Hybrid Water Quality Prediction With Multimodal Low-Rank Fusion and Localized Attention

Jing Bi, *Senior Member, IEEE*, Yibo Li, *Graduate Student Member, IEEE*, Haitao Yuan, *Senior Member, IEEE*, Mengyuan Wang, Ziqi Wang, *Graduate Student Member, IEEE*, Jia Zhang, *Senior Member, IEEE*, and Mengchu Zhou, *Fellow, IEEE*

***Abstract*—Water quality prediction methods forecast the short- or long-term trends of its changes, providing proactive advice for preventing and controlling water pollution. Existing water quality prediction methods typically fail to capture water quality's nonlinear characteristics accurately and only consider historical time series data. However, meteorology and other factors also significantly impact water quality indicators. Therefore, considering only historical data of water quality time series is not feasible. To solve this problem, this work proposes a hybrid water quality prediction model called CMLIP, which integrates convNeXt V2, multimodal bottleneck transformer, low-rank multimodal fusion, iTransformer, and PatchTST. CMLIP inputs water quality time series and meteorological remotely sensed rainfall images into a multimodal fusion module before prediction. Specifically, CMLIP integrates the model of ConvNeXt V2 to extract image features. Its multimodal fusion module combines a multimodal bottleneck transformer and the low-rank multimodal fusion to fuse the time series and images. Furthermore, CMLIP combines iTransformer and PatchTST to form an improved prediction module that realizes the prediction of fused features. Experimental results with real-life water quality time series and remotely sensed rainfall images demonstrate that CMLIP when fusing meteorological data, achieves an average improvement of 17% in water quality forecasting accuracy compared to forecasts using only water quality time series. Moreover, CMLIP outperforms other state-of-the-art algorithms in both data fusion and prediction, with an average enhancement of 6% in fusion effectiveness and an average improvement of 22% in prediction accuracy.**

***Index Terms*—iTransformer, low-rank fusion, multimodal bottleneck transformer (MBT), multimodal fusion, PatchTST, time series prediction, water quality.**

Received 14 January 2025; revised 9 February 2025; accepted 23 February 2025. Date of publication 4 March 2025; date of current version 9 June 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473014 and Grant 62173013; in part by the Beijing Natural Science Foundation under Grant L233005 and Grant 4232049; and in part by the Beihang World TOP University Cooperation Program. *(Corresponding author: Haitao Yuan.)*

Jing Bi, Yibo Li, and Ziqi Wang are with the College of Computer Science, and the Beijing Laboratory of Smart Environmental Protection, Beijing University of Technology, Beijing 100124, China (e-mail: bijing@bjut.edu.cn).

Haitao Yuan is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail:yuan@buaa.edu.cn).

Mengyuan Wang is with the School of Energy and Power Engineering, Beihang University, Beijing 100191, China (e-mail: mengyuanwang@buaa.edu.cn).

Jia Zhang is with the Department of Computer Science, Southern Methodist University, Dallas, TX 75206 USA (e-mail: jiazhang@smu.edu).

Mengchu Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: zhou@njit.edu).

Digital Object Identifier 10.1109/JIOT.2025.3547357

I. Introduction

WATER is the source of life, a precious resource on which all civilizations depend, and at the heart of many of today's social issues. With the progress of civilization in human society and the enhancement of public awareness of environmental protection, scientific usage and systematic protection of water resources have become an inevitable choice for the sustainable development of all countries in the world. Water quality prediction methods can obtain future short- or long-term water quality change trends. Thus, it can guide water pollution prevention and provide technical support for water environmental control. Water quality prediction [1] is essentially a time series prediction problem, which refers to predicting changes in water quality indicators in the future based on their values at historical time points. Current studies on water quality prediction can be divided into mechanistic and data-driven models. Mechanistic models require many parameters to be preset in advance and the training process is complex, requiring large computational resources and a long time. Data-driven models can be divided into statistical, machine learning, and deep learning methods. Autoregression [2] in statistical methods is one of the most typical and basic time series models, and autoregressive integrated moving average [3] is one of the most famous and widely used prediction methods. Meanwhile, higher-order supervised models in machine learning are available for time-series prediction. For example, extreme gradient boosting [4] can efficiently process complex data by gradient boosting, support vector machine [5], DeepForest [6] and other models can also realize the prediction. Machine learning is built upon statistical learning, and deep learning is a subfield of machine learning. They have achieved good results in time series prediction in recent years. Compared with machine learning, which requires complex feature engineering, deep learning can automatically learn patterns and trends in the time series data. Moreover, a neural network involves important parameters such as the number of hidden layers and that of neurons. For example, Transformer [7] treats time steps of the input sequence as positional information, designs the features of each time step as a vector, and adopts the encoder-decoder framework for prediction. FEDformer [8] introduces a local attention mechanism and a reversible one to convert the time domain into the frequency domain, and it better captures local features in the time series data and has higher computational efficiency.

However, many other factors affect water quality indicators in the water environment, e.g., meteorology, pollutants, and others. Thus, only considering the historical data on water quality is not sufficient to make an accurate prediction [9]. Other multimodal data, such as remotely sensed meteorological data, need to be jointly considered [10], moreover, the fusion of data information from different modalities is imperative. Multimodal fusion is a hot research direction in artificial intelligence, which is committed to utilizing different types of input data to complement various information. It obtains a more comprehensive semantic expression and improves the depth of the model's understanding of the target task, thereby enhancing the understanding of the complex scene and decision-making ability. Usually, multimodal fusion methods are divided into four types: 1) early fusion; 2) late fusion; 3) hybrid fusion; and 4) model-level fusion [11]. Early fusion integrates high-dimensional features immediately after feature extraction. Late fusion performs integration only after the output results of each modality. Hybrid fusion combines the advantages of the former types to decrease the model's structural complexity and the difficulty of training. Model-level fusion mainly depends on the fusion model, which is dedicated to learning the joint feature representation of different modalities.

To improve the accuracy of water quality prediction with the meteorological data, a hybrid prediction model combining the ConvNeXt V2 [12], multimodal bottleneck transformer (MBT) [13], low-rank multimodal fusion (LMF) [14], ITransformer [15], and PatchTST [16], called CMLIP for short, is proposed. The main contributions of this work are summarized as follows.

1) Considering the influence of meteorological factors on water quality indicators, the multimodal fusion of water quality time series and remotely sensed rainfall images is innovatively proposed. The multimodal fusion-based prediction model, CMLIP, is proposed with both time series and images as the input.
2) CMLIP integrates ConvNeXt V2, MBT, LMF, iTransformer, and PatchTST to extract remotely sensed rainfall image features, realize multimodal fusion of time series and images, and predict future information with the fused information, respectively.
3) Experimental results with real-world water quality and remotely sensed rainfall images demonstrate that CMLIP outperforms other models in prediction and fusion. The prediction accuracy of fusing time series and rainfall images is 17% higher than that of unfused ones on average.

The remaining sections of this article are organized as follows. Section II discusses the related work in recent years and summarizes the contributions. Section III introduces the model structures in detail. Section IV conducts various experiments with real water quality data to verify the model's performance. Section V summarizes this article and gives the future direction.

II. Related Work

A. Time Series Forecasting

Water quality prediction is essentially a time series prediction problem. The algorithms and deep learning models are powerful tools for solving complex and variable time series prediction problems [17]. The time series prediction model based on deep learning can be roughly summarized into three categories. *First*, recurrent neural networks (RNNs) [18], [19], e.g., long short-term memory models [20], [21], are good at dealing with nonlinear sequences and effectively capture temporal dependencies in the time series. However, the computational complexity is high and time-consuming when facing long time series with large periods. *Second*, convolutional neural networks (CNNs) [22] transform the time series data into a 2-D matrix and automatically extract its features through operations such as convolutional and pooling to realize the prediction. For example, temporal convolutional networks [23] solve problems of gradient vanishing and high computational complexity of traditional RNNs when dealing with long series. However, they are more demanding on the training data, and if the dataset is small or uneven, it can lead to insufficient generalization ability or overfitting. *Third*, Transformer and its variants, which are currently a hot research topic in the field of time series prediction, utilize the attention mechanism to weight various parts of the input data adaptively, thus making the model pay more attention to the key information while reducing the influence of irrelevant information and fully capturing the potential connections among time nodes. For example, Autoformer [24] uses a decomposition architecture with an autocorrelation mechanism to discover and represent dependencies at the subseries level for long-term forecasting of time series data. The rapid iterative development of time series prediction models has also facilitated the water quality prediction research process. Qiao et al. [25] adopted the attention mechanism and spatial-temporal map convolution to extract nonlinear features of water quality and spatial dependence of the river network, respectively, to improve the accuracy of long-term series prediction.

However, the above models only take historical time series data as input and lack the capability for data fusion, which limits their ability to incorporate additional information sources to enhance prediction accuracy. To address this limitation, this work innovatively combines multimodal fusion with time series prediction, utilizing image features to complement the time series information, aiming to enrich the model's understanding of temporal patterns. The prediction module of CMLIP employs a distinct embedding technique that first applies an inversion method to map the time series data into a suitable feature space. Following this, the data is processed with a patching strategy, which segments the time series into smaller and manageable tokens. These patched indicator tokens are then fed into an encoder network. Unlike traditional models that might only consider temporal dependencies, CMLIP pays more attention to correlations among indicators and adjacent regions in the multi-indicator time series.

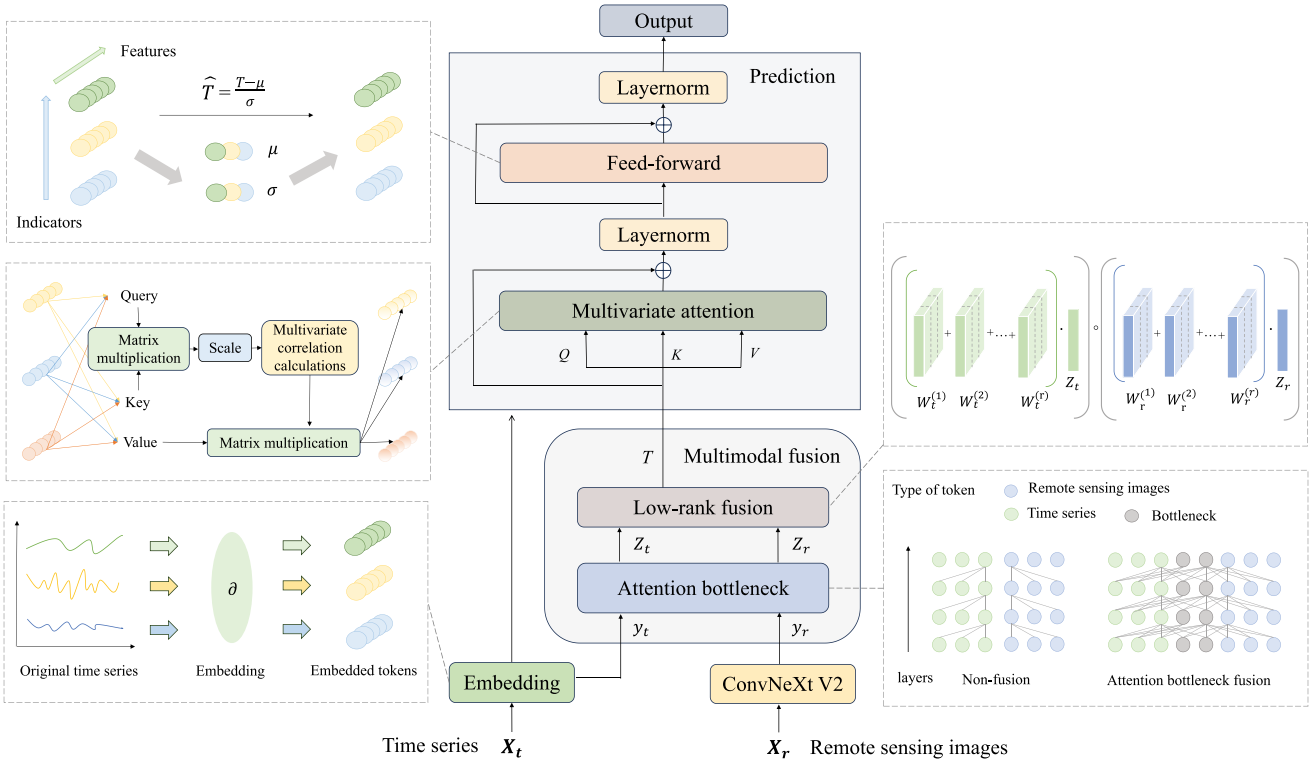

Fig. 1. Structure of CMLIP.

B. Multimodal Fusion

Data fusion methods based on deep learning have been widely adopted to solve many complex environment monitoring problems, including water quality prediction. The work in [26] analyzes the relationship between meteorological elements and wind power, and proposes a prediction method fusing wind speeds from multiple sources to predict the wind power generation. Guo et al. [27] proposed a temporal fusion method based on a spectral and temporal fusion of remotely sensed geostationary ocean color imager and Himawari images of inland waters. In addition, multimodal fusion methods based on the attention mechanism have become popular. The work in [28] designs a cross-modal skip connection method that allows visual modalities to skip cross-attention and directly perform self-attention, realizing efficient fusion. Liu et al. [29] designed a loss function for medical image fusion in different dimensions and propose a multimodal feature fusion module to preserve modality information better. The work in [30] extracts and transforms features into the same feature space and uses cross-attention for feature fusion to achieve 3-D object detection. Shvetsova et al. [31] trained and generated a fusion Transformer that can jointly process any number of modalities and allow modalities to focus on each other, without changing the Transformer. Tang et al. [32] designed multimodal dynamic augmentation blocks to capture within modality sentiment contexts, and bi-directional attention blocks to capture fine-grained multimodal sentiment contexts for multimodal sentiment analysis. The work in [33] proposes a generalized multimodal image fusion algorithm, which combines the Transformer and encoder structures to achieve global and local information fusion with a composite attention fusion strategy.

Unlike the above studies, the multimodal fusion module of CMLIP introduces a novel approach by incorporating new tokens as attention bottlenecks. These tokens play a crucial role in facilitating the sharing and interaction of information across different modalities. Meanwhile, CMLIP employs a low-rank fusion approach to multimodal data. This approach takes into account the weights of each modality in the fusion process, thus capturing the complex relationships between different modalities more accurately. The technique of low-rank fusion effectively improves the expressive power and computational efficiency of the model by reducing the redundant information between modalities and weighting the key information. The introduction of attention bottlenecks and the application of low-rank fusion techniques not only improves the predictive performance of CMLIP but also ensures that it better handles intricate multimodal data.

III. Model Framework

This section presents the overall structure of the CMLIP model. CMLIP includes three main components, i.e., data feature processing module, multimodal data fusion module, and prediction module. As shown in Fig. 1, the time series is encoded with the embedding block, which inverts normal time tokens into indicator tokens, and therefore, a token contains the variations of an indicator. We then slice them into multiple patches, reducing the length of the input sequence. This approach allows for learning both global and local temporal dependencies across multiple metrics in the time series

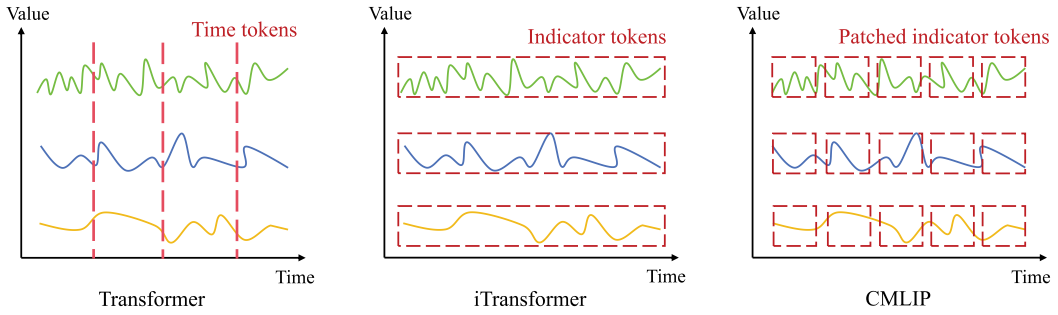

Fig. 2. Comparison of embeddings among transformer, iTransformer, and CMLIP.

while maintaining low complexity. Moreover, remotely sensed rainfall images are dimension-aligned and feature-learned with the ConvNeXt V2 network. ConvNeXt V2 builds upon the strengths of traditional CNNs, making it highly effective for image processing, especially for remote-sensing images. Compared to Transformer networks, ConvNeXt V2 has a more compact design, achieving good performance with fewer parameters. The fusion module employs the attention bottlenecks, the information of different modalities interacts and passes through these bottleneck tokens. Applying a low-rank fusion strategy, the fusion weights of each modality are fully calculated to fuse water quality time series and remotely sensed rainfall images. Finally, the fused data is adopted as input to the hybrid water quality prediction module. The multimodal attention, the feed-forward neural network, and the normalization layers in the prediction module are used to compute the attention distribution, introduce nonlinearity, and mitigate gradient vanishing, respectively.

Specifically, the input time series X_t is embedded into patched indicator tokens by the embedding block. The images X_r are extracted by ConvNeXt V2 to generate uniform-format features. The resulting tokens for the time series and the remotely sensed rainfall images are denoted as y_t and y_r, respectively. y_t and y_r are updated as Z_t and Z_r after interacting with each other through the MBT block. Z_t and Z_r are fed into the LMF block to produce the T, which is finally fed into the prediction module.

A. Data Feature Processing

1) Embedding: The embedding structure of CMLIP combines the idea of variate tokens in iTransformer and the idea of patching in PatchTST to embed the water quality time series, which is different from the embedding in the traditional Transformer. Transformer embeds all the indicators of the same time node in the sequence into a time token. Thus, it does not differentiate between single and multiple indicators and pays more attention to the correlations among time nodes. A time series of length c generates c multidimensional tokens, which have the complexity of $O(c^2)$ in both time and space. When CMLIP deals with multiple indicators, each indicator in the time series is embedded independently into the indicator token. Thus, it does not require the same time node, which enables clearer learning of the correlations among the indicators when the attention mechanism is utilized to describe inter-relationships between tokens. On this basis, overlapping or nonoverlapping new tokens are generated by patching the indicator tokens according to windows of a certain size and steps. The values of adjacent time points are close, and therefore, the tokens can capture the local information, which also makes the model focus on the features of different regions. Indicator tokens have positional logic within themselves, which can be implicitly stored in the neurons of the feed-forward neural network. Thus, CMLIP does not need the positional embedding in Transformer. The comparison of embeddings among Transformer, iTransformer, and CMLIP is shown in Fig. 2.

2) ConvNeXt V2: ConNeXt V2 is adopted to extract features of remotely sensed rainfall images. It is built upon ConNeXt by designing a fully convolutional masked autoencoder framework, which consists of a sparse convolution-based ConvNeXt encoder and a lightweight ConvNeXt block decoder. The feature collapse occurs when training ConNeXt directly on masked inputs. Therefore, a global response normalization layer is added to address this issue to enhance the feature competition among ConvNeXt block channels and promote feature diversity during the training.

B. Multimodal Fusion Module

CMLIP adopts the idea of the attention bottleneck fusion in the MBT and makes improvements based on it to fuse water quality time series data and remotely sensed rainfall images data. MBT is essentially a Transformer applied to the multimodal case, and it introduces multiple new tokens $\boldsymbol{y}_f = [y_f^1, y_f^2, \ldots, y_f^B]$ as attention bottlenecks in the input data. B is the number of tokens in the attention bottleneck. The input sequence $\boldsymbol{y}$ becomes $[\boldsymbol{y}_t \| \boldsymbol{y}_f \| \boldsymbol{y}_r]$. Different modalities can only share information and interact with each other through these bottleneck tokens. In this case, $\boldsymbol{y}_t$ and $\boldsymbol{y}_r$ can only exchange information through $\boldsymbol{y}_f$. To reduce the computational complexity, the model requires that each modality's information flow be organized and condensed before passing through the bottleneck tokens, and the necessary information must be shared to ignore the redundant information. The number of attention bottleneck tokens must be restricted to be much smaller than that of input data tokens. The bottleneck markers are updated separately according to different modes according to the time series and remotely sensed rainfall images. Finally, the bottleneck markers of each mode are averaged to yield the

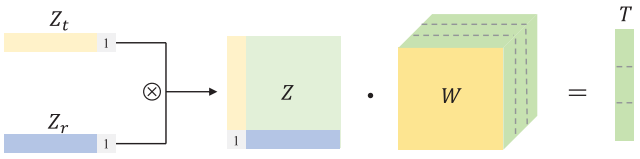

Fig. 3. Tensor fusion *via* the tensor outer product.

final fusion markers. The process can be defined as

$$\left[\boldsymbol{y}_t^{l+1} \| \hat{\boldsymbol{y}}_{f_t}^{l+1}\right] = \text{Transformer}\left(\left[\boldsymbol{y}_t^{l} \| \boldsymbol{y}_f^{l}\right]; \theta_t\right) \tag{1}$$

$$\left[\boldsymbol{y}_r^{l+1} \| \hat{\boldsymbol{y}}_{f_r}^{l+1}\right] = \text{Transformer}\left(\left[\boldsymbol{y}_r^{l} \| \boldsymbol{y}_f^{l}\right]; \theta_r\right) \tag{2}$$

$$\boldsymbol{y}_f^{l+1} = \text{Avg}_i\left(\hat{\boldsymbol{y}}_{f_i}^{l+1}\right) \tag{3}$$

where $\boldsymbol{y}_t^l$ denotes a vector of tokens of the time series in fusion layer l, $\boldsymbol{y}_r^l$ denotes a vector of tokens of the remotely sensed rainfall images in fusion layer l, θ_t represents a parameter vector of the time series, and θ_r denotes a parameter vector of the remotely sensed rainfall images.

In terms of the fusion location, a medium-term fusion strategy is employed. Fusion is performed at the nth layer, and each modality in the first $n-1$ layers learns its features with the self-attention mechanism, i.e.,

$$\boldsymbol{y}_t^{l+1} = \text{Transformer}\left(\boldsymbol{y}_t^{l}; \theta_t\right) \tag{4}$$

$$\boldsymbol{y}_r^{l+1} = \text{Transformer}\left(\boldsymbol{y}_r^{l}; \theta_r\right). \tag{5}$$

Then, $\boldsymbol{y}_t^l$ and $\boldsymbol{y}_r^l$ interact with each other through the attention bottlenecks in the fusion layer in a self-learning manner. $\boldsymbol{y}^l$ denotes a set of $\boldsymbol{y}_t^l$ and $\boldsymbol{y}_r^l$, which is given as

$$\boldsymbol{y}^l = \left[\boldsymbol{y}_t^l \| \boldsymbol{y}_r^l\right] \tag{6}$$

$$\boldsymbol{y}^{l+1} = \text{Multimodal-Transformer}\left(\boldsymbol{y}^l; \theta_t, \theta_r\right). \tag{7}$$

We improve MBT with low-rank fusion to achieve a better fusion of time series and remotely sensed rainfall images. Traditional MBT directly sums the multimodal tokens after interaction learning and calculates the average of the fusion results. Instead, we adopt LMF to perform a low-rank fusion of multimodal tokens after interaction learning by fully considering the fusion weights of each modality. A common fusion method considering the weights is the tensor fusion network (TFN) [34], where the input tensor Z is passed through the linear layer P to obtain the output tensor T. The process is shown in Fig. 3. M is the number of modalities. Z is the $M-$order tensor. W is the weight of the layer, which is the $(M+1)$th-order tensor. The exceeding $(M+1)$th dimension is the magnitude of the output tensor T, and b is the bias. Then

$$T = p(\mathcal{Z}; \mathcal{W}, b) = \mathcal{W} \cdot \mathcal{Z} + b. \tag{8}$$

However, this method has too many parameters. It is computationally complex and has a high risk of overfitting. To solve this problem, CMLIP decomposes the weight tensor W into M sets of modality-specific factors $\widetilde{\mathcal{W}}$, each of which is T-dimensional. In the case of the effective decomposition, the smallest value of R is the rank of the tensor. The process can be defined as

$$\widetilde{\mathcal{W}} = \sum_{i=1}^{R} \bigotimes_{m=1}^{M} w_m^{(i)}. \tag{9}$$

In fact, the tensor Z also needs to be decomposed into $\{z_m\}_{m=1}^M$ in the computation process, which is parallel to the modality-specific factors, and the output tensor T can be obtained, thus reducing the computational complexity of the fusion. The process is shown in Fig. 4

$$\begin{aligned} T &= \left(\sum_{i=1}^{R} \boldsymbol{w}_t^{(i)} \otimes \boldsymbol{w}_r^{(i)}\right) \cdot \mathcal{Z} \\ &= \left(\sum_{i=1}^{R} \boldsymbol{w}_t^{(i)} \cdot z_t\right) \circ \left(\sum_{i=1}^{R} \boldsymbol{w}_r^{(i)} \cdot z_r\right). \end{aligned} \tag{10}$$

C. Water Quality Prediction Module

This work adopts the encoder structure of the traditional transformer to implement hybrid water quality prediction, and its module functions have been changed because of different embedding methods and the changes of operands from time tokens to patched indicator tokens. In CMLIP, a feed-forward neural network learns the nonlinear characteristics of each patched indicator token, which encodes an individual token and decodes the future representation. A normalization layer is used to normalize patched indicator tokens, which keeps different indicator variables in the same interval and reduces differences in numerical properties among different indicators.

In Transformer, the attention mechanism performs the attention computation on different positions of the input sequence to learn its contextual relationships and dependencies. In CMLIP, the attention mechanism is used to capture the correlations among different indicator variables, as shown in Fig. 1. The attention mechanism module performs a linear map from indicator variables to yield the query (q), key (k), and value (v) of indicator tokens since indicator variables are normalized in their feature dimensions [35]. The attention mechanism computes the correlations, Atten($\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}$) of q, k, and v with the following formula:

$$\text{Atten}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \text{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{k}^\top}{\sqrt{d_k}}\right)\boldsymbol{v} \tag{11}$$

where d_k denotes the dimension of k.

IV. Experimental Evaluation

A. Dataset

Two datasets are adopted in this experiment to verify the performance of CMLIP. The first dataset is the real-time data of national surface water quality automatic monitoring from the China Environmental Monitoring Station [36] in Wucun, Langfang City, Hebei Province, China. It is recorded by the sensor every four hours, from August 2018 to December 2023. This dataset includes nine water quality indicators, i.e., dissolved oxygen, ammonia (AN), total nitrogen, the potential of hydrogen (PH), temperature, conductivity, turbidity, potassium

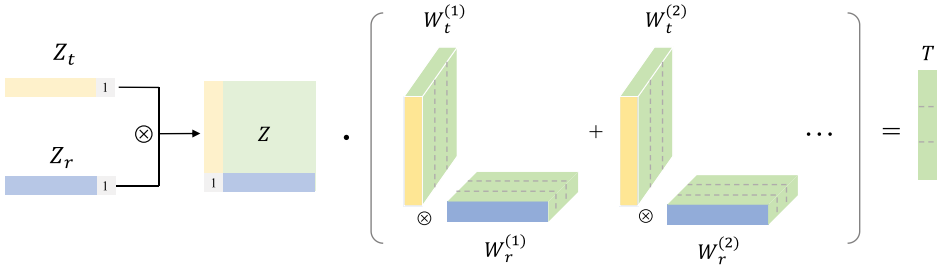

Fig. 4. Low-rank fusion *via* weight decomposition.

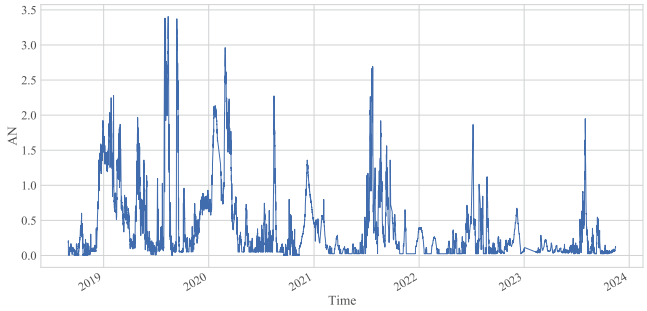

Fig. 5. In the water data in Hebei, China.

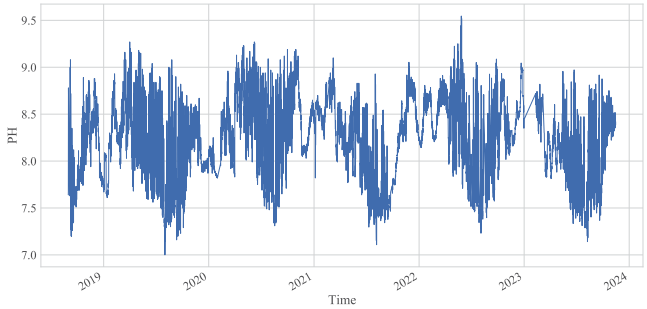

Fig. 6. PH in the water data in Hebei, China.

permanganate index, and total phosphorus. The time series of AN and PH are shown in Figs. 5 and 6. Another dataset uses the satellite remote sensing data published in the Global Satellite Precipitation Program mission [37]. It includes multisensor and multisatellite information in satellite networks. Moreover, the remote sensing data is recorded every 30 min with a spatial resolution of 0.1°x 0.1°. The period is also from August 2018 to December 2023, and the variables include latitude, longitude, time, and rainfall. A typical remotely sensed rainfall image in Beijing–Tianjin–Hebei of China is shown in Fig. 7.

B. Evaluation Metrics

To test the prediction accuracy of CMLIP, mean absolute error (MAE) [38] and mean squared error (MSE) [39] are adopted. MAE and MSE are calculated as

$$\text{MAE} = \frac{1}{a}\sum_{j=1}^{a}\left|\hat{h}_j - h_j\right| \tag{12}$$

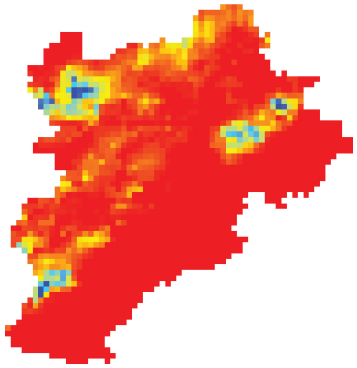

Fig. 7. Typical remotely sensed rainfall image in Beijing–Tianjin–Hebei of China.

TABLE I
PREDICTION RESULTS FOR CMLIP WITH DIFFERENT S

S	MSE	MAE	Prediction Time
48	$0.418\pm2.98\times10^{-8}$	$0.451\pm3.44\times10^{-8}$	70.812±2.448
72	$0.403\pm1.72\times10^{-8}$	$0.439\pm2.98\times10^{-8}$	67.368±3.282
96	$\mathbf{0.377}\pm1.53\times10^{-8}$	$\mathbf{0.434}\pm2.02\times10^{-8}$	**62.452**±2.251
120	$0.415\pm2.35\times10^{-8}$	$0.442\pm3.17\times10^{-8}$	69.744±2.863

$$\text{MSE} = \frac{1}{a}\sum_{i=1}^{a}\left|\hat{h}_j - h_j\right|^2 \tag{13}$$

where a denotes the number of samples. h_j and $\hat{h}_j$ denote the ground truth and predicted values of data point j.

C. Parameter Tuning

The selection of the hyperparameters greatly affects the prediction accuracy. CMLIP's hyperparameters include the length of the input sequence (S), the dimension of embedding (D), the number of fusion bottleneck tokens (B), the batch size, and the optimizer. The prediction accuracy varies significantly with S. If S is too short, the attention mechanism cannot capture the information, yielding lower prediction accuracy. However, if S is longer, there is too much noise or periodic information in the sequence, leading to overfitting that reduces the prediction accuracy. Table I shows the MAE, MSE, and the prediction time of CMLIP for different input sequences, and the results prove that the prediction accuracy of CMLIP is the best when $S = 96$.

Too small D does not capture enough information, and larger D yields a more expressive model. However, it requires more training time, computational resources, and overfitting.

TABLE II
PREDICTION RESULTS FOR CMLIP WITH DIFFERENT D

D	MSE	MAE	Prediction Time
128	0.311±0.00	0.346±0.00	65.331±2.262
256	$0.269\pm1.72\times10^{-8}$	$0.322\pm1.72\times10^{-8}$	63.545±4.429
512	$\mathbf{0.257}\pm4.55\times10^{-8}$	$\mathbf{0.322}\pm1.72\times10^{-8}$	**63.113**±0.772
1024	$0.266\pm1.37\times10^{-7}$	$0.332\pm1.49\times10^{-7}$	67.684±5.032

TABLE III
PREDICTION RESULTS FOR CMLIP WITH DIFFERENT B

B	MSE	MAE	Prediction Time
1	$\mathbf{0.372}\pm2.52\times10^{-8}$	$\mathbf{0.412}\pm1.93\times10^{-8}$	**59.234**±1.343
2	$0.423\pm3.69\times10^{-8}$	$0.439\pm2.51\times10^{-8}$	63.258±1.947
3	$0.427\pm2.52\times10^{-8}$	$0.450\pm3.31\times10^{-8}$	65.371±2.375
4	$0.425\pm2.91\times10^{-8}$	$0.452\pm3.94\times10^{-8}$	69.223±3.427
5	$0.458\pm3.76\times10^{-8}$	$0.473\pm4.27\times10^{-8}$	72.352±2.759

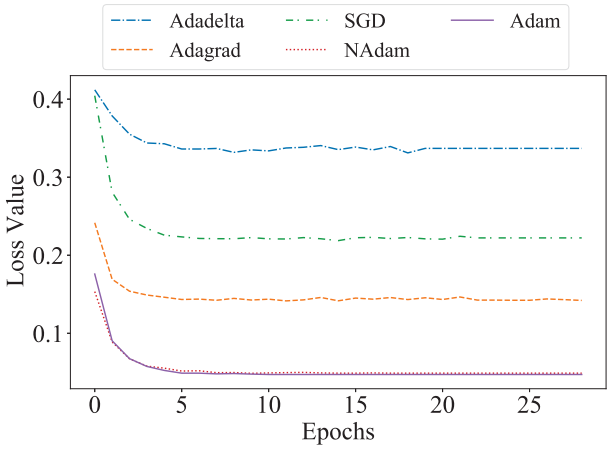

Fig. 8. Loss values for different optimizers.

During the tuning process, D greatly impacts the prediction accuracy. Table II shows MAE, MSE, and the prediction time of CMLIP when $D\in[128, 256, 512, 1024]$. The results prove that CMLIP achieves the best prediction accuracy when $D = 512$.

The number of bottleneck tokens is the most important hyperparameter in the fusion part. To avoid too large a computational complexity of the fusion, the number of bottleneck tokens needs to be much smaller than that of input data tokens. Table III shows MAE, MSE, and the prediction time of CMLIP with different B, and the result proves that the fusion performance is the best and the prediction result is the most accurate when $B = 1$.

We compare five optimization algorithms: 1) stochastic gradient descent (SGD); 2) adaptive delta (Adadelta); 3) adaptive gradient algorithm (Adagrad); 4) adaptive moment estimation (Adam); and 5) Nesterov-accelerated Adam (Nadam). As shown in Fig. 8, the results indicate that Adam exhibits the fastest convergence rate and achieves the lowest loss compared to other optimizers. Consequently, we select Adam as the optimizer.

The choice of batch size impacts both the stability and speed of training. A smaller batch size allows for more frequent updates but introduces instability, whereas a larger batch size provides more stable updates and faster per-step training, potentially at the cost of limiting the final model performance. Table IV shows the MAE, MSE, and the prediction time of CMLIP with different batch sizes, indicating that the most accurate prediction results are obtained when the batch size is 48. Based on these parameter tuning experiments, the best values of the model input parameters are given in Table V.

TABLE IV
PREDICTION RESULTS FOR CMLIP WITH DIFFERENT BATCH SIZES

Batch size	MSE	MAE	Prediction Time
16	$0.333\pm2.86\times10^{-6}$	$0.293\pm4.95\times10^{-6}$	**59.334**±0.042
24	$0.334\pm9.58\times10^{-8}$	$0.293\pm1.79\times10^{-8}$	70.892±0.769
32	$0.344\pm7.88\times10^{-8}$	$0.322\pm7.50\times10^{-7}$	63.972±1.298
48	$\mathbf{0.327}\pm1.72\times10^{-8}$	$\mathbf{0.270}\pm3.86\times10^{-8}$	62.892±0.575
56	$0.328\pm5.77\times10^{-8}$	$0.275\pm3.77\times10^{-8}$	71.870±1.292

TABLE V
PARAMETER SETTING OF CMLIP

Parameter	Values	Description
S	96	Length of input sequence
D	512	Dimension of embedding
B	1	Number of fusion bottleneck
Optimizer	Adam	Optimizer algorithm
Batch size	48	Number of training samples
Learning rate	0.0001	Initial learning rate of optimizer

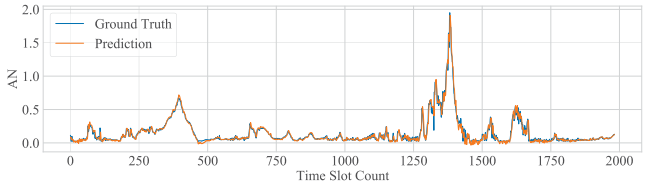

Fig. 9. Comparison of ground-truth values and predicted ones for AN.

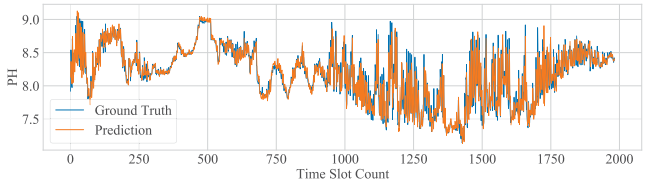

Fig. 10. Comparison of ground-truth values and predicted ones for PH.

D. Comparison of Experimental Results

We first realize single-indicator prediction for CMLIP with only water quality time series. The ground-truth values and predicted ones of AN and PH are shown in Figs. 9 and 10, respectively. The red line indicates the predicted values and the blue line indicates the ground-truth values. To verify the prediction accuracy of CMLIP, we choose four state-of-the-art models for comparison, including iTransformer, Autoformer, PatchTST, and Crossformer [40]. These models encompass a variety of methodologies and technical approaches. iTransformer is a significant advancement in transformer-based architectures for time series forecasting and serves as an

TABLE VI
COMPARISON OF PREDICTION RESULTS OF ITRANSFORMER, AUTOFORMER, PATCHTST, CROSSFORMER, AND CMLIP

Prediction Indicator	Models	48 Steps		128 Steps		256 Steps	
		MSE	MAE	MSE	MAE	MSE	MAE
AN	iTransformer	0.056±0.004	0.129±0.006	0.065±0.000	0.145±0.001	0.083±0.001	0.168±0.001
	Autoformer	0.098±0.006	0.193±0.009	0.123±0.002	0.249±0.010	0.174±0.009	0.274±0.013
	PatchTST	0.051±0.001	0.125±0.003	0.061±0.003	0.139±0.005	0.071±0.001	0.155±0.002
	Crossformer	0.063±0.006	0.182±0.012	0.104±0.021	0.230±0.039	0.321±0.035	0.363±0.066
	CMLIP	**0.041**±0.001	**0.113**±0.001	**0.059**±0.002	**0.137**±0.001	**0.069**±0.003	**0.153**±0.002
PH	iTransformer	0.099±0.005	0.240±0.005	0.140±0.001	0.294±0.002	0.176±0.002	0.330±0.002
	Autoformer	0.115±0.004	0.266±0.006	0.148±0.002	0.311±0.002	0.160±0.008	0.322±0.010
	PatchTST	0.087±0.001	0.255±0.002	0.116±0.004	0.270±0.005	0.163±0.002	0.320±0.003
	Crossformer	0.084±0.007	0.216±0.009	0.101±0.015	0.252±0.018	0.125±0.033	0.297±0.046
	CMLIP	**0.075**±0.001	**0.205**±0.000	**0.098**±0.002	**0.247**±0.001	**0.119**±0.001	**0.280**±0.001

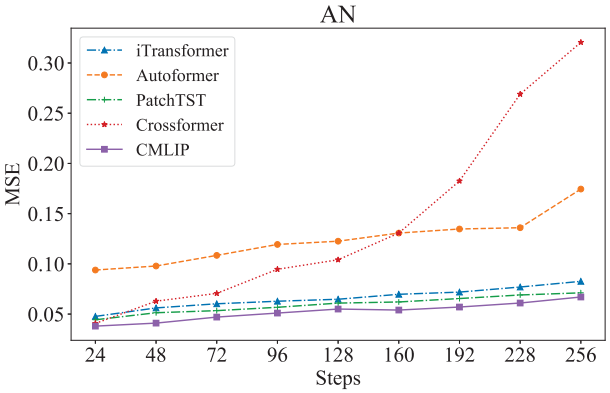

Fig. 11. MSE values of different models for AN.

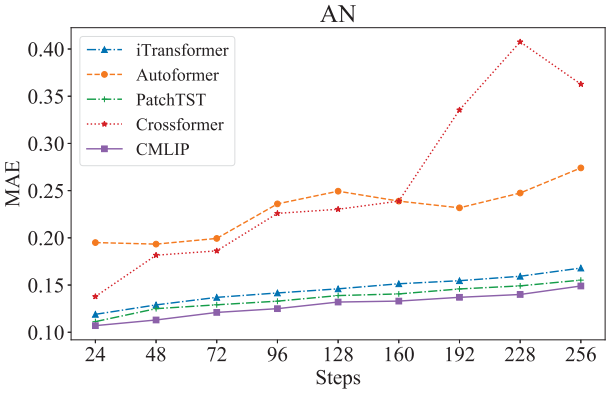

Fig. 12. MAE values of different models for AN.

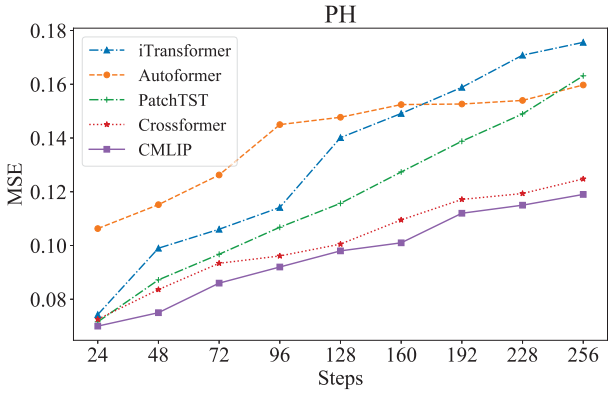

Fig. 13. MSE values of different models for PH.

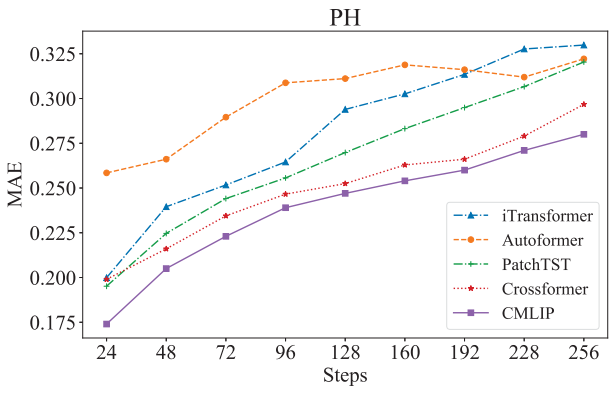

Fig. 14. MAE values of different models for PH.

essential benchmark for evaluating CMLIP's capability to handle complex temporal patterns. Autoformer is renowned for its automated feature learning capabilities and introduces self-correlation mechanisms that effectively capture long-term dependencies. PatchTST enhances sensitivity to local time patterns through its patching strategy, which directly corresponds to the patching strategy in CMLIP, making it a highly suitable comparator. Crossformer excels at establishing associations across multiple time series by leveraging cross-time and cross-variable attention mechanisms, enhancing the modeling of complex dependencies. Given that CMLIP's embedding design also adeptly captures features in multivariable time series, Crossformer is an appropriate contrastive model. Table VI shows the MSE and MAE of each model in predicting AN and PH. Each result is presented as the mean ± standard deviation, derived from five independent replicate experiments. Figs. 11–14 show MSE and MAE values of different models when the prediction steps are in a set of {24, 48, . . . , 256}, respectively. Table VII displays the floating point operations (FLOPs), the number of parameters, and the prediction duration for different models. The results show that the prediction accuracy of CMLIP is higher than the other four models and the computation time is shorter with more FLOPs and the number of parameters.

We investigate the fusion effect of water quality time series and remotely sensed rainfall data on the prediction accuracy

TABLE VII
COMPARISON OF COMPUTATIONAL COSTS OF iTRANSFORMER, AUTOFORMER, PATCHTST, CROSSFORMER, AND CMLIP

Prediction Length	Models	FLOPs	Number of Parameters	Prediction Time
48	iTransformer	7.04×10^{8}	2.44×10^{6}	20.731±0.062
	Autoformer	1.42×10^{10}	4.62×10^{6}	36.235±0.486
	PatchTST	4.61×10^{9}	2.67×10^{6}	20.858±0.491
	Crossformer	1.40×10^{10}	2.24×10^{7}	45.728±3.935
	CMLIP	$\mathbf{1.52}\times10^{10}$	$\mathbf{2.75}\times10^{7}$	**20.137**±0.421
128	iTransformer	7.15×10^{8}	2.49×10^{6}	19.572±0.372
	Autoformer	1.86×10^{10}	4.62×10^{6}	33.831±0.581
	PatchTST	4.69×10^{9}	3.16×10^{6}	19.581±0.537
	Crossformer	1.59×10^{10}	2.24×10^{7}	41.177±1,426
	CMLIP	$\mathbf{1.70}\times10^{10}$	$\mathbf{2.81}\times10^{7}$	**19.470**±0.349
256	iTransformer	7.34×10^{8}	2.55×10^{6}	19.001±0.252
	Autoformer	4.62×10^{6}	4.62×10^{6}	33.740±0.383
	PatchTST	4.82×10^{9}	3.95×10^{6}	20.012±0.569
	Crossformer	1.60×10^{10}	2.24×10^{7}	43.454±3.431
	CMLIP	$\mathbf{1.83}\times10^{10}$	$\mathbf{2.93}\times10^{7}$	**18.875**±0.312

TABLE VIII
COMPARISON OF CMLIP'S PREDICTION RESULTS OF WATER QUALITY TIME SERIES FUSED WITH REMOTELY SENSED RAINFALL IMAGES AND ONLY THE TIME SERIES

Prediction Length	X_t+X_r		X_t	
	MSE	MAE	MSE	MAE
96	$\mathbf{0.388}\pm3.57\times10^{-8}$	$\mathbf{0.438}\pm1.36\times10^{-8}$	0.429±0.013	0.456±0.021
192	$\mathbf{0.564}\pm5.88\times10^{-5}$	$\mathbf{0.562}\pm2.97\times10^{-5}$	0.602±0.026	0.571±0.029
256	**0.645**±0.045	**0.605**±0.019	0.694±0.037	0.624±0.049
320	**0.712**±0.038	**0.628**±0.035	0.801±0.048	0.683±0.024
384	**0.742**±0.025	**0.650**±0.008	0.961±0.039	0.759±0.037
450	**0.796**±0.056	**0.674**±0.061	1.030±0.044	0.796±0.049
512	**0.793**±0.047	**0.678**±0.038	1.096±0.049	0.819±0.051

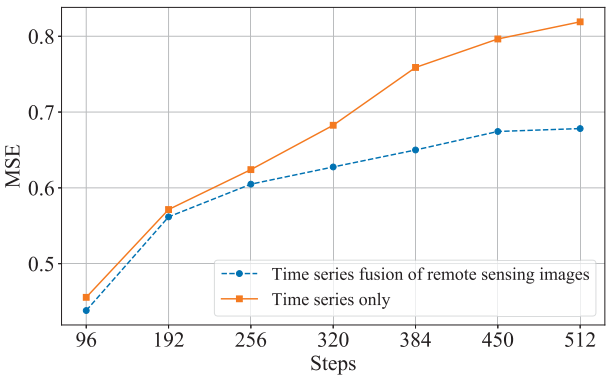

Fig. 15. MSE values of CMLIP for different inputs.

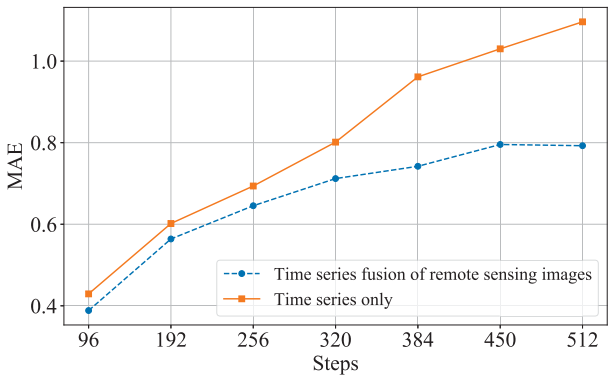

Fig. 16. MAE values of CMLIP for different inputs.

of multi-indicator in Table VIII, where column X_t is the result of the hybrid prediction with CMLIP for each indicator with only the input of time series, and column $X_t + X_r$ is that for each indicator with both the water quality time series and spatiotemporally aligned remotely sensed rainfall images. The comparison of MSE and MAE is shown in Figs. 15 and 16. The results show that CMLIP's prediction accuracy of the water quality time series fused with remotely sensed rainfall images is 17% higher than that with only the time series on average. Moreover, the prediction outcomes after fusion are more stable. The reason is that the fusion module in CMLIP complements the water quality time series features with the features of remotely sensed rainfall images, which increases the effective information of inputs and improves the prediction.

In addition, to verify the performance of CMLIP for fusion, we compare CMLIP with three commonly used fusion models, including MBT, LMF, and TFN. These models are representative of the multimodal data fusion field and adopt different technical routes and design philosophies, each showcasing unique advantages. MBT emphasizes flexibility and adaptability, utilizing self-attention mechanisms to capture dependencies. LMF reduces redundant information through dimensionality reduction techniques, effectively minimizing resource consumption. TFN excels at capturing complex

TABLE IX
COMPARISON OF FUSION RESULTS OF MBT, LMF, TFN, AND CMLIP

Models	128 Steps			256 Steps			384 Steps		
	MSE	MAE	Prediction Time	MSE	MAE	Prediction Time	MSE	MAE	Prediction Time
MBT	0.472±0.008	0.491±0.011	60.223±2.781	0.682±0.016	0.628±0.025	61.452±1.340	0.786±0.029	0.679±0.040	59.145±2.768
LMF	0.468±0.013	0.487±0.022	73.259±2.655	0.670±0.035	0.627±0.036	70.447±3.121	0.832±0.047	0.707±0.041	71.878±2.546
TFN	0.509±0.010	0.522±0.003	75.354±1.139	0.667±0.029	0.614±0.025	72.877±1.348	0.831±0.033	0.678±0.025	74.583±1.054
CMLIP	**0.442**±3.51 × 10^{-7}	**0.472**±1.49 × 10^{-7}	**57.567**±2.718	**0.645**±0.045	**0.605**±0.019	**54.661**±2.820	**0.742**±0.025	**0.650**±0.008	**43.457**±0.127

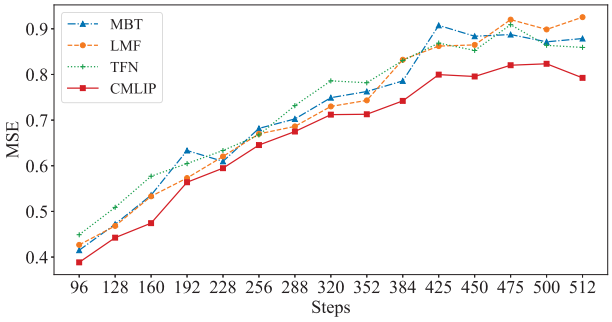

Fig. 17. MSE values of MBT, LMF, TFN, and CMLIP.

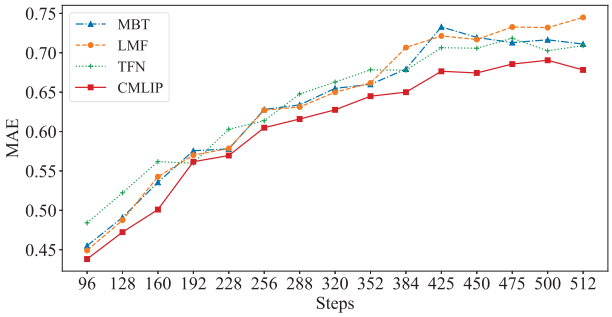

Fig. 18. MAE values of MBT, LMF, TFN, and CMLIP.

interactions between modalities, providing richer feature representations. By comparing these models, we can better reflect CMLIP's effectiveness in multimodal data fusion. The comparison of MSE, MAE, and Prediction Time of different models is shown in Table IX. Figs. 17 and 18 show each model's MSE and MAE values when the prediction steps are in a set of {96, 128, . . . , 512}, respectively. Table X displays the FLOPs and the number of parameters for different models. Results demonstrate that CMLIP outperforms MBT, LMF, and TFN in terms of fusion effectiveness, achieving superior results with a shorter processing time, higher FLOPs, and a greater number of parameters. This is because the fusion module in CMLIP combines attention bottlenecks and a low-rank fusion strategy to achieve shared interaction and dynamic fusion of different modal features, which more accurately takes into account the weights of each modality, efficiently manages computational complexity, and reduces the risk of overfitting. This makes the fusion of multimodal data more accurate and efficient.

V. CONCLUSION

Water quality is affected by meteorological factors in addition to the water environment. Existing water quality prediction methods only take water quality historical indicator

TABLE X
COMPARISON OF COMPUTATIONAL COSTS OF MBT, LMF, TFN, AND CMLIP

Models	FLOPs	Number of Parameters
MBT	1.29×10^{12}	2.54×10^{8}
LMF	5.37×10^{9}	8.69×10^{6}
TFN	1.25×10^{10}	9.02×10^{7}
CMLIP	$\mathbf{1.63}\times10^{12}$	$\mathbf{4.31}\times10^{8}$

data as the input. However, there are many other factors that affect water quality indicators, such as meteorology and pollutants. Therefore, considering only historical time series data on water quality is not sufficient for accurate prediction, and a fusion of data from different modalities is needed. This work proposes a novel hybrid water quality prediction model called CMLIP, which combines the ConvNeXt V2, multimodal bottleneck transformer (MBT), low-rank Multimodal Fusion (LMF), itransformer, and PatchTST. ConvNeXt V2 is integrated to learn features of remotely sensed rainfall images and align them with the feature dimensions of time series. The combination of MBT and LMF is used as a multimodal fusion module to learn the influences of the time series and rainfall images and fuse their respective features. Finally, the fused features are fed into the prediction module, which combines iTransformer and PatchTST for prediction. Experimental results with real-life water quality time series and remotely sensed rainfall images prove that CMLIP outperforms other state-of-the-art algorithms in fusion and prediction. CMLIP's accuracy of water quality prediction by fusing time series and rainfall images is 17% higher on average than that with only water quality time series.

The performance of CMLIP relies on the quality of the input data, including water quality time series and remotely sensed rainfall images. If these data have more noise or missing values, it may lead to a decrease in its prediction performance. To improve data quality, in the future, we intend to introduce techniques such as data cleaning, denoising, and missing value completion. Currently, CMLIP fuses water quality time series with remotely sensed rainfall images and demonstrates promising results for specific datasets. However, its generalization capability across different datasets still needs further validation. To enhance the prediction accuracy of the model and its generalization ability, we plan to fuse other modal data in the future, such as pollutant concentration data [41], to more comprehensively capture various factors affecting water quality. This will not only help to improve the prediction accuracy but also verify its applicability in diverse environments.

REFERENCES

[1] S. Yang et al., "Hybrid approach for early warning of mine water: Energy density-based identification of water-conducting channels combined with water inflow prediction by SA-LSTM," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Apr. 2024, Art. no. 5911312.

[2] A. Gupta and A. Kumar, "Mid term daily load forecasting using ARIMA, wavelet-ARIMA and machine learning," in *Proc. IEEE Int. Conf. Environ. Elect. Eng. IEEE Ind. Commer. Power Syst. Europe*, Madrid, Spain, 2020, pp. 1–5.

[3] J. Bi, H. Yuan, S. Li, K. Zhang, J. Zhang, and M. Zhou, "ARIMA-based and multiapplication workload prediction with wavelet decomposition and Savitzky–Golay filter in clouds," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 4, pp. 2495–2506, Apr. 2024.

[4] K. Guo et al., "Traffic data-empowered XGBoost-LSTM framework for infectious disease prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 774–785, Jan. 2024.

[5] S. Fong and S. Narasimhan, "An unsupervised Bayesian OC-SVM approach for early degradation detection, thresholding, and fault prediction in machinery monitoring," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, Dec. 2022.

[6] I. Harmon et al., "Injecting domain knowledge into deep neural networks for tree crown delineation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 4415419.

[7] W. Guan et al., "Egocentric early action prediction via multimodal transformer-based dual action prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4472–4483, Sep. 2023.

[8] J. Bi, Y. Li, X. Chang, H. Yuan, and J. Qiao, "Hybrid water quality prediction with frequency domain conversion enhancement and seasonal decomposition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2024, pp. 5200–5205.

[9] J. Xu and Z. Liu, "Improving the accuracy of MODIS near-infrared water vapor product under all weather conditions based on machine learning considering multiple dependence parameters," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Mar. 2023, Art. no. 4101115.

[10] M. Xu et al., "Implementation strategy and spatiotemporal extensibility of multipredictor ensemble model for water quality parameter retrieval with multispectral remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 4200616.

[11] H. Wang, B. Wang, A. M. Alharbi, D. W. Gao, H. Ma, and P. Luo, "Correlation-based multimodal fusion method for icing degree monitoring of transmission lines within Internet of Things," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 25413–25424, Jul. 2024.

[12] S. Woo et al., "ConvNeXt V2: Co-designing and scaling convNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 16133–16142.

[13] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 14200–14213.

[14] M. Das, D. Gupta, P. Radeva, and A. M. Bakde, "Optimized multimodal neurological image fusion based on low-rank texture prior decomposition and super-pixel segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, Apr. 2022.

[15] Y. Liu et al., "iTransformer: Inverted transformers are effective for time series forecasting," 2023, *arXiv:2310.06625*.

[16] Y. Nie, H. Nam, S. Phanwadee, and K. Jayant, "A time series is worth 64 words: Long-term forecasting with transformers," 2022, *arXiv:2211.14730*.

[17] H. Yuan, S. Wang, J. Bi, J. Zhang, and M. Zhou, "Hybrid and spatiotemporal detection of cyberattack network traffic in cloud data centers," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 18035–18046, May 2024.

[18] Z. Ma, H. Zhang, and J. Liu, "MM-RNN: A multimodal RNN for precipitation nowcasting," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 4101914.

[19] J. Geng, C. Yang, Y. Li, L. Lan, and Q. Luo, "MPA-RNN: A novel attention-based recurrent neural networks for total nitrogen prediction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6516–6525, Oct. 2022.

[20] R. Jin, Z. Chen, K. Wu, M. Wu, X. Li, and R. Yan, "Bi-LSTM-based two-stream network for machine remaining useful life prediction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, Apr. 2022.

[21] J. Bi, Z. Guan, H. Yuan, and J. Zhang, "Improved network intrusion classification with attention-assisted bidirectional LSTM and optimized sparse contractive autoencoders," *Expert Syst. Appl.*, vol. 244, pp. 1–13, Jun. 2024.

[22] C. Ma, Y. Zhao, G. Dai, X. Xu, and S.-C. Wong, "A Novel STFSA-CNN-GRU hybrid model for short-term traffic speed prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 3728–3737, Apr. 2023.

[23] Y. Li et al., " A TCN-based hybrid forecasting framework for hours-ahead utility-scale PV forecasting," *IEEE Trans. Smart Grid*, vol. 14, no. 5, pp. 4073–4085, Sep. 2023.

[24] C. Yang, C. Yang, X. Zhang, and J. Zhang, "Multisource information fusion for autoformer: Soft sensor modeling of FeO content in iron ore sintering process," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11584–11595, Dec. 2023.

[25] J. Qiao, Y. Lin, J. Bi, H. Yuan, G. Wang, and M. Zhou, "Attention-based spatiotemporal graph fusion convolution networks for water quality prediction," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 1–10, Mar. 2024, doi: 10.1109/TASE.2023.3285253.

[26] J. An, F. Yin, M. Wu, J. She, and X. Chen, "Multisource wind speed fusion method for short-term wind power prediction," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 5927–5937, Sep. 2021.

[27] Y. Guo, C. Huang, Y. Zhang, Y. Li, and W. Chen, "A novel multitemporal image-fusion algorithm: Method and application to GOCI and Himawari images for inland water remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4018–4032, Jun. 2020.

[28] C. Li et al., "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," 2022, *arXiv:2205.12005*.

[29] Y. Liu, Y. Shi, F. Mu, J. Cheng, C. Li, and X. Chen, "Multimodal MRI volumetric data fusion with convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, Jun. 2022.

[30] Y. Li et al., "DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 17161–17170.

[31] N. Shvetsova et al., "Everything at once–multi-modal fusion transformer for video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 19988–19997.

[32] J. Tan et al., "BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1966–1978, Apr. 2023.

[33] Y. Zhao, Q. Zheng, P. Zhu, X. Zhang, and W. Ma, "TUFusion: A transformer-based universal fusion algorithm for multimodal images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1712–1725, Mar. 2024.

[34] K. Wang, Y. Wang, X. L. Zhao, J. C. W. Chan, Z. Xu, and D. Meng, "Hyperspectral and multispectral image fusion via nonlocal low-rank tensor decomposition and spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7654–7671, Nov. 2020.

[35] S. Li, C. Zhang, L. Liu, and X. Zhang, "Gated transient fluctuation dual attention unit network for long-term remaining useful life prediction of rotating machinery using IIoT," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 18593–18604, May 2024.

[36] J. Bi, Z. Chen, H. Yuan, and J. Zhang, "Accurate water quality prediction with attention-based bidirectional LSTM and encoder–decoder," *Expert Syst. Appl.*, vol. 238, pp. 1–10, Mar. 2024.

[37] D. Jin, L. Oreopoulos, D. Lee, J. Tan, and N. Cho, "Cloud–precipitation hybrid regimes and their projection onto IMERG precipitation data," *J. Appl. Meteorol. Climatol.*, vol. 60, no. 6, pp. 733–748, Jun. 2022.

[38] H. Yuan, J. Bi, S. Li, J. Zhang, and M. Zhou, "An improved LSTM-based prediction approach for resources and workload in large-scale data centers," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 22816–22829, Jun. 2024.

[39] S. Cai and V. K. N. Lau, "MSE tail analysis for remote state estimation of linear systems over multiantenna random access channels," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 2046–2061, May 2020.

[40] W. Wang et al., "CrossFormer++: A versatile vision transformer hinging on cross-scale attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3123–3136, May 2024.

[41] R. K. Amineh, M. Ravan, and D. Tandel, "Detection of water pollutants with a nonuniform array of microwave sensors," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, Apr. 2023.

Jing Bi (Senior Member, IEEE) received the B.S., and Ph.D. degrees in computer science from Northeastern University, Shenyang, China, in 2003 and 2011, respectively.

From 2013 to 2015, she was a Postdoctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2018 to 2019, she was a Visiting Research Scholar with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA. She is currently a Professor with the College of Computer Science, Beijing University of Technology, Beijing. She has over 200 publications in international journals and conference proceedings, and she is named in the world's top 2% of Scientists List. Her research interests include distributed computing, cloud and edge computing, large-scale data analytics, machine learning, industrial Internet, and performance optimization.

Prof. Bi is currently an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS: SYSTEMS.

Yibo Li (Graduate Student Member, IEEE) received the B.E. degree in Internet of Things from Dalian Nationalities University, Dalian, China, in 2022. She is currently pursuing the master's degree with the College of Computer Science, Beijing University of Technology, Beijing, China.

Her research interests include time series forecasting, multimodal data fusion, and machine learning.

Haitao Yuan (Senior Member, IEEE) received the Ph.D. degree in computer engineering from New Jersey Institute of Technology, Newark, NJ, USA, in 2020.

He is currently a Deputy Director with the Department of Science and Technology Innovation, Wenchang International Aerospace City, Hainan, China. He is currently an Associate Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China, and he is named in the world's top 2% of Scientists List. His research interests include the Internet of Things, edge computing, deep learning, data-driven optimization, and computational intelligence algorithms.

Dr. Yuan received the Chinese Government Award for Outstanding Self-Financed Students Abroad, the 2021 Hashimoto Prize from NJIT, the Best work Award in the 17th ICNSC, and the Best Student work Award Nominees in 2024 IEEE SMC. He is an Associate Editor for IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, and *Expert Systems With Applications*.

Mengyuan Wang received the Ph.D. degree in mechanical engineering from the University of Connecticut (UConn), Storrs, CT, USA in 2021.

She was a Research Specialist/a Scholar with the Center for Clean Energy Engineering, UConn from 2021 to 2023, and joined Beihang University, Beijing, China, as an Associate Professor in December 2022. Her research interests focus on the chemical kinetic studies of alternative transportation fuels, including the experimental measurements of combustion characteristics, as well as the development, numerical simulation, and analysis of chemical kinetic mechanisms.

Dr. Wang was involved in as one of the core members of five projects from Lawrence Livermore National Laboratory and Department of Energy in U.S., and served as a Reviewer for American Chemical Society Publications and Proceedings of Combustion Institute.

Ziqi Wang (Graduate Student Member, IEEE) received the B.E. degree in Internet of Things from Beijing University of Technology, Beijing, China, in 2022, where he is currently pursuing the master's degree with the College of Computer Science.

His research interests include cloud computing, task scheduling, intelligent optimization algorithms, and machine learning.

Jia Zhang (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois at Chicago, Chicago, IL, USA, in 2000.

She is currently the Cruse C. and Marjorie F. Calahan Centennial Chair in Engineering, a Professor with the Department of Computer Science, Lyle School of Engineering, Southern Methodist University, Dallas, TX, USA. Her research interests emphasize the application of machine learning and information retrieval methods to tackle data science infrastructure problems, with a recent focus on scientific workflows, provenance mining, software discovery, knowledge graph, and interdisciplinary applications of all of these interests in earth science.

Mengchu Zhou (Fellow, IEEE) received the Ph.D. degree from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

Then, he joined New Jersey Institute of Technology, Newark, NJ, USA, where he is currently a Distinguished Professor. His research interests include Petri nets, automation, the Internet of Things, and big data. He has over 900 publications, including 12 books, more than 600 journal articles (more than 500 in IEEE Transactions), 29 patents, and 29 book-chapters.

Dr. Zhou is a Fellow of IFAC, AAAS, CAA, and NAI.