Adapting Video Foundation Models for Spatiotemporal Wildfire Forecasting via Cross-Modal Progressive Fine-Tuning

Wenwen Li, *Member, IEEE*, Chia-Yu Hsu, and Sizhe Wang

***Abstract*—Wildfires pose escalating threats to ecosystems, communities, and climate systems, highlighting the urgent need for accurate, high-resolution spatiotemporal forecasting. In this work, we explore the untapped potential of video foundation models for advancing wildfire spread prediction using multimodal satellite data. While large-scale foundation models have transformed artificial intelligence (AI) and show promise in geospatial AI (GeoAI), their direct application to domain-specific tasks like wildfire forecasting faces two major hurdles: 1) a substantial domain gap between general pretraining data (e.g., natural images and videos) and geospatial data (e.g., multispectral satellite imagery) and 2) limited labeled data for fine-tuning in real-world GeoAI tasks. To address these challenges, we introduce a cross-modal progressive fine-tuning (CMPF) strategy tailored for wildfire forecasting. CMPF combines: 1) Informed cross-modal architectural alignment, leveraging video-based Transformers pretrained on spatiotemporal tasks to better capture wildfire dynamics and 2) progressive fine-tuning, which gradually adapts models to wildfire-specific representations through intermediate domain adaptation before task-specific tuning. We evaluate CMPF using multiple Transformer backbones, including ViT, MViTv2, and VideoMAEv2, on wildfire spread forecasting benchmarks. Our results show that video foundation models, especially when fine-tuned progressively, outperform conventional convolutional neural networks (CNNs) and static vision Transformers in modeling wildfire evolution. These findings also validate the effectiveness of the proposed CMPF approach for adapting general-purpose AI foundation models to complex spatiotemporal geospatial tasks.**

***Index Terms*—Deep learning, Earth observation (EO), foundation models, geospatial artificial intelligence (GeoAI), progressive fine-tuning, satellite imagery, spatiotemporal forecasting, Vision Transformer (ViT), wildfire spread prediction.**

I. Introduction

THE field of artificial intelligence (AI) has been significantly reshaped by recent breakthroughs in foundation models [1]. These models, known for their ability to learn complex patterns from very large and varied multimodal datasets including text, images, and videos [2], are showing powerful abilities in many different areas [3], [4], [5]. At the same time, geospatial AI (GeoAI) has become a quickly growing field that combines AI methods with geospatial science to study and understand Earth observation (EO) data [6]. The combination of foundation models with GeoAI offers a great potential to fundamentally change how we monitor the Earth's surface, speed up geographic knowledge discovery, and help achieve important sustainability goals [7]. This combined effect would greatly influence decision-making in areas important for society, such as humanitarian mapping, natural disaster management (like floods and wildfires), and climate change studies [8]. The potential of GeoAI is further increased by fast developments in multimodal sensing technologies. These technologies are constantly improving the collection of geospatial data with higher resolutions and more data flow, providing a very high level of detail for understanding complex Earth system changes [9].

However, using these general foundation models effectively for specific geospatial tasks faces some important challenges [10]. A major difficulty is the domain discrepancy. This is the significant difference between the general datasets used for initial model training and the special characteristics of geospatial data [11]. For instance, foundation models are typically pretrained on vast and diverse datasets, which can include text, general imagery, videos, and other types of information, often learned through self-supervised methods from publicly available sources [12]. This general pretraining data differs greatly from the specific nature of geospatial data, which typically includes overhead perspectives from satellites or aerial platforms, specific spectral bands not visible to the human eye, various spatial resolutions, inherent geographic referencing, and often involves multiple types of sensor data [7]. Another key challenge relates to the data requirements for effective adaptation. While foundation models aim for broad generalizability, effectively tailoring them to specialized GeoAI tasks often requires more than just adding a task-specific prediction head. To optimally align the large, pretrained backbone of the foundation model with the unique characteristics of geospatial data, and to ensure the new components learn effectively without issues like overfitting, a sufficient amount of domain-specific labeled data for fine-tuning is often necessary. Yet, in many specialized GeoAI applications, obtaining such labeled datasets remains difficult because their creation can be expensive and require significant expert effort [13]. These difficulties currently limit how fully we can use the power of foundation models in GeoAI applications [10]. Furthermore, while geospatial foundation models (GFMs), such as NASA's Prithvi [7], [14], and international

Received 20 June 2025; revised 24 October 2025 and 30 November 2025; accepted 2 January 2026. Date of publication 12 January 2026; date of current version 29 January 2026. This work was supported in part by the National Science Foundation under Award 2120943. *(Corresponding author: Wenwen Li.)*

Wenwen Li and Chia-Yu Hsu are with the School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85281 USA (e-mail: wenwen@asu.edu).

Sizhe Wang is with the School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281 USA.

Digital Object Identifier 10.1109/TGRS.2026.3652453

business machines–european space agency (IBM-ESA's) TerraMind [15], are specifically designed to address geospatial problems, there are still important requirements regarding the domain data used for model adaptation. In addition, without proper adaptation, the performance of foundation models often falls short of that achieved by task-specific AI models [16].

To bridge the domain-adaptation gaps, this study proposes and evaluates a new domain adaptation and training methodology termed cross-modal progressive fine-tuning (CMPF), which comprises two key strategies designed to enhance the adaptation of foundation models to GeoAI tasks.

1) *Informed Cross-Modal Architectural Choice:* This strategy posits that selecting a foundation model pretrained on tasks and data modalities analogous to the target geospatial application leads to enhanced performance. This careful architectural selection is crucial for initially bridging the domain gap by leveraging relevant learned representations and architectural biases.
2) *Progressive Fine-Tuning:* This strategy involves a multistep adaptation process, which introduces an intermediate fine-tuning stage using an auxiliary geospatial dataset that shares common features or data modalities with the primary target dataset. By first adapting the model to this related geospatial data, it gains broader exposure to relevant geospatial patterns, and its representations become better aligned with the general characteristics of the target domain. This intermediate step acts as a critical stepping stone before the final task-specific fine-tuning on the target application's limited labeled data. This progressive sequence aims to lessen the dependency on extensive labeled data for the final task and ultimately enhance overall model adaptation and predictive performance.

This research systematically evaluates the CMPF strategy defined above, focusing on its effectiveness in adapting foundation models, with a particular emphasis on video-based architectures, to complex spatiotemporal geospatial tasks. Wildfire spread forecasting serves as an application for evaluating the effectiveness of our proposed adaptation strategy for foundation models. A core component of this evaluation involves investigating the informed cross-modal architectural choice component of CMPF. Specifically, we conduct a comparative analysis of several mainstream Transformer architectures, including Vision Transformer (ViT), Multiscale ViT (MViTv2), and VideoMAEv2, to determine their suitability for bridging the domain gap from their general pretraining domain (e.g., on images or videos) to the specific characteristics of multimodal and multitemporal geospatial data. This exploration of architectural suitability is coupled with assessing the progressive fine-tuning aspect of CMPF, which employs staged adaptation using intermediate, related geospatial datasets. Our findings demonstrate that CMPF, integrating these considerations of architectural adaptation and progressive fine-tuning, serves as a validated and effective training methodology. Applying this strategy to wildfire spread forecasting has led to notable improvements in predictive performance.

The remainder of this article is organized as follows. Section II reviews recent literature on foundation models, fine-tuning techniques, and current methodologies for wildfire spread forecasting. Section III describes the data, preprocessing steps, and the proposed CMPF approach. Section IV presents and discusses experimental results that demonstrate the effectiveness and generalizability of the CMPF strategy in enhancing model predictions for wildfire spread forecasting. Section V summarizes the key findings and methodological advances of our work and outlines future research directions. Section VI concludes the article, highlights the major contributions, and discusses the applicability of the work beyond wildfire forecasting.

II. Related Work

A. Foundation Models

Foundation models have become a key area of development in AI [17]. These models are generally based on the Transformer architecture [3] and are pretrained on very large, broad datasets, often using self-supervised learning methods [18]. A major advantage of foundation models is their ability to learn generalizable representations, which can then be adapted to different downstream tasks. This adaptability often allows them to perform well with limited task-specific data, a capability known as few-shot or zero-shot learning [19].

Several types of foundation models have been developed. Large language models (LLMs) were influential in demonstrating the effectiveness of large-scale pretraining on vast amounts of text data [20]. This success inspired vision foundation models (VFMs) in computer vision. For example, the segment anything model (SAM) [21] was trained on over 1 billion segmentation masks from 11 million images via model-in-the-loop annotation. For tasks that involve changes over time, video-based foundation models have been created. These models address challenges inherent in video data, such as its high dimensionality due to the additional temporal dimension, the need to capture temporal variations and dependencies across frames, and the associated computational overhead. Models like VideoMAE [22] use strategies such as spatiotemporal embedding and efficient dual masking to tackle these challenges, which makes them particularly relevant for analyzing dynamic geospatial events. In addition, multimodal foundation models are designed to integrate information from multiple data types simultaneously. A prominent example is contrastive language-image pretraining (CLIP) [23], which learns joint representations of images and text by contrasting positive image-text pairs against negative ones from a large dataset.

The potential of foundation models is now being actively investigated within the GeoAI community [7]. Recognizing that EO data has unique characteristics, such as multispectral and hyperspectral bands beyond the visible spectrum, inherent geospatial referencing, and varying spatial resolutions that differ from general visual data, specialized GFMs have been developed. Examples include NASA's Prithvi [7] and TerraMind from IBM [15]. By pretraining on large archives of EO data, these GFMs learn representations more attuned to specific spectral, spatial, and temporal patterns prevalent in satellite and aerial imagery. This provides a more domain-specific starting point that can, in principle, lead to better

performance and faster convergence when fine-tuning for downstream geospatial applications. Such applications where GFMs are being applied include crop biophysical parameters estimation [24], flood and landslide mapping for disaster response [16], [25].

B. Transfer Learning and Fine-Tuning

Transfer learning is a machine learning technique where knowledge from a pretrained model, typically trained on a large source dataset, is leveraged to improve performance and reduce data requirements for a related target task [26]. Fine-tuning is a prevalent transfer learning strategy, especially for deep neural networks, involving further training of a pretrained model's parameters on the specific target dataset. This enables the model to adapt its learned general features to the specifics of a new task, accelerating learning and often improving generalization.

Foundation models, by design, inherently rely on transfer learning for adaptation to diverse downstream applications [26]. For LLMs, the core architecture often remains unchanged, as their inherent text-generation ability handles diverse language tasks with the same core structure. Adaptation is typically achieved via prompts for task context, alongside fine-tuning existing parameters or using parameter-efficient techniques [27]. In contrast, VFMs usually undergo explicit architectural modifications, commonly adding a new, task-specific head (e.g., a segmentation head) to the pretrained backbone. This leads to the "large backbone, small head" challenge, where the backbone's rich representations can cause the small, randomly initialized head to overfit or underperform. To address this, some studies describe a two-stage adaptation process that first trains the VFM with its new head on a large, task-aligned auxiliary dataset, and then fine-tunes it on the target dataset [28]. However, this approach relies heavily on a large amount of labeled data for effective adaptation.

In the GeoAI domain, challenges with transfer learning and fine-tuning foundation models become greater. Geospatial data's vast variety, spanning different sensors, spectral bands, and modalities, often creates large gaps between pretraining data and the target data. This requires adapting not only the output head but often the model's input processing stages. For instance, Hsu et al. [10] adapted Prithvi GFM (pretrained on six-band data) to accept three-band geospatial imagery by modifying its initial patch embedding or input data layers. Furthermore, data complexity (e.g., up to 23 modalities for wildfire prediction, as relevant to this study's focus) worsens the scarcity of large, high-quality labeled datasets in GeoAI. This scarcity makes even VFM-style intermediate adaptation strategies particularly challenging. Consequently, standard fine-tuning approaches often prove inefficient for bridging these domain gaps and data complexities, showing the need for advanced, tailored adaptation strategies such as the proposed CMPF.

C. AI-Based Wildfire Forecasting

Accurate prediction of wildfire spread is essential for effective disaster response and risk mitigation. Forecasting how a fire evolves in space and time allows emergency managers to allocate resources, plan evacuations, and implement containment strategies. Given the increasing frequency and intensity of wildfires due to climate change, improving predictive capabilities has become a critical area of research in environmental modeling and applied machine learning [29], [30].

Traditionally, wildfire spread forecasting has relied heavily on physics-based simulation tools such as FARSITE [31] and Prometheus [32]. These models simulate fire behavior based on environmental inputs, including topography, fuel type, and weather conditions. While widely used in operational settings, their accuracy is often constrained by assumptions in fire dynamics and the difficulty of obtaining precise, real-time input data [33]. As a result, models trained solely on synthetic data may struggle to generalize to satellite observations or capture the full complexity of real-world fire behavior. To bridge this gap, researchers have increasingly adopted empirical approaches to enhance wildfire forecasting using data-driven methods [34]. Building on this momentum, deep learning models have emerged as powerful tools for predicting wildfire spread by learning underlying patterns and spatiotemporal relationships from observational data.

For example, FireCast uses 8 imagery and a convolutional neural network (CNN) to predict next-day fire spread, achieving better accuracy than simulation-based models like FARSITE [35]. However, a key challenge with deep learning is preparing high-quality training data, as model performance is highly dependent on dataset size and quality. While FireCast demonstrates the potential of deep learning, it uses a relatively simple architecture (a six-layer CNN) and is limited to the Rocky Mountain region of the United States.

Gerard et al. [36] compared the use of multiple CNN models, including U-Net [37], ConvLSTM [38], and U-net with temporal attention encoder (UTAE) [39]. U-Net is an encoder-decoder-based architecture with skip connections to perform image segmentation. Its unique design allows the model to capture both global context localization, and therefore has become one of the most popular models in image segmentation. The ConvLSTM model captures spatiotemporal patterns in sequential data by combining convolutional operations with long short-term memory (LSTM) units to model both spatial and temporal dependencies. UTAE is a deep learning model designed for multitemporal satellite image analysis, combining a U-Net-like architecture with temporal attention mechanisms to capture both spatial features and temporal dynamics effectively. Results show that the UTAE model achieves the best performance due to its strong ability to handle time-sequenced satellite imagery.

It is important to note that wildfire forecasting involves time-sequenced climate and environmental variables in addition to satellite imagery. Capturing the spatiotemporal relationships within these multimodal datasets poses significant challenges for traditional image-based AI models. However, foundation models, known for their adaptability, along with those designed for time-sequenced data (e.g., video and satellite time series), offer promising solutions for accurate wildfire spread prediction. In this article, we explore how foundation

models can be adapted and enhanced for this critical application.

Section III details the dataset used in our analysis, the proposed foundation model architecture, and the CMPF strategy.

III. Methodology

A. Datasets and Preprocessing

1) Target Task Dataset: WildfireSpreadTS (WTS): In recent years, several comprehensive AI-ready datasets [33], [36], [40], [41], [42] have been developed to support research in wildfire forecasting. Among these, WTS [36], an extension of NextDayWildfireSpread (NDWS) [41], stands out due to its inclusion of time-series inputs, diverse variables, recent observations, and high-resolution data. Therefore, we use the WTS dataset as the primary data source for our study. WTS is a multimodal, multitemporal remote sensing dataset specifically designed for predicting active wildfire spread with a 24-h lead time. WTS encompasses $N_{\text{WTS}} = 607$ fire events across the United States, recorded from January 2018 to October 2021, totaling 13 607 daily image sets. Each fire event i is represented as a time series of T_i daily observations. A single daily observation consists of an input tensor $X_{\text{WTS}} \in \mathbb{R}^{H_{\text{WTS}} \times W_{\text{WTS}} \times M_{\text{WTS}}}$ with $M_{\text{WTS}} = 23$ multimodal input channels related to fuel conditions, topography, and weather, and a corresponding target active fire map $S_{\text{WTS}} \in \{0, 1\}^{H_{\text{WTS}} \times W_{\text{WTS}}}$. A key characteristic of this dataset is its high spatial resolution of 375 m for both the active fire maps and other resampled features, defining the dimensions $H_{\text{WTS}} \times W_{\text{WTS}}$.

The development of WTS was heavily inspired by the NDWS dataset (introduced in Section III-A2). WTS extends its predecessor primarily by incorporating a full-time series structure for each fire event, $\{(X_{\text{WTS},t}, S_{\text{WTS},t})\}_{t=1}^{T_i}$, which allows for multitemporal modeling, a core contribution differentiating it from the single-day input-output pairs of NDWS. Further enhancements include the integration of more recent daily observations for vegetation conditions [from visible infrared imaging radiometer suite (VIIRS)], additional input features such as weather forecasts, topographic aspect, and slope, and a finer spatial resolution for active fire data (375-m VIIRS versus 1-km MODIS).

2) Auxiliary Geospatial Dataset—NDWS for Intermediate Fine-Tuning: The auxiliary dataset employed for the intermediate stage of our progressive fine-tuning approach is NDWS [41]. This dataset was curated as a large-scale, multivariate collection of historical wildfires, primarily to serve as a benchmark for developing machine learning models that predict wildfire propagation with a one-day lead time from remote sensing data. It aggregates nearly a decade of data (2012–2020) from across the contiguous United States and notably encompasses $N_{\text{NDWS}} = 18\,545$ distinct fire events. The large number of events in this dataset offers a broad and diverse set of samples, making NDWS particularly well-suited for the initial adaptation phase in a progressive fine-tuning strategy.

NDWS combines 2-D fire data with $M_{\text{NDWS}} = 12$ explanatory variables. For each fire event, it generally provides two snapshots: an input observation at time t comprising the previous fire mask and environmental conditions, $X_{\text{NDWS},t} \in \mathbb{R}^{H_{\text{NDWS}} \times W_{\text{NDWS}} \times M_{\text{NDWS}}}$, and a target fire mask for the next day, $S_{\text{NDWS},t+1} \in \{0, 1\}^{H_{\text{NDWS}} \times W_{\text{NDWS}}}$. Both X_{NDWS} and S_{NDWS} are aligned over 2-D regions at a 1km resolution, defining the spatial dimensions $H_{\text{NDWS}} \times W_{\text{NDWS}}$. The dataset focuses on predicting $S_{\text{NDWS},t+1}$ based on the single prior day's observations $X_{\text{NDWS},t}$. It was designed to provide a feature-rich resource for machine learning by combining fire data with multiple relevant environmental variables.

3) Shared Data Characteristics and Splitting Strategy: A key rationale for selecting NDWS as the auxiliary dataset is the significant overlap in data modalities and the nature of the wildfire phenomena captured, despite differences in temporal structure (full time series in WTS versus single-day pairs in NDWS) and spatial resolution (375 m versus 1 km). Fig. 1 provides a visual comparison of the input features available in WTS (with $M_{\text{WTS}} = 23$ channels) and NDWS (with $M_{\text{NDWS}} = 12$ channels).

As illustrated by the modalities with red titles in Fig. 1, both datasets share several critical input features essential for wildfire modeling. These common modalities include fuel, weather, humidity, and topography. Let $M_{\text{shared}} = 11$ denote the number of these shared modalities. While WTS incorporates additional features (e.g., EVI2, various VIIRS reflectance bands), the foundational shared $M_{\text{shared}} = 11$ modalities provide a basis for transferring learned representations. This allows the model to learn fundamental relationships relevant to wildfire behavior during the intermediate fine-tuning on the larger NDWS dataset, before being further specialized on WTS.

For the data splitting strategy, all $N_{\text{NDWS}} = 18\,545$ events from NDWS (2012–2020) are utilized during the intermediate progressive fine-tuning stage. For WTS, a specific temporal split is employed: data from 2018–2020 are used for training, while data from 2021 ($N_{\text{WTS,val}}$) is reserved as a distinct testing set. This ensures that the test data comes from a time period not seen during training, preventing data leakage.

B. Cross-Modal Progressive Fine-Tuning

This section outlines our proposed CMPF methodology, a framework developed to more effectively adapt foundation models for GeoAI tasks. The overall architecture of the CMPF approach is illustrated in Fig. 2. CMPF aims to address common challenges such as domain gaps and data scarcity through two strategies: 1) informed cross-modal architectural choice and 2) progressive fine-tuning.

1) Informed Cross-Modal Architectural Choice: The "informed cross-modal architectural choice" strategy posits that aligning a foundation model's original pretraining with the target geospatial application's characteristics is key to optimal performance. For spatiotemporal applications like wildfire spread forecasting from multitemporal satellite imagery, we hypothesize that models pretrained on general video data will demonstrate stronger adaptation than models pretrained solely on spatial images, due to their inherent ability to learn spatiotemporal dynamics.

For investigating this hypothesis, we evaluated three Transformer models: ViT [43], MViTv2 [44], and VideoMAEv2

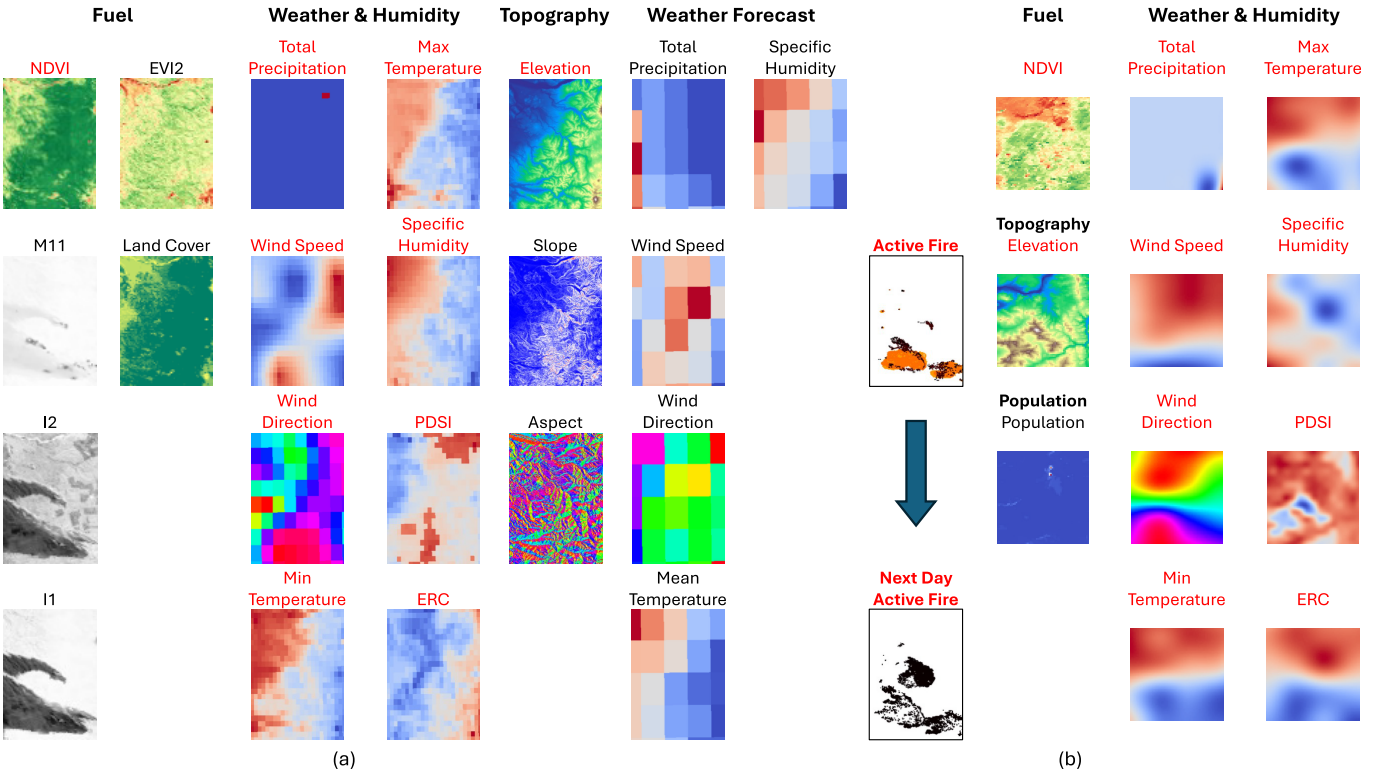


Fig. 1. Comparison of input data modalities for (a) WTS dataset and (b) NDWS dataset. Modalities with titles highlighted in red are common to both datasets.

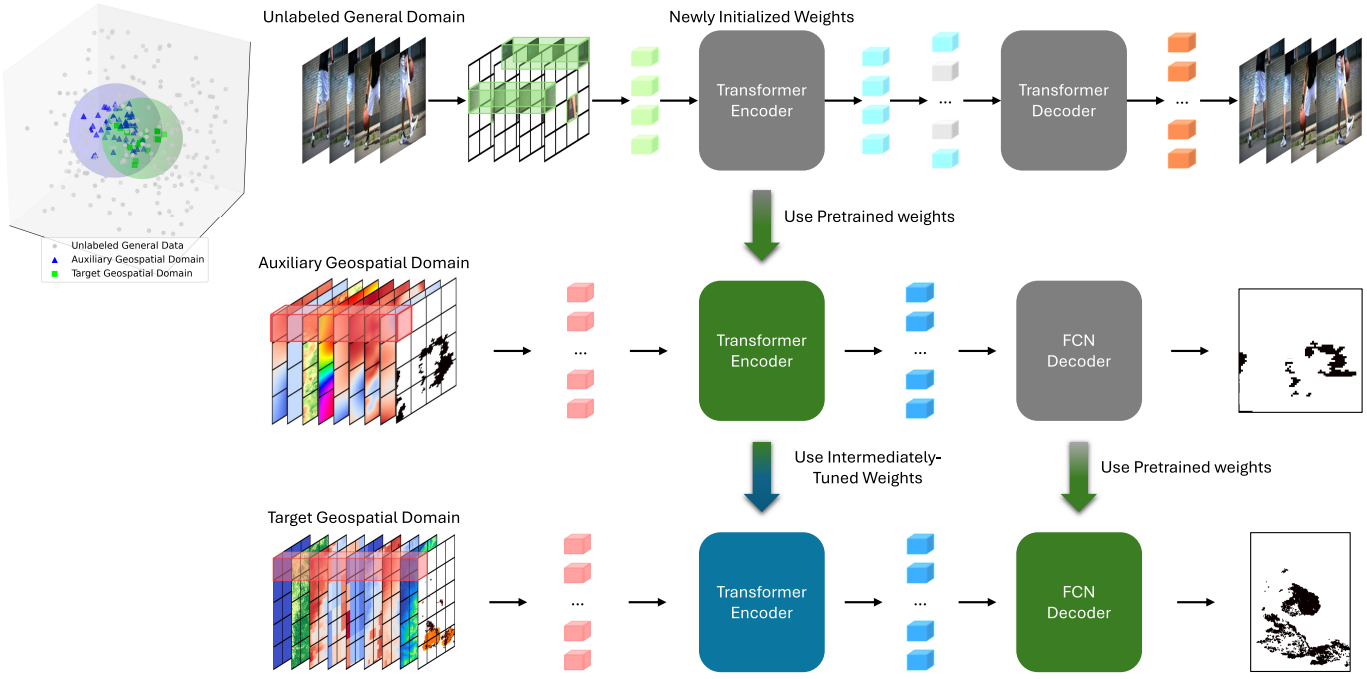


Fig. 2. Overall architecture of the proposed CMPF framework. The process consists of three main stages. (Top) Foundation model, such as a video transformer, is pretrained on a large-scale, unlabeled, general domain dataset (e.g., natural videos). (Middle) Pretrained Transformer Encoder is adapted for an intermediate geospatial task using an auxiliary dataset (e.g., NDWS). This step uses the pretrained weights and fine-tunes the model, including a new fully convolutional network (FCN) decoder, to learn general geospatial features. (Bottom) Final stage involves task-specific fine-tuning on the target geospatial dataset (e.g., WTS). The model is initialized with the intermediate-tuned weights from the previous stage and is further fine-tuned to produce the final predictions for the specific task.

[45]. Variants of these architectures with approximately 24 layers of depth were selected to allow a focused comparison of their architectural and pretraining differences, rather than model scale. The specifics of each model are as follows.

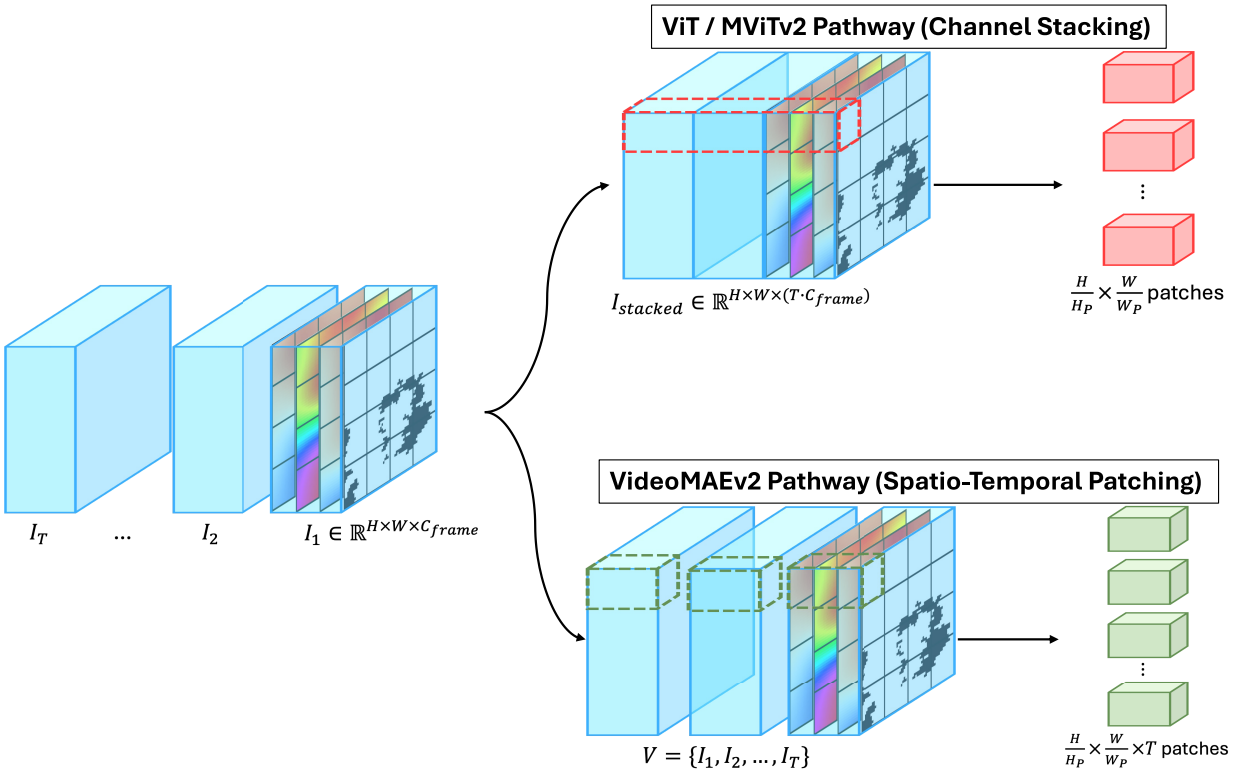


Fig. 3. Spatiotemporal tokenization strategies for ViTs. The top path shows the channel stacking and 2-D patching method used by ViT and MViTv2. The bottom path illustrates the direct 3-D spatiotemporal patching from a video volume, as used by VideoMAEv2.

1) *ViT:* This model serves as our baseline, representing a standard transformer architecture adapted for image data. To handle multitemporal data, such as a sequence of T daily satellite image sets (frames), where each frame $I_t \in \mathbb{R}^{H\times W\times C_{\text{frame}}}$ (with C_{frame} being the number of modalities for that frame from the specific dataset, e.g., M_{shared} or M_{WTS}), we employ a channel-stacking strategy, as illustrated in the top pathway of Fig. 3. The T frames are concatenated along the channel dimension, forming a single, thicker input tensor $I_{\text{stacked}} \in \mathbb{R}^{H\times W\times(T\cdot C_{\text{frame}})}$. This tensor is then divided and embedded into nonoverlapping patches with the number of patches is $H/H_P \times W/W_P$, where H_P, W_P are patch dimensions. A critical implication of this approach is that the input channel dimension $(T \cdot C_{\text{frame}})$ differs from the one used during ImageNet pretraining (typically three channels). Consequently, the initial patch-embedding layer, which projects these patches into tokens, cannot use the pretrained weights and must be reinitialized and trained from scratch for our specific temporal depth T. After this custom embedding, the tokens are processed by a series of identical transformer blocks, as shown in Fig. 4(a), which maintain a uniform token resolution throughout the network.
2) *MViTv2:* MViTv2 refines the ViT architecture by introducing a hierarchical structure that processes features at multiple scales. It begins with the same channel-stacking and 2-D patching tokenization method as ViT (top pathway of Fig. 3). As such, it faces the same challenge: the change in input channel depth requires that its patch-embedding layer must also be reinitialized and retrained. Where it deviates is in its network body, as depicted in Fig. 4(b). Instead of ViT's uniform processing, MViTv2 contains several stages where pooling layers progressively downsample the spatial resolution of tokens while increasing their feature channel depth. This pyramidal design allows the model to capture both fine-grained and global features simultaneously.
3) *VideoMAEv2:* In contrast to the image-centric models, VideoMAEv2 is explicitly designed for learning from video data, which provides a significant advantage for weight initialization. As illustrated in the bottom pathway of Fig. 3, it bypasses channel-stacking in favor of direct spatiotemporal patching. From an input sequence $V = \{I_1, I_2, \ldots, I_T\}$, it extracts 3-D spatiotemporal patches with the number of patches equals $H/H_P \times W/W_P \times T$. The architecture's explicit temporal design offers a distinct advantage through an intermediate training stage. By first continuing the pretraining of VideoMAEv2 on a large, domain-specific dataset with the correct C_{frame} channels, we create a specialized backbone. The key benefit is that for the final, downstream supervised task, we can fully leverage the weights from this intermediate stage, including the now-compatible patch-embedding layer. Unlike ViT and MViTv2, which require reinitializing this crucial first layer based on the sequence length T, our adapted VideoMAEv2 retains its powerful, pretrained understanding of spatiotemporal

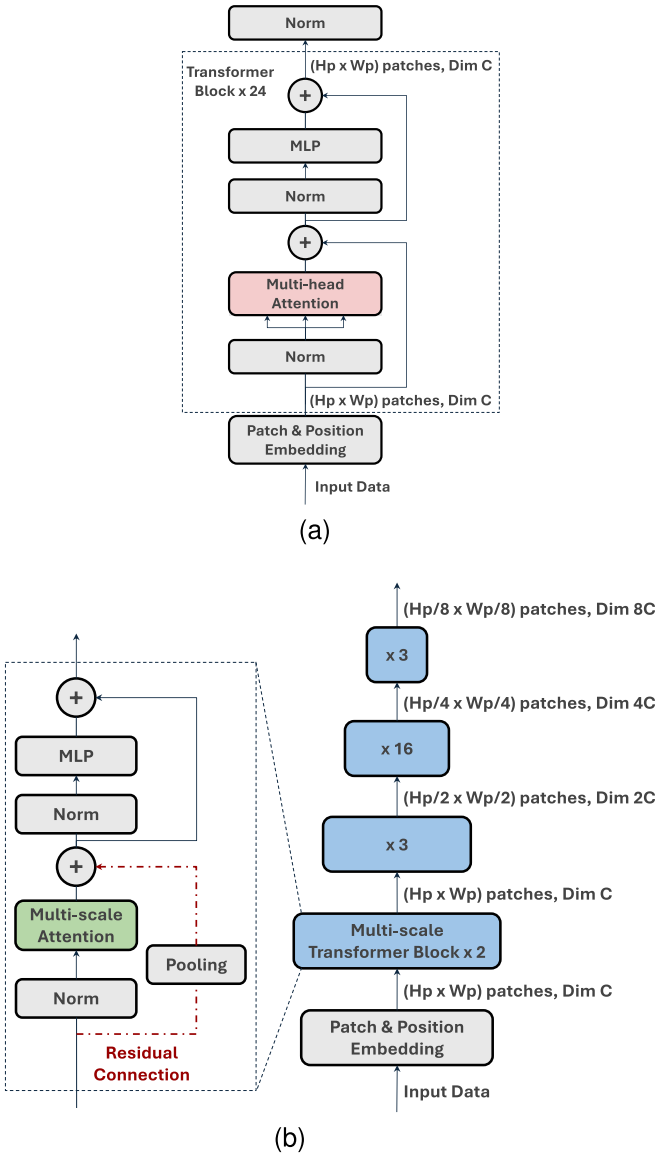


Fig. 4. Comparison of transformer architectures. (a) ViT. (b) MViTv2.

structures. This allows its self-attention mechanism [in a block structure like Fig. 4(a)] to effectively model relationships across both space and time from the very first layer, making it inherently superior for capturing temporal evolution.

The selection of ViT, MViTv2, and VideoMAEv2 directly supports the evaluation of our "informed cross-modal architectural choice" hypothesis. These widely used architectures in image analysis and computer vision form the basis of current GFMs. Their differing input modalities (image versus video) also allow us to assess the impact of pretraining modality when adapting to GeoAI tasks with highly specific multimodal inputs. Wildfire forecasting, for instance, integrates dynamic weather variables with satellite imagery, presenting unique adaptation requirements that may differ significantly even from the standard EO data used to pretrain many GFMs. This highlights the need for robust adaptation strategies regardless of the initial foundation model.

Therefore, by comparing the performance of the image-centric ViT and MViTv2 against the video-centric VideoMAEv2 on the spatiotemporal wildfire forecasting task, we can directly test whether pretraining on an analogous spatiotemporal domain (i.e., general video) provides benefits. This comparison aims to validate the principle of aligning a foundation model's architectural and pretraining characteristics with the inherent nature (e.g., spatial image versus spatiotemporal data sequence) of the target GeoAI application.

2) Progressive Fine-Tuning: The second core strategy of CMPF is "progressive fine-tuning," a multistage adaptation process designed to incrementally bridge the domain gap from general pretrained models to the specific target GeoAI application. Let Θ_{FM} be the initial parameters of a selected pretrained foundation model M_{FM}. This protocol involves the following steps.

1) *Initialization and Input Adaptation:* The process begins with M_{FM} initialized with Θ_{FM}. Two key modifications are made: 1) the initial patch-embedding layer, E_{patch}, is modified to E'_{patch} to accept the specific number of input channels C_{in} present in the geospatial data tensors. For instance, if using ViT/MViTv2 with T temporal frames stacked, each having C_{frame} channels, then $C_{\text{in}} = T \cdot C_{\text{frame}}$. For VideoMAEv2, $C_{\text{in}} = C_{\text{frame}}$ for the 3-D patch embedding. The parameters of E'_{patch} become part of the model's learnable parameters and 2) a task-specific segmentation head, H_{seg}, with initially randomized parameters Θ_{head}, is appended to the Transformer backbone to produce pixelwise fire spread predictions.
2) *Intermediate Geospatial Domain Adaptation on Auxiliary Data:* The architecturally adapted model undergoes an intermediate fine-tuning stage on the auxiliary NDWS dataset (depicted in the middle row of Fig. 2), $D_{\text{aux}} = \{(x_i^{\text{aux}}, y_i^{\text{aux}})\}_{i=1}^{N_{\text{aux}}}$. Here, x_i^{aux} represents input tensors constructed using exclusively the M_{shared} common modalities (so $C_{\text{frame}} = M_{\text{shared}}$ for defining C_{in} in E'_{patch} for this stage), and y_i^{aux} are the corresponding target fire masks. The objective is to learn updated parameters for both the foundation model's backbone ($\Theta_{\text{FM}}^{(1)}$) and the downstream head ($\Theta_{\text{head}}^{(1)}$).
The task of predicting active fire spread is characterized by a significant class imbalance, with active fire pixels typically representing a very small fraction of the total area (e.g., approximately 0.1% of pixels in our datasets). To address this challenge, the loss function $\mathcal{L}_{\text{aux}}$ used for this segmentation task is the focal loss ($\mathcal{L}_{\text{Focal}}$), which is specifically designed to handle such severe class imbalance by downweighting the loss assigned to well-classified examples (abundant background pixels) and thereby focusing learning on the rare positive class (e.g., active fire pixels in our study) [46].
3) *Final Task-Specific Fine-Tuning on Target Data:* The model with parameters $\Theta_{\text{FM}}^{(1)}$ and $\Theta_{\text{head}}^{(1)}$ obtained from the intermediate adaptation is then fine-tuned on the target WTS dataset (depicted in the bottom row of Fig. 2), $D_{\text{target}} = \{(x_i^{\text{target}}, y_i^{\text{target}})\}_{i=1}^{N_{\text{target}}}$. For this stage, inputs x_i^{target} utilize all available M_{WTS} modalities from WTS (so

TABLE I

PERFORMANCE OF DIRECTLY FINE-TUNED MODELS ON WTS. AP IS REPORTED FOR NEXT-DAY WILDFIRE SPREAD PREDICTION USING ONE-DAY ($T = 1$) AND FIVE-DAY ($T = 5$) TEMPORAL INPUTS WITH ALL FEATURES

Model	AP ($T = 1$)	AP ($T = 5$)	Parameters ($T = 1$)	Parameters ($T = 5$)
ConvLSTM	-	0.292	-	2.40×10^5
UTAE	-	0.321	-	1.10×10^6
Res18 U-Net	0.341	0.325	1.44×10^7	1.47×10^7
ViT	0.351	0.364	3.28×10^8	3.52×10^8
MViTv2	0.359	0.371	6.67×10^7	6.72×10^7
VideoMAEv2	0.370	0.377	3.27×10^8	3.27×10^8

$C_{\text{frame}} = M_{\text{WTS}}$ for defining C_{in} in E'_{patch}, which might involve adapting E'_{patch} again if $M_{\text{WTS}} \neq M_{\text{shared}}$). The objective is to obtain the final model parameters for the backbone $\Theta^{(2)}_{\text{FM}}$ and for the forecasting head $\Theta^{(2)}_{\text{head}}$. Consistent with the intermediate stage and to continue addressing the inherent class imbalance in active fire prediction, the loss function $\mathcal{L}_{\text{target}}$ is also the focal loss ($\mathcal{L}_{\text{Focal}}$), as defined in the previous section. This final step specializes the model to the specific characteristics of the target WTS data.

To mitigate potential overfitting given the large parameterization of transformer backbones, our framework integrates multiple forms of regularization. The VideoMAEv2 encoder is pretrained with a masked-autoencoding objective that randomly removes up to 90% of spatiotemporal patches and learns to reconstruct the missing content from the remaining visible ones. This reconstruction-based learning encourages the encoder to capture semantically meaningful and noise-tolerant representations rather than memorizing low-level details. During fine-tuning, the CMPF procedure further regularizes training through an intermediate geospatial adaptation stage that bridges the gap between general foundation model pretraining and wildfire-specific specialization. This staged progression limits overfitting to the relatively small labeled dataset. In addition, dropout, weight decay, and the class-balanced focal loss are applied to improve robustness. Together, these mechanisms yield stable validation behavior and consistent generalization gains.

IV. EXPERIMENTS AND RESULTS

A. Evaluating Cross-Modal Architectural Choice

This experiment directly evaluates the "informed cross-modal architectural choice" component of our CMPF strategy. The goal is to assess if foundation models pretrained on spatiotemporal data (i.e., videos) inherently outperform those pretrained on spatial images when applied to a spatiotemporal GeoAI task like wildfire spread forecasting. To this end, we compare the direct fine-tuning performance of ViT and MViTv2, both pretrained on ImageNet-1k, against VideoMAEv2, pretrained on a large unlabeled video dataset. We also compared the performance of large, transformer-based foundation models with that of popular CNN models adapted for spatiotemporal data processing, including U-Net, ConvLSTM, and UTAE.

The experimental procedure followed steps 1 (Initialization and Input Adaptation) and 3 (final task-specific fine-tuning) outlined in the progressive fine-tuning protocol (see Section III-B2), with the omission of step 2 (intermediate geospatial adaptation). Models were adapted for the $M_{\text{WTS}} = 23$ channels of the WTS dataset and directly fine-tuned on its 2018–2020 training data. This direct fine-tuning was tested using both single-day ($T = 1$) and five-day ($T = 5$) temporal inputs to predict the next day's active fire spread. Performance was evaluated on the 2021 validation data from WTS, primarily using average precision (AP) for the fire class as the metric.

The results of this direct fine-tuning experiment are summarized in Table I. As the results show, VideoMAEv2 consistently achieved the highest AP in both temporal input scenarios. For single-day inputs ($T = 1$), VideoMAEv2 obtained an AP of 0.370, outperforming MViTv2 (0.359) and ViT (0.351). This performance advantage was maintained with five-day inputs ($T = 5$), where VideoMAEv2 reached an AP of 0.377, compared with 0.371 for MViTv2 and 0.364 for ViT.

All three transformer-based models demonstrated improved performance with the extended temporal context provided by five days of input data. Specifically, ViT's AP increased by 0.013, MViTv2's by 0.012, and VideoMAEv2's by 0.007 when increasing the input from $T = 1$ to $T = 5$. Although VideoMAEv2 showed the smallest absolute improvement with more temporal data, it started from and maintained the highest AP scores.

Regarding model size, MViTv2 operated with substantially fewer parameters (approximately 67 million) than ViT (approximately 328–352 million) and VideoMAEv2 (approximately 327 million). It is also observed that the parameter counts for ViT and MViTv2 slightly increased with the longer five-day input sequence due to the channel-stacking approach for handling temporal data. In contrast, VideoMAEv2's parameter count remained constant, reflecting its architecture designed for native processing of spatiotemporal sequences.

In addition to the transformer-based model predictions, we also include a comparison with other CNN-based models, including U-Net, ConvLSTM, and UTAE. These results are based on those reported in [36]. The comparison serves to illustrate the performance level of smaller, task-specific models. As shown, the ConvLSTM and UTAE models, both with significantly smaller parameter spaces and no need for pretraining, achieve substantially lower predictive performance compared with the transformer-based models. U-Net, using a ResNet18 backbone, achieves an AP score of 0.325 when the input sequence length is five days, which is still

considerably lower than the best-performing VideoMAEv2 model introduced for wildfire data analysis.

These initial findings lend support to the "informed cross-modal architectural choice" strategy. The superior AP scores of VideoMAEv2, pretrained on general video data, suggest that its learned representations of temporal and spatial dynamics are more readily adaptable to the spatiotemporal nature of wildfire spread forecasting compared with the image-centric ViT and MViTv2 models. This advantage holds even when MViTv2 offers greater parameter efficiency.

B. Optimizing Intermediate Adaptation

Building on the insights from the architectural comparisons in Experiment 1, this second experiment focuses on the "progressive fine-tuning" component of our CMPF methodology. This experiment investigates optimal intermediate fine-tuning strategies within our CMPF framework. We compare three strategies using the NDWS auxiliary dataset. As NDWS only provides single-day inputs, the intermediate fine-tuning stage inherently processes inputs with a temporal dimension equivalent to $T = 1$. VideoMAEv2, the top performer from Experiment 1, served as the primary architecture for this initial strategy comparison.

The three intermediate fine-tuning strategies evaluated were as follows.

1) *Strategy 1 (S1)—Shared Bands (Intermediate and Final Fine-Tuning):* In this approach, the model is first fine-tuned on the NDWS dataset, using its single-day inputs ($T = 1$) and only the $M_{\text{shared}} = 11$ input variables common to both NDWS and WTS, as identified in Section III-A3. Subsequently, for the final stage, the model is fine-tuned and evaluated on WTS using these same 11 shared bands, tested separately for $T = 1$ and $T = 5$ input scenarios; unique WTS features are deliberately excluded. This strategy assesses progressive transfer with a minimal and consistent feature space. A key aspect of this strategy is that the input processing layers do not require modification or reinitialization between the intermediate and final stages, allowing the model to leverage continuously learned weights.
2) *Strategy 2 (S2)—Auxiliary and Target Data Used Separately (Sequential, Reinitialize Input):* This strategy aims for full data utilization at each stage by sequentially adapting the model to the richest available feature sets. Initially, the model undergoes intermediate fine-tuning on the NDWS dataset, using its single-day inputs and full set of 12 input variables. Following this, for the final stage, the model is fine-tuned and evaluated on WTS using its complete set of $M_{\text{WTS}} = 23$ input variables, with separate evaluations conducted for $T = 1$ and $T = 5$ input scenarios. To accommodate the differing number of input channels between the 12 NDWS variables and the 23 WTS variables, the model's initial patch-embedding layer (E'_{patch} as described in Section III-B2, step 1) is necessarily reinitialized or adapted before the final fine-tuning. This approach tests the model's ability to adapt to different rich feature sets in sequence, leveraging the maximum information from each dataset.
3) *Strategy 3 (S3)—Shared Bands (Intermediate Training) → Full Bands (Final Fine-Tuning):* In this approach, we aim to build a robust intermediate representation by leveraging a combined dataset focused on shared patterns before final specialization. Intermediate fine-tuning is conducted using only the $M_{\text{shared}} = 11$ common input variables. The training data for this stage is a combination of the NDWS dataset and, crucially, only the training split portion of the WTS dataset. This design choice is intended to better align the model with the target data distribution early in adaptation and to mitigate abrupt domain shifts between the coarser, single-day NDWS data and the finer, multiday WTS sequences. By jointly exposing the model to both datasets, we encourage shared feature representations across related modalities and create a smoother curriculum from general pretraining to target-specific fine-tuning, consistent with the CMPF framework. Following this enriched intermediate stage, the model is fine-tuned on the WTS dataset using its complete set of $M_{\text{WTS}} = 23$ input variables, with separate evaluations performed for $T = 1$ and $T = 5$ input scenarios. The goal is to assess if a more comprehensive intermediate training on shared features translates to improved performance when adapting to the full-featured target task.

All other fine-tuning parameters, such as loss and the data splits, remain consistent with the overall methodology. For each strategy, models were trained for 60 epochs during the intermediate fine-tuning phase and subsequently for another 60 epochs during each instance of the final fine-tuning phase on WTS.

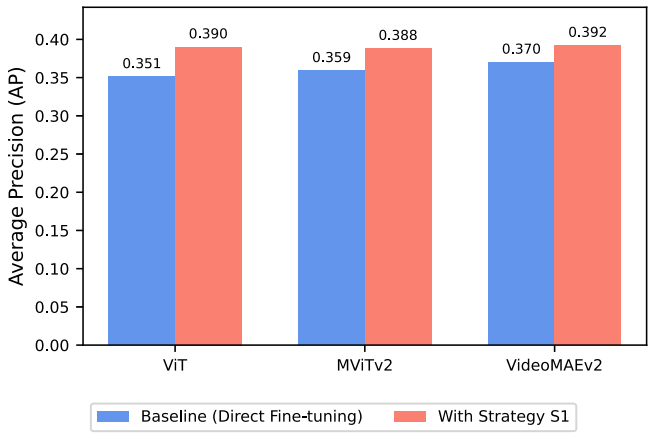
Table II details video foundation model's (i.e., VideoMAEv2) performance with the three intermediate fine-tuning strategies (S1, S2, S3), benchmarked against direct fine-tuning results from Experiment I (Baseline AP: 0.370 for $T = 1$, 0.377 for $T = 5$). All three strategies enhanced AP over this baseline for both $T = 1$ and $T = 5$ final WTS inputs, affirming the value of an intermediate adaptation stage.

Among the evaluated approaches, strategy S1 (Shared bands used at both stages) was most effective, achieving the highest AP scores: 0.392 for $T = 1$ WTS inputs (2.2% improvement over baseline) and 0.407 for $T = 5$ WTS inputs (3% improvement over the baseline). This superior performance may be attributed, in part, to S1's design, where the input processing layers remain consistent between the intermediate and final stages, allowing for uninterrupted learning and leveraging of weights for these shared features, unlike S2, which required input layer adaptation, or S3, which transitioned to a much larger feature set in the final stage. Although S2 and S3 also surpassed the baseline, with S3 showing stronger $T = 5$ performance than S2 (0.393 versus 0.391), neither matched the consistent superiority of S1 (0.407). Consequently, S1 is identified as the most effective intermediate fine-tuning strategy and is used in the subsequent experiments.

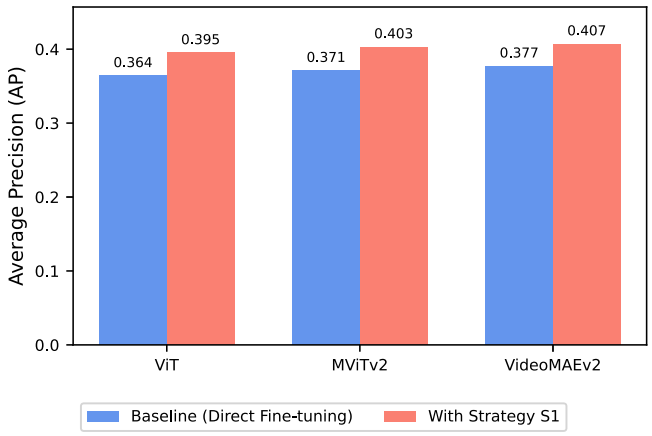
TABLE II

PERFORMANCE OF THE VIDEO FOUNDATION MODEL WITH DIFFERENT INTERMEDIATE FINE-TUNING STRATEGIES FOR WILDFIRE SPREAD FORECASTING. THE "BASELINE" REFERS TO THE MODEL DIRECTLY FINE-TUNED ON WTS WITHOUT ANY INTERMEDIATE TRAINING STAGE (RESULTS FROM EXPERIMENT I). ALL INTERMEDIATE FINE-TUNING STAGES USE AUXILIARY INPUTS EFFECTIVELY EQUIVALENT TO $T = 1$. FINAL PERFORMANCE IS REPORTED AS AP ON WTS INPUTS WITH $T = 1$ AND $T = 5$. STAGE 1: INTERMEDIATE TRAINING; STAGE 2: FINAL FINE-TUNING ON TARGET DATA

No.	Intermediate Fine-Tuning Strategy	AP ($T = 1$)	AP ($T = 5$)
Baseline	No intermediate training	0.370	0.377
S1	Shared bands used at both stages	0.392	0.407
S2	Auxiliary and target data used separately at Stage 1 and Stage 2	0.385	0.391
S3	Shared bands used at Stage 1, and full bands used at Stage 2	0.383	0.393



(a)



(b)

Fig. 5. Generalizability of the intermediate fine-tuning strategy S1 (shared bands at both stages) across different transformer architectures. Performance is compared between the baseline (direct fine-tuning) from Experiment 1 and the results "with strategy S1." (a) Performance using one-day input for wildfire spread forecasting. (b) Performance using five-day input for wildfire spread forecasting. The target dataset used is WTS (a) Predictive performance with one-day input data. (b) Predictive Performance with five-day input data.

C. Generalizability of Intermediate Fine-Tuning Across Different Transformer Architectures

Following the identification of strategy S1 (shared bands at both stages) as the most effective intermediate fine-tuning approach for the video foundation model (i.e., VideoMAEv2), we further investigated its generalizability by applying it to other mainstream transformer models (ViT and MViTv2) using the same progressive fine-tuning settings. Fig. 5 presents the predictive performance of each model when fine-tuned with strategy S1, compared with the baseline direct fine-tuning results, for both the one-day [see Fig. 5(a)] and five-day [see Fig. 5(b)] wildfire forecasting scenarios. A clear overall trend is that strategy S1 consistently improves performance over the direct fine-tuning baseline across all three architectures and both temporal input conditions. This demonstrates the robustness of the progressive fine-tuning approach (S1) in bridging domain gaps and enhancing predictive accuracy for spatiotemporal GeoAI tasks, regardless of the specific ViT backbone, thereby validating the generalizability of our CMPF methodology.

D. Impact of Auxiliary Data Volume

In this experiment, we assess how the amount of data used from the auxiliary NDWS dataset during intermediate fine-tuning affects final model performance. The goal is to understand the relationship between auxiliary data volume and the effectiveness of progressive fine-tuning. For this analysis, we used the best-performing combination from previous experiments: the VideoMAEv2 architecture with strategy S1 (shared bands at both stages). The intermediate fine-tuning stage was performed five times, each time using a different subset of the auxiliary dataset: 0% (baseline, no intermediate tuning), 25%, 50%, 75%, and 100%.

The results of this experiment are illustrated in Fig. 6. A primary finding is that increasing the volume of auxiliary data generally improves final model performance, but with clear diminishing returns. For both the one-day and five-day final tasks, performance rises steadily as the data percentage increases from 0% to 75%. The most significant performance gains are observed when introducing the first 50% of the auxiliary data.

Interestingly, the performance for both tasks begins to plateau after the 75% mark, showing diminishing returns. While using 100% of the auxiliary data yields the highest AP in both scenarios, the improvement from 75% to 100% is marginal. For the one-day task, the AP increases minimally from 0.391 to 0.392. Similarly, the five-day model's performance continues to climb from 0.404 to 0.407, but the rate of improvement slows considerably. This suggests that using the full auxiliary dataset may not be the most efficient approach, as a 75% subset achieves near-optimal performance and could offer a better tradeoff between computational cost and final accuracy.

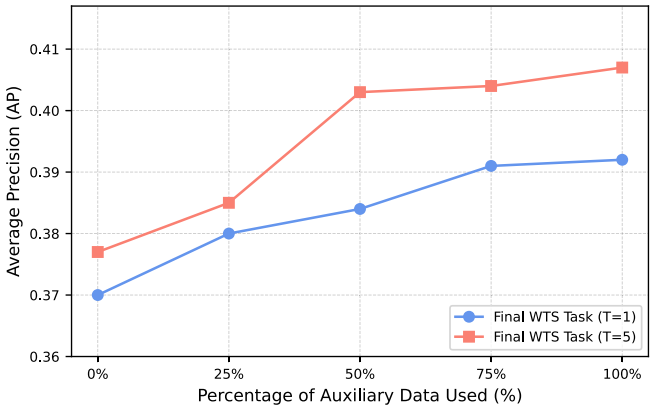


Fig. 6. Impact of auxiliary data volume on final model performance. The chart shows the model prediction accuracy on the WTS by VideoMAEv2 with varying percentages of the NDWS auxiliary dataset. Results are shown for models evaluated on both one-day ($T = 1$) and five-day ($T = 5$) input scenarios. AP is used as the performance measure. The 0% data point represents the direct fine-tuning baseline without an intermediate stage.

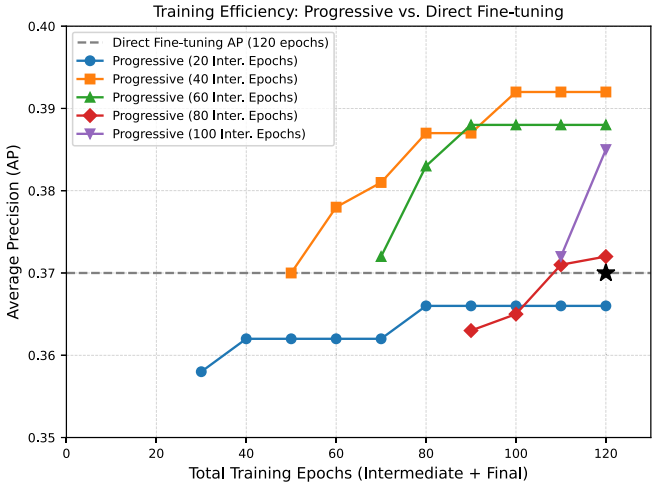


Fig. 7. Training efficiency: progressive versus direct fine-tuning. The plot compares AP against the total number of training epochs (intermediate + final). The dashed gray line marks the performance of the direct fine-tuning model after 120 epochs (with no intermediate training). Each colored line shows the performance of the model after applying a progressive fine-tuning strategy (S1), where final fine-tuning is initiated after a different number of intermediate training epochs (20, 40, 60, 80, or 100). For example, the blue line represents the progressive fine-tuning strategy with 20 epochs of intermediate training followed by 100 epochs of final fine-tuning. Since model performance is recorded every 10 additional epochs, the first point reflecting the final model predictions appears at epoch 30 (20 + 10), rather than at epoch 20 (for the blue line). The results demonstrate that progressive fine-tuning, particularly with 40 or 60 epochs at the intermediate stage, surpasses the baseline performance in fewer total epochs.

E. Computational Efficiency of Progressive Fine-Tuning

In this experiment, we compare the training efficiency of the optimal progressive fine-tuning strategy (VideoMAEv2 with strategy S1) against a direct fine-tuning baseline. The baseline model was trained directly on the WTS dataset for a total of 120 epochs. For the progressive fine-tuning approach, five separate models were first trained on the auxiliary NDWS dataset for varying durations: 20, 40, 60, 80, and 100 intermediate epochs. Following this intermediate stage, each of the five models was then fine-tuned on the final WTS dataset. During this final fine-tuning phase, performance (AP) was evaluated and recorded every 10 epochs to track the learning trajectory. The total training epochs, plotted on the x-axis in Fig. 7, represent the sum of the intermediate and subsequent final fine-tuning epochs.

The results, presented in Fig. 7, demonstrate a clear efficiency gain from using progressive fine-tuning. The direct fine-tuning baseline required 120 epochs to achieve its best AP of 0.370. In contrast, most progressive fine-tuning strategies not only reached higher peak performance but did so in fewer total epochs. Notably, the model with 40 intermediate epochs (orange line) reached 0.370 AP after just 10 epochs of final fine-tuning. This means that, with the intermediate training strategy applied, the same performance is achieved in less than half the number of epochs (50 versus 120) compared with direct fine-tuning. Similarly, the model with 60 intermediate epochs surpassed the baseline at a total of 70 epochs (green line). This indicates that intermediate adaptation using an auxiliary dataset enables the model to learn relevant geospatial features, allowing for much faster convergence during the final task-specific tuning.

The experiment also reveals a tradeoff related to the duration of intermediate training. While a very short intermediate stage (20 epochs) was insufficient to build a strong enough foundation, leading to a performance below the direct fine-tuning baseline, models with substantial intermediate training (e.g., 80 or 100 epochs) began the final fine-tuning stage with a high AP. This also means that it takes longer to reach the overall best model performance. When total training time is constrained (as in the 120-epoch limit shown in Fig. 7), the most efficient balance between performance and training cost is achieved at around one-third to one-half of the total epochs (as seen in the orange and green lines).

These findings confirm that progressive fine-tuning is not only a method for boosting final model accuracy but also a computationally efficient training strategy. By leveraging knowledge from an auxiliary dataset, it can reduce the total number of training epochs required to reach a high-performing state, offering a practical advantage for resource-constrained applications.

F. Stratified Evaluation by Fire Characteristics and Regions

To further assess the robustness and generalization capability of the CMPF framework, we conducted a stratified evaluation that decomposes model performance across different fire characteristics and spatial domains. Specifically, we analyzed results by fire size, season, and geographic location (state) using the 2021 WTS validation set. Table III summarizes AP for the baseline model (direct fine-tuning) and the CMPF-enhanced model under both one-day ($T = 1$) and five-day ($T = 5$) temporal input settings.

1) Fire Size: The CMPF model shows consistent improvements across all fire-size categories, with the largest relative gains observed for small fires (45 pixels on average), where AP increases from 0.257 to 0.291 for $T = 1$ and from 0.260 to 0.278 for $T = 5$. This improvement indicates that CMPF

TABLE III

DECOMPOSED EVALUATION OF WILDFIRE FORECASTING PERFORMANCE BY FIRE SIZE, SEASON, AND GEOGRAPHIC LOCATION. RESULTS ARE REPORTED AS AP FOR BOTH THE BASELINE (DIRECT FINE-TUNING) AND CMPF MODELS USING ONE-DAY ($T = 1$) AND FIVE-DAY ($T = 5$) TEMPORAL INPUTS. THE SPATIAL RESOLUTION IS 375 M PER PIXEL. AVERAGE FIRE SIZE IS EXPRESSED IN PIXEL COUNT

Category	Subset	Baseline (T=1)	Baseline (T=5)	CMPF (T=1)	CMPF (T=5)	Average Size (Pixels)	Number of Events
Fire Size	Small	0.257	0.260	0.291	0.278	45	-
	Medium	0.439	0.448	0.462	0.482	486	-
	Large	0.449	0.465	0.465	0.486	1256	-
Season	Winter (12-2)	0.005	0.015	0.030	0.032	8	-
	Spring (3-5)	0.003	0.008	0.004	0.004	10	-
	Summer (6-8)	0.353	0.374	0.380	0.402	152	-
	Fall (9-11)	0.448	0.418	0.457	0.440	208	-
Location (State)	California	0.366	0.383	0.378	0.397	161	32
	Colorado	0.437	0.330	0.404	0.332	82	7
	Idaho	0.377	0.394	0.411	0.416	160	32
	Montana	0.365	0.379	0.391	0.460	135	18
	Nebraska	0.394	0.485	0.390	0.459	144	2
	Nevada	0.073	0.074	0.108	0.090	25	2
	New Mexico	0.290	0.186	0.332	0.288	65	12
	North Dakota	0.001	0.001	0.001	0.001	10	3
	Oregon	0.358	0.374	0.415	0.407	176	27
	Utah	0.274	0.269	0.263	0.321	50	2
	Washington	0.435	0.438	0.410	0.442	208	14
	Wyoming	0.309	0.268	0.285	0.349	67	5

enhances sensitivity to small and fragmented fire events that are typically more challenging to detect due to class imbalance and weaker spatial signals. Medium and large fires, which occupy larger pixel areas (486–1256 pixels on average), also benefit modestly from CMPF, achieving stable or slightly higher AP values compared with the baseline. Overall, CMPF demonstrates reliable performance across scales, suggesting that progressive fine-tuning effectively transfers representations for both minor and extensive events.

2) Seasonal Variation: The decomposed seasonal results highlight how fire prediction difficulty varies with climatological conditions. Both baseline and CMPF models perform poorly in spring (AP $\approx$ 0.003–0.008) and winter (AP $\approx$ 0.005–0.032) when active fires are infrequent and spatially limited. Performance increases sharply during summer and fall, consistent with the concentration of wildfire activity in these periods. CMPF yields the most noticeable advantage in fall, improving AP from 0.418 to 0.440 for $T = 5$. The modest but consistent improvements across all seasons imply that intermediate geospatial adaptation allows the model to generalize better across distinct temporal regimes without overfitting to a particular seasonal pattern.

3) Geographic Variation: Across all three stratification dimensions, CMPF consistently matches or surpasses the baseline, confirming that the proposed progressive fine-tuning framework enhances generalization across event scales, seasons, and landscapes. The model's improved stability for small fires and geographically diverse regions is particularly notable, suggesting that intermediate adaptation helps retain transferable spatial and temporal features relevant under varying conditions. These results reinforce CMPF's robustness for operational wildfire forecasting and demonstrate its potential scalability to other spatiotemporal GeoAI applications characterized by heterogeneous input domains.

4) Summary: Across all three stratification dimensions, CMPF consistently matches or surpasses the baseline, confirming that the proposed progressive fine-tuning framework enhances generalization across event scales, seasons, and landscapes. The model's improved stability for small fires and geographically diverse regions is particularly notable, suggesting that intermediate adaptation helps retain transferable spatial and temporal features relevant under varying conditions. These results reinforce CMPF's robustness for operational wildfire forecasting and demonstrate its potential scalability to other spatiotemporal GeoAI applications characterized by heterogeneous input domains.

G. Feature Relevance-Reliance Analysis Using Mutual Information (MI)

To further interpret how the proposed CMPF model leverages multisource geospatial inputs, we conduct a quantitative feature relevance-reliance analysis using MI. MI measures the amount of shared information between two variables, capturing both linear and nonlinear statistical dependencies without assuming any specific functional relationship. In this context, MI serves as a model-agnostic diagnostic tool to identify how strongly each input feature contributes to predicting wildfire spread and whether the model's learned dependencies align with the intrinsic information structure of the data. Specifically, we compute the normalized MI (NMI) between each input feature channel X_c and 1) the ground-truth wildfire spread mask Y, referred to as feature–ground truth dependence and 2) the model prediction $\widehat{Y}$, referred to as feature–model prediction dependence. Higher NMI values indicate stronger statistical dependence, reflecting either the inherent predictive power of a feature or the degree to which the model relies on it during inference.

This analysis is performed using the best-performing CMPF-trained VideoMAEv2 model with five-day input sequences ($T = 5$), which achieved an AP of 0.407 on the WTS dataset. As shown in Fig. 8 (left), both the ground truth and model exhibit high dependence on physically meaningful drivers such as active fire, wind direction, and maximum temperature, confirming that CMPF effectively captures the

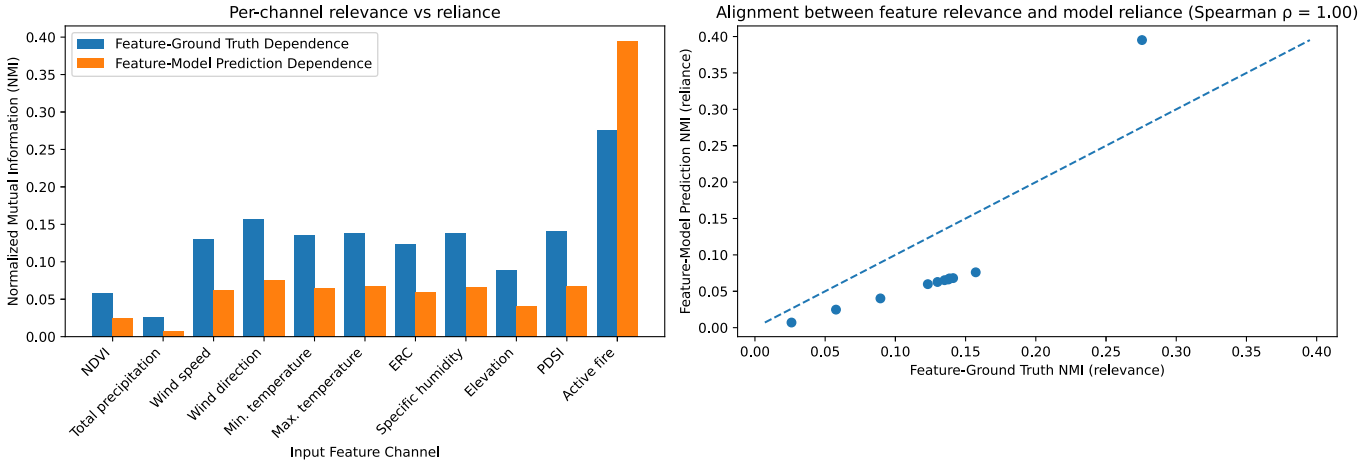


Fig. 8. MI-based interpretability analysis of the best-performing CMPF model ($T = 5$). (Left) NMI between each input feature channel and the ground-truth fire mask (feature–ground truth dependence, blue) and between each input feature and model prediction (feature–model prediction dependence, orange). Higher NMI indicates stronger statistical dependence. (Right) Scatter plot showing the alignment between feature relevance and model reliance across all input channels, with the Spearman rank correlation coefficient (ρ) indicating the degree of correspondence between the two. Points near the 1:1 dashed line suggest that the model's learned dependencies align closely with the intrinsic information content of the input features.

dominant environmental controls of wildfire spread. The per-channel trend of model reliance (orange) closely follows the intrinsic relevance distribution (blue), though with slightly reduced magnitudes, an expected effect of model regularization and feature interaction smoothing during fine-tuning. Notably, the active fire variable exhibits a higher model-side NMI than its ground-truth relevance, suggesting that the CMPF model emphasizes this channel more strongly than the feature–ground truth dependence would imply. This overreliance likely stems from the variable's strong spatial correspondence to current fire boundaries, which serve as a direct indicator of near-future spread, making it a high-signal but potentially self-reinforcing feature in predictive learning.

To quantify the overall consistency between the two dependency patterns, we compute the Spearman rank correlation coefficient ($\rho = 1.00$) between per-channel NMI values (see Fig. 8, right). The near-perfect monotonic alignment indicates that the CMPF model's learned feature dependencies are strongly coherent with the true statistical structure of the input–target relationship. Importantly, the close correspondence between relevance and reliance provides strong evidence that the CMPF model grounds its predictions in physically meaningful signals rather than artifacts of data distribution or noise. This interpretability analysis thus reinforces CMPF's core objective: enabling GFMs to achieve high predictive accuracy while maintaining transparent, domain-consistent reasoning.

In addition to the MI analysis, we incorporate a spatial interpretability analysis to quantify each input variable's localized contribution to the predicted fire spread. The spatial attribution/attention map was generated using integrated gradients (IG) [47] for all 11 input variables on a per-sample basis. Because attribution patterns naturally vary across different fire events, we select the top five variables for each example by ranking their mean absolute IG attribution within that specific sample. This samplewise ranking ensures that the visualization highlights the variables most influential for that particular event. Fig. 9 presents two wildfire events; for each event, the top five input variables (top row) are shown alongside their corresponding IG attribution maps (bottom row). The resulting visualizations reveal several physically consistent patterns.

In the first event [see Fig. 9(a)], the CMPF model places its highest attribution on active fire, confirming its strong reliance on the current fire perimeter as the primary indicator of next-day spread. PDSI and temperature variables show broad, spatially smooth contributions, consistent with their roles in shaping fuel moisture and combustibility. Elevation also exhibits meaningful attribution, particularly along terrain gradients, reflecting known topographic influences on fire acceleration and spread direction. In the second event [see Fig. 9(b)], the ordering of influential variables shifts, illustrating that CMPF adapts its reliance to event-specific conditions. NDVI emerges as a key contributor where vegetation contrasts are strong, while PDSI, Specific Humidity, and ERC highlight local moisture and fuel-dryness conditions that shape spread potential. Active fire again remains a dominant signal. Despite variation across events, the IG maps consistently form smooth, interpretable structures rather than noisy or artifact-driven patterns.

This result verifies that the model effectively balances immediate fire-state cues with environmental controls, and that the relative contributions vary naturally across events in response to local conditions. This interpretability analysis reinforces that the performance gains achieved through CMPF are accompanied by domain-consistent reasoning rather than overfitting to spurious correlations.

H. Visual Analysis of Forecasting Results

This section provides a qualitative visual analysis of wildfire forecasting performance, highlighting how the CMPF

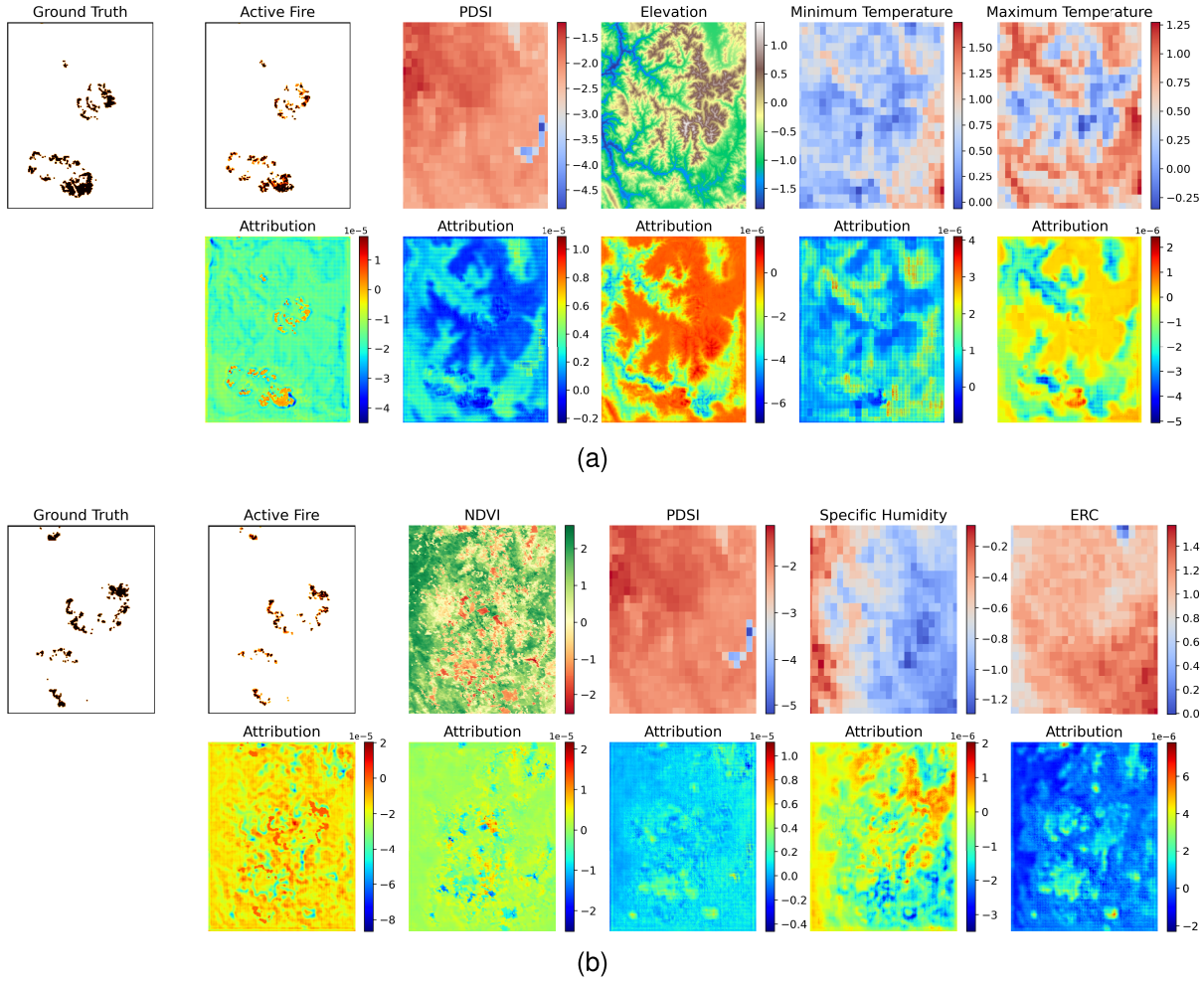


Fig. 9. Spatial attribution/attention maps generated using IGs for two wildfire events. For each event, IG scores are computed for all input variables, and the top five variables, ranked by samplewise mean absolute attribution, are shown in the top row. The bottom row visualizes the corresponding IG attribution/attention maps, highlighting where each variable most strongly influences the predicted fire spread.

framework improves predictive accuracy and spatial coherence. Fig. 10 presents three representative wildfire events from the 2021 validation set, comparing ground-truth fire spread masks with predictions from the baseline model (VideoMAEv2 with direct fine-tuning) and the CMPF-trained model (VideoMAEv2 with strategy S1) under both single-day ($T = 1$) and five-day ($T = 5$) temporal inputs. For each sample, the AP scores from different methods are presented to establish a direct quantitative–qualitative link between model performance and visual accuracy.

Across all three samples, the CMPF model consistently produces higher AP scores and visibly more complete fire predictions. In the first event, CMPF markedly reduces the number of missed interior pixels (red) and reconstructs the full fire footprint with greater coherence (AP increasing from 0.492 to 0.606 for $T = 1$). The second event shows that CMPF better preserves the detailed contours of the burned area, filling fragmented gaps that remain in the baseline (AP improving from 0.471–0.516 to 0.511–0.543). In the third case, the baseline underestimates the elongated fire front, while CMPF captures its continuous shape and overall spread direction more accurately (AP rising from 0.443 to 0.503 for $T = 5$).

Although both models yield occasional false positives (blue), CMPF's overpredictions typically appear as small extensions along active fronts rather than isolated noise. This pattern suggests a more realistic representation of fire growth dynamics. Notably, even with a single-day input ($T = 1$), CMPF outperforms the baseline using a five-day input ($T = 5$) in nearly all cases, which shows that the progressive fine-tuning procedure, rather than the temporal window length, is the primary driver of improved generalization. Overall, these visual results reinforce CMPF's ability to enhance both the precision and spatial coherence of wildfire spread forecasts, in agreement with the quantitative trends reported earlier.

V. Discussion

This study demonstrates the critical importance of architectural design and domain adaptation in tailoring foundation models for domain-specific geospatial tasks. Specifically, we find that video foundation models excel at capturing the complex spatiotemporal dynamics of wildfire spread, outperforming other transformer-based architectures that are limited to processing static satellite imagery.

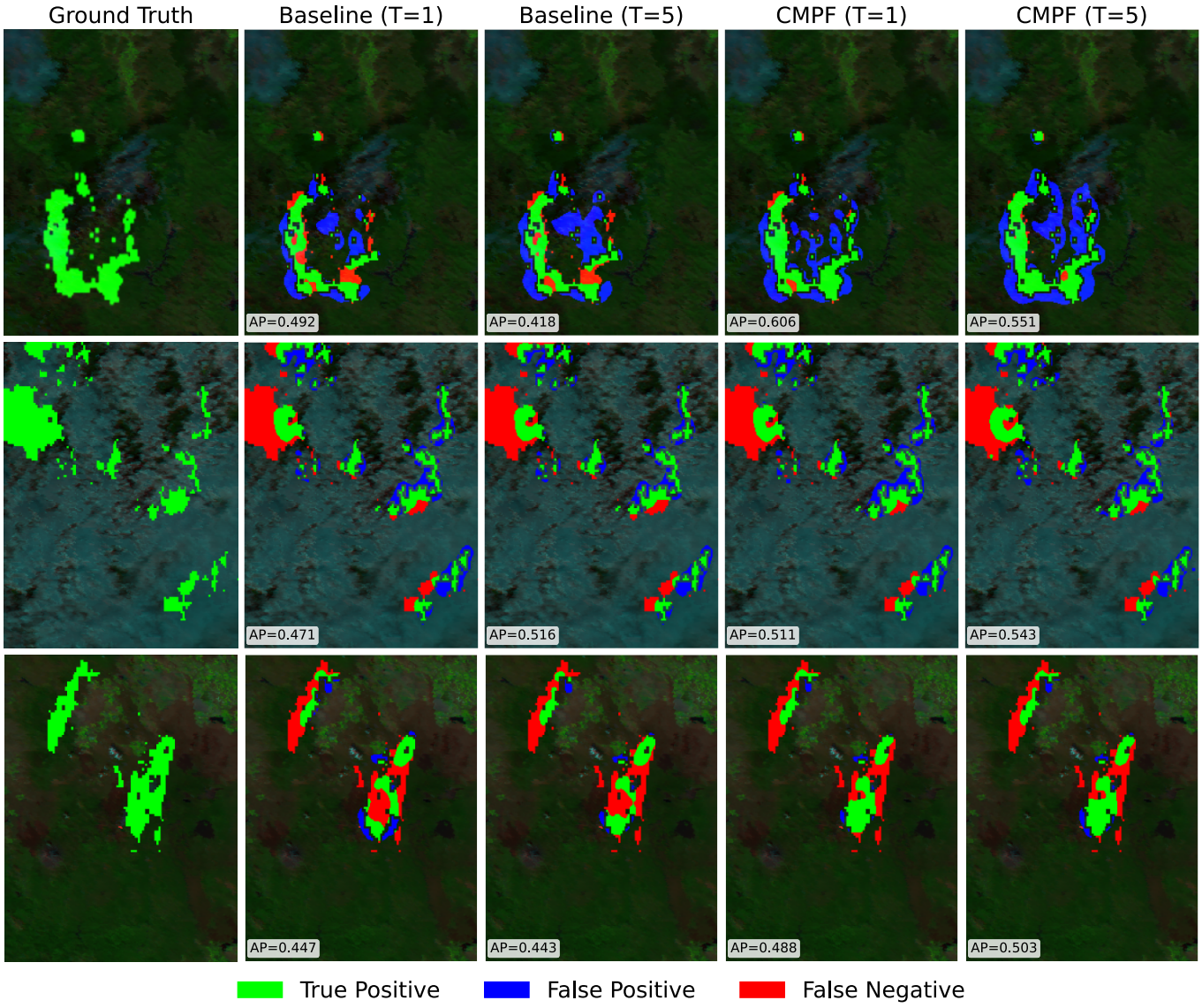


Fig. 10. Visual comparison of wildfire spread forecasts. The figure displays forecasting results for three distinct wildfire events from the 2021 validation set. Each row compares the "ground truth" fire mask against predictions from the "baseline" model (direct fine-tuning) and the proposed "CMPF" model, for both single-day ($T = 1$) and five-day ($T = 5$) temporal inputs. Colors indicate: true positive (green), false positive (blue), and false negative (red). AP: average precision.

A key contribution of this work is the development and validation of the CMPF strategy. CMPF addresses two major challenges in geospatial machine learning: domain mismatch and limited labeled data. By aligning architectures across modalities and introducing a progressive fine-tuning pathway, CMPF enables more effective transfer from general-purpose models to specialized geospatial applications. Our results show that direct fine-tuning without intermediate adaptation consistently leads to suboptimal performance, affirming the importance of structured domain adaptation.

Beyond improving accuracy, CMPF offers a more efficient training process, achieving stronger performance in fewer epochs. This efficiency is particularly advantageous for real-world scenarios where computational resources are limited and timely inference is essential. Moreover, the robustness of CMPF across multiple transformer backbones highlights its general applicability for diverse downstream spatiotemporal tasks in GeoAI.

An additional finding from our experiments is that strategy S1, which maintains a consistent input and feature space across intermediate and final stages, outperforms S2 and S3 that expand the feature set late in training. This result highlights an important tradeoff between feature-space consistency and richness. When labeled data are limited, preserving a stable input interface can yield stronger generalization than introducing additional variables after intermediate adaptation. This phenomenon echoes the classic Hughes effect in remote sensing, where increasing dimensionality without sufficient samples degrades performance [48], [49], [50].

From a representation standpoint, domain-adaptation research similarly emphasizes the benefits of feature-space alignment between stages [51], [52]. In our case, S1 preserves

the same patch-embedding pathway, ensuring representational continuity, whereas S2 and S3 require reconfiguring the input embeddings to accommodate new bands, potentially disrupting previously learned representations. Recent studies in GFMs [10], [53] also report that maintaining consistent channel semantics across modalities improves adaptation efficiency. Together, these results suggest that when auxiliary and target datasets share a strong common feature core, maintaining input consistency may be preferable to late-stage feature expansion. Future extensions could explore channel-aware adapters that enable incremental feature inclusion without reinitializing the shared-band pathway.

These findings reinforce the importance of selecting architectures capable of modeling temporal dynamics and applying domain-aware fine-tuning strategies. They also provide actionable guidance for the development of future GFMs, especially in multimodal contexts.

Unlike foundation models in general computer vision, GFMs often face additional challenges due to the heterogeneity of geospatial data (e.g., differences in data types, channels, modalities, and spatial–temporal resolutions) that are critical for tasks such as water monitoring, land cover classification, or disaster response. Bridging the gap between pretraining domains and target geospatial applications requires more adaptable model designs. Future GFMs should therefore be equipped to learn from and generalize across diverse geospatial modalities by incorporating a broader range of training data.

On the application side, our findings suggest that incorporating intermediate adaptation stages is essential for maximizing performance. Our current implementation integrates weather, fuel, and topographic information alongside multispectral satellite observations, enabling the model to capture key environmental drivers of wildfire behavior. These variables together serve as practical proxies for broader landscape controls, allowing the model to generalize across heterogeneous environments even without explicit landscape or fire-type categorizations.

Through the CMPF strategy, the model progressively adapts to wildfire-specific spatiotemporal dynamics and is therefore capable of scaling across diverse fire regimes and landscape settings. Future work will explore the transferability of foundation models across varying geographic regions and investigate new intermediate training strategies that account for both shared and localized characteristics of geospatial phenomena. Additional enhancements, such as incorporating finer-grained fuel or landscape descriptors and explicitly encoding location information, represent promising opportunities to further strengthen the model's ability to capture dynamic, region-specific fire behavior.

Beyond wildfires, CMPF serves as a general framework for spatiotemporal GeoAI research, beginning with the selection of an architecture aligned with the sensing modality of the geospatial application. For example, video-centric Transformers are well-suited for forecasting spatiotemporal phenomena because they can process sequences of related frames over time. Introducing an intermediate training stage on an auxiliary dataset that shares core variables with the target dataset can further improve model performance by reducing domain shift during final fine-tuning.

VI. Conclusion

This work advances the integration of foundation models in GeoAI by presenting a validated strategy for adapting video-based architectures to domain-specific forecasting tasks. Focusing on wildfire spread prediction as a case study, we demonstrate that coupling an informed architectural choice such as VideoMAEv2 with a tailored adaptation approach through CMPF yields notable improvements in both predictive accuracy and training efficiency.

The results affirm the promise of video foundation models for spatiotemporal geospatial tasks. Building on these results, CMPF's structured adaptation, validated across multiple backbones, provides a principled strategy for leveraging foundation models under conditions of label scarcity and domain mismatch. By aligning architecture and fine-tuning methodology with the nature of geospatial data, this study lays foundational work for scalable, high-performing AI systems in environmental monitoring and Earth system modeling.

Our proposed CMPF framework differs from existing adaptation approaches, such as domain-adaptive pretraining and standard multistage fine-tuning in several important ways. Unlike domain-adaptive pre-training, which typically requires retraining a large general-purpose foundation model on substantial amounts of domain-specific geospatial data, CMPF offers a data-efficient adaptation stage that progressively bridges natural-video pretraining with multispectral wildfire forecasting. This is especially valuable given the limited availability of labeled wildfire spread data.

In addition, CMPF introduces a cross-modal architectural alignment step that explicitly adapts video foundation models to heterogeneous geospatial inputs, rather than assuming red, green, blue (RGB)-like input consistency. This also contrasts with conventional multistage fine-tuning pipelines, which generally assume that all stages share identical or closely matched input modalities (for example, RGB to RGB) or consistent feature sets. In practice, wildfire forecasting datasets for the same task often differ in their available variables across regions and years. CMPF explicitly accommodates this real-world GeoAI scenario by enabling both full-modality and shared-modality strategies across stages, providing a more flexible and representative multistage adaptation framework for geospatial applications. Looking ahead, the CMPF approach holds potential for broader applications beyond wildfire forecasting. Its adaptability across transformer backbones and modalities makes it a compelling strategy for future research in climate modeling, disaster response, and other complex geospatial phenomena. For example, the same approach can support weather nowcasting and short-term prediction [54], where the model captures interrelationships within a short timeframe, such as 24-h wildfire spread forecasting. It can also extend to seasonal forecasting problems, such as predicting sea ice extent with lead times of one to six months using time series of climate variables [55]. In addition, our work provides a strong testbed for gradual-change phenomena such as Arctic permafrost thaw, which is influenced by both climatic drivers and

seasonal hazards like wildfires. This will enable seasonal to subseasonal forecasting of abrupt thaw events, which remains a significant gap in permafrost research [56].

This work serves as a step toward more intelligent, transferable, and resource-efficient GeoAI systems.

Data Availability

The data, models, and experimental code supporting the findings of this study are publicly available in the project GitHub repository at https://asucicilab.github.io/wildfire-forecasting/

References

[1] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[2] C. Li et al., "Multimodal foundation models: From specialists to general-purpose assistants," *Found. Trends Comput. Graph. Vis.*, vol. 16, nos. 1–2, pp. 1–214, 2024.

[3] M. Awais et al., "Foundation models defining a new era in vision: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2245–2264, Apr. 2025.

[4] Y. He et al., "Foundation model for advancing healthcare: Challenges, opportunities and future directions," *IEEE Rev. Biomed. Eng.*, vol. 18, pp. 172–191, 2025.

[5] G. Mai, C. Cundy, K. Choi, Y. Hu, N. Lao, and S. Ermon, "Towards a foundation model for geospatial artificial intelligence (vision paper)," in *Proc. 30th Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2022, pp. 1–4.

[6] W. Li, "GeoAI: Where machine learning and big data converge in GIScience," *J. Spatial Inf. Sci.*, vol. 2020, no. 20, pp. 71–77, Jun. 2020.

[7] D. Szwarcman et al., "Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth observation applications," 2024, *arXiv:2412.02732*.

[8] W. Li et al., "GeoAI for science and the science of GeoAI," *J. Spatial Inf. Sci.*, vol. 2024, no. 29, pp. 1–17, Sep. 2024.

[9] R. Lam et al., "Learning skillful medium-range global weather forecasting," *Science*, vol. 382, no. 6677, pp. 1416–1421, Dec. 2023.

[10] C.-Y. Hsu, W. Li, and S. Wang, "Geospatial foundation models for image analysis: Evaluating and enhancing nasa-ibm Prithvi's domain adaptability," *Int. J. Geographical Inf. Sci.*, vol. 39, no. 9, pp. 1–30, 2024.

[11] W. Li et al., "Segment anything model can not segment anything: Assessing AI foundation model's generalizability in permafrost mapping," *Remote Sens.*, vol. 16, no. 5, p. 797, Feb. 2024.

[12] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[13] W. Li and C.-Y. Hsu, "GeoAI for large-scale image analysis and machine vision: Recent progress of artificial intelligence in geography," *ISPRS Int. J. Geo-Information*, vol. 11, no. 7, p. 385, Jul. 2022.

[14] J. Jakubik et al., "Foundation models for generalist geospatial artificial intelligence," 2023, *arXiv:2310.18660*.

[15] J. Jakubik et al., "TerraMind: Large-scale generative multimodality for Earth observation," 2025, *arXiv:2504.11171*.

[16] W. Li, H. Lee, S. Wang, C.-Y. Hsu, and S. T. Arundel, "Assessment of a new GeoAI foundation model for flood inundation mapping," in *Proc. 6th ACM SIGSPATIAL Int. Workshop AI Geographic Knowl. Discovery*, Nov. 2023, pp. 102–109.

[17] J. Schneider, C. Meske, and P. Kuss, "Foundation models: A new paradigm for artificial intelligence," *Bus. Inf. Syst. Eng.*, vol. 66, no. 2, pp. 221–231, Apr. 2024.

[18] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[19] J. Chen et al., "Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey," *Proc. IEEE*, vol. 111, no. 6, pp. 653–685, Jun. 2023.

[20] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. individual differences*, vol. 103, Nov. 2023, Art. no. 102274.

[21] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[22] T. Zhan, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10078–10093.

[23] W. Tu, W. Deng, and T. Gedeon, "A closer look at the robustness of contrastive language-image pre-training (CLIP)," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 13678–13691.

[24] M. G. Z. Hashemi et al., "Estimating crop biophysical parameters from satellite-based SAR and optical observations using self-supervised learning with geospatial foundation models," *Remote Sens. Environ.*, vol. 327, Sep. 2025, Art. no. 114825.

[25] W. Li et al., *Landslide hazard mapping with geospatial foundation models: Geographical generalizability, data scarcity, and band adaptability*, 2025, *arXiv:2511.04474*.

[26] K. R. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[27] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey," 2024, *arXiv:2403.14608*.

[28] C. Bodnar et al., "A foundation model for the Earth system," *Nature*, vol. 641, no. 8065, pp. 1180–1187, 2025.

[29] A. L. Westerling, "Increasing western U.S. forest wildfire activity: Sensitivity to changes in the timing of spring," *Phil. Trans. Roy. Soc. B: Biol. Sci.*, vol. 371, no. 1696, Jun. 2016, Art. no. 20150178.

[30] M. W. Jones et al., "Global and regional trends and drivers of fire under climate change," *Rev. Geophysics*, vol. 60, no. 3, p. 2020, Sep. 2022.

[31] M. A. Finney, *FARSITE, Fire Area Simulator–Model Development and Evaluation*. Washington, DC, USA: U.S. Department of Agriculture, 1998.

[32] C. Tymstra et al., *Development and structure of prometheus: The Canadian wildland fire growth simulation model*, Forest Service, Northern Forestry Centre, Edmonton, AB, Canada, Inf. NOR-X-417, 2010.

[33] Y. Zhao, S. Gerard, and Y. Ban, "TS-SatFire: A multi-task satellite image time-series dataset for wildfire detection and prediction," 2024, *arXiv:2412.11555*.

[34] H. S. Andrianarivony and M. A. Akhloufi, "Machine learning and deep learning for wildfire spread prediction: A review," *Fire*, vol. 7, no. 12, p. 482, Dec. 2024.

[35] D. Radke, A. Hessler, and D. Ellsworth, "FireCast: Leveraging deep learning to predict wildfire spread," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4575–4581.

[36] S. Gerard, Y. Zhao, and J. Sullivan, "WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp.74515–74529.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[38] X. Shi, Z. Chen, H. Wang, D. Yeung, W. K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[39] V. S. Fare Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4852–4861.

[40] S. Singla et al., "Wildfiredb: An open-source dataset connecting wildfire occurrence with relevant determinants," in *Proc. 35th Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 74515–74529.

[41] F. Huot, R. L. Hu, N. Goyal, T. Sankar, M. Ihme, and Y.-F. Chen, "Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412513, doi: 10.1109/TGRS.2022.3192974.

[42] Q. E. Barber et al., "The Canadian fire spread dataset," *Sci. Data*, vol. 11, no. 1, p. 764, Jul. 2024.

[43] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[44] Y. Li et al., "MViTv2: Improved multiscale vision transformers for classification and detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4804–4814.

[45] L. Wang et al., "VideoMAE v2: Scaling video masked autoencoders with dual masking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14549–14560.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[47] C.-Y. Hsu and W. Li, "Explainable GeoAI: Can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection," *Int. J. Geographical Inf. Sci.*, vol. 37, no. 5, pp. 963–987, May 2023.
[48] L. Jimenez and D. Landgrebe, "High dimensional feature reduction via projection pursuit," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 2, Jun. 1994, pp. 1145–1147.
[49] M. C. Alonso, J. A. Malpica, and A. M. de Agirre, "Consequences of the Hughes phenomenon on some classification techniques," in *Proc. Annu. Conf.*, 2011, pp. 1–5.
[50] S. H. A. Moghaddam, M. Mokhtarzade, and B. A. Beirami, "A feature extraction method based on spectral segmentation and integration of hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 89, Jul. 2020, Art. no. 102097.
[51] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. ECCV Workshops*. Cham, Switzerland: Springer, 2016, pp. 443–450.
[52] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Proc. Domain adaptation Comput. Vis. Appl.*, 2017, pp. 153–171.
[53] H. Si, Y. Wan, M. Do, D. Vasisht, H. Zhao, and H. F. Hamann, "Towards scalable foundation model for multi-modal and hyperspectral geospatial data," 2025, *arXiv:2503.12843*.
[54] S. Conti, "Artificial intelligence for weather forecasting," *Nature Rev. Electr. Eng.*, vol. 1, no. 1, p. 8, Jan. 2024.
[55] S. Wang, W. Li, and C.-Y. Hsu, "STEPNet: A spatial and temporal encoding pipeline to handle temporal heterogeneity in climate modeling using AI: A use case of sea ice forecasting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 4921–4935, 2025.
[56] W. Li, "Artificial intelligence in Earth science: A GeoAI perspective," *J. Geophys. Res., Mach. Learn. Comput.*, vol. 2, no. 3, p. 2025, Sep. 2025.

Wenwen Li (Member, IEEE) received the Ph.D. degree in Earth system and geoinformation science from George Mason University, Fairfax, VA, USA, in 2019.

She is currently a Professor of geographic information science with the School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA, where she also directs the Spatial Analysis Research Center and the Cyberinfrastructure and Computational Intelligence Lab. Her research interests include cyberinfrastructure, geospatial big data, geospatial artificial intelligence (GeoAI), and their applications in data-intensive environmental and social sciences.

Chia-Yu Hsu received the master's degree in computer science from Arizona State University, Tempe, AZ, USA, in 2018.

He is currently a Research Professional at Arizona State University. His research interests focus on applying machine learning and artificial intelligence techniques to address geographical big data challenges. In recent years, his work has emphasized geospatial artificial intelligence (GeoAI), including developing foundation models for Earth observation, advancing Arctic science with AI, and enhancing explainability in deep learning for geospatial applications.

Sizhe Wang received the master's degree in geography from Arizona State University, Tempe, AZ, USA, in 2016, where he is currently pursuing the Ph.D. degree in computer science.

His research focuses on geospatial artificial intelligence (GeoAI), machine learning, and geospatial data analysis, with applications in environmental monitoring, terrain classification, and permafrost feature detection. He is also exploring the integration of knowledge graphs and spatial data fusion to enhance AI models in various geospatial contexts.