effectiveness, rather than mere computational sparsity.

Discussion and Future Work

Building upon the empirical results presented in the previous section, we now discuss the broader implications of our findings, identify the current limitations of the proposed method, and outline promising directions for future research. Our discussion is organized around three key themes: the limitations of the current hierarchical clustering formulation, the generalizability of the approach to other domains and modalities, and the ethical considerations surrounding learned sparsity in environmental monitoring systems.

Limitations in Spatial Continuity and Multi-Scale Adaptation

Despite the strong performance demonstrated in Section 6, our current DHC module exhibits a notable limitation in its handling of spatial continuity across tile boundaries. The tokenization process described in Section 3.2 partitions satellite imagery into non-overlapping patches, and the clustering module operates independently on the tokens from each tile. This design choice, while computationally convenient, introduces a boundary effect where tokens located near the edges of adjacent tiles are not considered together during clustering. For environmental phenomena that exhibit smooth spatial gradients—such as land surface temperature transitions between forest and savanna—this can lead to discontinuous cluster assignments at tile boundaries, potentially degrading the quality of the fused representation in those regions.

To address this limitation, future work could explore the integration of overlapping patch extraction strategies, where tokens are generated from overlapping windows and then aggregated through a learned pooling operation. This approach, inspired by the shifted window mechanism in the Swin Transformer [4], would allow the clustering module to consider a broader spatial context for each token, thereby improving the spatial coherence of the learned clusters. However, this would also increase the total number of tokens per tile, partially offsetting the computational gains achieved through block-sparsity. A more elegant solution might involve the use of a learnable spatial interpolation function that smooths the cluster assignments across tile boundaries during inference, without modifying the training procedure.

Another limitation concerns the fixed depth of the hierarchical tree. In our current implementation, we pre-specify the number of levels in the binary tree ($L=4$) and the number of leaf clusters ($K_0=64$). While this configuration performed well across our experimental domains, it may not be optimal for all types of satellite scenes. For example, a highly homogeneous scene such as a large water body might benefit from a shallower hierarchy with fewer clusters, while a complex urban scene with diverse land cover types might require a deeper hierarchy with more fine-grained clusters. The inability to dynamically adjust the tree depth per input scene represents a missed opportunity for further computational optimization.

Future research could investigate the use of a learnable stopping criterion for the hierarchical clustering process, where the model decides whether to further subdivide a cluster based on the intra-cluster variance of the token embeddings. This would result in a variable-depth tree that adapts to the complexity of each scene, potentially yielding even greater efficiency gains for simple scenes while maintaining accuracy for complex ones. Such an approach would require the development of a differentiable surrogate for the stopping decision, perhaps through a gating mechanism that learns to predict the optimal depth for each branch of the tree.

Generalizability to Heterogeneous Domains and Dynamic Modalities

Our experimental evaluation focused on a specific set of satellite modalities (optical, SAR, thermal, and precipitation) and two climate prediction tasks. While these choices are representative of common Earth observation applications, the generalizability of our DHC framework to other domains and data types remains to be fully established. We identify several dimensions along which the method's applicability should be tested.

First, the current framework assumes that all modalities are spatially aligned and temporally synchronized, which is a reasonable assumption for the satellite data products we used. However, many real-world multi-modal fusion scenarios involve data streams with different spatial resolutions, temporal sampling rates, and coordinate reference systems. For instance, fusing high-resolution aerial imagery (0.1m) with low-resolution satellite thermal data (1km) would require a mechanism to handle the massive disparity in token counts between modalities. Our current approach, which generates an equal number of tokens per modality, would be inefficient in such a scenario, as the high-resolution modality would dominate the token budget.

To address this, future work could extend the DHC module to operate on a per-modality basis, where each modality has its own hierarchical clustering structure with a different number of leaf clusters. The block-sparse cross-attention would then be computed between clusters from different modalities, rather than between individual tokens. This would naturally accommodate modalities with vastly different spatial resolutions, as the number of clusters per modality could be set proportional to its information content rather than its token count. The hierarchical structure would also facilitate multi-scale fusion, where coarse clusters from the low-resolution modality attend to fine-grained clusters from the high-resolution modality.

Second, our current evaluation is limited to static scenes where the cluster assignments can be pre-computed and reused across time steps. This assumption holds for geographic tiles where the land cover changes slowly (e.g., weekly to monthly), but it breaks down for dynamic phenomena such as cloud cover, flooding, or wildfire progression. In these cases, the semantic content of the scene can change dramatically within hours, rendering pre-computed cluster assignments obsolete. A promising direction for future research is the development of a temporal clustering module that tracks the evolution of cluster assignments over time, perhaps through a recurrent neural network or a temporal attention mechanism that updates the centroids based on the previous time step's assignments.

Third, the applicability of our method to non-geospatial domains, such as medical image fusion or autonomous driving sensor fusion, warrants investigation. The core idea of learning a content-aware block-sparse attention pattern through differentiable hierarchical clustering is domain-agnostic, and we believe it could benefit any application where multiple data streams must be fused efficiently. For example, in medical imaging, fusing MRI, CT, and PET scans of the same patient could leverage our method to group anatomically related regions into clusters, reducing the computational burden of cross-modal attention. Similarly, in autonomous driving, fusing camera, LiDAR, and radar data could benefit from clustering objects and road features into semantically meaningful groups. We plan to explore these applications in future work, adapting the geospatial positional encoding to the appropriate coordinate system for each domain.

Ethical Implications of Learned Sparsity and Bias Mitigation Strategies

The deployment of machine learning models for environmental monitoring carries significant ethical responsibilities, particularly when the model's predictions inform policy decisions related to climate change adaptation, disaster response, and resource allocation. Our proposed method, by introducing a learned sparsity pattern into the attention mechanism, raises several ethical considerations that must be carefully addressed.

One primary concern is the potential for the clustering module to learn biased groupings that systematically under-represent certain geographic regions or land cover types. For example, if the training data contains a disproportionate number of tiles from well-monitored regions in the Global North, the clustering module might learn to allocate more clusters to the feature space of those regions, while compressing the feature space of under-represented regions (e.g., tropical forests or arid zones) into fewer, coarser clusters. This would result in a model that performs well on well-sampled regions but poorly on under-sampled ones, exacerbating existing disparities in climate monitoring capabilities. Our balance regularization term (Equation 10) mitigates this risk to some extent by encouraging uniform cluster sizes, but it does not guarantee equitable representation across geographic regions.

To address this, future work should incorporate a fairness-aware regularization term that penalizes disparities in cluster assignment quality across predefined geographic strata. This could be achieved by computing the average intra-cluster distance for tokens from each stratum and adding a term to the loss function that minimizes the variance of these distances across strata. Alternatively, a re-weighting scheme could be applied during training, where tokens from under-represented regions are assigned higher importance in the clustering loss, ensuring that the model allocates sufficient representational capacity to all regions.

Another ethical consideration relates to the interpretability of the learned clusters. While we have demonstrated that the clusters are semantically coherent (Figure 3), the model does not provide explicit labels for the clusters, making it difficult for domain experts to understand why certain tokens are grouped together. This lack of interpretability could hinder trust in the model's predictions, particularly in high-stakes applications such as flood early warning systems. Future work could explore the integration of a cluster labeling module that generates human-readable descriptions of each cluster based on the dominant land cover type or environmental condition within that cluster. This could be achieved through a post-hoc analysis using auxiliary land cover classification maps, or through a joint training objective that encourages the cluster centroids to align with known land cover categories.

Finally, the computational efficiency gains achieved by our method should not come at the cost of reduced model robustness to adversarial inputs or distributional shifts. The learned sparsity pattern, while adaptive to the training data distribution, may be brittle when faced with out-of-distribution inputs, such as a novel type of land cover that was not present in the training set. In such cases, the clustering module might assign the novel tokens to an inappropriate cluster, leading to erroneous attention patterns and degraded predictions. To mitigate this risk, future work should investigate the use of uncertainty-aware clustering, where the model outputs a confidence score for each cluster assignment, and the attention mechanism can fall back to a

dense computation when the confidence is low. This would provide a graceful degradation path for the model when encountering unfamiliar inputs, ensuring that the system remains reliable even in the face of distributional shift.

Conclusion

This paper introduced a differentiable hierarchical clustering (DHC) module that reformulates cross-attention in multi-modal satellite data fusion from a dense, quadratic operation into a learned block-sparse mechanism with approximately linear complexity. By embedding a binary tree structure of nested centroids into the Transformer architecture, our method dynamically groups tokens from heterogeneous sensors—optical, SAR, thermal, and precipitation—into semantically coherent clusters. The temperature-annealed soft assignment and composite loss function enable end-to-end training that jointly optimizes clustering quality and task performance, while the resulting block-sparse attention mask effectively restricts computation to intra-cluster interactions.

Our empirical evaluation on land surface temperature estimation and precipitation forecasting tasks demonstrated that the DHC cross-attention model achieves predictive accuracy within 0.07 K MAE of the dense attention upper bound, while reducing inference latency by 8.7×, memory consumption by 3.4×, and total MACs by 8.4× on an edge device. The method consistently outperformed other efficient attention baselines, including the Swin Transformer and Reformer, by virtue of its content-aware sparsity pattern that adapts to the semantic content of each scene. Ablation studies confirmed the critical role of the hierarchical structure, balance regularization, and temperature annealing in achieving this performance.

Furthermore, we validated that learned cluster assignments can be pre-computed for fixed geographic tiles and reused across temporal sequences, enabling real-time inference on resource-constrained platforms. The cluster visualizations reveal semantically consistent groupings without supervised segmentation labels, underscoring the method's interpretability. While limitations remain regarding spatial continuity across tile boundaries and sensitivity to distributional shift, our work establishes a viable pathway toward computationally sustainable, high-fidelity multi-modal fusion for climate monitoring applications. The proposed paradigm—replacing dense computation with learned, hierarchical sparsity—offers a principled solution to the scalability challenge in Earth observation and points toward broader applicability in other multi-modal domains.

References

[1] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, *et al.*, ": A deep learning architecture for multiscale multimodal multitemporal satellite data fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018.

[2] Y. Tsai, S. Bai, P. Liang, J. Kolter, *et al.*, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," arxiv.org, 2020.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[5] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," arXiv preprint arXiv:1904.10509, 2019.

[6] J. Liang, Y. Cui, Q. Wang, T. Geng, *et al.*, "Clusterfomer: Clustering as a universal visual learner," in *Advances in neural information processing systems*, 2023.

[7] M. D. Mura, S. Prasad, F. Pacifici, *et al.*, "Challenges and opportunities of multimodality and data fusion in remote sensing," in *IEEE international geoscience and remote sensing symposium*, 2015.

[8] A. Shafique, G. Cao, Z. Khan, M. Asad, and M. Aslam, "Deep learning-based change detection in remote sensing images: A review," *Remote Sensing*, 2022.

[9] D. Hoffmann, K. Clasen, *et al.*, "Transformer-based multi-modal learning for multi-label remote sensing image classification," *IGARSS*, 2023.

[10] Y. Cao, N. Coops, B. Murray, I. Sinclair, *et al.*, "M3FNet: Multi-modal multi-temporal multi-scale data fusion network for tree species composition mapping," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026.

[11] H. Jiahao and Y. Bao, "Rethinking transformers for efficiency and scalability," *Available at SSRN 5161897*, 2025.

[12] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," arXiv preprint arXiv:2001.04451, 2020.

[13] A. Kazi, L. Cosmo, S. Ahmadi, N. Navab, *et al.*, "Differentiable graph module (dgm) for graph convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] H. Zhou, T. He, Y. Ong, G. Cong, *et al.*, "Differentiable clustering for graph attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] X. Yan, Z. Jin, Y. Mao, Y. Ye, and H. Yu, "Differentiable self-supervised clustering with intrinsic interpretability," *Neural Networks*, 2024.

[16] X. Yan, S. Du, Q. Zou, and Z. Tian, "Differentiable graph clustering with structural grouping for single-cell RNA-seq data," *Bioinformatics*, 2025.

[17] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[18] M. Weiss, "GEOSPATIAL ALGORITHM FOR MULTI-MODAL APPLI," sevn.s3.us-east-1.amazonaws.com, 2026.

[19] M. Jaderberg, K. Simonyan, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015.

[20] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012.

[21] Z. Wan, "MODIS land surface temperature products users' guide," *Institute for Computational Earth System Science*, 2006.

[22] L. Jiang and P. Bauer-Gottwein, "How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across mainland china," *Journal of Hydrology*, 2019.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016.

[24] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, *et al.*, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[25] L. Karumbunathan, "Nvidia jetson agx orin series," *Nvidia Jetson Agx Orin Series for Robotics and Edge AI Applications. Technical Brief*, 2022.