ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XII-4/W1-2026
21st International 3D GeoInfo Conference 2026, 28–29 September 2026, Sofia, Bulgaria

An Exploratory Study of Transformer-Based Generative AI Surrogate Modeling for 3D Wildfire Spread Simulation in Voxelized City

Haowen Xu[1], Sisi Zlatanova[1], Ruiyu Liang[2], Ismet Canbulat[2]

[1] GRID, School of Built Environment, UNSW Sydney, NSW 2052 Australia - (haowen.xu1, s.zlatanova)@unsw.edu.au
[2] School of Minerals and Energy Resources Engineering, UNSW Sydney, NSW 2052 Australia - (ruiyu.liang, i.canbulat)@unsw.edu.au

Keywords: Voxel Simulation, Surrogate Modeling, Voxels, Fire Spread Visualization, Generative AI

Abstract

Accurate and efficient simulation of wildfire spread in complex wildland–urban interface (WUI) environments remains a significant challenge due to the high computational cost of physics-based models and the limited fidelity of simplified approaches. This study presents an exploratory generative AI surrogate model for 3D wildfire simulation in voxelized urban environments. The proposed framework leverages a patch-based Transformer architecture with convolutional embedding and autoregressive modeling to learn spatiotemporal fire dynamics from data generated by a physics-based simulation model. Environmental and physical drivers, including fuel properties, wind conditions, and terrain characteristics, are encoded as aligned multi-channel voxel features to provide structured inputs for learning. The model predicts voxel-wise combustion states over time, enabling efficient temporal rollout of fire spread without explicitly simulating the full physical process. A pilot study using over 800 simulated scenarios demonstrates that the proposed approach achieves strong predictive performance, with voxel-wise accuracy ranging from 76% to 100%, while maintaining inference times on the order of 90–220 milliseconds per timestep. The results indicate that the model can effectively capture both spatial propagation and temporal evolution of wildfire dynamics, closely approximating the behavior of the underlying physics-based simulations. This work demonstrates the feasibility of using generative AI as an efficient surrogate for high-resolution 3D wildfire modeling, providing a foundation for scalable, real-time fire simulation and potential integration into urban-scale digital twin systems.

1. Introduction

Wildfire simulation plays a critical role in risk mitigation and emergency response, particularly in complex Urban–Wildland Interface (WUI) environments where fire interacts with dense infrastructure, heterogeneous fuels, and human activities (Ager et al., 2010; Radeloff et al., 2018). As climate change drives increasing wildfire frequency and intensity, impacts on human life, ecosystems, and urban systems have become more severe and widespread (Goss et al., 2020; Abatzoglou and Williams, 2016; Li et al., 2025). In this context, accurately simulating 3D fire spread and anticipating evolving risks is essential for supporting real-time decision-making, including evacuation planning, resource allocation, and infrastructure protection (Xu et al., 2026; Barton et al., 2024). However, wildfire dynamics remain highly complex and stochastic, governed by nonlinear interactions among atmospheric conditions, vegetation, built environments, topography, and human activities across multiple scales (Kondylatos et al., 2022; Xu et al., 2025). These challenges underscore that reliable and timely 3D wildfire simulation is both a scientific challenge and a critical operational necessity for protecting communities and infrastructure at scale (Thompson et al., 2013; Radeloff et al., 2018).

This work targets fast, high-resolution (sub-meter) 3D fire spread simulation at the urban-block scale (200 m–1 km) to support time-sensitive decision-making in complex urban–wildland interface environments. Existing wildfire models often face a trade-off between physical fidelity and computational efficiency: physics-based approaches are typically too computationally demanding for real-time or large-scale use, while simplified or empirical models may lack interpretability and transparency (Xu et al., 2025; Sullivan, 2009a). In addition, many methods rely on 2D representations or coarse spatial resolutions, limiting their ability to capture fine-grained interactions among fire, built structures, and heterogeneous fuels in urban settings (Xu et al., 2026; Baptiste Filippi et al., 2009). To address these limitations, this work explores a physically informed surrogate modeling approach for 3D wildfire simulation, enabling rapid inference while preserving key fire dynamics to support scenario exploration, risk assessment, and emergency response planning.

Recent advances in generative artificial intelligence (GenAI) offer a promising direction for addressing these limitations by enabling data-driven surrogate modeling of complex spatiotemporal processes (Willard et al., 2020). In particular, autoregressive models provide a flexible framework for learning the temporal evolution of wildfire dynamics, capturing dependencies across space and time through sequential prediction (Xu et al., 2025; Rasul et al., 2021). Unlike traditional physics-based simulators, such models can achieve significantly faster inference once trained, making them well-suited for real-time or interactive applications (Kochkov et al., 2021; Willard et al., 2020). At the same time, incorporating physics-informed training strategies,such as embedding physical constraints, conservation principles, or guidance from established fire models,allows these approaches to approximate underlying fire processes while maintaining physical plausibility and consistency (Raissi et al., 2019). This combination of generative AI modeling and physics-informed learning forms the foundation and plausibility for developing efficient, scalable, and interpretable surrogate models for 3D wildfire simulation.

Building upon this perspective, this paper presents an exploratory generative AI-based surrogate model for 3D wildfire simulation in voxelized environments. The proposed approach leverages autoregressive generative modeling to learn spatiotemporal fire dynamics from data generated by physics-based simula-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-317-2026

tions, while incorporating theory-guided constraints to preserve key physical behaviors across the voxel space. As a pilot study, this work focuses on assessing the feasibility of such a surrogate modeling paradigm, rather than replacing established and validated physical fire models. Particular emphasis is placed on accuracy and computational efficiency, with the goal of enabling rapid inference for large-scale and high-resolution scenarios. The detailed contributions of this work include: (1) the development of a generative AI surrogate model based on a 3D convolutional Transformer architecture for voxel-based wildfire simulation; (2) the incorporation of physically relevant variables and domain-informed design considerations to guide the learning of fire dynamics; and (3) an initial evaluation demonstrating the feasibility of the approach for efficient, scenario-driven 3D fire spread simulation and decision support.

2. Related Work

Existing wildfire simulation approaches are broadly categorized into physics-based and cellular automata (CA)-based models, both widely used to represent fire dynamics under varying conditions (Pastor et al., 2003; Sullivan, 2009a). Physics-based models, typically grounded in computational fluid dynamics (CFD), simulate combustion, heat transfer, and fire–atmosphere interactions through coupled conservation equations, enabling high-fidelity fire representation (McGrattan et al., 2013; Kirkpatrick and Kuo, 2024). However, they are computationally intensive and often limited to small spatial domains or short time simulation (Linn et al., 2002; Mell et al., 2007). In contrast, CA-based models approximate fire spread through rule-based or probabilistic transitions across grid cells, offering greater efficiency and scalability for large-scale simulations (Alexandridis et al., 2008; Sullivan, 2009b; Singh et al., 2024). Nevertheless, their reliance on simplified or empirically calibrated rules limits physical interpretability and generalizability (Sullivan, 2009b). As noted in prior studies, both paradigms are primarily applied to 2D simulation and exhibit inherent trade-offs between physical realism and computational scalability, posing challenges for large-scale, high-resolution 3D fire modeling (Xu et al., 2026).

Following advances in physics-informed surrogate modeling, data-driven approaches to wildfire simulation have gained increasing attention with the development of statistical learning, machine learning, and deep learning techniques (Wu et al., 2022; Jain et al., 2020). Early studies focused on correlations between fire occurrence and environmental variables, while more recent models—such as convolutional neural networks (CNNs) and ConvLSTM—demonstrate strong capability in learning 2D spatiotemporal fire dynamics from data (Marjani et al., 2023, 2024; Masrur et al., 2024). These approaches enable efficient 2D fire spread prediction and are well-suited for large-scale or near-real-time applications. For example, Sun et al. (2024) proposes a hybrid framework that combines CA, a semi-empirical fire spread model, and machine learning to jointly simulate fire spread and predict burned area, achieving improved accuracy over traditional methods.

More recent work has advanced wildfire simulation by integrating generative AI model with physical formulation to improve the extraction of fire spread rules (Xu et al., 2025). Cheng et al. (2025) presents a GAN-based surrogate modeling framework that leverages CFD-generated data to approximate fire dynamics, enabling fast and accurate prediction of temperature and smoke fields in 2D building floor plans, with substantial computational speedup over traditional physics-based simulations. Zhou et al. (2025) proposes a novel Transformer-enhanced cellular automata (Transformer-CA) model that learns 2D fire spread transition rules directly from historical data, enabling more accurate simulation of spatiotemporal wildfire dynamics. The results show that the approach outperforms traditional methods (e.g., LSSVM-CA) in capturing fire patterns and improving prediction accuracy and spatial consistency.

Despite their value, many existing AI-based surrogate models for wildfire simulation primarily rely on local spatiotemporal patterns and do not explicitly incorporate key driving factors such as fuel types, terrain slopes, and wind fields (Jain et al., 2020; Filippi et al., 2018). In addition, most current AI-based surrogate models for wildfire simulation are predominantly developed for 2D fire spread, with limited capability to represent detailed 3D fire dynamics at sub-meter (Xu et al., 2025, 2024). Consequently, modeling large-scale three-dimensional fire behavior with strong global coupling remains a significant challenge, particularly in complex WUI environments.

3. Generative AI Surrogate Modeling Framework

In the proposed generative AI surrogate framework, we first define the problem and design objectives, followed by the voxel-based feature representation of environmental variables. We then introduce a patch-based 3D Transformer to capture complex spatial dependencies and fuel–wind interactions, and finally describe the training and inference pipeline for efficient fire dynamics prediction.

3.1 Problem Formulation and Design Objectives

We build upon our previously developed physics-based 3D voxel wildfire simulation model for wildland–urban interface (WUI) environments, which leverages parallel computing to resolve detailed fire dynamics (Xu et al., 2025). While this model achieves strong physical fidelity, it exhibits limited scalability due to computational and memory constraints, requiring approximately 3 minutes of runtime on a standard laptop for a $300 \times 300 \times 150$ m urban space. These limitations become more pronounced as spatial extent and fuel and wind complexity increase.

To address these challenges, we reformulate the problem as learning a generative AI-based surrogate model that approximates spatiotemporal fire dynamics in voxel space. The proposed framework adopts an autoregressive formulation, in which the model predicts the next-step fire state conditioned on the current and recent system states, rather than explicitly simulating the full fire progression from ignition in a single run, as is common in many physics-based models. This autoregressive design significantly reduces computational overhead and memory requirements, enabling scalable, high-resolution, and near real-time wildfire prediction across large and heterogeneous environments.

Computational Efficiency and Scalability: The model is designed to significantly reduce computational cost compared to physics-based simulations, enabling scalable modeling over large spatial domains and high-resolution voxel grids.

Autoregressive Spatiotemporal Modeling: The framework adopts an autoregressive formulation to predict future fire

https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-317-2026 |

states conditioned on current and recent system states, avoiding full trajectory simulation from ignition and reducing memory overhead.

Representation of Heterogeneous Physical Drivers: The model supports multi-resolution voxel representations to encode heterogeneous environmental and physical variables, including fuel properties, wind fields, and terrain characteristics.

Modeling Long-range Spatial Dependencies: A Transformer-based architecture is employed to capture both local interactions and long-range spatial dependencies across complex fuel–wind configurations in 3D space.

Physical Consistency: The surrogate model is designed to approximate underlying fire dynamics with physically plausible behavior, despite not explicitly solving governing physical equations.

Near Real-time Inference: The framework enables rapid inference suitable for interactive simulation, decision support, and integration into digital twin environments.

Motivated by the identified design requirements and recent advances in spatiotemporal deep learning, we propose a patch-based 3D Transformer architecture with convolutional patch embedding for voxel-based wildfire prediction.

$$\mathbf{Y}_{t+1} = f_\theta\left(\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots, \mathbf{Y}_{t-k}, \mathbf{X}\right) \qquad (1)$$

The autoregressive formulation models the temporal evolution of fire dynamics by predicting the next-step combustion state as a function of prior states and environmental drivers (Equation 3.1). In this formulation, $\mathbf{Y}_t$ denotes the voxel-wise combustion state at time t, while $\mathbf{X}$ represents voxel-level environmental variables, including fuel characteristics and wind parameters. The model learns to predict future combustion states to simulate fire spread, with each voxel implicitly influenced by its spatial neighborhood through the Transformer architecture. Consequently, the proposed autoregressive model is conditioned not only on combustion state transitions but also on physically meaningful environmental and fuel-related factors that govern fire behavior. Rather than learning wildfire dynamics directly from real-world observations, however, the model learns a surrogate mapping from the outputs of a validated physics-based wildfire simulator. Its objective is therefore to efficiently approximate the numerical behavior of the underlying simulator and accelerate wildfire simulation, rather than to discover new fire physics.

3.2 Training Data Preparation

The training data are derived from simulations generated by the previously developed physics-based 3D voxel wildfire model (Xu et al., 2026). Specifically, the inputs consist of voxel-level fuel type, combination condition (e.g., burning age and heat transfer), and their associated characteristics (e.g., wind conditions). The outputs correspond to the temporal evolution of fire dynamics, represented by the combustion state of each voxel at successive time steps, along with the influence of neighboring voxel states. This paired input–output structure enables supervised learning of spatiotemporal fire propagation.

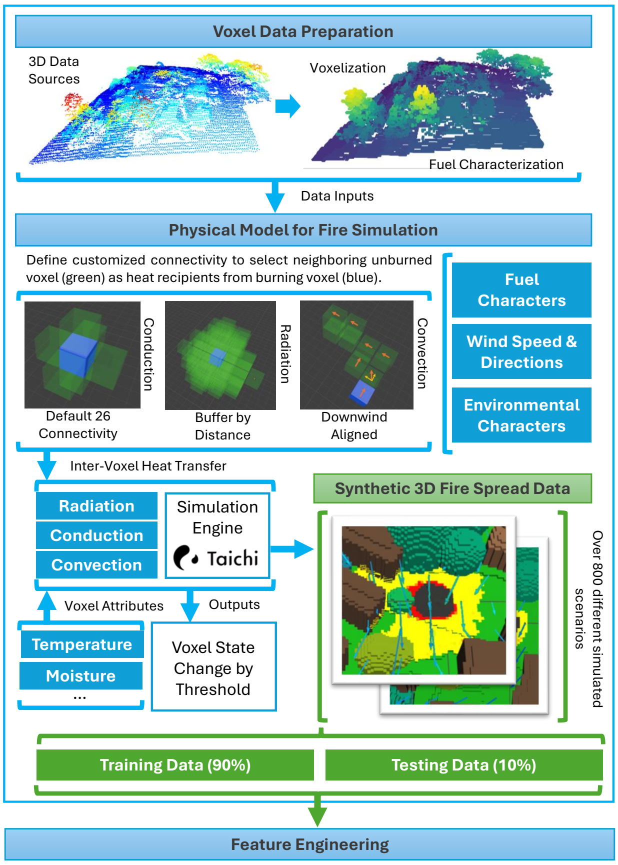

Figure 1. Training data generation pipeline. 3D data are voxelized and enriched with fuel attributes and wind direction. A physics-based model simulates fire spread via conduction, radiation, and convection, producing spatiotemporal combustion states used .

To ensure diversity and generalization, a large set of simulation scenarios is generated by varying key physical conditions. In particular, we consider 20 distinct ignition locations, each simulated under 8 wind directions and 5 wind speed levels, as well as a no-wind scenario. This results in a comprehensive set of spatiotemporal fire evolution patterns under heterogeneous environmental configurations, enabling the model to learn robust and transferable representations of wildfire dynamics across varying fuel and wind conditions.

3.3 Feature Engineering

Given the heterogeneity and high dimensionality of the raw simulation data, feature engineering, which is the process of transforming raw variables into structured numerical representations suitable for machine learning, is performed to convert these variables into a deep learning-compatible format. Environmental and physical attributes are encoded as aligned multichannel voxel features, where each voxel is associated with a D-dimensional feature vector and each channel corresponds to a distinct variable (e.g., environmental variables such as wind speed and direction, and physical variables such as fuel properties and terrain slope). These channels are spatially aligned across the voxel grid, forming a structured tensor representation that preserves key parameters of fire dynamics while enabling efficient processing by deep learning models. The overall data engineering process is depicted in Figure 2.

At a high level, the training data are organized as voxelized

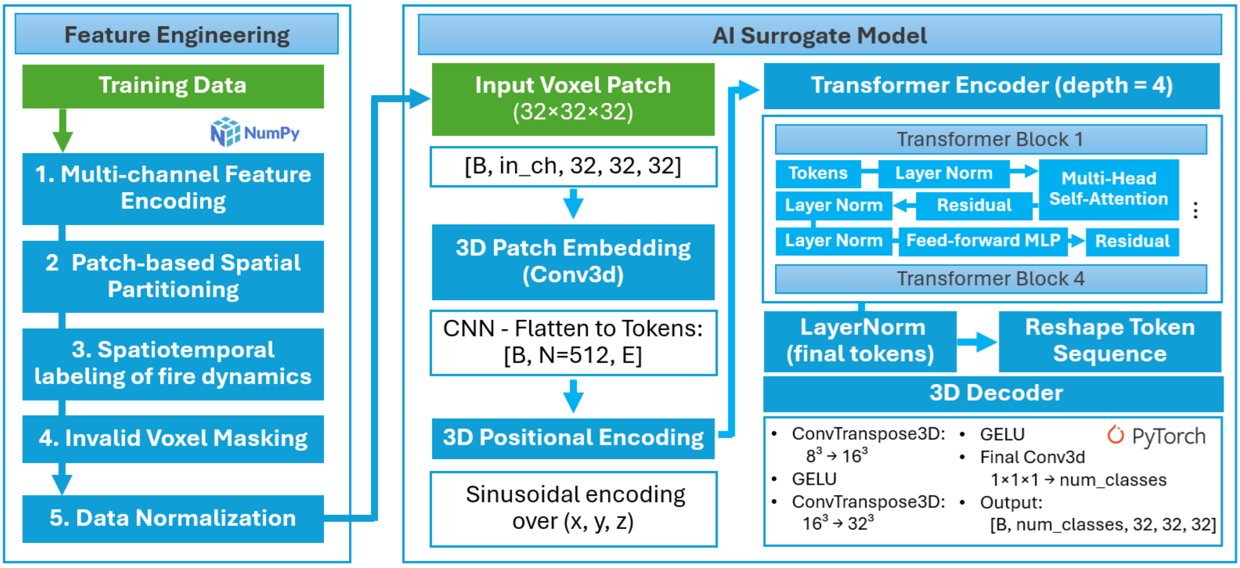

Figure 2. Proposed 3D patch-based Transformer surrogate model, integrating voxel feature engineering with encoder–decoder attention mechanisms for predicting spatiotemporal fire dynamics.

samples, each consisting of a multi-channel input tensor $\mathbf{X} \in \mathbb{R}^{D\times S\times S\times S}$, a voxel-wise label tensor $\mathbf{Y} \in \mathbb{Z}^{S\times S\times S}$, and a binary mask $\mathbf{M} \in \{0,1\}^{S\times S\times S}$ indicating valid voxels. Here, the multi-channel input tensor $\mathbf{X}$ is composed of D aligned feature channels, where each channel represents a distinct variable (e.g., environmental variables such as wind speed and direction, and physical variables such as fuel properties and terrain slope). The label tensor represents the spatiotemporal evolution of voxel combustion states. To support scalable learning, the voxel grid is partitioned into fixed-size 3D patches (e.g., 32^3), which serve as the fundamental training units. This patch-based representation enables efficient computation while maintaining localized spatial structure. During training, invalid or non-relevant voxels are excluded using the mask $\mathbf{M}$, improving robustness under sparse or incomplete observations. Additionally, feature values are normalized to ensure numerical stability and compatibility with the learning framework. The feature engineering procedures are summarized as follows:

1. **Multi-channel Feature Encoding:** Environmental variables (e.g., wind speed and direction, ambient temperature, and humidity) and physical variables are encoded as aligned voxel-wise feature channels to form a structured multi-dimensional input.

2. **Patch-based Spatial Partitioning:** The voxel grid is divided into fixed-size 3D patches to enable scalable training and efficient computation over large spatial domains.

3. **Spatiotemporal Labeling of Fire Dynamics:** The combustion states of each voxel are assigned as labels across time steps, capturing the temporal evolution of fire spread.

4. **Invalid Voxel Masking:** Binary masks are applied to exclude empty or non-relevant voxels, improving training stability and model robustness.

5. **Data Normalization:** Feature values are standardized to ensure numerical stability and compatibility with deep learning models.

Together, these procedures establish a unified patch-based voxels representation that captures complex interactions among fuel, wind, and terrain, providing a structured foundation for training AI surrogate models.

3.4 AI surrogate model

The overall architecture of the proposed AI surrogate model is illustrated in Figure 2. The model adopts a patch-based 3D Transformer design (Vaswani et al., 2017) to efficiently capture both short-range and long-range spatial dependencies in voxelized environments. By combining convolutional inductive bias with global self-attention, the architecture enables effective modeling of complex interactions among fuel, wind, and terrain in three-dimensional space.

3D Patch Embedding. An input voxel patch is produced via feature engineering, represented as $\mathbf{X} \in \mathbb{R}^{B\times C\times 32\times 32\times 32}$, in which C denotes the number of feature channels, we first apply a 3D convolutional patch embedding layer to extract localized spatial features. Specifically, a Conv3D operation with predefined kernel size and stride partitions the volumetric input into non-overlapping or partially overlapping 3D patches, while simultaneously projecting them into a latent embedding space. The resulting feature maps are flattened into a sequence of tokens $\mathbf{Z} \in \mathbb{R}^{B\times N\times E}$, where N is the number of patches and E is the embedding dimension. This step introduces locality and reduces the computational burden compared to operating on full-resolution voxel grids.

3D Positional Encoding. To retain spatial structure after tokenization, 3D positional encoding is incorporated into the token sequence. We adopt sinusoidal encoding over the (x, y, z) coordinates to inject explicit spatial information into each token. This allows the Transformer to distinguish spatial locations and reason over geometric relationships in the voxel grid, which is critical for modeling directional fire spread and anisotropic environmental effects.

Transformer Encoder. The encoded token sequence is processed by a stack of Transformer encoder blocks with depth L

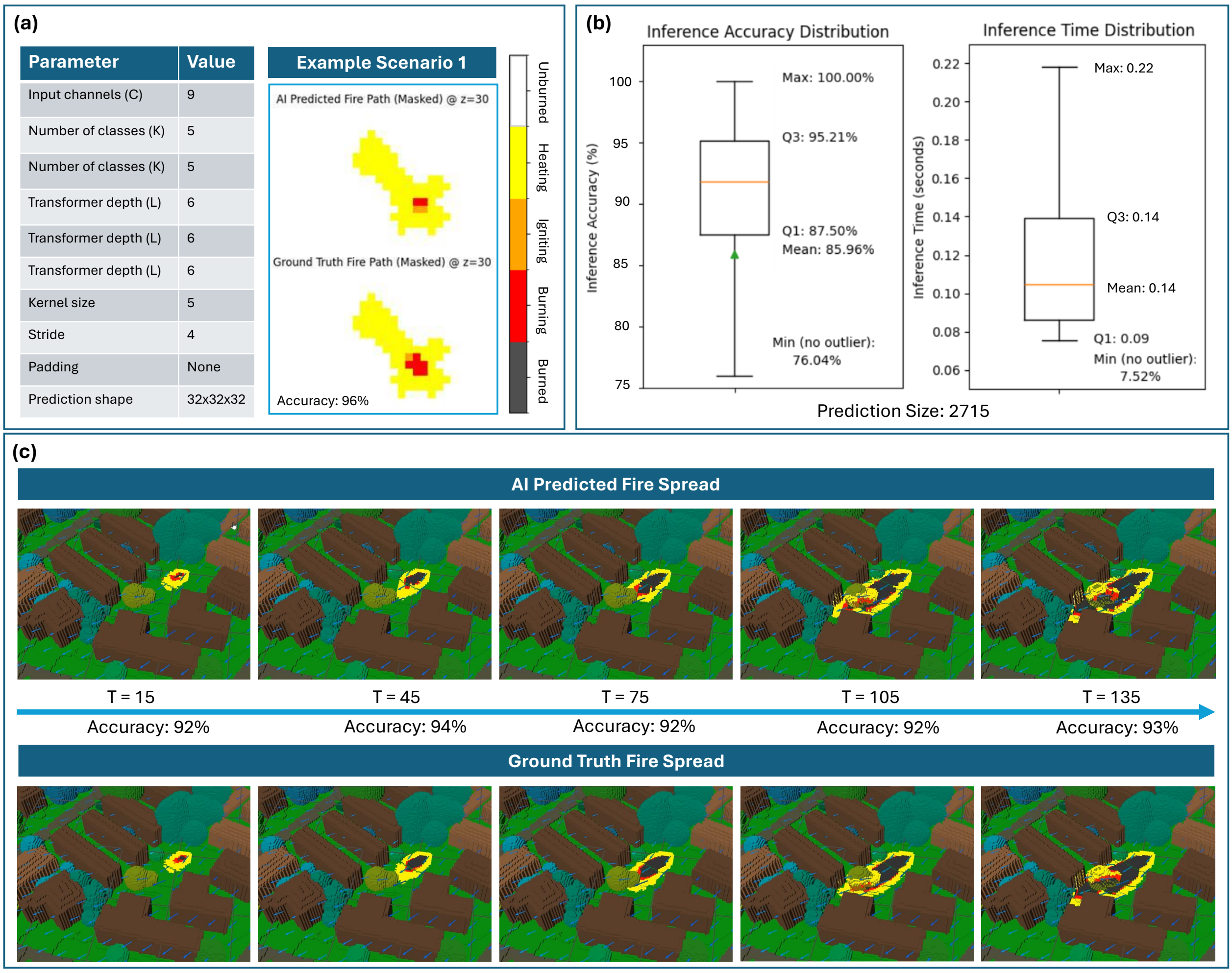

Figure 3. Overview and evaluation of the generative AI surrogate model for 3D wildfire simulation. (a) Model configuration and example prediction showing close agreement with ground truth. (b) Distributions of inference accuracy and inference time, indicating high accuracy and efficiency. (c) Temporal comparison of AI-predicted (top) and ground-truth (bottom) fire spread at selected timesteps (T=15, 45, 75, 105, 135), demonstrating consistent agreement in propagation patterns.

(e.g., $L = 4$), where L denotes the number of stacked Transformer blocks. Each block consists of two sub-layers: a multi-head self-attention (MHSA) module and a position-wise feed-forward multilayer perceptron (MLP), both equipped with residual connections and layer normalization.

The MHSA mechanism enables each token to attend to all other tokens, thereby capturing global interactions across distant voxel regions represented by different patches. This is particularly important for modeling long-range dependencies driven by wind fields and heterogeneous fuel distributions. The feed-forward MLP further refines token representations through non-linear transformations, enhancing feature expressiveness. Residual connections facilitate gradient flow and stabilize training, while layer normalization improves convergence and numerical stability.

Token Reshaping and 3D Decoder. After the Transformer encoder, the output token sequence is normalized and reshaped back into a structured 3D latent representation. This volumetric representation is then passed through a 3D decoder composed of transposed convolution layers (ConvTranspose3D) with non-linear activation functions (e.g., GELU). The decoder progressively upsamples the latent features from lower-resolution embeddings (e.g., $8^3 \rightarrow 16^3 \rightarrow 32^3$) to reconstruct the original spatial resolution.

For voxel-wise Prediction, a $1 \times 1 \times 1$ convolution layer is applied to map the decoded features to voxel-wise class logits, producing an output tensor $\mathbf{Y} \in \mathbb{R}^{B \times K \times 32 \times 32 \times 32}$, where K denotes the number of fire state classes. This enables dense prediction of combustion states across the entire voxel grid.

Overall, the proposed architecture leverages convolutional patch embedding for efficient local feature extraction and Transformer-based self-attention for global context modeling. This hybrid design significantly reduces computational complexity compared to full-resolution attention, while maintaining the ability to capture long-range spatial dependencies essential for accurate wildfire spread prediction.

3.5 Training and Inference Pipeline

The proposed feature engineering pipeline and Transformer-based AI surrogate model are deployed on the Katana high-performance computing (HPC) system. Training data are generated using the physics-based wildfire simulation model, and subsequently processed into voxelized patches stored as serialized tensor files. These samples are dynamically loaded during training, enabling scalable handling of large datasets.

The dataset is randomly partitioned into training and validation subsets, with a fixed validation fraction (e.g., 10%). Training is performed using mini-batch stochastic optimization with a batch size of 2, leveraging GPU acceleration. The model is implemented in PyTorch and trained using the AdamW optimizer with weight decay regularization and an initial learning rate of 3×10^{-4}. The objective function is defined as the voxel-wise cross-entropy loss, with invalid voxels excluded using a predefined ignore index.

To improve computational efficiency and numerical stability, mixed-precision training is employed using automatic mixed precision, with gradient scaling applied during backpropagation. The model parameters are updated iteratively across multiple epochs, while training progress is monitored using both loss and voxel-wise segmentation accuracy metrics.

A checkpointing mechanism is implemented to ensure training robustness and reproducibility. Model states, optimizer parameters, and training configurations are periodically saved, including epoch-level checkpoints, rolling checkpoints, and the best-performing model based on validation loss. Early stopping is optionally applied to prevent overfitting when validation performance no longer improves.

During inference, the trained model is applied in an autoregressive manner. Given the current voxel state and environmental inputs, the model predicts the next-step combustion state, which is iteratively fed back as input to generate subsequent time steps. This approach enables efficient temporal rollout of fire spread dynamics without simulating the full progression from ignition, significantly reducing computational cost compared to physics-based methods.

4. Pilot Study and Experiments

To evaluate the proposed AI surrogate model, a pilot study is conducted using simulation data generated from the physics-based wildfire model. A total of over 800 scenarios are used to train the model for 52 epochs on Katana HPC. A lightweight Python-based inference script, implemented in a Jupyter notebook environment, is used to perform step-by-step prediction and visualize the results.

The model configuration is depicted in Figure 3(a). The model takes 9 input feature channels and predicts 5 combustion state classes including unburned, heating, igniting, burning, and burned. In the visualization, the unburned state is not explicitly shown, as it corresponds to inactive background voxels. The Transformer backbone uses an embedding dimension of 256 with 6 stacked encoder layers and 4 attention heads, which follows a standard moderate-scale Vision Transformer configuration. This design provides sufficient capacity to capture complex spatial dependencies in 3D fire dynamics while maintaining computational efficiency for large-scale voxel inputs.

The feed-forward network adopts an MLP ratio of 4.0, a commonly used setting in Transformer architectures, which expands the hidden dimension to enhance non-linear feature representation without introducing excessive computational overhead. A 3D convolutional patch embedding is employed with a kernel size of 5 and stride of 4 to effectively aggregate local spatial context while reducing input resolution. This configuration enables overlapping receptive fields, improving the model's ability to capture fine-grained spatial interactions in fire spread, while also controlling memory and computational cost. The model produces voxel-wise predictions at a resolution of $32 \times 32 \times 32$.

The trained model demonstrates strong predictive performance in capturing voxel-level fire dynamics. To provide a comprehensive assessment of the surrogate model, performance was evaluated using voxel-level classification accuracy together with macro-averaged precision, recall, and F1-score, as well as the weighted F1-score. While accuracy reflects the overall proportion of correctly classified voxels, the macro-averaged metrics assign equal importance to each fire state and therefore provide a more balanced evaluation under class imbalance, whereas the weighted F1-score additionally accounts for the relative frequency of each class.

As shown in Figure 3(b), all metrics were computed only over fire-affected voxels included in the active evaluation mask, comprising the heating, igniting, burning, and burned states; unburned background voxels were excluded from the evaluation. The reported accuracy therefore measures combustion-state classification performance only among active voxels, rather than conventional voxel-wise accuracy computed over the complete voxel grid, which includes unburned voxels. Across the evaluated scenarios, the model achieves a median active-voxel classification accuracy of approximately 92%, with an interquartile range of 87.5% to 95.2% and a maximum of 100%. The minimum accuracy, excluding outliers, remains above 76.0%, indicating generally consistent predictive performance across the tested scenarios. When aggregated over all active voxels from the evaluation scenarios, the model achieves an overall active-voxel classification accuracy of 88.0%, together with a macro-averaged precision of 74.2%, recall of 63.8%, and F1-score of 63.4%, while the weighted F1-score reaches 87.3%. The lower macro-averaged metrics are primarily attributable to the transient igniting state, which represents a relatively small proportion of the dataset and is frequently confused with the adjacent heating and burning states because of its short temporal duration.

The surrogate model also maintains efficient inference, requiring an average of approximately 0.14 s to predict each future timestep. Since the surrogate operates autoregressively, each prediction is conditioned on the previously generated state, and the total inference time therefore scales with the simulation horizon. In contrast, the physics-based simulator performs a single integrated simulation of the complete fire evolution (typically 1000–2000 s of simulated time), requiring approximately 3–5 min on a standard laptop. The physics-based approach also requires substantial memory to initialize and maintain the voxel-based simulation domain and its associated physical state variables, which limits the feasible simulation duration and spatial scale on commodity hardware. Although the surrogate performs sequential inference, each timestep is generated through a single neural network forward pass rather than repeatedly solving the coupled physical processes governing fire spread. Consequently, the surrogate is well suited for repeated scenario generation, interactive planning, and uncertainty analysis, where a large number of wildfire simulations must be evaluated efficiently.

Qualitatively, Figure 3(c) shows that the predicted fire spread patterns closely match the ground-truth simulations across different timesteps (T=15 to 135). The model effectively captures both the spatial propagation and temporal evolution of fire, including the expansion and directional spread of the fire front. Minor discrepancies are observed in localized regions at later

https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-317-2026 | © Author(s) 2026. CC BY 4.0 License.

stages, but the overall fire perimeter and progression trends remain consistent. These results demonstrate that the proposed AI surrogate model can approximate complex wildfire dynamics with high fidelity while significantly improving computational efficiency compared to physics-based simulations.

5. Limitations and Future Work

This study represents an initial investigation of Transformer-based generative AI surrogate modeling for three-dimensional wildfire spread simulation rather than a comprehensive benchmark. The proposed surrogate was trained and evaluated using data generated from a single physics-based voxel fire simulator. Although the experiments considered different ignition locations and wind conditions, they were conducted within a single voxelized urban environment. Consequently, the generalizability of the proposed approach to different cities, urban morphologies, vegetation distributions, terrain characteristics, and unseen environmental conditions remains to be systematically investigated.

The surrogate operates autoregressively, with each prediction conditioned on the previously generated fire state. As with many autoregressive sequence models, prediction errors may accumulate over extended simulation horizons. Although stable fire propagation was observed throughout the evaluated prediction sequences without observable catastrophic error divergence, future work will investigate techniques such as scheduled sampling, physics-guided regularization, and hybrid correction strategies to further improve long-horizon robustness.

From a computational perspective, the current framework requires relatively time-consuming feature engineering and relies on fixed-size voxel patches (e.g., $32 \times 32 \times 32$), introducing additional preprocessing overhead and limiting flexibility for large or irregular domains. Future work will explore automated feature learning, adaptive voxel resolutions, hierarchical or multi-scale architectures, and patch-free formulations to improve scalability and efficiency.

Finally, the present work compares the proposed approach only with its underlying physics-based simulator. Future studies will conduct comprehensive benchmarking against alternative surrogate learning approaches, including convolutional neural networks (CNNs), ConvLSTM, U-Net, multilayer perceptrons, and other state-of-the-art architectures. Together with physics-guided learning and systematic ablation studies, these investigations will help quantify the benefits of the proposed Transformer architecture, improve model interpretability and robustness, and advance the development of efficient, reliable, and generalizable AI surrogate models for large-scale 3D wildfire simulation.

6. Conclusion

This study presents a generative AI surrogate model for three-dimensional voxel-based wildfire simulation, evaluated through a pilot study using more than 800 physics-based simulation scenarios. The proposed patch-based Transformer model demonstrates strong predictive performance, achieving a median active-voxel classification accuracy of approximately 92% across the evaluated scenarios, while maintaining an average inference time of approximately 0.14 s per predicted timestep. The results show that the model effectively captures both the spatial propagation and temporal evolution of wildfire dynamics, closely approximating the behavior of the underlying physics-based simulator. By adopting an autoregressive formulation, the proposed approach generates future fire states through efficient neural network inference rather than repeatedly solving the underlying physical processes, substantially reducing computational cost while preserving the key spatiotemporal characteristics of fire spread.

This work should be interpreted as a proof-of-concept demonstrating that Transformer-based generative AI surrogate models can efficiently approximate high-resolution three-dimensional wildfire simulations generated by physics-based models, rather than replacing validated physics-based fire simulation models. The proposed framework provides a promising foundation for scalable, data-driven wildfire modeling and rapid scenario exploration. With further validation across diverse urban environments, improved long-horizon robustness, and comprehensive benchmarking against alternative surrogate architectures, the approach has the potential to support large-scale urban fire risk assessment, emergency planning, and digital twin decision-support applications.

Acknowledgments This work was supported by the Australian Research Council (ARC) through the Resilient and Intelligent Infrastructure Systems (RIIS) program and was conducted in collaboration with Asia Air Survey Co., Ltd. (AAS). The authors also gratefully acknowledge the support of Fire and Rescue NSW and the Australasian Fire and Emergency Service Authorities Council (AFAC), whose collaboration has contributed to the development and practical relevance of this research.

References

Abatzoglou, J. T., Williams, A. P., 2016. Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of sciences*, 113(42), 11770–11775.

Ager, A. A., Vaillant, N. M., Finney, M. A., 2010. A comparison of landscape fuel treatment strategies to mitigate wildland fire risk in the urban interface and preserve old forest structure. *Forest Ecology and Management*, 259(8), 1556–1570.

Alexandridis, A., Vakalis, D., Siettos, C. I., Bafas, G. V., 2008. A cellular automata model for forest fire spread prediction: The case of the wildfire that swept through Spetses Island in 1990. *Applied Mathematics and Computation*, 204(1), 191–201.

Baptiste Filippi, J., Bosseur, F., Mari, C., Lac, C., Le Moigne, P., Cuenot, B., Veynante, D., Cariolle, D., Balbi, J.-H., 2009. Coupled atmosphere-wildland fire modelling. *Journal of Advances in Modeling Earth Systems*, 1(4).

Barton, J., Nishino, A., Kohtake, N., Westrin, H., Shimada, M., Guo, R., Ichikawa, R., Li, J., Canbulat, I., Zlatanova, S., 2024. Near-realtime Location Specific Messaging During Extreme Bushfire Events. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 9–14.

Cheng, Q., Wang, X., Zhuang, J., Liao, W., Xie, L., 2025. Fire scenario simulation method for residential buildings based on generative adversarial network. *Developments in the Built Environment*, 100720.

Filippi, J.-B., Bosseur, F., Mari, C., Lac, C., 2018. Simulation of a large wildfire in a coupled fire-atmosphere model. *Atmosphere*, 9(6), 218.

Goss, M., Swain, D. L., Abatzoglou, J. T., Sarhadi, A., Kolden, C. A., Williams, A. P., Diffenbaugh, N. S., 2020. Climate change is increasing the likelihood of extreme autumn wildfire conditions across California. *Environmental Research Letters*, 15(9), 094016.

Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., Flannigan, M. D., 2020. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4), 478–505.

Kirkpatrick, A. T., Kuo, K. K., 2024. *Principles of combustion*. John Wiley & Sons.

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., Hoyer, S., 2021. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21), e2101784118.

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-Torres, M.-Á., Carvalhais, N., 2022. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters*, 49(17), e2022GL099368.

Li, J., Zlatanova, S., Stouffs, R., 2025. Cooling effects of urban green spaces: A systematic review of methods applied in the past two decades. *Sustainable Cities and Society*, 106833.

Linn, R., Reisner, J., Colman, J. J., Winterkamp, J., 2002. Studying wildfire behavior using FIRETEC. *International journal of wildland fire*, 11(4), 233–246.

Marjani, M., Ahmadi, S. A., Mahdianpari, M., 2023. FirePred: A hybrid multi-temporal convolutional neural network model for wildfire spread prediction. *Ecological Informatics*, 78, 102282.

Marjani, M., Mahdianpari, M., Mohammadimanesh, F., 2024. CNN-BiLSTM: A novel deep learning model for near-real-time daily wildfire spread prediction. *Remote Sensing*, 16(8), 1467.

Masrur, A., Yu, M., Taylor, A., 2024. Capturing and interpreting wildfire spread dynamics: attention-based spatiotemporal models using ConvLSTM networks. *Ecological Informatics*, 82, 102760.

McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Weinschenk, C., Overholt, K., 2013. Fire dynamics simulator technical reference guide volume 1: mathematical model. *NIST special publication*, 1018(1), 175.

Mell, W., Jenkins, M. A., Gould, J., Cheney, P., 2007. A physics-based approach to modelling grassland fires. *International Journal of Wildland Fire*, 16(1), 1–22.

Pastor, E., Zárate, L., Planas, E., Arnaldos, J., 2003. Mathematical models and calculation systems for the study of wildland fire behaviour. *Progress in Energy and Combustion Science*, 29(2), 139–153.

Radeloff, V. C., Helmers, D. P., Kramer, H. A., Mockrin, M. H., Alexandre, P. M., Bar-Massada, A., Butsic, V., Hawbaker, T. J., Martinuzzi, S., Syphard, A. D. et al., 2018. Rapid growth of the US wildland-urban interface raises wildfire risk. *Proceedings of the national academy of sciences*, 115(13), 3314–3319.

Raissi, M., Perdikaris, P., Karniadakis, G. E., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378, 686–707.

Rasul, K., Seward, C., Schuster, I., Vollgraf, R., 2021. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. *International conference on machine learning*, PMLR, 8857–8868.

Singh, H., Ang, L.-M., Lewis, T., Paudyal, D., Acuna, M., Srivastava, P. K., Srivastava, S. K., 2024. Trending and emerging prospects of physics-based and ML-based wildfire spread models: A comprehensive review. *Journal of Forestry Research*, 35(1), 135.

Sullivan, A. L., 2009a. Wildland surface fire spread modelling, 1990–2007. 1: Physical and quasi-physical models. *International Journal of Wildland Fire*, 18(4), 349–368.

Sullivan, A. L., 2009b. Wildland surface fire spread modelling, 1990–2007. 3: Simulation and mathematical analogue models. *International Journal of Wildland Fire*, 18(4), 387–403.

Sun, X., Li, N., Chen, D., Chen, G., Sun, C., Shi, M., Gao, X., Wang, K., Hezam, I. M., 2024. A forest fire prediction model based on cellular automata and machine learning. *IEEe Access*, 12, 55389–55403.

Thompson, M. P., Calkin, D. E., Finney, M. A., Gebert, K. M., Hand, M. S., 2013. A risk-based approach to wildland fire budgetary planning. *Forest Science*, 59(1), 63–77.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2020. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 1(1), 1–34.

Wu, Z., Wang, B., Li, M., Tian, Y., Quan, Y., Liu, J., 2022. Simulation of forest fire spread based on artificial intelligence. *Ecological Indicators*, 136, 108653.

Xu, H., Omitaomu, F., Sabri, S., Zlatanova, S., Li, X., Song, Y., 2024. Leveraging generative AI for urban digital twins: a scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Informatics*, 3(1), 29.

Xu, H., Zlatanova, S., Liang, R., Canbulat, I., 2025. Generative AI as a Pillar for Predicting 2D and 3D Wildfire Spread: Beyond Physics-Based Models and Traditional Deep Learning. *Fire*, 8(8), 293.

Xu, H., Zlatanova, S., Liang, R., Canbulat, I., 2026. A Modular Light-weight Voxel-Based 3D Wildfire Propagation Simulator in Python Using LiDAR Data, High-Performance Computing (HPC), and Immersive Scientific Visualization. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 127–134.

Zhou, Z., Xu, Q., Yi, J., Li, Y., Zhang, S., Li, W., 2025. Modeling the spread of forest fires through cellular automata by leveraging deep learning to derive transition rules. *Ecological Informatics*, 88, 103150.