Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Environmental semantic clustering-guided multimodal fusion for enhanced interpretability in methane concentration prediction

Yang Xu[a], Hao Wang[b], Jude D. Kong[a,c,d,e,f,*]

[a] Artificial Intelligence and Mathematical Modelling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Office 662, Toronto, ON, M5T 3M7, Canada
[b] The Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2J5, Canada
[c] Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Canada
[d] Department of Mathematics, University of Toronto, Bahen Centre for Information Technology, Room 6291, 40 St. George Street, Toronto, Ontario, M5S 2E4, Canada
[e] Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Canada
[f] Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP), Canada

ARTICLE INFO

Keywords:
Methane monitoring
Multimodal fusion
Wavelet decomposition
Bidirectional cross-attention networks
Semantic cluster
Remote sensing
Dynamic fusion gate

ABSTRACT

Methane is a potent greenhouse gas with significant climate implications, being approximately 84 times more impactful than CO_2 over a 20-year timeframe. Accurately predicting the spatiotemporal distribution of methane concentrations, particularly near industrial sources, is essential for effective environmental monitoring and provides a critical foundation for subsequent emission source identification. This study introduces a novel Spatial-Temporal Cross-Attention Network (ST-CAN) to address the challenge of fusing sparse, high-frequency ground-based observations with spatially extensive but temporally infrequent satellite imagery. Using the Athabasca oil sands region as a case study, ST-CAN incorporates three synergistic innovations that work together to address data fusion challenges: (1) wavelet decomposition transforming high-frequency ground measurements into multi-scale temporal features capturing both long-term trends and short-term emission events; (2) environmental semantic clustering that identifies distinct atmospheric patterns from these wavelet features, providing interpretable contextual labels; and (3) a bidirectional cross-attention mechanism where these semantic cluster labels dynamically guide how ground temporal features query and fuse with satellite spatial information, adaptively prioritizing relevant features based on identified environmental states. The model is designed to leverage time-dense ground data to enhance the temporal resolution of weekly satellite-derived concentration maps, generating high-fidelity spatially representative methane concentration predictions by integrating information from four spatially distributed monitoring stations and satellite imagery, capturing regional-scale atmospheric dynamics. Extensive evaluations demonstrate ST-CAN significantly outperforms all the baseline models in predictive accuracy and robustness. The bidirectional mechanism notably improves interpolation during satellite data gaps, mitigating cloud cover and data sparsity challenges. By combining interpretability with advanced AI techniques, ST-CAN provides a transparent and scalable framework for high-resolution methane concentration modelling, advancing environmental monitoring capabilities and supporting targeted climate mitigation efforts.

1. Introduction

The increasing climate crisis requires immediate efforts on the mitigation of potent but short-lived greenhouse gas (GHG) emissions, of which methane is the most critical due to its short lifetime and substantial influence on near-term global warming. The efficacy of the methane molecular in trapping heat is much greater than CO_2, which is reported to be more than 80 times that of CO_2 over an important 20-year timeframe (Etminan et al., 2016). This characteristic makes methane mitigation not only beneficial, but essential for any feasible strategy aiming to rapidly slow down the planetary warming. The oil and gas industry is a primary contributor, tackling methane leaks all along its

* Corresponding author. Artificial Intelligence and Mathematic Modelling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Office 662, Toronto, ON, M5T 3M7, Canada.
E-mail address: jude.kong@utoronto.ca (J.D. Kong).

https://doi.org/10.1016/j.jenvman.2026.128845
Received 8 August 2025; Received in revised form 7 December 2025; Accepted 31 January 2026
Available online 6 February 2026
0301-4797/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

supply chain from wells to pipelines (Kong et al., 2019) to local delivery, which remains a difficult and ongoing task. Therefore, it is essential to develop accurate modelling approaches for effective environmental management and targeted climate action.

Traditional methane detection methods face significant limitations. Field measurements provide high accuracy with clear temporal patterns but suffer from limited spatial coverage, requiring interpolation for regional applications (Jacob et al., 2016). Satellite imagery offers broader spatial coverage through platforms like MethaneSAT (MethaneSAT, 2024), TROPOMI (Veefkind et al., 2012), and GHGSat (GHGSat, 2024a,b), but faces constraints from cloud interference, restricted data access, and infrequent revisit cycles. Multi-source data integration has emerged as necessary, yet cross-platform validation errors of 30–50% highlight fusion challenges (Fan et al., 2024).

However, recent advances in multimodal spatiotemporal fusion have demonstrated significant promise across various environmental monitoring applications (Li et al., 2022). provided a comprehensive review of deep learning approaches for multimodal remote sensing data fusion, highlighting the evolution from pixel-level to decision-level fusion strategies. For atmospheric and environmental applications specifically, multimodal frameworks integrating ground sensors with satellite imagery have shown remarkable improvements in prediction accuracy (Hameed et al., 2023). demonstrated an 18–20% enhancement in air quality forecasting by fusing ground sensor data with CCTV imagery through deep learning architectures, while (Lilhore et al., 2025) achieved superior air quality predictions using hybrid models combining CNNs, Bidirectional Long Short-Term Memory (Bi-LSTM), and attention mechanisms with multimodal data sources.

In recent years, the maturation of AI and Machine Learning (ML) has led to widespread consensus regarding their rapid emergence across multiple engineering fields. AI and ML have become transformative tools for methane emission modelling. For example, Xu conducted a systematic review of 110 papers showing that AI models have evolved from single-modal to multi-modal approaches: deep learning achieves over 98% accuracy in leak detection using hyperspectral imagery, random forest provides interpretable models for agricultural emissions, while physics-informed models enhance dynamic prediction capabilities by incorporating fluid dynamics equations (Xu et al., 2025). Notably, deep learning architectures, particularly those adapted from the field of computer vision, are demonstrating remarkable success in processing satellite imagery for methane plume identification. Vision Transformers (ViTs), for example, represent a significant breakthrough in detecting methane plumes as small as 0.01 km^2 from Sentinel-2 data (Rouet-Leduc and Hulbert, 2024). LSTM provides early warnings up to two weeks in advance using Sentinel-5P data (Chen et al., 2023). In addition, cross-attention mechanisms have emerged as particularly effective for multimodal fusion tasks (Wen et al., 2024). developed a Cross-Attention Spatio-Temporal Spectral Fusion (CASTSF) model for integrating Sentinel-2 and Sentinel-3 data, demonstrating superior performance over traditional methods like FSDAF and STARFM in capturing both spatial texture and spectral information. Similarly (Wang et al., 2024), proposed a spatial-temporal cross-attention fusion module for multimodal trajectory prediction, achieving significant improvements in capturing spatiotemporal interactions. The attention mechanism's ability to dynamically weight feature importance has proven essential for handling heterogeneous data sources with varying spatial and temporal resolutions (Nagrani et al., 2021). In terms of multi-source collaboration, a federated learning framework enables distributed model training, protecting the privacy of oilfield data while improving cross-regional detection accuracy by 25% (Quamar et al., 2023). However, the transition to operational systems faces critical hurdles.

Despite these advances, several architectural challenges remain inadequately addressed. Multimodal fusion framework must effectively handle the inherent trade-offs between spatial and temporal resolutions inherent in multi-sensor systems. Recent work in remote sensing has explored hierarchical fusion strategies (Lian et al., 2025). reviewed attention-based CNN methods that retain fine details from high-resolution images while capturing large-scale patterns from low-resolution data, while (Zhou et al., 2025) proposed bidirectional cross-fusion modules to model temporal changes and spatial details simultaneously. Current AI approaches employ simplistic fusion strategies with early concatenation or intermediate fusion layers that fail to capture complex interactions between temporally dense ground measurements and spatially extensive satellite imagery. These methods render models overly sensitive to noise or fail to leverage complementary strengths, leading to unsatisfactory performance. Second, deep learning models lack interpretability, creating “black box" problems that hinder scientific validation and operational trust (Li et al., 2020). Finally, satellite monitoring faces temporal gaps from cloud cover while ground networks provide sparse spatial coverage, with standard interpolation techniques often smoothing over critical “super-emitter" events (IEA, 2024).

To address the critical challenges of sophisticated data fusion, model interpretability, and robustness against spatio-temporal data sparsity and dynamics, this paper introduces the ST-CAN. ST-CAN is proposed as an innovative framework specifically designed for high-accuracy methane concentration prediction by synergistically integrating multi-source data streams of ground-based stations and satellite imagery data. Utilizing the Athabasca Valley oil sands region in Alberta, Canada, as a complex and relevant case study, where previous work has demonstrated the value of machine learning approaches for methane source detection and concentration prediction using ground-based monitoring data (Sysoeva et al., 2025), ST-CAN aims to advance methane monitoring through a novel architecture. While prior studies have successfully applied Random Forest models to identify methane sources and predict concentrations from weather station data, our work extends this foundation by incorporating satellite-based spatial information and developing physics-guided attention mechanisms for multi-source data fusion. Its core design focuses on employing physics-guided attention mechanisms to effectively fuse ground-based temporal patterns with satellite spatial information, enhancing interpretability through integrated pattern analysis, and incorporating strategies for robust handling of temporal misalignments and data gaps inherent in real-world monitoring scenarios. This approach seeks to provide a more reliable, transparent, and dynamically adaptive solution compared to existing methods, providing a transparent and reproducible framework for modelling the regionally representative methane concentration in complex industrial landscapes. This work addresses a critical data gap in the methane monitoring pipeline, paving the way for more reliable subsequent emission estimation.

2. Experimental and methodology

2.1. Case study

The Athabasca Valley oil sands (AVOS) in northeastern Alberta represent one of North America's most contentious energy frontiers. Those vast deposits have become both an economic lifeline and an environmental flashpoint for Canada's resource economy. AVOS extends beyond just the Athabasca formation, incorporating the less publicized Cold Lake and Peace River deposits. Together, this triumvirate commands a staggering 95% of Canada's proven oil reserves. Our research focuses specifically on the Athabasca region, which is by far the most extensively developed and economically significant of the three. Spanning roughly 142,200 km^2, this massive geological formation dwarfs many European countries in size (BOE Report, 2025).

Despite their importance to national energy security, the extraction process raises serious ecological concerns. The oil sands are a significant source of greenhouse gas emissions, including methane. Methane is released during various stages of oil extraction and processing, particularly from equipment leaks, process upsets, and tailings ponds (Government of Alberta, 2025).

2.2. Data Description

This study integrates ground-based air quality and meteorological measurements with satellite-derived atmospheric methane concentration data to develop and validate the proposed ST-CAN. The datasets were selected to provide comprehensive spatio-temporal coverage of the AOSR study area.

2.2.1. Ground monitoring data

This study utilizes data from the Wood Buffalo Environmental Association (WBEA) to capture ground-level methane concentrations within the AVOS. WBEA operates an extensive network of continuous ambient air quality monitoring stations throughout the region. Fig. 1 presents the spatial distribution of the four selected WBEA stations within the AVOS study area. The stations are strategically positioned to capture diverse emission sources and atmospheric conditions: AMS09 (Barge Landing) in the core industrial zone, AMS02 (Mildred Lake) at the industrial fenceline, AMS04 (Buffalo Viewpoint) near tailings ponds, and AMS07 (Athabasca Valley) in the Fort McMurray urban center. This spatial configuration enables comprehensive monitoring across approximately 220 km north-south extent, representing distinct exposure environments from near-source industrial emissions to community-level impacts.

AMS 09 (Barge Landing) represents the core industrial zone monitoring aspect in WBEA. The Barge Landing station is situated within the core industrial area, approximately 10 km SW of the CNRL Albian Sands plant and 20 km SE of the CNRL Horizon plant. It is near Barge Landing Road and Highway 63. AMS 09 is strategically selected to capture emissions signals close to major oil sands processing facilities. Its location within the densest industrial activity provides data minimally affected by atmospheric dilution or transport, serving as a crucial reference point for understanding near-source concentrations and validating emissions originating directly from the primary industrial complexes. It monitors key pollutants, including Total Hydrocarbon (THC), CH_4, Non-Methane Hydrocarbons (NMHC), SO_2, NO_x, $PM_{2.5}$, and meteorological parameters.

AMS 02 (Mildred Lake) represents the industrial fenceline and transport influence areas. It is located near the fenceline of the Syncrude Mildred Lake oil sands facility. It is also proximate to an airport runway (~20m) and Highway 63 (~300m). AMS 02 monitors air quality at the edge of a major industrial facility, capturing pollutants potentially dispersing from the Syncrude operations. It is close to significant transportation infrastructure allows for the assessment of combined industrial and traffic-related emissions influences. Data from this site helps characterize the transition from near-source to ambient conditions and understand the interactions between different emission types at the industrial edge zone.

AMS 04 (Buffalo Viewpoint) represents tailings pond influence. It is positioned to specifically monitor emissions from tailings ponds, which are known significant sources of methane in the AVOS. Data from this station is vital for understanding the magnitude and variability of emissions from these large area sources, which differ significantly from fugitive emissions from processing plants. Monitoring at this location helps isolate the contribution of tailings ponds to regional CH_4 levels and assess the effectiveness of mitigation efforts targeting these specific sources.

The last selected station is AMS 07 (Athabasca Valley), which represents community and urban influence. It is situated in the downtown area of Fort McMurray within the Athabasca River valley and provides insights into the air quality experienced by the urban population of Fort McMurray. This site is crucial for assessing the potential impact of regional industrial emissions on community health and understanding pollutant transport dynamics from the industrial core towards populated areas.

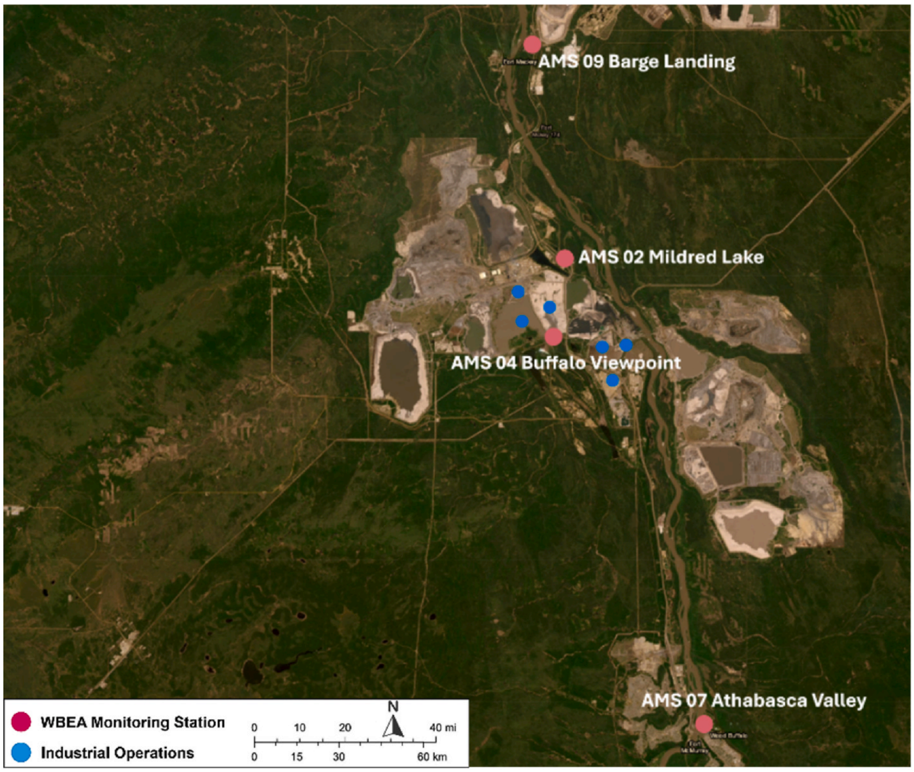

Fig. 1. Spatial distribution of WBEA air monitoring stations in the AVOS. AMS09 (Barge Landing, 57.198°N, 111.600°W) monitors the core industrial zone near CNRL facilities; AMS02 (Mildred Lake, 57.050°N, 111.564°W) captures fenceline emissions from Syncrude operations; AMS04 (Buffalo Viewpoint, 56.996°N, 111.594°W) assesses tailings pond influence at the South Mine; and AMS07 (Athabasca Valley, 56.733°N, 111.390°W) evaluates community exposure in downtown Fort McMurray.

Together, these four stations provide spatially distributed, continuous ground-based measurements representing distinct source influences within the AVOS. This multi-perspective dataset forms the foundation for training and validating the spatio-temporal models developed in this study, particularly when fused with satellite observations. The WBEA data portal can be found at https://wbea.org/data/continuous-monitoring-data/.

2.2.2. Satellite imagery data from GHGSAT.SPECTRA

To complement the point-source ground measurements with broader spatial context, this study incorporated atmospheric methane data from GHGSat's SPECTRA platform. Specifically, we utilized the 'Spectra Basic' tier, which provides weekly heatmaps of area-averaged, cloud-free atmospheric methane concentrations. This dataset is publicly available and free of charge. (GHGSat, 2024a,b). Researchers can access the data by visiting the product page at https://www.ghgsat.com/en/products-services/spectra/. From this page, users can navigate to the registration portal and create a free Spectra Basic account by providing a valid email address. Upon registration, users are granted immediate access to the global methane map interface, where the specific weekly datasets for the Athabasca region can be visualized and extracted. This open-access policy ensures that our methodology can be reproduced by other researchers without concerns over restrictive data access policies or costs.

Specifically, the data represent the column-averaged dry-air mole fraction of methane (X_{CH4}). This metric quantifies the average concentration of methane in a vertical column of air extending from the ground to the top of the atmosphere, effectively representing the total atmospheric methane burden over a given area, rather than a surface-level measurement. These data are provided at a spatial resolution of approximately 10 km and a weekly temporal resolution. For this study, satellite data were acquired for the same one-year interval as the ground observations (early January 2024 to early January 2025). The weekly X_{CH4} heatmaps were then manually cropped to a geographic extent that encompasses the four WBEA ground monitoring stations, ensuring direct relevance to the regions under investigation.

It is critical to explicitly define the prediction target given this heterogeneity of our input data. The WBEA ground station data, which provides in-situ, ground-level CH_4 concentration in ppm, serves as the ground truth and the sole prediction target for our model. The satellite-derived X_{CH4} data, representing the column-averaged dry-air mole fraction, are used exclusively as an input feature to provide spatial context. Therefore, the model's task is to learn the complex, non-linear relationship between the column-averaged atmospheric state (from satellite) and the specific, in-situ concentration at the surface (from ground stations). All performance metrics, Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are calculated by comparing the model's output directly against these ground-level station measurements.

2.3. Data pre-processing

Before model development, the collected ground station and satellite datasets must undergo several pre-processing steps to ensure data quality, consistency, and suitability for input into the spatio-temporal fusion framework. These steps included treatment of meteorological variables, imputation of missing values, feature engineering through wavelet decomposition, and temporal pattern identification via cluster analysis.

2.3.1. Transformation of wind direction feature

Wind direction, inherently a periodic variable (0–360°), presents challenges for direct input into the machine learning models, as the numerical proximity for examples 359° and 1° does not reflect their actual physical proximity. To address this, the raw wind direction data were decomposed into their zonal components, which are vector U representing the east-west component and vector V representing the north-south component. This transformation converts the circular wind direction into two continuous linear variables, which are more amenable to standard modelling techniques and accurately represent the directional flow of air masses.

2.3.2. Imputation of missing data

The continuous hourly monitoring data from the WBEA stations contained instances of missing values. For isolated or short sequences of missing data points across most variables and stations, linear interpolation was employed as a primary imputation strategy. This method estimates missing values based on a linear trend between the nearest available data points before and after the gap. However, a significant period of continuous missing data was identified for the wind speed and wind direction features at the AMS04 station. Direct application of linear interpolation over such an extended gap would likely introduce substantial discrepancy. Therefore, an alternative substitution-based approach was adopted. Based on a correlation analysis of wind patterns and the relative geographical positioning of the stations, wind data from AMS02 corresponding to the missing time intervals at AMS04 were used as a substitute. To rigorously validate this substitution strategy, a comparative analysis of wind characteristics between AMS02 and AMS04 was conducted. This analysis utilized concurrent valid hourly observations from July and August 2024, encompassing the periods immediately adjacent to the specific data gap (July 26, 19:00 to July 31, 10:00). This temporal selection ensures that the validation captures the specific seasonal wind dynamics relevant to the imputed period. This involved calculating the correlation between the respective wind speed and the derived U/V vector components, alongside a visual comparison of their wind patterns using wind rose diagrams. The results of this analysis are presented in Figs. 2 and 3.

2.3.3. Wavelet decomposition for feature extraction

We have applied wavelet decomposition to capture both long-term trends and short-term momentary dynamics within the ground station time series data. This technique is well-suited for analyzing non-stationary environmental time series, as it can effectively separate signals into different frequency components at various time scales. For this study, each original time series feature, including pollutant concentrations and meteorological variables from all four ground stations, was decomposed into multiple wavelet coefficients. We employed the Daubechies 4 (db4) wavelet as the basis function based on several reasons. First, we have collected the quantitative evidence from multi-feature sensitivity analysis. The results could be found in the subsequent section 3.1.1. In addition, from the literature precedent, db4 is standard for environmental time series analysis (Percival and Walden, 2000). Regarding the decomposition level selection, we applied five-level wavelet decomposition (A5, D1-D5). This selection is primarily based on energy distribution validation results, which are also demonstrated in Section 3.1.1. To optimize the feature set for model training by reducing redundancy and potential noise often present in higher-order detail coefficients (Feng et al., 2019), only the A5, D1, and D2 components were used to replace each original feature.

This selection was not arbitrary but was based on a rigorous quantitative energy distribution analysis conducted during our preliminary experiments. We analyzed the percentage of total signal energy contained in each wavelet component (A5, D1-D5) for all four stations. The results, which are included in the Supplementary Material (Fig. S3a–S3d), were definitive. For all stations, the A5, D1, and D2 components cumulatively accounted for the vast majority of the signal energy (e.g., 99.5% for the AMS02 and AMS04 stations). Conversely, the higher-level detail coefficients (D3, D4, and D5) combined contained less than 1% of the total energy. This quantitative analysis illustrates that the D3-D5 components are statistically insignificant and primarily represent high-frequency noise. Given this evidence, we concluded that including these components would introduce noise and unnecessary complexity without adding meaningful predictive information.

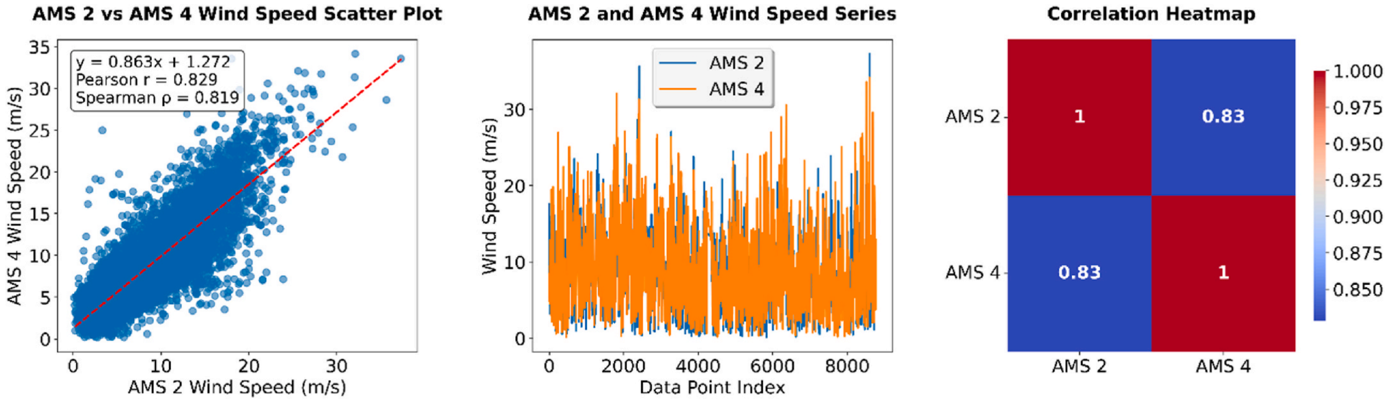

Fig. 2. Comparison between AMS 02 and AMS 04 on wind speed. Scatter plot showing the correlation between wind speeds measured at AMS02 and AMS04 stations during periods with valid data from both stations. The analysis reveals a strong positive correlation with Pearson correlation coefficient r = 0.83 and Spearman rank correlation coefficient ρ = 0.82, indicating high concordance in wind speed measurements between the two locations. The regression line demonstrates the linear relationship used for data substitution, with statistical parameters displayed. Time series comparison and correlation heatmap provide additional validation of the strong relationship between stations.

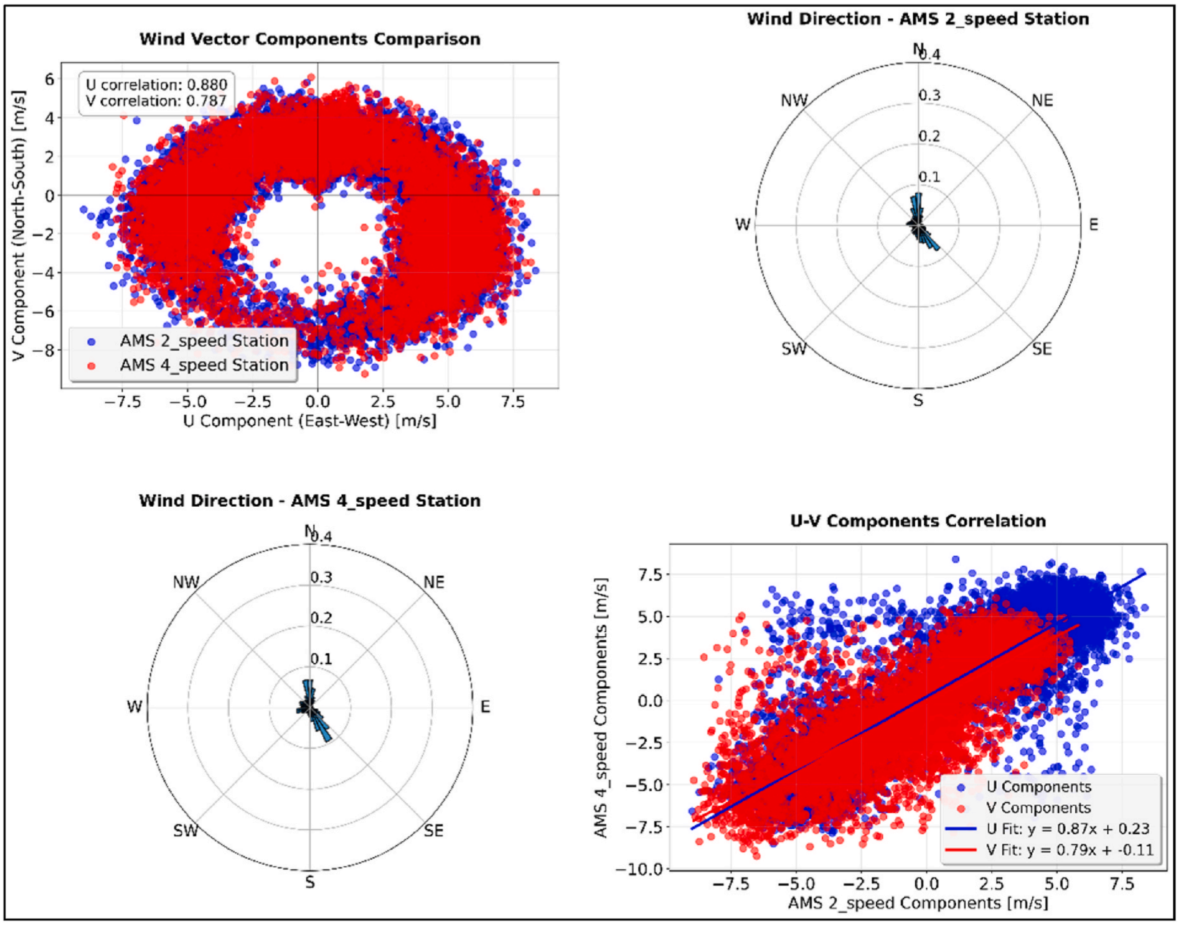

Fig. 3. Comparison Analysis between AMS 02 and AMS 04 on wind direction. Wind vector analysis comparing directional patterns between AMS02 and AMS04 stations. Vector component scatter plot showing U and V wind components from both stations. The wind rose diagrams for AMS02 and AMS04 exhibit considerable similarity in dominant wind directions. These quantitative and qualitative assessments confirm a high degree of concordance in wind patterns between the two locations, supporting the decision to use AMS02 data as a reliable substitute for the missing wind measurements at AMS04 during the identified period. Correlation analysis of U and V components with fitted regression lines, demonstrating consistent wind directional behaviour between stations. The high correlation in both U and V components supports the reliability of using AMS02 data to substitute missing wind direction measurements at AMS04.

Furthermore, consistent with findings in the literature, higher-level detail coefficients (D3-D5), although capable of capturing finer-scale features, often predominantly represent noise or irrelevant high-frequency vibrations that may compromise model performance (Alfaouri and Daqrouq, 2008). This selection of A5, D1, and D2 aims to provide a rich, multi-scale representation of the temporal data while mitigating the risk of overfitting and reducing computational complexity due to high dimensionality (Sahoo et al., 2024).

2.3.4. Temporal pattern labelling through clustering analysis

To identify distinct operational or emission patterns within the ground station data and subsequently guide the attention mechanisms of the ST-CAN, a cluster analysis was performed on the wavelet-decomposed features. Because high-dimensional data can pose challenges for clustering algorithms, we have evaluated several clustering algorithms, including K-Means, DBSCAN, Hierarchical clustering, and Gaussian Mixture Models (GMM), which combined with several dimensionality reduction techniques, such as Principal Component Analysis (PCA), Kernel PCA (KPCA), and t-distributed Stochastic Neighbour Embedding (t-SNE) to visualize the combination results of clustering.

This process was separated into two parts to capture different temporal scales of atmospheric behaviour, aiming to assign two distinct sets of categorical labels to each time point in the ground station dataset. First, clustering was applied to the A5 approximation coefficients for all features. The A5 components, representing the low-frequency baseline trends, are well-suited for identifying longer-term patterns, such as seasonal variations in background pollution levels or persistent meteorological situations. The labels generated from A5 clustering, for example, “Seasonal High Pollution Baseline," or “Seasonal Low Pollution Baseline," are intended to provide the ST-CAN with context about the broader, slowly evolving environmental state. Second, a separate clustering analysis was performed on the D2 detail coefficients. The D2 components, capturing higher-frequency, short-term fluctuations, can point out the moment events like potential methane leaks, rapid pollutant dispersion, or sudden meteorological shifts. Labels derived from D2 clustering aim to alert the model to more immediate, rapidly changing conditions.

By generating the combinations of these two complementary sets of temporal pattern labels, we aim to provide the ST-CAN with a richer and deeper understanding of the ground-level emission patterns. This dual-labelling strategy is supposed to enhance the model's ability to precisely assign attentions to satellite imagery features and to improve both its overall prediction accuracy and its responsiveness to anomalous events.

2.3.5. Temporal alignment strategy and its limitations

A significant challenge in this study is the temporal misalignment between the high-frequency (hourly) ground station data and the low-frequency (weekly) satellite-derived features. To enable the fusion of these different data streams within a single deep learning framework, an up-sampling of the satellite features to an hourly frequency was necessary. For this purpose, we employed a time-weighted interpolation method. This approach calculates a value for each hourly timestamp based on a weighted average of the two nearest weekly data points, with weights being inversely proportional to the temporal distance. This gives greater influence on the closer weekly observation.

We explicitly acknowledge that this temporal up-sampling is a primary limitation of our study. This interpolation process, by its nature, cannot recreate true atmospheric variability that occurs at sub-weekly timescales and risks introducing artificial temporal smoothness. This study should therefore be viewed as an exploration of the potential of a deep learning architecture to fuse multi-modal data under conditions of significant data sparsity. The model's performance and the resulting concentration fields should be interpreted within the context of this inherent data limitation. Overcoming this challenge will require future research using data from next-generation satellites with higher revisit rates.

2.4. Baseline models training

To comprehensively evaluate the performance of the proposed ST-CAN, a series of baseline models was implemented and trained. These models were strategically selected to serve two key purposes: 1) to establish the performance of single-modality models, using a single data stream, like only ground or only satellite, and 2) to compare ST-CAN against other well-known models with dual data streams. Additionally, to ensure a fair and rigorous comparison, all baseline models described in this section were trained and evaluated on the same pre-processed dataset used by ST-CAN. This includes the wavelet-decomposed ground features and the time-weighted interpolated hourly satellite features. For the dual-stream baselines, the ground and satellite features were combined into a single input vector at each hourly time step via simple concatenation.

2.4.1. Single data stream baseline models

Two baseline models were developed to assess the predictive power of each data stream. LSTM was designed due to its powerful prediction performance on time-series data. For ground data, its input consists of the pre-processed time-series features from the WBEA stations, including meteorological variables and the wavelet-decomposed components of CH_4 concentrations. The architecture comprises a bidirectional LSTM layer with 64 hidden units followed by a fully connected prediction layer. This baseline serves to quantify the predictive accuracy achievable with high-frequency temporal data alone. For the satellite data stream, its input is a time series derived from the satellite features, which are extracted from the pre-trained model ResNet18 corresponding to the pixel closest to the target ground station. The architecture is identical to the ground-only model, featuring a bidirectional LSTM layer with 64 hidden units. This baseline helps to understand the inherent predictive value, or lack thereof, of the temporally sparse satellite data.

2.4.2. Dual data stream baseline models

Three distinct baseline models were implemented and trained to compare with the performance of the proposed ST-CAN Support Vector Regression (SVR), LSTM, and a Transformer-based model. These models were carefully selected to provide a rigorous and fair benchmark. The SVR serves as a strong classical machine learning baseline. The LSTM was chosen as a powerful standard for time-series forecasting. Crucially, the Transformer model was included as a state-of-the-art deep learning baseline specifically because its self-attention mechanism provides the most relevant and fair comparison for evaluating the novel cross-attention framework of ST-CAN. To ensure a rigorous and fair evaluation, the performance of all baseline models was optimized prior to comparison. We conducted a systematic hyperparameter tuning process for each baseline model. For example, adjusting kernel parameters for SVR, hidden layers for LSTM and Transformer. The final configurations selected for comparison represent the optimal settings that achieved the lowest error on the validation dataset, ensuring that the proposed ST-CAN is benchmarked against the strongest possible versions of these established methods.

The SVR model was selected as a strong classical machine learning baseline, valued for its efficiency in capturing non-linear relationships, with relatively few parameters to reduce overfitting risk. Key configuration parameters for the SVR included the use of a Radial Basis Function (RBF) kernel, chosen for its ability to effectively model the non-linear interaction between methane emissions and various meteorological and atmospheric variables. A regularization parameter C of 1.0 was used, allowing the model to more closely fit the training data, which is beneficial for capturing critical emission patterns, including sporadic events. An epsilon value of 0.1 was set to enhance the model's sensitivity to subtle changes in methane levels, enabling it to respond to both gradual trends and short-term emission incidents. To ensure robustness and the temporal nature of the methane emission data, a time-series cross-validation approach (5-fold) was implemented, preventing future observations from leaking into past predictions. An ensemble model was then created by averaging the predictions from all cross-validation folds, enhancing stability and mitigating the influence of anomalous periods on the final forecast. Input features were scaled using Min-Max normalization. This configuration positions the SVR as a capable baseline, adept at handling non-linear environmental data influenced by multiple factors.

An LSTM network was selected as a representative deep learning architecture, well known for its ability to learn long-range dependencies in sequential data. The architecture comprised a two-layer LSTM with 64 units in the first layer and 32 units in the second. This design allows the model to initially capture complex patterns and subsequently refine them into higher-level features. This structure is well-suited for modelling the intricate non-linear relationships in methane time series. The input to the LSTM consisted of sequences with a 24-h time step, enabling the model to learn daily cycles and human activity-related periodicities in emissions. The training process was conducted with a batch size of 32 and an initial learning rate of 0.001, utilizing the Adam optimizer. The LSTM layers employed the tanh activation function, while dense layers used ReLU. A dropout rate of 0.2 was applied to mitigate overfitting, particularly important for potentially noisy environmental data, while retaining sufficient network capacity.

A Transformer model was included as a state-of-the-art deep learning baseline, recognized for its efficiency in processing long sequences via its self-attention mechanism. A more comprehensive architecture was adopted, featuring a three transformer encoder layer with eight attention heads, an embedding dimension of 128, and a feed-forward network dimension of 256. This streamlined design aimed to balance performance with computational efficiency and reduce the risk of overfitting on the available dataset. A relatively high dropout rate of 0.1 was applied, crucial for mitigating overfitting given the potential noise and variability in methane data. The model processed inputs with a 24-h time step, the same as the LSTM configuration.

2.5. Spatio-temporal cross-attention network

To effectively integrate the distinct yet complementary information streams from ground-based monitoring stations and satellite observations, we propose the ST-CAN. This novel deep learning architecture, illustrated schematically in Figs. 4 and 5, is designed to explicitly model the complex interactions between high-frequency temporal data and spatially extensive, lower-frequency imagery, while adapting to the inherent dynamics and potential data gaps in environmental monitoring. The ST-CAN processes the pre-processed ground station data, including the multi-scale wavelet features and temporal pattern cluster labels, and the satellite-derived methane concentration maps through parallel pathways initially, before employing specialized fusion mechanisms.

The core innovations of ST-CAN are its bidirectional cross-attention mechanism, which facilitates deep information exchange, and its semantic-guided dynamic fusion gate, which provides an interpretable and adaptive weighting strategy.

2.5.1. Bidirectional cross-attention mechanism

To effectively model the interplay between ground-based time-series data and satellite imagery, we employ a bidirectional cross-attention mechanism. This process, illustrated in the Stage 4 panel of Fig. 5, consists of two parallel streams where each modality's feature representation is enhanced by the context of the other before the final fusion. The first stream is the Ground to Satellite (G-S) attention. In this stream, the ground temporal features (h_g) from the Bi-LSTM act as the Query, while the satellite spatial features (h_s) from ResNet18 serve as the Key and Value. The calculation is defined as follows:

$$Q_g = h_g W^{Q_g} K_s, V_s = h_s W^{K_s}, h_s W^{V_s} h_{g\rightarrow s} = Attention\left(Q_g, K_s, V_s\right)$$
$$= softmax\left(\frac{Q_g K_s^T}{\sqrt{d_k}} V_s\right)$$

Here, h_g and h_s are the input feature representations from the ground and satellite encoders, respectively. W^{Q_g}, W^{K_s}, and W^{V_s} are learnable weight matrices used to project the inputs into the Query (Q), Key (K), and Value (V) spaces. This results in Q_g (the query derived from ground features), and K_s and V_s (the key and value derived from satellite features). d_k is the dimension of the key, used for scaling the dot product. The final output, $h_{g\rightarrow s}$, represents the satellite spatial features as attended to by the ground temporal context. This stream produces a representation that determines the most relevant spatial context from the satellite, given the ground station's temporal patterns. For example, if the ground station detects a sudden spike, the model can query the satellite image to confirm if a large-scale emission plume is spatially present.

The second, complementary stream is the Satellite to Ground (S-G) attention. Here, the satellite spatial features (h_s) conversely act as the Query, while the ground temporal features (h_g) serve as the Key and Value. The calculation is:

$$Q_s = h_s W^{Q_s} K_g, V_g = h_g W^{K_g}, h_g W^{V_g} h_{s\rightarrow g} = Attention\left(Q_s, K_g, V_g\right)$$
$$= softmax\left(\frac{Q_s K_g^T}{\sqrt{d_k}} V_g\right)$$

This stream produces $h_{s\rightarrow g}$, which identifies the most relevant high-frequency temporal dynamics from the ground stations given the satel-

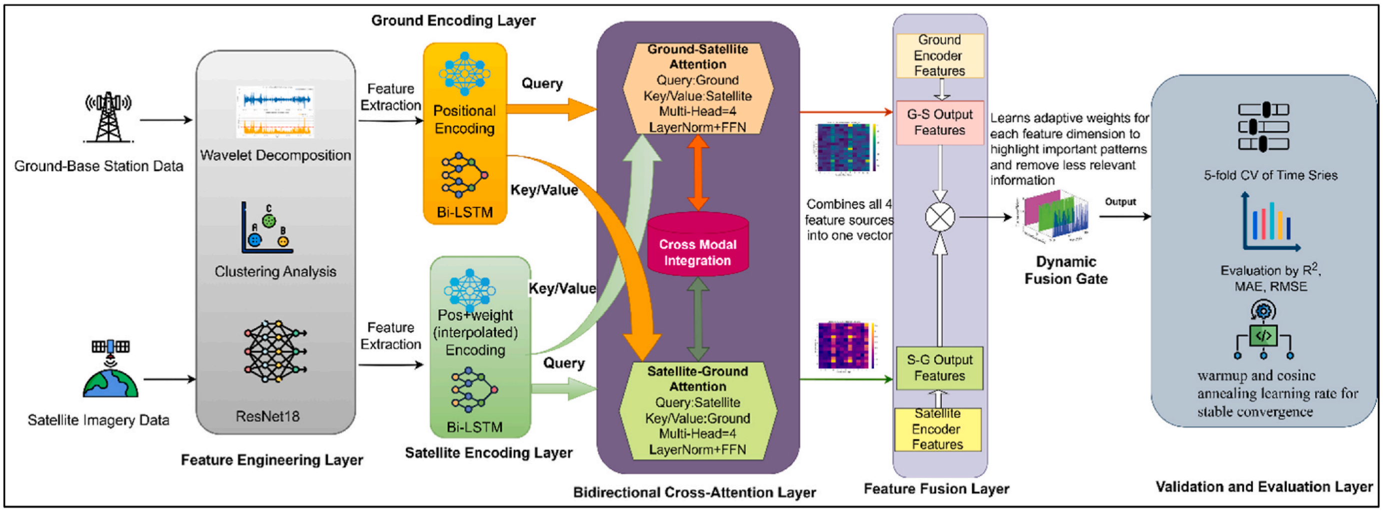

Fig. 4. Schematic Diagram of the Proposed ST-CAN for Multi-Source Methane Data Fusion. Inputs include hourly ground station data, processed via wavelet decomposition and clustering analysis, and weekly satellite methane maps. Data streams are encoded using LSTMs and CNNs before entering the Bidirectional Cross-Attention module, guided by temporal pattern weights. Fused features are adaptively weighted by the Dynamic Fusion Gate before passing to the prediction head.

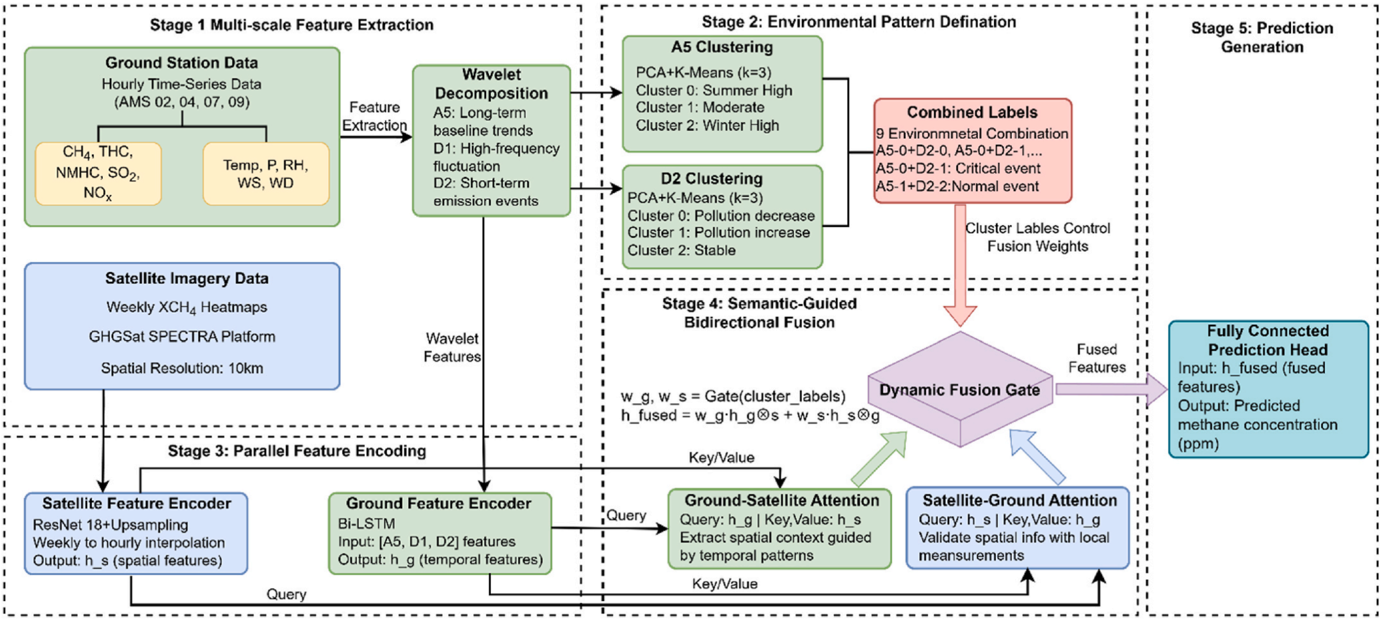
Stage 1 Multi-scale Feature Extraction
Ground Station Data
Hourly Time-Series Data
(AMS 02, 04, 07, 09)
CH4, THC, NMHC, SO2, NOx
Temp, P, RH, WS, WD
Feature Extraction
Wavelet Decomposition
A5: Long-term baseline trends
D1: High-frequency fluctuation
D2: Short-term emission events
Satellite Imagery Data
Weekly XCH4 Heatmaps
GHGSat SPECTRA Platform
Spatial Resolution: 10km
Wavelet Features
Stage 2: Environmental Pattern Defination
A5 Clustering
PCA+K-Means (k=3)
Cluster 0: Summer High
Cluster 1: Moderate
Cluster 2: Winter High
D2 Clustering
PCA+K-Means (k=3)
Cluster 0: Pollution decrease
Cluster 1: Pollution increase
Cluster 2: Stable
Combined Labels
9 Environmnetal Combination
A5-0+D2-0, A5-0+D2-1,...
A5-0+D2-1: Critical event
A5-1+D2-2:Normal event
Cluster Lables Control Fusion Weights
Stage 5: Prediction Generation
Stage 4: Semantic-Guided Bidirectional Fusion
w_g, w_s = Gate(cluster_labels)
h_fused = w_g·h_g⊗s + w_s·h_s⊗g
Dynamic Fusion Gate
Fused Features
Fully Connected Prediction Head
Input: h_fused (fused features)
Output: Predicted methane concentration (ppm)
Stage 3: Parallel Feature Encoding
Key/Value
Satellite Feature Encoder
ResNet 18+Upsampling
Weekly to hourly interpolation
Output: h_s (spatial features)
Ground Feature Encoder
Bi-LSTM
Input: [A5, D1, D2] features
Output: h_g (temporal features)
Query
Ground-Satellite Attention
Query: h_g | Key,Value: h_s
Extract spatial context guided by temporal patterns
Satellite-Ground Attention
Query: h_s | Key,Value: h_g
Validate spatial info with local measurements
Query
Key/Value

Fig. 5. Five-stage Workflow Pipeline of ST-CAN for Methane Predictions. (1) Multi-scale feature extraction via wavelet decomposition (A5, D1, D2) and satellite imagery processing; (2) Environmental pattern definition through separate clustering of A5 (long-term baseline) and D2 (short-term dynamics), generating nine combined environmental states; (3) Parallel encoding of ground features via Bi-LSTM and satellite features via ResNet18; (4) Semantic-guided bidirectional fusion where cluster labels dynamically control the fusion gate weights; and (5) Prediction generation through fully connected layers. This workflow demonstrates how wavelet decomposition, clustering, and cross-attention work synergistically to enable physics-informed adaptive fusion.

lite's spatial overview. For example, if the satellite sees a faint anomaly over a region, the model can query the ground station's time series to confirm if this corresponds to a genuine, dynamic emission event. These two outputs, $h_{g\to s}$ and $h_{s\to g}$, represent features from both modalities that are now mutually aware of each other's context. They are then passed to the Dynamic Fusion Gate.

2.5.2. Semantic-guided dynamic fusion gate

Another critical innovation of ST-CAN is that the final fusion of these two enhanced feature streams is not static. It is dynamically controlled by the environmental semantic labels derived from our clustering analysis. This mechanism allows the model to adaptively trust one modality more than the other based on the real-world physical context. The gate mechanism (as shown at stage 4 in Fig. 5) is a learnable neural network module that takes three inputs. These are the G-S enhanced features $h_{g\to s}$, the S-G enhanced features $h_{s\to g}$, and the current environmental semantic label with a categorical value from 0 to 8, representing the nine cluster combinations. The working principle of this gate involves several steps. First, the categorical cluster label is converted into a dense vector representation. Specifically, we utilize a learnable embedding layer (dimension $d_{emb} = 16$) rather than simple one-hot encoding. This allows the network to learn a continuous representation of the environmental states, capturing latent similarities between different semantic conditions. Second, the two enhanced feature vectors and the new cluster embedding vector are concatenated. This combined vector is then passed through a small feed-forward network that terminates in a Softmax function. The network outputs two scalar weights, w_g and w_s, which sum to one. Because the cluster label is a direct input to this network, the resulting weights are dynamically generated based on the environmental state. Finally, the fused feature representation, h_{fused}, is computed as a dynamically weighted sum of the two mutually aware feature streams, as shown in the following equation:

$$h_{fused} = w_g \cdot h_{g\to s} + w_s{\cdot}h_{s\to g}$$

This semantic guidance is physically interpretable. For instance, during a sudden, high-emission event, for example, Cluster A5-0 + D2-1, the model learns to assign a higher weight w_g ($w_g = 0.8$, $w_s = 0.2$). This reflects a learned strategy to trust the high-frequency ground data more, as the weekly satellite data is too coarse and likely misses the peak. In contrast, during a large-scale regional transport event with Cluster A5-2 + D2-0, the model learns to assign a higher weight w_s ($w_g = 0.3$, $w_s = 0.7$). This reflects a strategy to trust the satellite's wide-area spatial coverage more, as a single ground station cannot capture the full spatial extent of the event. This adaptive, interpretable fusion mechanism moves beyond simple concatenation or static attention, providing a robust representation that is passed to the final prediction layer.

Finally, to ensure stable and effective learning, the ST-CAN is trained end-to-end using a robust optimization strategy, specifically the optimization combined with learning rate warmup and cosine annealing schedules. This approach helps stabilize training in the early phases and promotes convergence towards optimal model parameters. The adaptively fused spatio-temporal representation generated through these steps is then fed into a final prediction layer of a fully connected network to output the target variable for methane regression prediction.

By integrating these components, ST-CAN aims to provide a robust and adaptive framework for methane monitoring that leverages the strengths of both ground-based and satellite observations while addressing key challenges in multi-source data fusion, interpretability, and temporal dynamics.

3. Results and discussion

3.1. Results of data pre-processing and clustering analysis

3.1.1. Multi-scale feature extraction using wavelet decomposition

After the data cleaning and interpolation, wavelet decomposition was applied to all hourly time series variables, including pollutants and meteorological parameters from the four selected WBEA stations. As detailed in Section 2.3.3, this technique transferred the original signal into components representing different time scales, effectively separating underlying trends from transient fluctuations. The low-frequency approximation component (A5) captures the smoother, long-term variations, while the detail components (D1-D5) isolate higher-frequency fluctuations corresponding to short-term events like emission spikes, plume transport, or rapid meteorological changes. The selection criteria are based on the energy distribution validation as shown in Fig. S1a and S1b in the Supplement Materials. For CH_4, the approximation component A5 represents the dominant energy component, and cumulative A5-D2 energy exceeds 97%, confirming that five levels sufficiently capture meaningful information while minimizing noise. Furthermore, to rigorously validate db4's suitability, we conducted a comprehensive sensitivity analysis comparing four common wavelets (db2, db4, db6, sym4) across all 17 features in our dataset. As shown in Fig. 6, db4 achieves the highest average approximation coefficient energy concentration of 68.2h% with the most stable performance across different feature types, demonstrating its superior applicability to our environmental dataset.

Fig. 7 presents an example of this decomposition applied to the CH_4 time series data at the AMS09 (Barge Landing) station. It displays the smoothed A5 component onto the original signal, visually demonstrating how A5 captures the underlying baseline trend. This multi-scale representation, separating trends from fluctuations and enabling the identification of anomalous periods via components like D2, formed the basis for the subsequent clustering analysis and provided enhanced feature inputs for the ST-CAN.

3.1.2. Clustering analysis for temporal pattern identification

We applied a clustering analysis to identify distinct seasonal emission patterns and short-term leakage events within the ground station data and subsequently provided a guiding temporal context to the ST-CAN. We have evaluated combinations of four clustering algorithms (K-Means, DBSCAN, Hierarchical, GMM) and three dimensionality reduction techniques (PCA, KPCA, t-SNE) applied to the wavelet features A5 and D2 derived from all measured variables to extract emission patterns from the high-dimensional wavelet-decomposed ground station data, and to subsequently provide guiding temporal context to the ST-CAN.

This comprehensive comparison aimed to select the most effective combination for classifying the data into meaningful groups. We evaluated 12 combinations of four clustering algorithms, including K-Means Clustering (K-Means), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Hierarchical Clustering (Hierarchical) and Gaussian Mixture Model (GMM) and three dimensionality reduction techniques (PCA, KPCA, t-SNE). As demonstrated in Supplementary Fig. S4a and S4b, the PCA + K-Means combination yielded the most distinct and well-separated clusters, which also aligned best with subsequent physical interpretation. In addition, as detailed in Supplementary Table S1, this combination achieved the highest Calinski-Harabasz Index (12,845.2 for A5 and 7476.2 for D2) among all tested methods, indicating superior cluster definition and compactness. Furthermore, unlike density-based methods such as DBSCAN, which exhibited poor structural scores (CH Index <300) and failed to provide complete temporal coverage (98.1%), K-Means ensured 100% data availability with distinct, robust boundaries required for the semantic gating mechanism. Moreover, the selection of $k = 3$ was quantitatively validated using the Elbow Method, Silhouette Score, and Bayesian Information Criterion (BIC). The results of this analysis shown in Fig. 8 below clearly indicate that $k = 3$ is the optimal number of clusters, as evidenced by a distinct peak in the Silhouette Score, representing the best balance of cluster cohesion and separation.

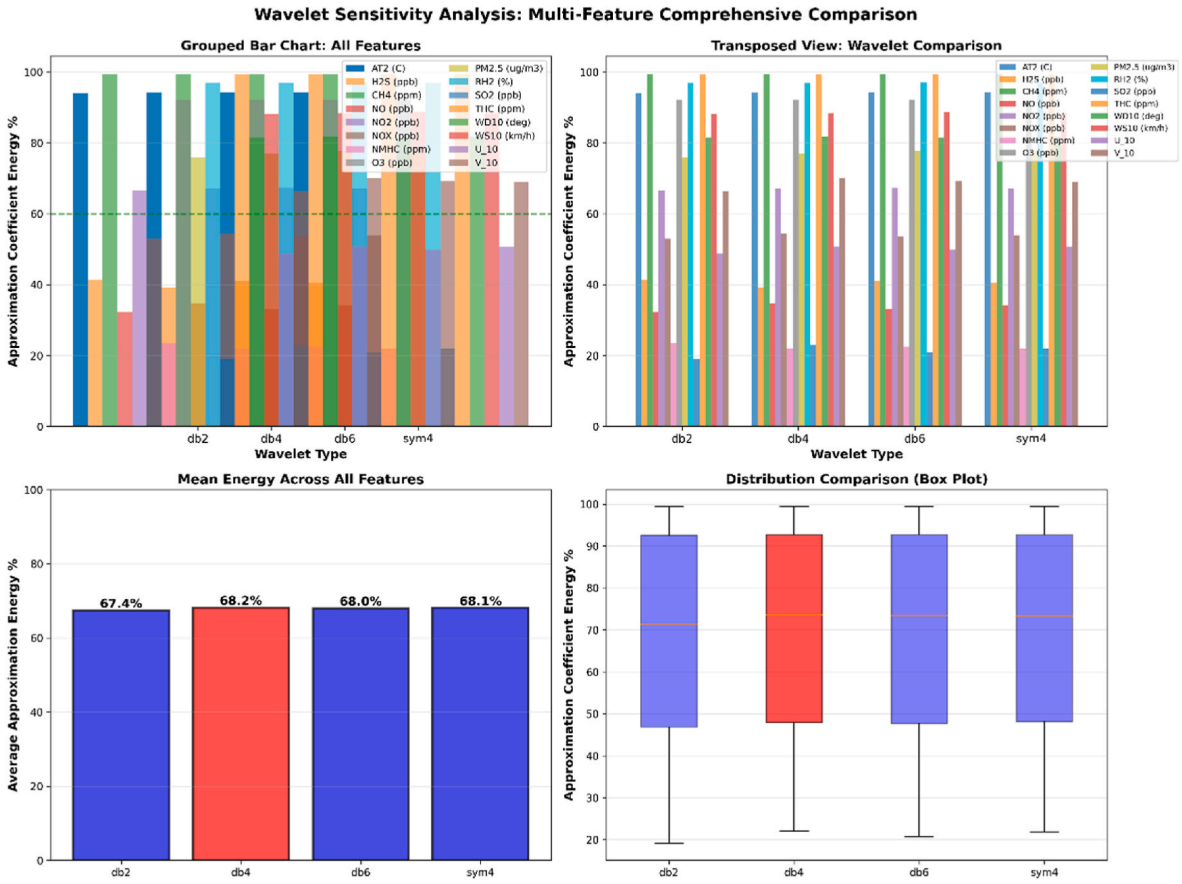

Fig. 6. Wavelet Basis Function Sensitivity Analysis. Four panels show complementary perspectives of approximation coefficient energy concentration: (a) grouped comparison across features; (b) energy distribution by wavelet type; (c) mean energy showing db4 (red) achieves the highest average concentration; (d) box plot demonstrating db4's superior stability (smallest variance). Results collectively validate db4 as the optimal wavelet basis for environmental time series decomposition.

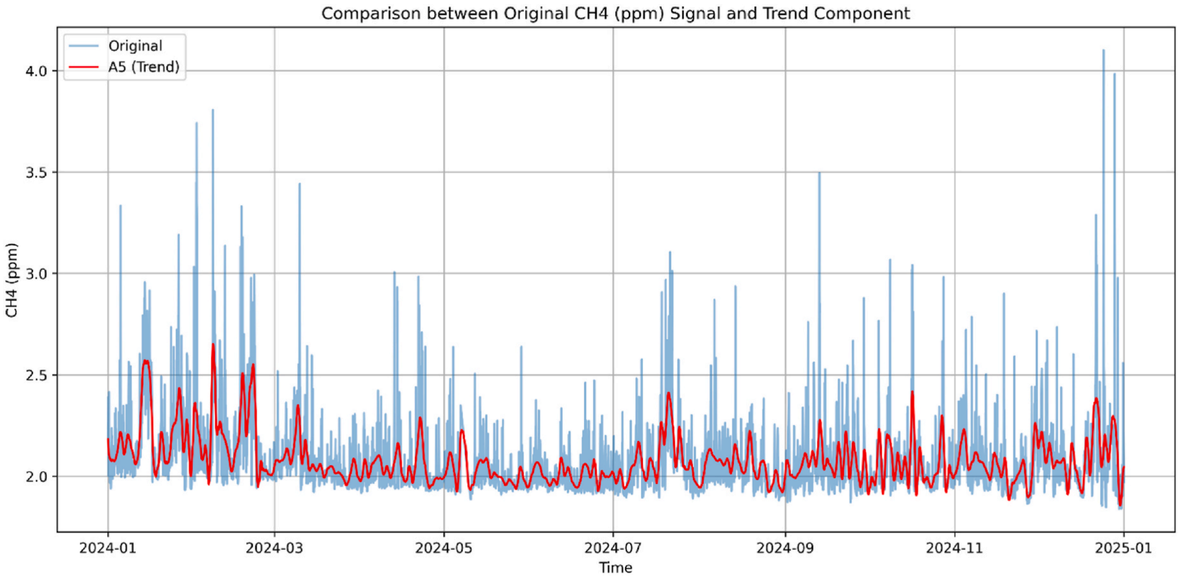

Fig. 7. Example of Wavelet Decomposition applied to hourly CH_4 concentration data from AMS09.

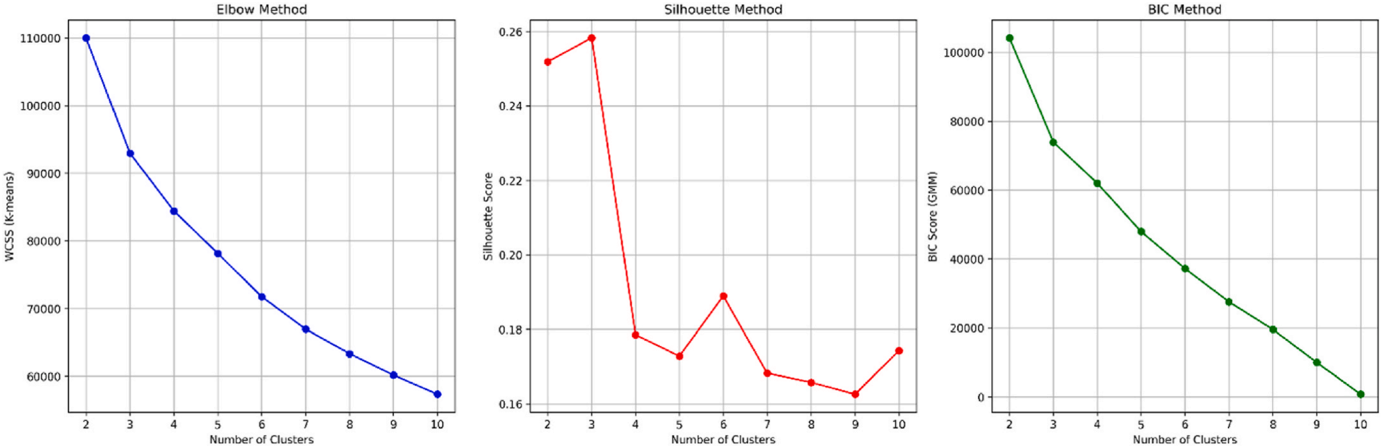

Fig. 8. Quantitative Validation for Optimal Cluster Number Selection for A5 Features. Objective metrics validate the selection of k = 3 for A5 feature clustering. The Elbow Method (left) shows a diminishing return in WCSS reduction after k = 3. The Silhouette Method (center) provides the strongest evidence, displaying a clear optimal peak at k = 3. The BIC Method (right) also shows an inflection point at k = 3, collectively confirming k = 3 as the optimal choice for cluster separation.

Fig. 9 visually presents the resulting cluster distributions in the PCA-reduced feature space by using K-Means clustering methods for both the A5 and D2 components, illustrating the separation achieved for these two distinct types of temporal characteristics. The resulting cluster boundaries show minimal overlap, suggesting robust segregation of different emission patterns.

Characterization of these three temporal patterns was performed by examining the distribution of the A5 and D2 wavelet features within each cluster. Fig. 10 presents radar charts summarizing the standardized coefficients for all atmospheric parameters within each cluster.

These nine distinct environmental states, which capture both the underlying baseline and the immediate dynamics, form the core of our semantic guidance. Each state corresponds to a specific, interpretable environmental event. We have defined and provided detailed examples for all nine combinations in Table 1. For instance, 'State 6' (A5-1 + D2-2) represents normal stable conditions, while 'State 2' (A5-0 + D2-1) represents a critical emission spike during a summer stagnation period. These dual labels were subsequently used as guidance to assign the appropriate attention weights to the ST-CAN, enabling it to adapt its fusion strategy with greater precision based on a comprehensive understanding of the current environmental context.

3.2. Baseline models training and performance evaluation

The three baseline models, including SVR, LSTM, and a Transformer-based model, were trained and evaluated to establish a benchmark for the proposed ST-CAN. The performance of these models provides a reference point against which the capabilities of the more complex ST-CAN architecture can be assessed.

The overall predictive accuracy of the baseline models that evaluated on all cross-validation folds, is initially assessed. Here, we provided an example result of SVR as a reference. Fig. 11 below summarizes a comprehensive prediction performance on the SVR model. While the SVR model demonstrates a certain level of predictive capability, its performance, for example, a relatively lower R^2 of 0.68, may not be entirely satisfactory. One potential explanation for this could be the inherent limitations of SVR in effectively processing and integrating complex spatial features. The training dataset, which combines ground station time series with satellite image features which are extracted using ResNet18 and subsequently interpolated, presents a multi-modal

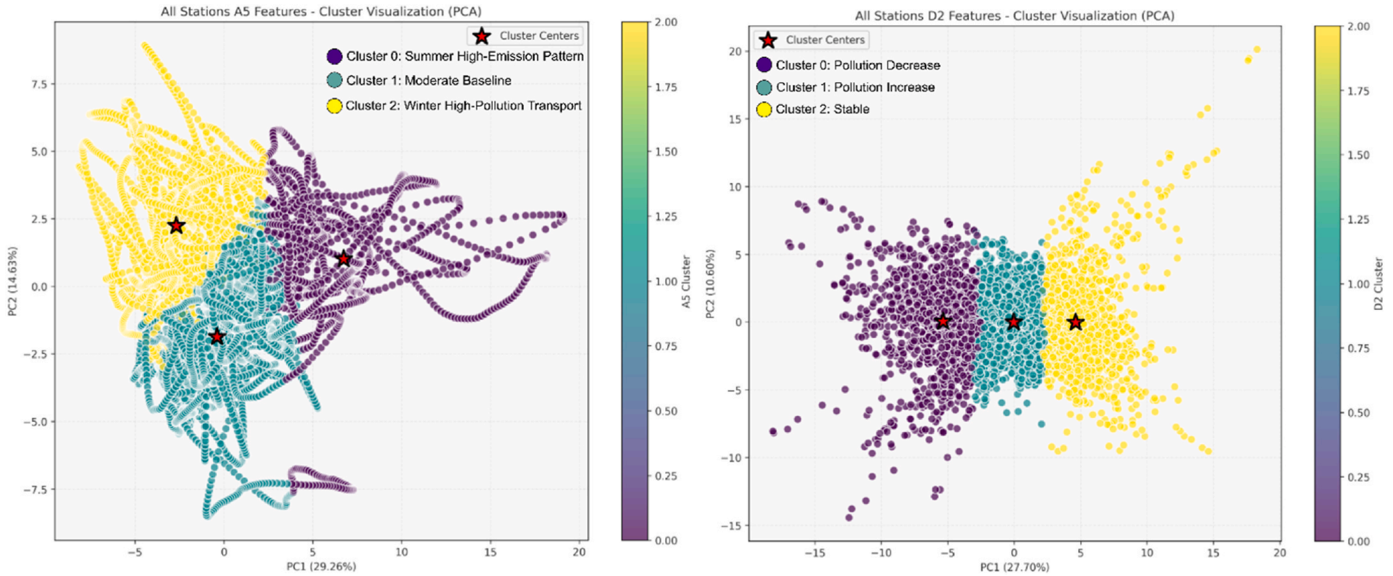

Fig. 9. Visualization of Environmental Semantic Clusters in PCA Feature Space. The left panel shows the clustering of A5 (long-term trend) features, and the right panel shows the clustering of D2 (short-term dynamic) features. The clusters are labelled with their derived environmental meanings: A5-0 = Summer High-Emission, A5-1 = Moderate Baseline, A5-2 = Winter High-Transport; D2-0 = Pollution Decrease, D2-1 = Pollution Increase, D2-2 = Stable. These figures demonstrate clear separation in the feature space, forming the basis for the semantic labels. The detailed environmental interpretation of each cluster is presented in Fig. 4 and Table 1.

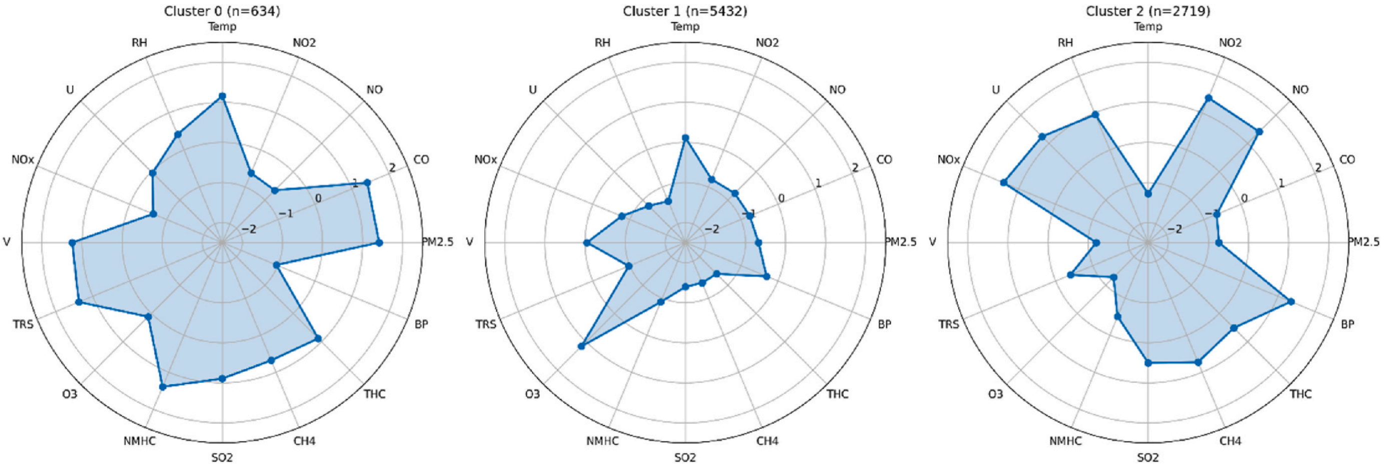

Fig. 10. Radar chart characterizing the three clusters based on the A5 trend. Each radar chart displays the measured atmospheric parameters. The radial distance from the center represents the magnitude of standardization, with positive values (extending outward) indicating above-average levels and negative values (closer to the center) indicating below-average levels for each parameter. The shadow areas represent the characteristic “fingerprint" of each cluster: Cluster 0 (n = 634) shows summer high-emission stagnant pattern, Cluster 1 (n = 5432) represents moderate photochemical baseline conditions, and Cluster 2 (n = 2719) indicates winter high-pollution transport pattern. Parameters include pollutants, including CO, NO, NO_2, NO_x, O_3, $PM_{2.5}$, SO_2, CH_4, Total Hydrocarbons (THC), Non-Methane Hydrocarbons (NMHC), Total Reduced Sulfur (TRS), meteorological variables (temperature, barometric pressure, relative humidity, wind speed), and wind components (U: easterly, V: southerly).

challenge. SVR models, while adept at handling non-linear temporal relationships, might struggle to fully capture and leverage the rich spatial information contained within the numerous satellite-derived features. This could lead to a poor representation of the spatio-temporal interaction on CH_4 concentrations, potentially explaining some of the observed discrepancies between predicted and actual values. The following evaluation of the deep learning baselines (LSTM and Transformer) and the ST-CAN will explore whether architectures more explicitly designed for spatio-temporal and multi-modal fusion can offer improved performance.

From the performance metrics, we can observe that the Deep Learning models, including LSTM and Transformer, dominated the SVR in this prediction problem. The Transformer model showed the smallest RMSE of 0.0479, which means that it effectively reduced the large prediction error values. The LSTM model came in second place with an RMSE of 0.0508. The superior performance of the Transformer, particularly in achieving the lowest RMSE and MAE and the highest R^2, highlights the potential benefits of attention-based architectures for this type of environmental forecasting. This comprehensive evaluation of the baseline models establishes a clear performance benchmark, paving the way for assessing the advancements offered by the proposed ST-CAN, which is specifically designed for enhanced spatio-temporal fusion.

The performance of the single data stream baseline models reveals critical insights into the nature of the data modalities. The Ground-Only model achieved an acceptable R^2 of 0.64, indicating that ground-based time-series data contains significant predictive power. In contrast, the

Table 1
Definition and environmental interpretation of the nine semantic cluster combinations.

A5 (Long-term) Cluster	D2 (Short-term) Cluster	Combination (State)	Environmental Meaning & Specific Event Example
A5-0 (Summer High-Emission)	**D2-0** (Pollution Decrease)	State 1	**High-Emission (Cooling Down):** A high-pollution event (e.g., from stagnant air) that is beginning to disperse.
A5-0 (Summer High-Emission)	**D2-1** (Pollution Increase)	State 2	**Critical Emission Spike (Summer):** A sudden, acute emission event (e.g., industrial leak) occurring during an already high-pollution summer period. **(Critical Event)**
A5-0 (Summer High-Emission)	**D2-2** (Stable)	State 3	**Persistent Stagnation:** A prolonged period of high pollution, typically caused by stagnant air masses in summer, with no significant short-term changes.
A5-1 (Moderate Baseline)	**D2-0** (Pollution Decrease)	State 4	**Post-Event Recovery:** The environment is returning to normal background levels after a minor pollution event.
A5-1 (Moderate Baseline)	**D2-1** (Pollution Increase)	State 5	**Minor Emission Event:** A typical, short-term emission spike (e.g., from operational upsets) under normal background conditions.
A5-1 (Moderate Baseline)	**D2-2** (Stable)	State 6	**Normal Stable Conditions:** The most frequent state, representing typical background concentrations with no significant events. **(Baseline State)**
A5-2 (Winter High-Transport)	**D2-0** (Pollution Decrease)	State 7	**Regional Transport (Ending):** The tail-end of a large-scale pollution transport event, where concentrations are decreasing but still high.
A5-2 (Winter High-Transport)	**D2-1** (Pollution Increase)	State 8	**Regional Transport (Arriving):** The arrival of a large-scale plume from an external source, causing a rapid increase in local pollution during winter.
A5-2 (Winter High-Transport)	**D2-2** (Stable)	State 9	**Persistent Transport Event:** A prolonged period where the region is sitting within a large, stable high-pollution air mass, often associated with winter conditions.

Satellite-Only model performed exceptionally poorly ($R^2 \approx 0.01$), confirming that the weekly satellite features, when interpolated to an hourly frequency, do not possess sufficient temporal resolution to act as a standalone predictor and behave largely as noise at this scale. This result highlights the severe challenge of temporal data sparsity.

3.3. Comparative analysis and ST-CAN evaluation

The ST-CAN, with its architecture designed for enhanced multi-source data fusion and contextual understanding, was trained and evaluated using the same five-fold cross-validation procedure applied to the baseline models to compare the differences. A direct visual comparison of the predictive precision across all models is presented in Fig. 12, which displays density scatter plots of predicted versus actual CH_4 concentrations. Each panel illustrates the linear fit between predictions and observations, with the colour intensity representing the density of data points. The ST-CAN exhibits a notably tight clustering of points, indicating a strong linear relationship and high agreement between its predictions and the actual CH_4 concentrations. In contrast, while the SVR, LSTM, and Transformer models show varying degrees of correlation, their scatter plots generally display a greater dispersion of points and potentially more deviation from the ideal line compared to ST-CAN, suggesting a less precise fit to the observed data.

A more comprehensive quantitative and qualitative assessment is provided in the dashboard presented in Fig. 13. The bar charts in the upper section directly compare the key evaluation metrics with R^2, RMSE, and MAE for ST-CAN against the three baseline models. Across all three metrics, ST-CAN demonstrates markedly superior performance. It achieves the highest R^2 value of 0.96, the lowest RMSE of 0.0214, and the lowest MAE of 0.0141. These figures represent a significant improvement over the best-performing baseline model of Transformer, with $R^2 = 0.79$, RMSE $= 0.0479$, and MAE $= 0.0316$, indicating that ST-CAN explains a substantially larger proportion of the variance in CH_4 predictions with considerably smaller errors on average.

Finally, the full-year CH_4 prediction comparison at the bottom illustrates enhanced predictive capability of ST-CAN. The plot presents the actual observed CH_4 concentrations alongside the predictions from all four models over the entire year. It is visually apparent that the ST-CAN's prediction line tracks the actual CH_4 dynamics much more closely than those of the SVR, LSTM, and Transformer. ST-CAN demonstrates a superior ability to capture both the baseline fluctuations and, crucially, the peaks and troughs in CH_4 concentration, which often represent significant emission events or dispersion phenomena. The baseline models, while capturing some general trends, tend to oversmoothed these variations in their response to rapid changes.

To provide a definitive assessment of the ST-CAN, a comprehensive comparative analysis was conducted against all baseline models. The results are presented in Fig. 14, offering clear insights into the models' performance. The quantitative metrics in Fig. 14 establish the superiority of the ST-CAN. It achieved the highest R^2 of 0.95, the lowest RMSE of 0.020, and the lowest MAE of 0.014. This represents a substantial 34.76% improvement in R^2 over the best-performing single-modality model of Ground-Only ($R^2 = 0.714$), validating the significant benefits of multimodal fusion. The poor performance of the Satellite-Only model ($R^2 = 0.006$) further underscores that neither data stream is sufficient in isolation. Notably, the time-series plot in Fig. 14 provides strong visual confirmation of the quantitative results. The prediction from the ST-CAN (green line) demonstrates an excellent fit, closely tracking the dynamics of the actual observed concentrations (black line) from September to December, including both baseline fluctuations and peak events. In contrast, while the Ground-Only model (red line) captures the general trend, it exhibits significantly higher volatility and larger prediction errors, especially during peak periods.

Therefore, the combined quantitative and qualitative evidence leads to a clear conclusion: the ST-CAN, through its semantic-guided cross-attention mechanism, effectively fuses the high-frequency temporal information from ground stations with the spatial context from satellite data. This fusion process successfully leverages the strengths of both modalities to produce predictions that are not only more accurate on average but also more stable and reliable than those from single-data-stream models.

To provide a qualitative validation of ST-CAN's fusion mechanism, we conducted a case study on the model's behaviour during two distinct peak events in late 2024, where the Ground-Only model failed. This analysis moves beyond R^2 scores to demonstrate how the satellite data provides practical, corrective value. As a preliminary step, a review of public reports for the study period (Jan 2024–Jan 2025) confirmed that no major, acute leakage events were reported. Instead, literature confirms emissions are dominated by persistent, ongoing sources, for example, tailings ponds and operational fluctuations, which are known to be systemically under-reported. This finding highlights the value of

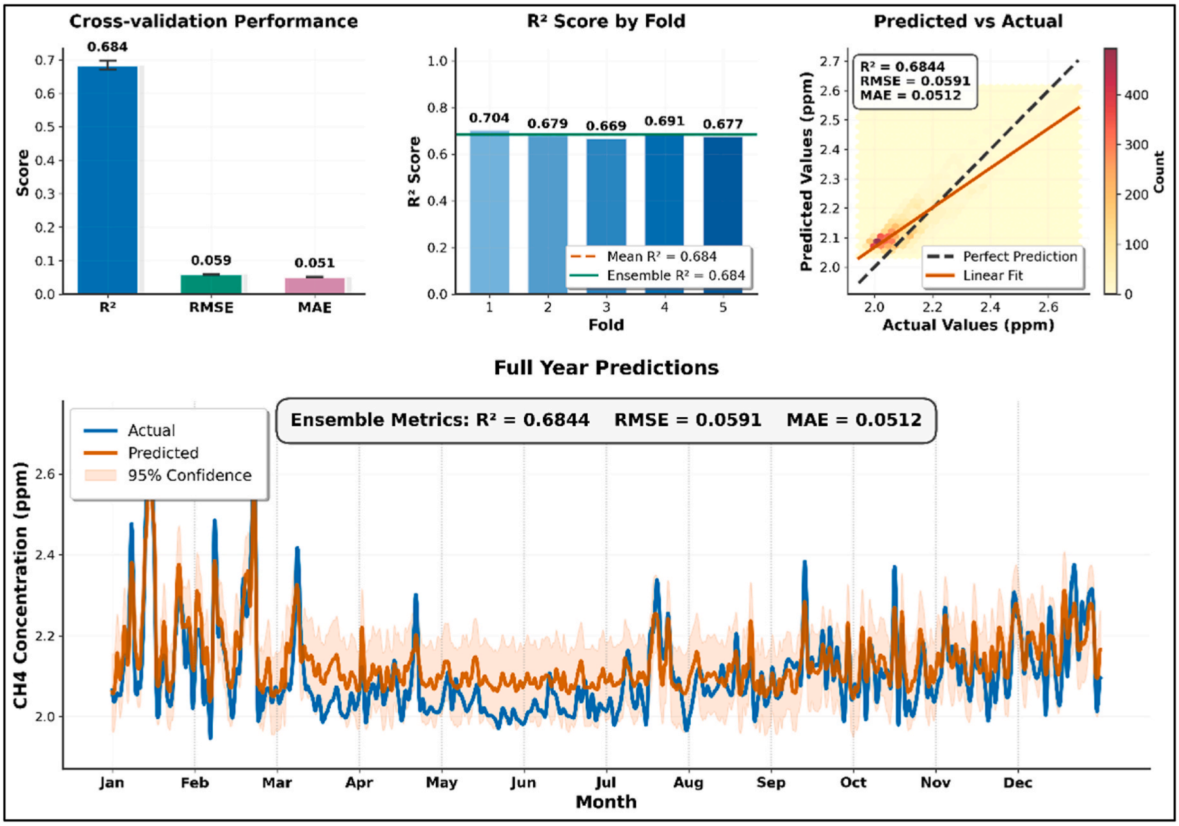

Fig. 11. Comprehensive performance evaluation of the SVR model for all stations on CH_4 concentration prediction. (a) Cross-validation performance metrics showing R^2, RMSE, and MAE with error bars representing standard deviation across 5 folds. (b) R^2 scores for individual CV folds with mean performance indicated by the dashed line. (c) Density scatter plot comparing predicted versus actual CH_4 concentrations, with darker colours indicating higher data density; the dashed line represents perfect prediction, and the solid line shows the linear fit. (d) Full-year time series comparison of actual (blue) and predicted (red) CH_4 concentrations with 95% confidence intervals within the shaded area.

ST-CAN in modelling complex dynamics beyond simple event detection. The case study, presented in Fig. 15, compares the performance of the full ST-CAN (green) against the Ground-Only model (red) and the actual ground truth (black) during two challenging periods.

Case 1. Early November 2024. As shown in Fig. 15a, the Ground-Only model significantly over-predicted a peak, spiking to ~2.34 ppm while the actual value was ~2.27 ppm. The full ST-CAN, however, correctly suppressed this temporal spike, holding its prediction near the ground truth.

Case 2. Early December 2024. Conversely, in Fig. 15c, the Ground-Only model under-predicted a peak, rising only to ~2.30 ppm while the actual event peaked at ~2.34 ppm. ST-CAN correctly boosted this prediction, accurately matching the true peak.

The explanation for ST-CAN's success in both cases lies in the satellite data. The GHGSat heatmaps for both the November (Fig. 15b) and December (Fig. 15d) periods show high regional X_{CH4} concentrations. ST-CAN's fusion mechanism learned to use this consistent spatial context as a critical "corrector" for the volatile, time-only predictions. This is particularly evident in Case 2, which highlights the specific contribution of satellite data in capturing spatial diffusion events. The satellite heatmap revealed a widespread high-concentration field (purple zones), characteristic of large-scale spatial diffusion rather than a localized point source. The Ground-Only model, lacking this spatial context, treated the signal as localized noise and failed to capture the full magnitude of the event. By integrating this spatial diffusion signal, the ST-CAN framework correctly identified the regional scale of the transport. This analysis provides definitive qualitative proof that the satellite data is not a minor contributor; it is an active, intelligent component that makes the model more robust by capturing regional transport dynamics invisible to the ground-only stream. This behaviour corresponds to the model semantically identifying these periods as "State 9: Persistent Transport Event" (Winter High-Transport + Stable) in Table 1, validating the physical interpretability of our framework.

To rigorously quantify the contribution of satellite data across these contrasting scenarios, Table 2 presents a direct performance comparison between the Ground-Only and full ST-CAN models during these critical peak moments. The results demonstrate that incorporating satellite spatial context yields substantial quantifiable gains. In the November localized scenario, ST-CAN achieved an 86% error reduction, effectively corrected the Ground-Only model's false overshoot (2.34 ppm) and brought the prediction within 0.01 ppm of the ground truth. Even more critical is the result for the December spatial diffusion event. In this scenario, the Ground-Only model failed to capture the full magnitude of the event due to a lack of spatial awareness (under-predicting at 2.30 ppm). In contrast, the ST-CAN model, guided by the high-concentration regional signal from the satellite, achieved a near-complete error reduction (>95%), closely tracking the observed peak at 2.34 ppm with negligible residual error. This result confirms that in the presence of large-scale spatial diffusion, satellite data acts as an indispensable spatial regularizer, providing the necessary context to capture regional transport dynamics that are mathematically inaccessible to temporal-only models.

3.4. *Interpretability of semantic clustering-guided attention*

A foundational innovation of the ST-CAN framework is its enhanced interpretability, addressing the "black-box" nature common in many

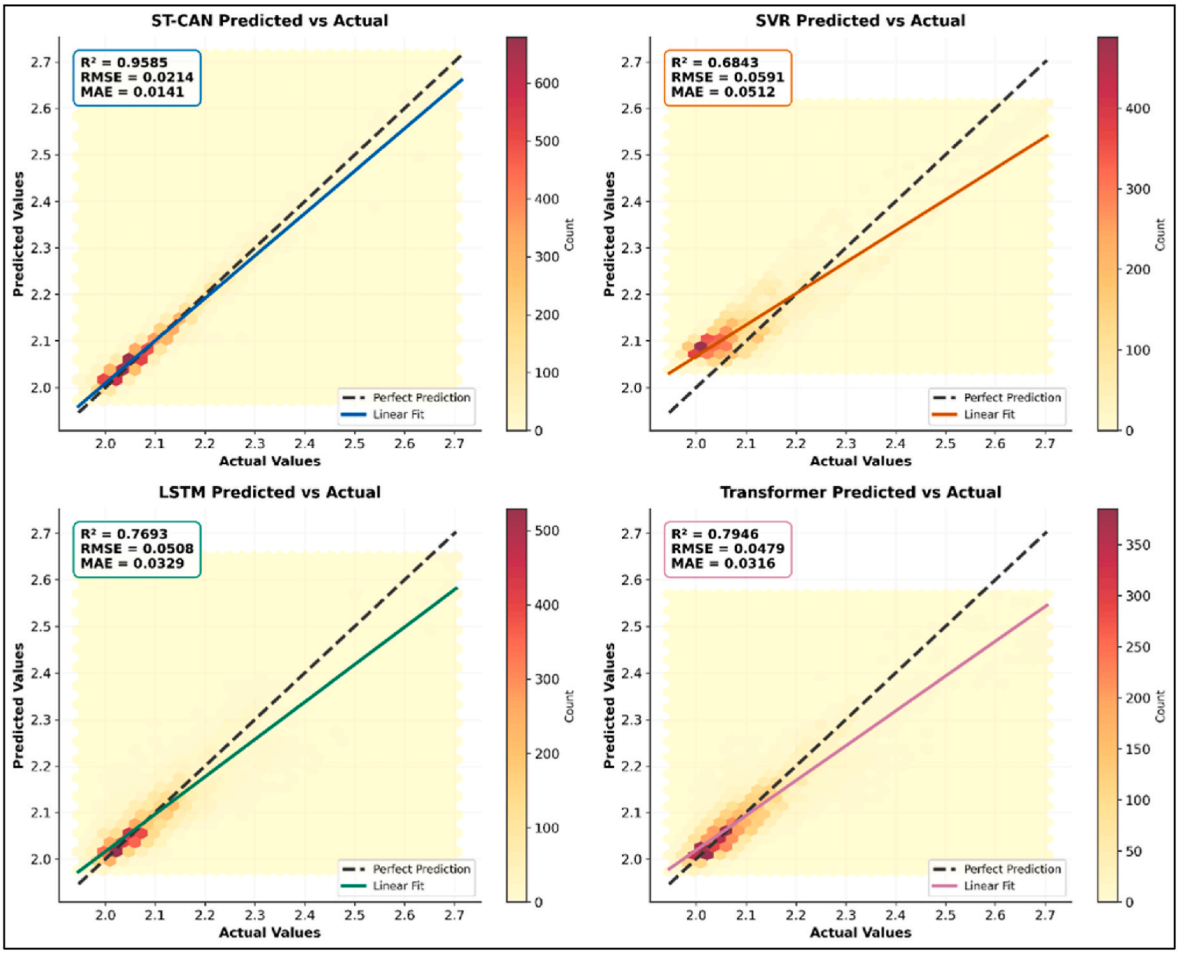

Fig. 12. Comparison of linear fit between baseline models and ST-CAN.

deep learning models. This interpretability is primarily rooted in a model-inherent explanation, rather than a post-hoc analysis. The environmental semantic labels, derived from the interpretable A5 and D2 wavelet components, are not merely used for analysis after prediction; they serve as direct, real-time control signals to the dynamic fusion gate. This design fundamentally integrates environmental understanding into the model's adaptive information processing, allowing the semantic context to actively steer how information from ground and satellite sources is weighted and combined.

This model-based interpretability is particularly valuable for domain experts. Unlike the traditional machine learning features, our semantic labels, for example, “Summer High-Emission “or “Pollution Increase," are clearly linked to physically meaningful environmental conditions, making them directly comprehensible and trustworthy to atmospheric scientists and environmental managers. The mechanism allows domain experts to trace the model's adaptive strategies, for instance, observing how the model “prioritizes" satellite spatial context during a regional transport event, or ground temporal dynamics during a localized emission spike. The subsequent visualization and analysis of attention weights, conditioned on these familiar environmental states, provides a clear and intuitive manifestation of this enhanced interpretability. This approach transcends simple input-output correlations, offering insights into why the model makes certain fusion decisions under specific environmental contexts.

To provide insight into this adaptive fusion process, Fig. 16 visualizes the learned attention importance within the bidirectional cross-attention mechanism, specifically showing how different combinations of long-term (A5) and short-term (D2) environmental states influence the information flow between the two data modalities.

The heatmaps in Fig. 16 reveal that the model learns a sophisticated trust strategy, dynamically deciding when to trust or suppress information from each modality based on the environmental context. The left panel, illustrating Ground-to-Satellite (G-S) attention, shows the trust placed on satellite data when queried by ground data. A high positive weight (0.41) emerges during the “High Pollution + Stable" state. This suggests that when the model receives confirmation of a stable, high-pollution baseline from ground sensors, it learns to “High Trust" and heavily query the satellite's spatial context. The model's logic is similar to a learned policy: when ground sensors confirm a severe pollution event, the model adaptively increases the importance of the satellite's spatial view to confirm and locate that event. Conversely, a strong negative weight (−0.42) appears during the “Low Pollution/Decreasing" state (A5-Low, D2-Decreasing). This demonstrates that the model learns to actively distrust or suppress the satellite data. The physical interpretation is that during these low-concentration periods, the satellite's column-averaged data (X_{CH4}) is likely dominated by background atmospheric concentrations and is not sensitive to the tiny local variations measured at the ground level. In this state, the satellite data is likely considered irrelevant or as potential noise, and its influence is therefore actively minimized.

The right panel, showing Satellite-to-Ground (S-G) attention, reveals a different dynamic where satellite data queries the ground time series. The highest weight (0.59) occurs during the “Medium Pollution + Increasing" state. Here, the model learns to “High Trust" the satellite data to act as the Query, guiding its search for specific information within the ground station's temporal data. This demonstrates a learned strategy where the model leverages the satellite's broad spatial context to guide its attention. This guidance allows the model to precisely

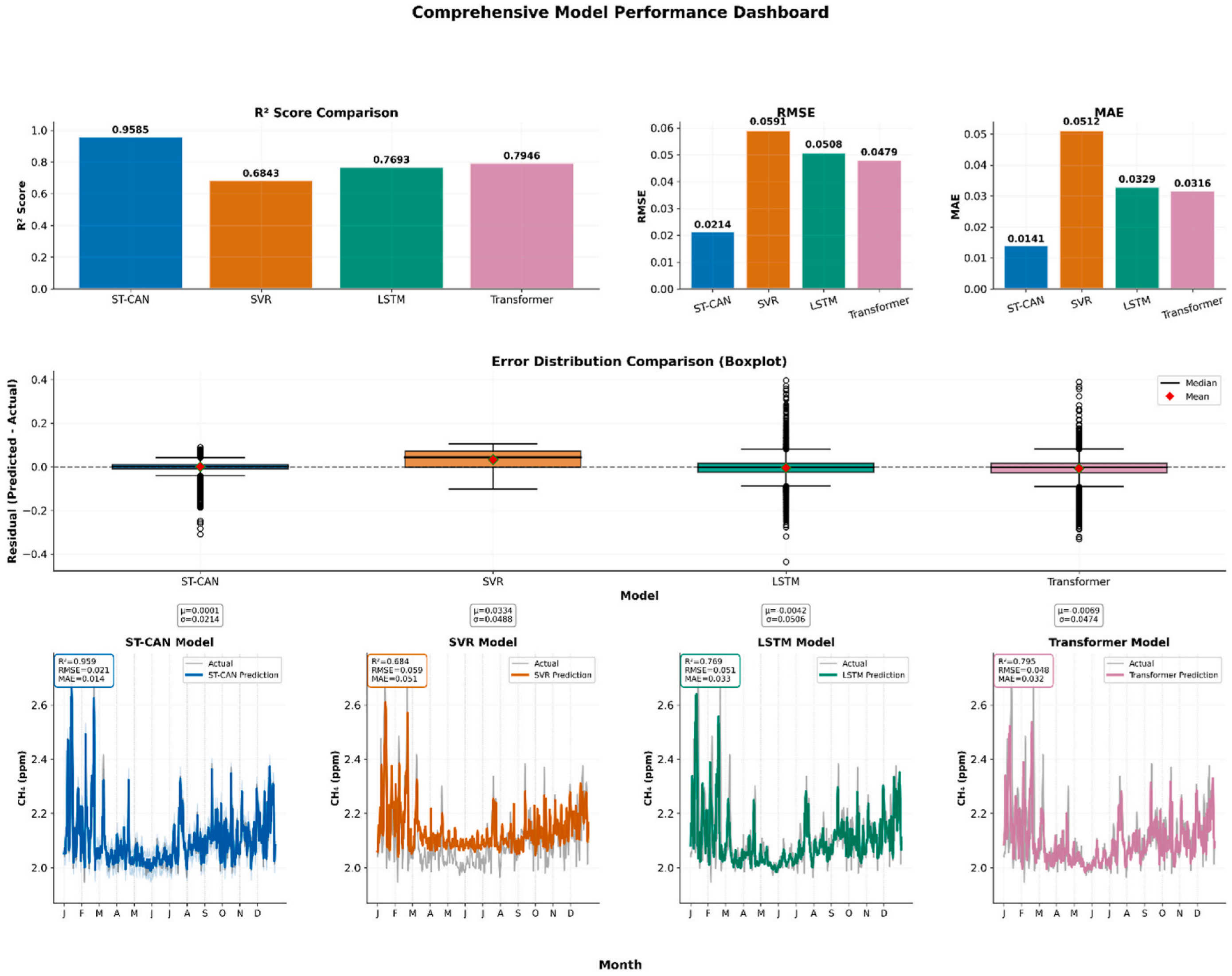

Fig. 13. Comprehensive Performance Dashboard for ST-CAN and Baseline Models in CH_4 Concentration Prediction. The dashboard presents a multi-faceted comparison of model performance: (Top row) Bar charts comparing R^2, RMSE, and MAE metrics across ST-CAN, SVR, LSTM, and Transformer models. (Middle row) Boxplot visualization displaying error distribution characteristics for each model, including median (horizontal line), mean (red diamond), quartile ranges (box), and whiskers; annotated with mean (μ) and standard deviation (σ) values below each distribution. (Bottom row) Individual time series subplots for each model showing predicted CH_4 concentrations (colored lines) against actual observations (gray reference lines) across a full year, with unified y-axis scales for direct comparison; performance metrics (R^2, RMSE, MAE) are embedded in each subplot.

identify and extract the most relevant temporal details of an emerging pollution event from the high-frequency ground-station data.

Together, these visualizations demonstrate the adaptive nature of ST-CAN's fusion mechanism. Unlike traditional “black-box" attention models, where the drivers of attention weights are unclear, this semantic clustering-guided approach offers a degree of interpretability by linking attention patterns to recognizable environmental conditions. This not only contributes to understanding the model's decision-making process but also highlights a key factor in its enhanced predictive accuracy and resilience to real-world data imperfections.

3.5. Ablation study on ST-CAN components

To validate the independent contribution of each novel component in the ST-CAN, we conducted a comprehensive ablation study. We trained four ablated versions of the model: (1) Ablation No Fusion Gate (using simple concatenation); (2) Ablation No Gate Guidance (a “blind" gate without cluster labels), and two unidirectional variants (S-G Only and G-S Only). Both ablation models are trained and validated on the same dataset and compared their performance (R^2, RMSE) against the full ST-CAN. The results are presented in Table 3.

The results in Table 3 confirm that every component of ST-CAN is necessary for its high performance, as the removal of any part leads to a measurable drop in accuracy. The analysis reveals three key findings. First, bidirectional attention is critical. Both unidirectional models performed poorly. The most significant finding is the dramatic performance drop with +63.9% increase in RMSE when the S-G attention stream is removed. This quantitatively proves that using the satellite's spatial context to query the high-frequency temporal data is the single most important mechanism in the framework. Secondly, the semantic guidance is a key performance driver. The No Gate Guidance model with RMSE = 0.02652 displayed a +14.0% increase in RMSE. This is a substantially larger error increase than that of the No Fusion Gate model with +4.1%, proving that the interpretable cluster labels are a major contributor to the model's predictive accuracy. Finally, the dynamic gating outperforms concatenation. The No Fusion Gate model, which use simple concatenation features as input, performed worse than the full model, confirming that the learnable, adaptive weighting of the

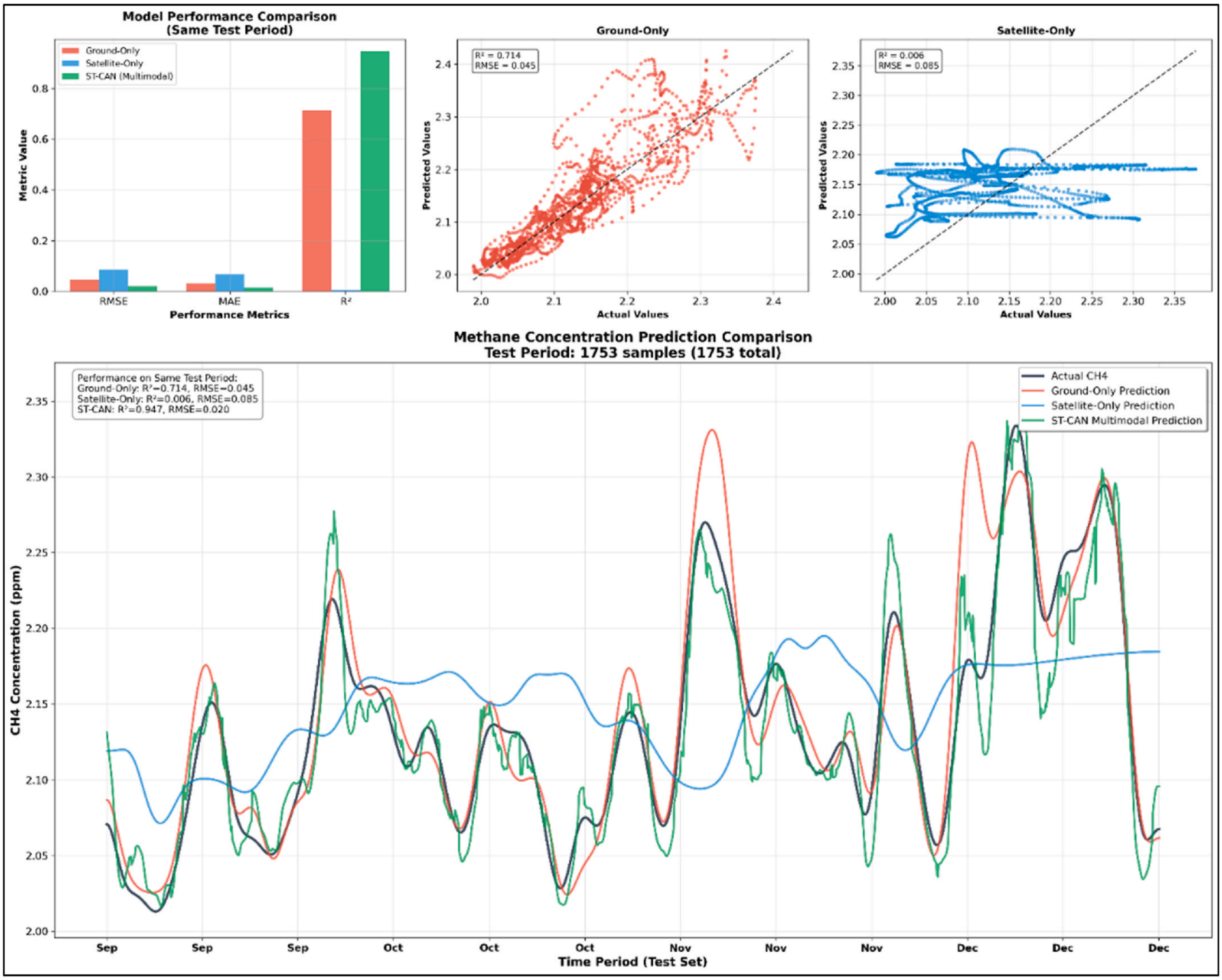

Fig. 14. Comprehensive Performance Comparison of ST-CAN against Single-Modality and Multimodal Baseline Models. The figure presents a detailed evaluation of all models after correcting a data processing artifact in the visualization code. (a) Bar charts comparing three key statistical metrics: R^2, RMSE, and MAE. (b) The time-series plot from September to December compares the hourly predicted CH_4 concentrations from all models against the actual observed values (black line). The models evaluated include the proposed ST-CAN (green line) and single-modality baselines, Ground-Only in red, Satellite-Only in blue. Both the quantitative metrics and the qualitative time-series visualization consistently demonstrate the superior accuracy and stability of the ST-CAN.

dynamic gate is a superior fusion strategy.

4. Conclusions

The ST-CAN demonstrates significant advancement in methane concentration prediction through its novel environmental semantic clustering-guided approach and bidirectional cross-attention mechanism. The superior performance over established baselines validates the effectiveness of integrating semantic understanding of environmental states with adaptive fusion strategies. The framework's ability to identify distinct long-term baseline patterns and short-term dynamic states enables more accurate predictions of the regionally representative concentration, while simultaneously enhancing interpretability. This approach addresses persistent challenges in multi-source environmental data fusion, particularly the heterogeneity and temporal misalignment of diverse data streams. Unlike traditional data-driven deep learning models that operate attention mechanisms as “black boxes," the proposed architecture achieves not only superior performance but also provides meaningful interpretability through environmental semantic understanding.

As noted in our methodology, the temporal up-sampling of weekly satellite data to an hourly resolution remains a key limitation of this work. While our ST-CAN demonstrates a strong capability to fuse the available information, the fidelity of its predictions at fine temporal scales is inherently constrained by the low frequency of the satellite observations. This underscores a critical need within the environmental monitoring field for next-generation satellite missions with higher revisit rates, which would fully unlock the potential of the data fusion framework proposed here. Furthermore, we must note that our performance evaluation metrics of R^2, RMSE, and MAE are necessarily validated against the in-situ measurements from the ground stations, as these provide the only available ground truth. As this study serves primarily as pilot research to develop and test the advantages of multimodal fusion modelling, we have focused on demonstrating the methodology's effectiveness using available ground truth data. In future research, we will conduct comprehensive validation across all available ground stations in the study area to more thoroughly assess the model's spatial predictive performance. Evaluating the model's accuracy at

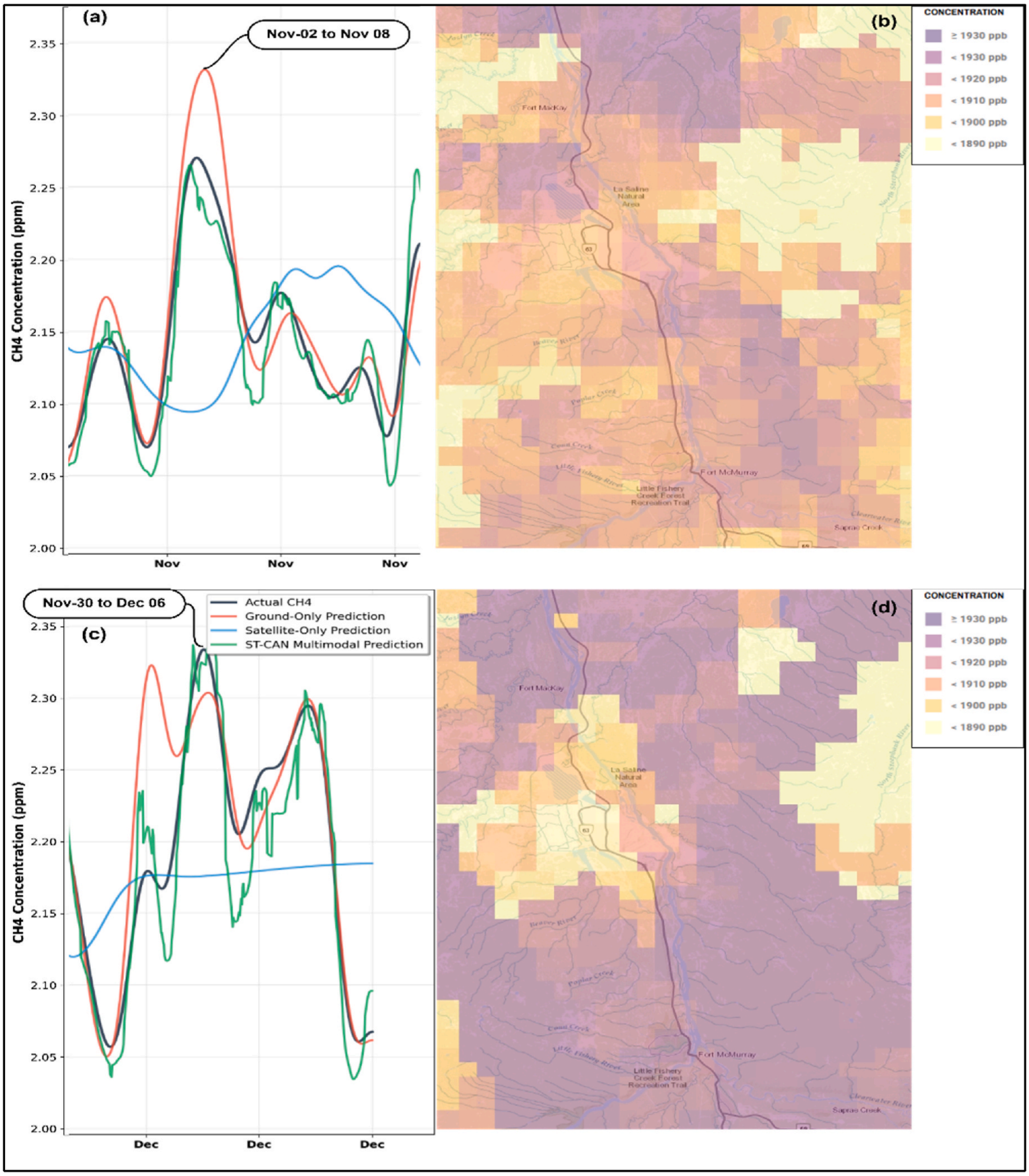

Fig. 15. Case Study on Satellite Data Correction of Temporal Prediction Errors. (a) In early November, the Ground-Only model (red) overshoots the actual peak (black), while ST-CAN (green) correctly suppresses it. (b) The corresponding GHGSat map (Nov 2–8) shows high regional X_{CH4}. (c) In early December, the Ground-Only model undershoots the actual peak, while ST-CAN correctly boosts it. (d) The corresponding GHGSat map (Nov 30-Dec 6) also shows high regional X_{CH4}. This demonstrates how ST-CAN uses spatial context to correct errors in the temporal-only prediction.

Table 2
Quantitative comparison of satellite data contribution in correction of peak estimation errors.

Event Scenario	Ground Truth	Ground-Only Model		ST-CAN		Satellite Contribution
		Prediction	Abs. Error	Prediction	Abs. Error	**Error Reduction (%)**
Nov: Localized (False Peak)	2.27	2.34	0.07	**2.26**	0.01	85.70%
Dec: Spatial Diffusion	2.34	2.3	0.04	**~2.34**	<0.01	**> 95.0%**

unobserved, non-ground station locations remains a significant challenge that requires a dedicated spatial validation dataset, which we plan to address in subsequent work. Additionally, we must acknowledge the inherent physical heterogeneity of our input modalities. Our model fuses surface-level in-situ measurements (ppm) with satellite-derived column-averaged dry-air mole fractions (X_{CH4}). While X_{CH4} provides invaluable

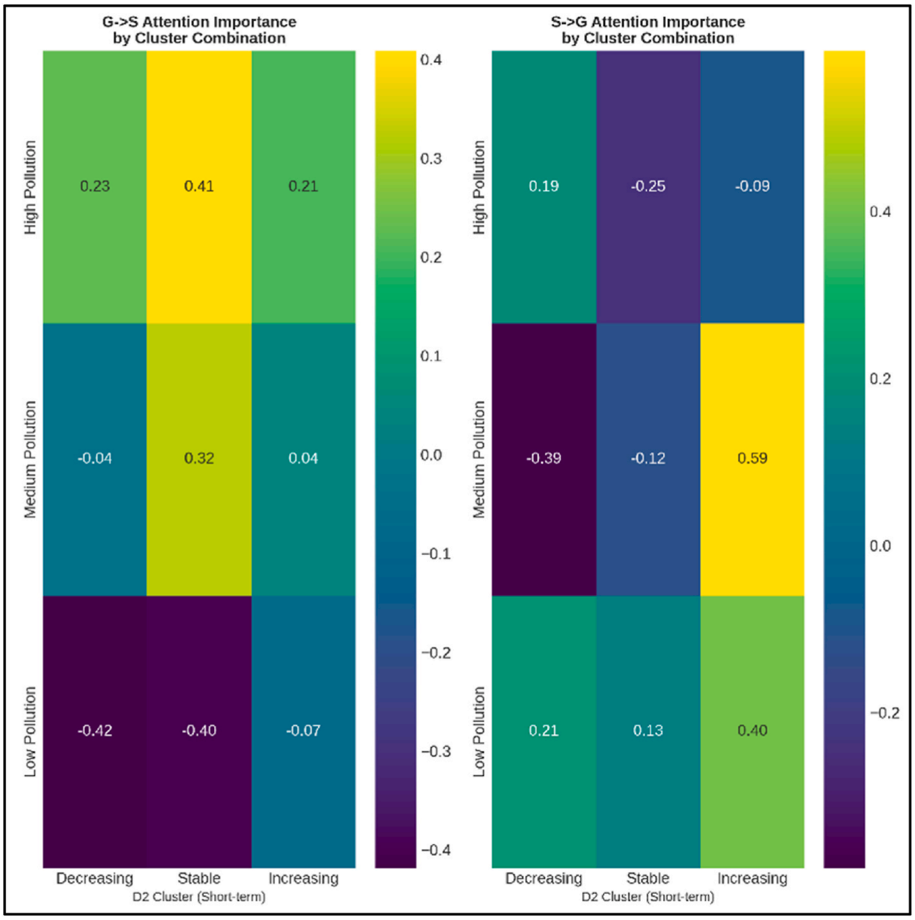

Fig. 16. Visualization of Average Learned Attention Importance in ST-CAN. Heatmaps illustrate the aggregated attention weights for G-S and S-G interactions, conditioned on the nine environmental states. The value in each cell represents the average attention importance across all time steps belonging to that specific environmental state, demonstrating the model's consistent and systematic learned fusion strategy rather than an instantaneous snapshot.

Table 3
Ablation study results comparing the full ST-CAN model to Abridged variants.

Model Variant	Description	R^2	RMSE	% Increase in RMSE (vs. Full)
ST-CAN (Full Model)	The completely proposed model.	**0.9513**	**0.02325**	
Ablation: No Fusion Gate	Replaces the Dynamic Gate with simple concatenation.	0.9468	0.02420	+4.1%
Ablation: No Gate Guidance	The Dynamic Gate runs “blind" without cluster label inputs.	0.9342	0.02652	+14.0%
Ablation: S-> G Only	Unidirectional. Removes the G- > S attention stream.	0.9227	0.02896	+24.5%
Ablation: G-> S Only	Unidirectional. Removes the S- > G attention stream.	0.8675	0.03811	+63.9%

spatial context, it does not directly measure surface concentrations, and the relationship between the two can be complex and non-linear, influenced by boundary layer height, atmospheric mixing, and local emission sources. Our model's task is to learn this complex relationship. Consequently, all predictions from ST-CAN must be strictly interpreted as estimations of the ground-level concentration, as this is the target against which the model was trained and validated. Finally, a further limitation, related to the spatial sparsity of the ground network, is the scope of our model validation. Our validation was performed temporally, assessing the model's ability to predict the spatially averaged concentration over time. We did not perform spatial cross-validation. However, ST-CAN was designed for a different objective: to enhance the temporal fidelity of predictions at known, monitored locations by fusing multi-modal data, rather than to generalize to entirely unobserved coordinates. Given the sparse (n = 4) and non-uniformly distributed training network, applying spatial cross-validation would be a significantly different and challenging research task.

Looking forward, enhancing the model's spatial generalization capability to unobserved locations is the primary direction for our future work. To address the inherent sparsity of ground monitoring networks, we are currently developing a next-generation framework that moves beyond pure data-driven approach. Our strategy focuses on two synergistic advancements. First, we aim to integrate Graph Neural Networks (GNNs) to model the non-Euclidean spatial dependencies among monitoring sites. Unlike standard interpolation, GNNs can learn adaptive edge weights based on wind vectors and industrial source proximity, capturing the complex topology of pollution transport. Second, we plan to embed physical laws directly into the learning process using Physics-Informed Neural Networks (PINNs). By incorporating the atmospheric Advection-Diffusion partial differential equations (PDEs) as a regularization term in the loss function, we can constrain the model's predictions to be physically consistent. This approach aims to simulate realistic methane plume dispersion and chemical decay in unmonitored areas, effectively bridging the gap between sparse ground observations and continuous spatial fields without requiring a prohibitively dense sensor network.

Applications extend beyond the AVOS, offering potential for standardized methane monitoring protocols across diverse industrial contexts. Beyond the spatial expansions mentioned above, other future directions include exploring federated learning approaches for privacy-preserving collaborative modelling and further improving the interpretability of attention mechanisms. The ST-CAN represents a significant step toward more robust, transparent, and actionable methane monitoring systems essential for effective climate mitigation strategies.

CRediT authorship contribution statement

Yang Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Hao Wang:** Writing – review & editing, Supervision, Funding acquisition. **Jude D. Kong:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through an NSERC Alliance Missions grant (577242) “Predicting Methane Emissions from Alberta Oil Sands Territories Using a Holistic Model and Monitoring System" (HW as the PI and JDK as a co-PI). HW acknowledges support from the Natural Sciences and Engineering Research Council of Canada (Individual Discovery Grant RGPIN-2020-03911 and Discovery Accelerator Supplement Award RGPAS-2020-00090) and the Canada Research Chairs program (Tier 1 Canada Research Chair Award). JDK acknowledges support from Canada' s International Development Research Centre (IDRC) (Grant No. 109981-001), IDRC and the Foreign, Commonwealth and Development Office (FCDO) (Grant No. 110554-001), NSERC Discovery Grant (Grant No. RGPIN-2022-04559), NSERC Discovery Launch Supplement (Grant No: DGECR-2022-00454), New Frontier in Research Fund- Exploratory (Grant No. NFRFE-2021-00879) and Canada Research Chair in Community-Oriented Artificial Intelligence and Mathematical Modelling of Infectious Diseases (Award ID: CRC-2023-00234.). The authors also acknowledge GHGSat for making their weekly methane concentration data publicly available through the SPECTRA Basic platform, which was used in this study. No specific project collaboration or exclusive data agreement was required for the use of this public dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2026.128845.

Data availability

Data will be made available on request.

References

Alfaouri, M., Daqrouq, K., 2008. ECG signal denoising by wavelet transform thresholding. Am. J. Appl. Sci. 5 (3), 276–281. https://doi.org/10.3844/ajassp.2008.276.281.

BOE Report, 2025. Energy projects occupy less than three per cent of Alberta's oil sands region, report says. https://boereport.com/2025/05/05/energy-projects-occupy-less-than-three-per-cent-of-albertas-oil-sands-region-report-says/. (Accessed 16 May 2025).

Chen, Z., Ma, M., Li, T., Wang, H., Li, C., 2023. Long sequence time-series forecasting with deep learning: a survey. Inf. Fusion 97, 101819. https://doi.org/10.1016/j.inffus.2023.101819.

Etminan, M., Myhre, G., Highwood, E.J., Shine, K.P., 2016. Radiative forcing of carbon dioxide, methane, and nitrous oxide: a significant revision of the methane radiative forcing. Geophys. Res. Lett. 43 (24), 12–614. https://doi.org/10.1002/2016GL071930.

Fan, L., Wan, Y., Dai, Y., 2024. Development of a multi-source satellite fusion method for X_{CH4} product generation in oil and gas production areas. Applied Sciences 14 (23), 11100. https://doi.org/10.3390/app142311100.

Feng, T., Yang, S., Han, F., 2019. Chaotic time series prediction using wavelet transform and multi-model hybrid method. Journal of Vibroengineering 21 (7), 1983–1999. https://doi.org/10.21595/jve.2019.20579.

GHGSat, 2024a. About GHGSat. https://www.ghgsat.com/en/who-we-are/. (Accessed 16 May 2025).

GHGSat, 2024b. Spectra BASIC. https://spectra-basic.ghgsat.com/. (Accessed 16 May 2025).

Government of Alberta, 2025. Reducing methane emissions. https://www.alberta.ca/climate-methane-emissions. (Accessed 16 May 2025).

Hameed, S., Islam, A., Ahmad, K., Belhaouari, S.B., Qadir, J., Al-Fuqaha, A., 2023. Deep learning based multimodal urban air quality prediction and traffic analytics. Sci. Rep. 13 (1), 22181. https://doi.org/10.1038/s41598-023-49296-7.

IEA, 2024. Global Methane Tracker 2024. IEA, Paris. https://www.iea.org/reports/global-methane-tracker-2024. Licence: CC BY 4.0.

Jacob, D.J., Turner, A.J., Maasakkers, J.D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., Frankenberg, C., 2016. Satellite observations of atmospheric methane and their value for quantifying methane emissions. Atmos. Chem. Phys. 16 (22), 14371–14396. https://doi.org/10.5194/acp-16-14371-2016.

Kong, J.D., Wang, H., Siddique, T., Foght, J., Semple, K., Burkus, Z., Lewis, M.A., 2019. Second-generation stoichiometric mathematical model to predict methane emissions from oil sands tailings. Sci. Total Environ. 694, 133645. https://doi.org/10.1016/j.scitotenv.2019.133645.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep learning in multimodal remote sensing data fusion: a comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926. https://doi.org/10.1016/j.jag.2022.102926.

Li, J., Pei, Y., Zhao, S., Xiao, R., Sang, X., Zhang, C., 2020. A review of remote sensing for environmental monitoring in China. Remote Sens. 12 (7), 1130. https://doi.org/10.3390/rs12071130.

Lian, Z., Zhan, Y., Zhang, W., Wang, Z., Liu, W., Huang, X., 2025. Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images. Sensors 25 (4), 1093. https://doi.org/10.3390/s25041093.

Lilhore, U.K., Simaiya, S., Singh, R.K., Baqasah, A.M., Alroobaea, R., Alsafyani, M., Alhazmi, A., Khan, M.M., 2025. Advanced air quality prediction using multimodal data and dynamic modeling techniques. Sci. Rep. 15 (1), 27867. https://doi.org/10.1038/s41598-025-11039-1.

MethaneSAT, 2024. New data reveal previously undetectable methane emissions. https://www.methanesat.org/project-updates/new-data-reveal-previously-undetectable-methane-emissions. (Accessed 16 May 2025).

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C., 2021. Attention bottlenecks for multimodal fusion. Adv. Neural Inf. Process. Syst. 34, 14200–14213. https://dl.acm.org/doi/10.5555/3540261.3541349.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis, vol. 4. Cambridge university press. https://doi.org/10.1017/CBO9780511841040.

Quamar, M.M., Al-Ramadan, B., Khan, K., Shafiullah, M., El Ferik, S., 2023. Advancements and applications of drone-integrated geographic information system technology—A review. Remote Sens. 15 (20), 5039. https://doi.org/10.3390/rs15205039.

Rouet-Leduc, B., Hulbert, C., 2024. Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer. Nat. Commun. 15 (1), 3801. https://doi.org/10.1038/s41467-024-47754-y.

Sahoo, G.R., Freed, J.H., Srivastava, M., 2024. Optimal wavelet selection for signal denoising. IEEE Access 12, 45369–45380. https://doi.org/10.1109/ACCESS.2024.3377664.

Sysoeva, L., Bouderbala, I., Kent, M.H., Saha, E., Zambrano-Luna, B.A., Milne, R., Wang, H., 2025. Decoding methane concentration in Alberta oil sands: a machine learning exploration. Ecol. Indic. 170, 112835. https://doi.org/10.1016/j.ecolind.2024.112835.

Veefkind, J.P., Aben, I., McMullan, K., Förster, H., De Vries, J., Otter, G., et al., 2012. TROPOMI on the ESA Sentinel-5 precursor: a GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. Remote sensing of environment 120, 70–83. https://doi.org/10.1016/j.rse.2011.09.027.

Wang, X., Liu, J., Lin, H., Garg, S., Alrashoud, M., 2024. A multi-modal spatial–temporal model for accurate motion forecasting with visual fusion. Inf. Fusion 102, 102046. https://doi.org/10.1016/j.inffus.2023.102046.

Wen, Y., Chen, P., Zhang, Z., Li, Y., 2024. Cross-attention-based high spatial-temporal resolution fusion of Sentinel-2 and Sentinel-3 data for ocean water quality assessment. Remote Sens. 16 (24), 4781. https://doi.org/10.3390/rs16244781.

Xu, Y., Yazdinejad, A., Wang, H., Kong, J.D., 2025. From detection to decision: a systematic literature review of AI and machine learning evolution in methane modelling. https://dx.doi.org/10.2139/ssrn.5218753.

Zhou, D., Wu, K., Xu, G., 2025. A bidirectional cross spatiotemporal fusion network with spectral restoration for remote sensing imagery. Applied Sciences 15 (12), 6649. https://doi.org/10.3390/app15126649.