Petroleum Science 22 (2025) 5005–5019

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

LLMs-guided parameters prediction of tight sandstone reservoirs

Juan Wu [a], Ren-Ze Luo [a,*], Lei Luo [b], Can-Ru Lei [c], Xing-Ting Chen [a]

[a] School of Geoscience and Technology, Southwest Petroleum University, Chengdu, 610500, Sichuan, China
[b] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[c] Shale Gas Research Institute of PetroChina Southwest Oil & Gasfield Company, Chengdu, 610051, Sichuan, China

A R T I C L E I N F O

Article history:
Received 3 August 2024
Received in revised form
23 January 2025
Accepted 20 August 2025
Available online 25 August 2025

Edited by Meng-Jiao Zhou

Keywords:
Large language models
Tight sandstone reservoirs
Cross-modal alignment
Data augmentation
Petrophysical parameters prediction

A B S T R A C T

Accurate prediction of reservoir parameters is essential for reservoir evaluation. Recent machine learning methods have spurred significant advancements in reservoir prediction; however, limited well logging data and strong reservoir heterogeneity still hinder the accuracy and reliability of such predictions. Addressing these challenges requires methods capable of effectively predicting reservoir parameters under data scarcity and complex reservoir structures. In this study, we propose CAF^2, a feature-fusion cross-modal alignment framework guided by large language models (LLMs). CAF^2 integrates data augmentation, knowledge distillation, cross-modal alignment, and feature fusion. The data augmentation module employs the RealTabFormer model to generate synthetic datasets that mirror the distribution of real logging data, addressing the challenge of data scarcity. Knowledge distillation extracts essential domain knowledge from LLMs, guiding cross-modal alignment between well logging data and textual knowledge. This alignment mitigates modality gaps via token and sequence alignment, enhancing depth-domain feature representation. Finally, a cross-variable and cross-depth feature fusion strategy exploits both variable information and depth correlations, overcoming the difficulties in accurate reservoir parameter prediction posed by reservoir heterogeneity. Experimental results demonstrate that CAF^2 significantly outperforms existing models in predicting tight sandstone reservoir parameters, serving as an effective tool for oil and gas exploration and development.
© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Tight sandstone gas, as a major type of unconventional natural gas resource, plays a vital role in the field of natural gas exploration and development due to its abundant reserves and growing exploration potential (Meng et al., 2023; Tian et al., 2023). Reservoir evaluation parameters—such as porosity, permeability, and water saturation—serve as key indicators of a reservoir's capacity to store and transmit fluids (Bagheri and Rezaei, 2019; Moosavi et al., 2023). Accurate prediction of these parameters is essential for optimizing exploration strategies and carries significant implications for the efficient and economic development of tight sandstone reservoirs (Sang et al., 2022; Zhang and Wu, 2023; Zhou et al., 2019).

Traditionally, reservoir parameter estimation has relied primarily on rock physics, through the construction of response equations or physical models (Zare et al., 2020). Such approaches offer strong explanatory power, clearly revealing the inherent mechanisms within the data and providing effective solutions to conventional reservoir problems (Wang et al., 2023; Xu and Xiao, 2024). However, rock physics mechanisms under actual geological conditions are often highly complex. These methods are constrained by the limited availability of core experimental data, which hinders their ability to fully capture the nonlinear relationships among reservoir parameters, logging data, and spatial continuity. Measurement results in complex strata often exhibit substantial uncertainty, complicating the application of rock physics mechanism models (Shao et al., 2022). Meanwhile, tight gas reservoirs are characterized by “low porosity, low permeability, and low pressure” (Zhang et al., 2024). The strong heterogeneity of their geological conditions further limits the predictive power of conventional methods, posing significant challenges to

* Corresponding author.
E-mail address: lrzsmith@126.com (R.-Z. Luo).
Peer review under the responsibility of China University of Petroleum (Beijing).

https://doi.org/10.1016/j.petsci.2025.08.029
1995-8226/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the efficient development of gas fields (Feng et al., 2023; Moosavi et al., 2022).

To overcome these bottlenecks, the geological community has increasingly turned to machine learning (ML) and deep learning (DL) approaches (Bagheri and Riahi, 2015, 2017; Huang et al., 2023; Li et al., 2023; Tavakolizadeh and Bagheri, 2021). For example, Chen et al. (2024) constructed an ML-based model to predict adsorbed gas content, while Bai et al. (2022) and Goliatt et al. (2023) proposed integrated ML strategies for reservoir parameter prediction. However, shallow ML models are inherently limited by their small number of parameters, which constrains their ability to capture complex geological characteristics such as heterogeneity and anisotropy (Yang et al., 2023a, 2023b; Zhao et al., 2022). By contrast, DL methods—including Long Short-Term Memory (LSTM) (Huo et al., 2022), Generative Adversarial Networks (GANs) (Tang et al., 2021), and Transformer (Sun et al., 2024)—are capable of extracting multi-level abstract features from data, enabling reservoir parameter identification and prediction without manual feature engineering (Noh et al., 2023). Despite the significant advancements in reservoir parameter prediction made by DL methods, their performance depends heavily on high-quality core data. In practice, however, such data are often scarce and costly to acquire, which can diminish the accuracy of model predictions (Lin et al., 2024).

Compared with traditional ML and DL methods, large language models (LLMs) demonstrate superior potential for handling time/depth series data with complex nonlinear characteristics. LLMs also exhibit strong few-shot learning capabilities, making them less sensitive to limited data (Garza et al., 2023; Shao et al., 2024). Recent studies have explored the application of LLMs for time series prediction and the development of geology-specific models (Jin et al., 2024; Lin et al., 2023; Xue and Salim, 2024). For instance, Jin et al. (2024) employed LLMs to handle complex token sequences and predict time series, showing that these models achieve excellent performance in both zero-shot and few-shot learning. In the geosciences, Lin et al. (2023) introduced GeoGalactica, an LLM specifically tailored for geological applications.

Typically, vertical depth information derived from well logging data is treated as time series data, reflecting local responses. Despite this, there is a paucity of research efforts directed towards the development of LLMs in geosciences, particularly for reservoir parameter prediction. Moreover, most existing approaches neglect the integration of vertical depth information with lateral reservoir features, limiting their ability to achieve high-accuracy predictions (Challenge 1). In addition, the performance of these methods in scenarios with scarce labeled data remains suboptimal (Challenge 2).

In light of these challenges, we propose LLMs-guided cross-modal alignment with feature fusion (CAF^2) framework, designed to leverage the rich knowledge encapsulated in LLMs to guide parameter prediction in tight sandstone reservoirs. The framework begins by generating synthetic datasets from the original well logging data through a data augmentation model. These synthetic datasets, combined with the principal knowledge extracted from LLMs, are then used to train the CAF^2 model. During this process, cross-modal alignment is employed to mitigate the modal differences between textual knowledge and well logging data, followed by feature fusion to integrate depth-wise and variable-wise data features effectively. Finally, the trained model is applied to unseen logging datasets for reservoir parameter prediction. Overall, the CAF^2 framework enables a comprehensive analysis and exploitation of logging data, enhancing both the accuracy and reliability of parameter predictions in tight sandstone reservoirs.

2. Motivation

In this section, we attempt to analyze the challenges faced by LLMs-guided tight sandstone reservoir parameter prediction methods. Our primary research task is to improve the accuracy and reliability of reservoir parameter predictions by addressing the key shortcomings of current studies in this field.

2.1. How can data features be effectively extracted from highly heterogeneous reservoirs, and how can LLMs be applied to well logging interpretation? (Challenge 1)

The local physical properties of a reservoir are largely determined by its sedimentary facies. Given the relative stability of the sedimentary facies within a specific timeframe, sediments during the same period tend to exhibit high similarity. This results in spatial dependence among adjacent sedimentary rocks, causing the physical properties of rocks to usually show a pattern of change with depth. Each data point sampled from well logging data along the depth contains multiple rock attribute feature variables. Typically, these well logging observations are influenced by adjacent well logging responses, indicating intrinsic interconnections among the feature variables. Therefore, models for predicting petrophysical parameters from well logging data should consider both the correlations between logging feature variables and petrophysical parameters, as well as the cross-depth dependencies inherent in the data.

To effectively integrate well logging feature information across multiple dimensions, as shown in Fig. 1, we separately extract the depth-sequence information and feature-variable information from the logging data and construct a feature fusion module to address this challenge.

Additionally, LLMs are typically pre-trained on large-scale datasets, enabling them to accumulate extensive knowledge. However, as depicted in Fig. 2, the embedding vectors of pre-trained LLM tokens and well logging data are poorly aligned. This misalignment weakens the ability of LLMs to model cross-depth sequences and, in turn, undermines feature fusion. The key to tackling this challenge lies in identifying an effective way to overcome modality misalignment.

Recent studies in various fields, including action detection (Dai et al., 2021), semantic segmentation (Vobecky et al., 2022), and speech recognition (Jin et al., 2023), have proposed the application of cross-modal knowledge distillation techniques to mitigate

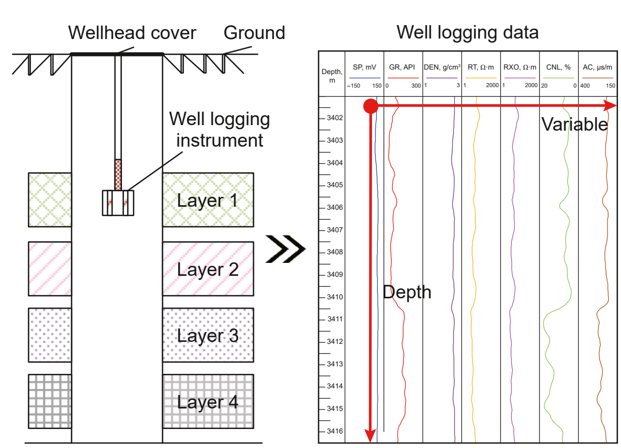

Fig. 1. The diagram of well logging.

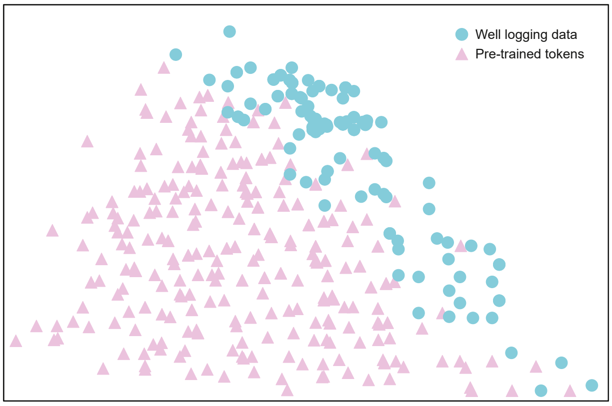

Fig. 2. The t-distributed stochastic neighbor embedding (t-SNE) visualization of well logging data and LLM-pretrained word tokens.

modality differences. However, most existing studies remain concentrated on natural language processing or computer vision, while cross-modal knowledge distillation for reservoir parameter prediction in tight sandstones has received little attention. To bridge this gap, we design a knowledge distillation and cross-modal alignment strategy that narrows the modality gap between textual knowledge and well logging data, aiding in capturing the cross-depth features of well logging data.

2.2. How can the model address insufficient implicit detailed features in the data? (Challenge 2)

In practice, the high costs of drilling and coring often result in a shortage of samples for well logging interpretation models that rely on core data. This presents a classic low-data modeling challenge, where the limited data significantly constrains the accuracy of reservoir parameter prediction. In response to this challenge, we adopt a data augmentation strategy to enrich the implicit detailed features of the data. More specifically, we employ the RealTabFormer (Solatorio and Dupriez, 2023) model, renowned for its ability to capture long-range dependences and intrinsic features within the data. RealTabFormer generates synthetic data that closely approximates the feature distribution of core samples, thereby expanding the training dataset with diverse and representative examples. This augmentation enhances the accuracy of reservoir parameter prediction.

Note that benefiting from the effective cross-modal alignment and feature fusion strategy, our CAF^2 model can achieve favorable performance in reservoir parameter prediction, even in scenarios of data scarcity without the implementation of data augmentation.

3. Methodology

In this section, we present a detailed description of the proposed CAF^2 model, highlighting its structure and functionality, as illustrated in Fig. 3. The CAF^2 model consists of four main components: data augmentation, knowledge distillation, cross-modal alignment, and feature fusion. The data augmentation module serves to enrich the implicit detailed features of the data. The knowledge distillation module extracts the principal knowledge from LLMs to guide reservoir parameter prediction. The cross-modal alignment module is designed to mitigate the modal differences between text knowledge and well logging data. Lastly, the feature fusion module is constructed to effectively extract data features from both depth and variable dimensions. Prior to detailing each component, we briefly introduce the basic principles of LLMs.

3.1. Basic principles of LLMs

LLMs are trained on vast amounts of data, including text and other modalities such as images and audio, enabling them to learn the intricacies of language, multimodal representations, and contextual relationships. The Transformer (Vaswani et al., 2017) architecture is the foundation of most modern LLMs. For instance, GPT-2, which we utilized in our work, is based on this architecture and consists solely of multiple decoder layers. Generally, LLMs follow a two-phase training process: pre-training and fine-tuning. During the pre-training phase, the model is trained on a large corpus of diverse data in an unsupervised manner. The objective is to learn general language and multimodal representations by predicting missing elements or the next element in a sequence. In the fine-tuning phase, the pre-trained model is further trained on a specific dataset with labeled examples to adapt it to a particular task or domain. Fine-tuning helps the model specialize in tasks such as text classification, time series prediction, or, in our case, reservoir parameter prediction.

To conduct reservoir parameter prediction, the LLM is fine-tuned using well logging data to bridge the modality gap between textual knowledge and well logging data. Specifically, this fine-tuning process involves utilizing a cross-modal alignment strategy to align the well logging data with the textual knowledge embedded in the LLM. Leveraging the rich knowledge within the LLM, this approach guides the prediction of reservoir parameters by identifying patterns and dependencies in the well logging data. This process involves adjusting the model's parameters to minimize prediction error. Such fine-tuning enables the model to learn complex patterns and relationships, thereby improving the accuracy and reliability of reservoir parameter predictions.

3.2. Data augmentation

RealTabFormer is a data augmentation model based on the Transformer architecture, designed to generate synthetic data that retain statistical properties similar to the original core data. Initially, the model converts the well logging data into a vocabulary form, effectively preventing the generation of invalid samples during data synthesis. It then employs an autoregressive GPT-2 model to capture the key information inherent in the well logging data. The multi-head attention mechanism of GPT-2 enables the model to focus on local details while simultaneously capturing features at varying scales, ensuring the high quality and fidelity of the synthetic data. To prevent the model from simply “memorizing” and replicating the training data, a target masking strategy is implemented to reduce the probability of data duplication, thereby bolstering the model's generalization ability. Given the original well logging data $\mathbf{X}^{\text{org}}$, the augmented data $\mathbf{X}^{\text{aug}} \in \mathbb{R}^{T\times C}$ generated by RealTabFormer can be represented as follows:

$$\mathbf{X}^{\text{aug}} = \text{RealTabFormer}\left(\mathbf{X}^{\text{org}}\right) \tag{1}$$

where T and C are the sequence length and the number of feature variables of well logging data, respectively.

RealTabFormer exhibits exceptional capability in capturing well logging data features across multiple scales, including both short-range local patterns and long-range global trends. Given the inherent long-range dependencies present in vertical-depth logging data, the model ensures that the synthetic data faithfully preserves the key characteristics of the original core data. Furthermore, it reflects the complexity and diversity inherent in

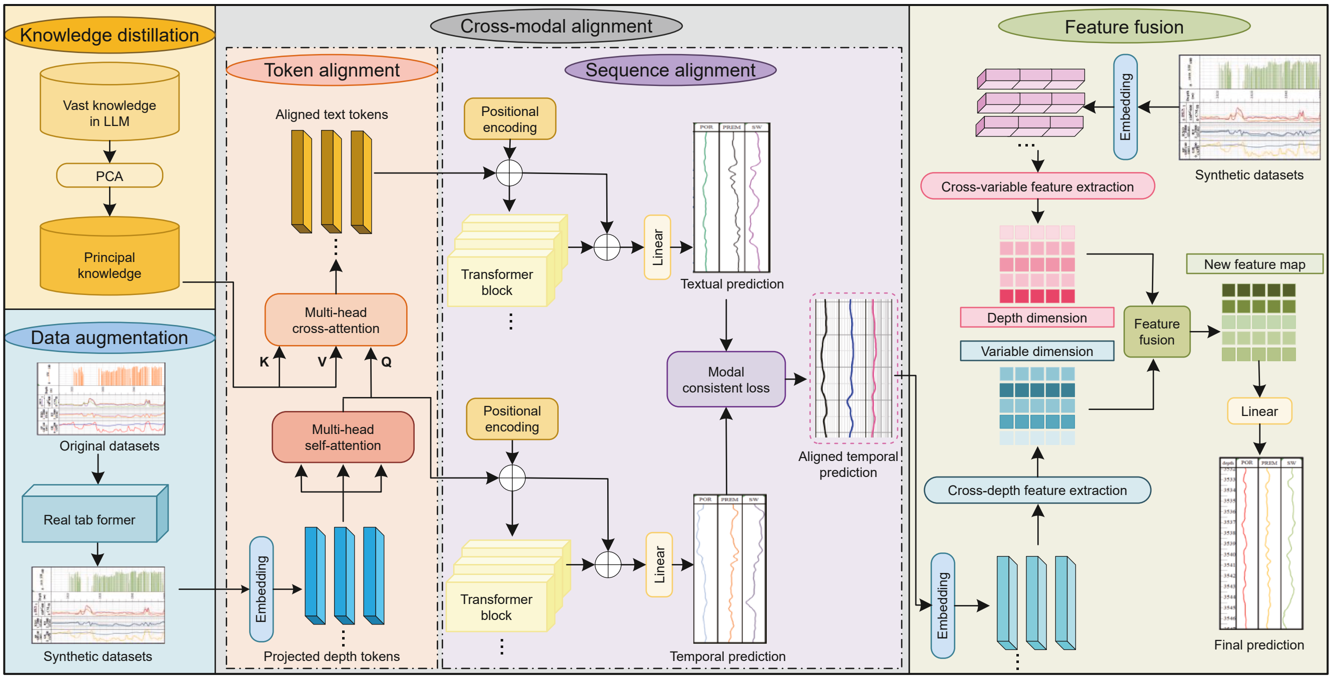

Fig. 3. Overview of the proposed CAF^2 model, comprising four main components: data augmentation, knowledge distillation, cross-modal alignment, and feature fusion.

the data, effectively addressing the challenge posed by insufficient implicit detailed features of the data.

3.3. Knowledge distillation

LLMs are typically pre-trained on large-scale datasets, thereby accumulating a vast repository of knowledge. Let the word embedding dictionary of an LLM be denoted as $E \in \mathbb{R}^{|\mathcal{A}| \times M}$, where $|\mathcal{A}|$ is the size of the vocabulary, and M is the feature dimension of the embeddings. A key challenge lies in $|\mathcal{A}|$, which is typically huge, e.g., 50257 in the GPT-2 model, and directly utilizing all vocabulary tokens for model computation entails substantial computational cost. In addition, the modal differences between LLMs' text knowledge and well logging data are considerable. Therefore, directly employing raw word embeddings to guide reservoir parameter prediction poses a significant challenge in achieving optimal performance.

To tackle these challenges, we propose a knowledge distillation strategy designed to extract the principal knowledge information to guide reservoir parameter prediction. Generally, semantically similar words form "synonym clusters". By selecting the cluster center to represent its neighboring words, the number of word entries involved in model computation can be significantly reduced. Specifically, we apply principal component analysis (PCA) to the word embedding dictionary E to perform dimensionality reduction, thereby obtaining the principal knowledge information, denoted as $\varepsilon \in \mathbb{R}^{d \times M}$:

$$\varepsilon = \mathrm{PCA}(E) \tag{2}$$

where d is the number of principal components, and $d \ll |\mathcal{A}|$.

3.4. Cross-modal alignment

The goal of cross-modal alignment is to capture the cross-depth features of well logging data. When applying LLMs for cross-depth sequence modeling, a key challenge is mitigating the inherent modality gap between the textual knowledge of LLMs and well logging data. To address this issue, we design a token alignment and sequence alignment strategy to facilitate the transfer of knowledge from LLMs, enabling the model to effectively capture the cross-depth dependencies present in well logging data.

Token alignment: Given the augmented well logging data $\mathbf{X}^{\mathrm{aug}}$ as input, it is initially fed into an embedding layer, yielding the projected depth tokens of the target variables (i.e., porosity, permeability, and water saturation). Following this, multi-head self attention (MHSA) mechanism is utilized to get the attention output representation $\mathbf{X}^{\mathrm{W}} \in \mathbb{R}^{T \times M}$ of the depth tokens sequence for the reservoir parameter:

$$\mathbf{X}^{\mathrm{W}} = \mathrm{MHSA}\left(\mathrm{Embedding}\left(\mathbf{X}^{\mathrm{aug}}\right)\right) \tag{3}$$

By leveraging $\mathbf{X}^{\mathrm{W}}$ and the principal knowledge ε extracted via knowledge distillation, with ε serving as key and value, and $\mathbf{X}^{\mathrm{W}}$ serves as query, we employ the Multi-head Cross Attention to align ε and $\mathbf{X}^{\mathrm{W}}$, thereby obtaining aligned text sequence tokens $\mathbf{X}^{\mathrm{t}} \in \mathbb{R}^{T \times M}$:

$$\mathbf{X}^{\mathrm{t}} = \mathrm{softmax}\left(\frac{\mathbf{QK}^{T}}{\sqrt{C}}\right)\mathbf{V}$$

$$\mathbf{Q} = \mathbf{X}^{\mathrm{W}}\mathbf{W}_{\mathrm{q}}, \mathbf{K} = \varepsilon\mathbf{W}_{\mathrm{k}}, \mathbf{V} = \varepsilon\mathbf{W}_{\mathrm{v}} \tag{4}$$

where $\mathbf{W}_{\mathrm{q}}$, $\mathbf{W}_{\mathrm{k}}$ and $\mathbf{W}_{\mathrm{v}} \in \mathbb{R}^{M \times M}$ denote the projection matrixes for the query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$), respectively.

Sequence alignment: The attention output $\mathbf{X}^{\mathrm{W}}$ of the depth tokens sequence and the aligned text tokens sequence $\mathbf{X}^{\mathrm{t}}$ are input into the Transformer block to calculate their hidden vector features, respectively:

$$\mathbf{H}_{l}^{\mathrm{W}} = \mathrm{Transformer}\left(\mathbf{X}^{\mathrm{W}} + \mathbf{X}_{\mathrm{PE}}^{\mathrm{W}}\right) \tag{5}$$

$$\mathbf{H}_l^{\mathrm{t}} = \mathrm{Transformer}\left(\mathbf{X}^{\mathrm{t}} + \mathbf{X}_{\mathrm{PE}}^{\mathrm{t}}\right) \tag{6}$$

where $\mathbf{H}_l^{\mathrm{W}}, \mathbf{H}_l^{\mathrm{t}}$ represent the hidden features of the depth tokens sequence and the aligned text sequence, respectively. $\mathbf{X}_{\mathrm{PE}}^{\mathrm{W}}, \mathbf{X}_{\mathrm{PE}}^{\mathrm{t}}$ represent the position encoding matrices of the depth tokens sequence and the aligned text sequence, respectively. $l \in [1, 2, \cdots, L]$ denotes the l-th layer of the Transformer.

Furthermore, the feature $\mathbf{H}_L^{\mathrm{W}}$ of the depth tokens sequence from the L-th layer and the text sequence feature $\mathbf{H}_L^{\mathrm{t}}$ are connected via residual connections with $\mathbf{X}^{\mathrm{W}}$ and $\mathbf{X}^{\mathrm{t}}$, respectively. These are then mapped through a linear layer. Ultimately, the predicted reservoir parameters sequence $\mathbf{D}^{\mathrm{W}}$, from cross-depth perspective, and the corresponding text sequence $\mathbf{D}^{\mathrm{t}}$ are calculated as follows:

$$\mathbf{D}^{\mathrm{W}} = \mathrm{Linear}\left(\mathbf{H}_L^{\mathrm{W}} + \mathbf{X}^{\mathrm{W}}\right) \tag{7}$$

$$\mathbf{D}^{\mathrm{t}} = \mathrm{Linear}\left(\mathbf{H}_L^{\mathrm{t}} + \mathbf{X}^{\mathrm{t}}\right) \tag{8}$$

The objective of the sequence alignment is to ensure the consistency of cross-modal outputs (i.e., $\mathbf{D}^{\mathrm{W}}$ and $\mathbf{D}^{\mathrm{t}}$). Therefore, the modal consistency loss is defined as:

$$\mathcal{L}^{\mathrm{seq}} = \mathrm{sim}\left(\mathbf{D}^{\mathrm{W}}, \mathbf{D}^{\mathrm{t}}\right) \tag{9}$$

where sim $(\cdot, \cdot)$ is a similarity function.

3.5. Feature fusion

Traditional well logging methods predominantly predict reservoir parameters by analyzing multiple conventional curves. However, reservoir physical properties and geological conditions are depth-dependent, and sedimentary rocks are formed through the gradual accumulation or deposition of mineral or organic particles over time. Therefore, we incorporate both the vertical depth information and the lateral feature variables inherent in well logging data, and introduce a feature fusion mechanism to enhance the accuracy of reservoir parameter prediction.

For the augmented well logging data $\mathbf{X}^{\mathrm{aug}}$ and the predicted reservoir parameters sequence $\mathbf{D}^{\mathrm{W}}$ from a cross-depth perspective, they are independently fed into an embedding layer to obtain their corresponding embedding representations. Then, cross-variable and cross-depth feature extraction are performed separately to derive the feature representations of the feature variable $\mathbf{F}^{\mathrm{aug}}$ and the reservoir parameter $\mathbf{F}^{\mathrm{W}}$:

$$\mathbf{F}^{\mathrm{aug}} = MLP(\mathrm{Embedding}(\mathbf{X}^{\mathrm{aug}})) \tag{10}$$

$$\mathbf{F}^{\mathrm{W}} = MLP\left(\mathrm{Embedding}\left(\mathbf{D}^{\mathrm{W}}\right)\right) \tag{11}$$

where *MLP* is the fully-connected neural network. Note that $\mathbf{F}^{\mathrm{aug}}$ is a shared feature representation of all feature variables at each respective depth, while $\mathbf{F}^{\mathrm{W}}$ is the feature representation of the reservoir parameter across all depths. To predict the reservoir parameter at depth t, the feature representations $\mathbf{F}_t^{\mathrm{aug}}$ and $\mathbf{F}_t^{\mathrm{W}}$ at depth t are fed into the feature fusion module, which facilitates the integration of information from these two distinct aspects. Then, the final prediction output $\mathbf{Y}_t^{\mathrm{W}}$ is derived via a linear layer:

$$\mathbf{Y}_t^{\mathrm{W}} = \mathrm{Linear}\left(Concat\left(\alpha \cdot \mathbf{F}_t^{\mathrm{aug}}, \beta \cdot \mathbf{F}_t^{W}\right)\right) \tag{12}$$

where α and β represent the learnable weight parameters, and $Concat(\cdot)$ denote the concatenation operation.

4. Experiments and results

This section provides a brief overview of the experimental data background, the workflow of the model prediction, evaluation metrics, and finally analyzes the results of the experiment.

4.1. Data source

The experimental data for our work are derived from two core wells in the Sulige gas field, where the research wells S1 and S2 are located at a considerable distance from each other, as depicted in the pentagonal area in Fig. 4. The strata under study are the 8th member of the Shihezi Formation and the 1st member of the Shanxi Formation from the Upper Paleozoic Permian system, with burial depths ranging from 3500 to 3650 m. These strata are characterized by a strongly heterogeneous, water-rich tight sandstone reservoir.

Specifically, the experimental data comprises a set of core analysis data and a corresponding set of conventional well logging data, aligned with the depth of the cores. Given that the response characteristics of formation logging are affected by the conductivity of mineral components, several factors including grain size, fractures, formation water mineralization degree, and rock pore structure, collectively impact the triple porosity curve and resistivity. Therefore, SP, GR, DEN, AC, CNL, RT, and RXO were selected as input parameters for the reservoir evaluation prediction model. Following data cleaning and outlier processing, there are more than 500 sets of data, with over 300 and 200 from well S1 and S2, respectively.

Fig. 5 presents the probability density distribution of the original and synthetic data. The comparison shows that the synthetic data closely matches the statistical characteristics of the original dataset, with consistent overall shape and patterns between the two types of data. The distribution characteristics of the synthetic data have not changed significantly, confirming that data augmentation maintains a consistent scale without altering inherent data patterns. Overall, the results reveal that the correlation patterns in the synthetic data closely mirror those in the original data, indicating that the augmentation process preserved the essential features of the original data, thus guaranteeing the authenticity of the synthetic data.

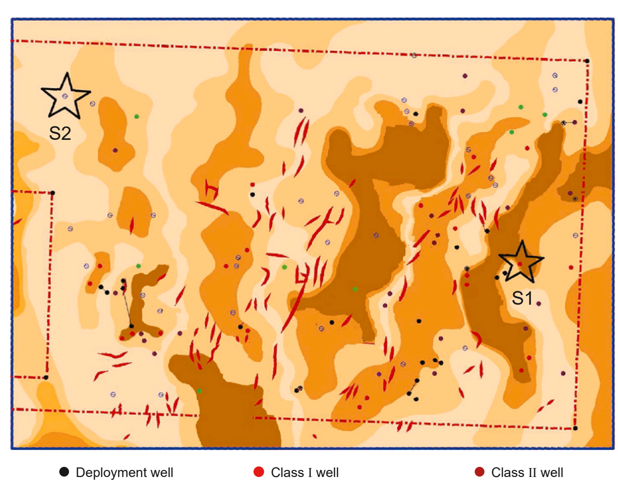

Fig. 4. The diagram of the studied core well location.

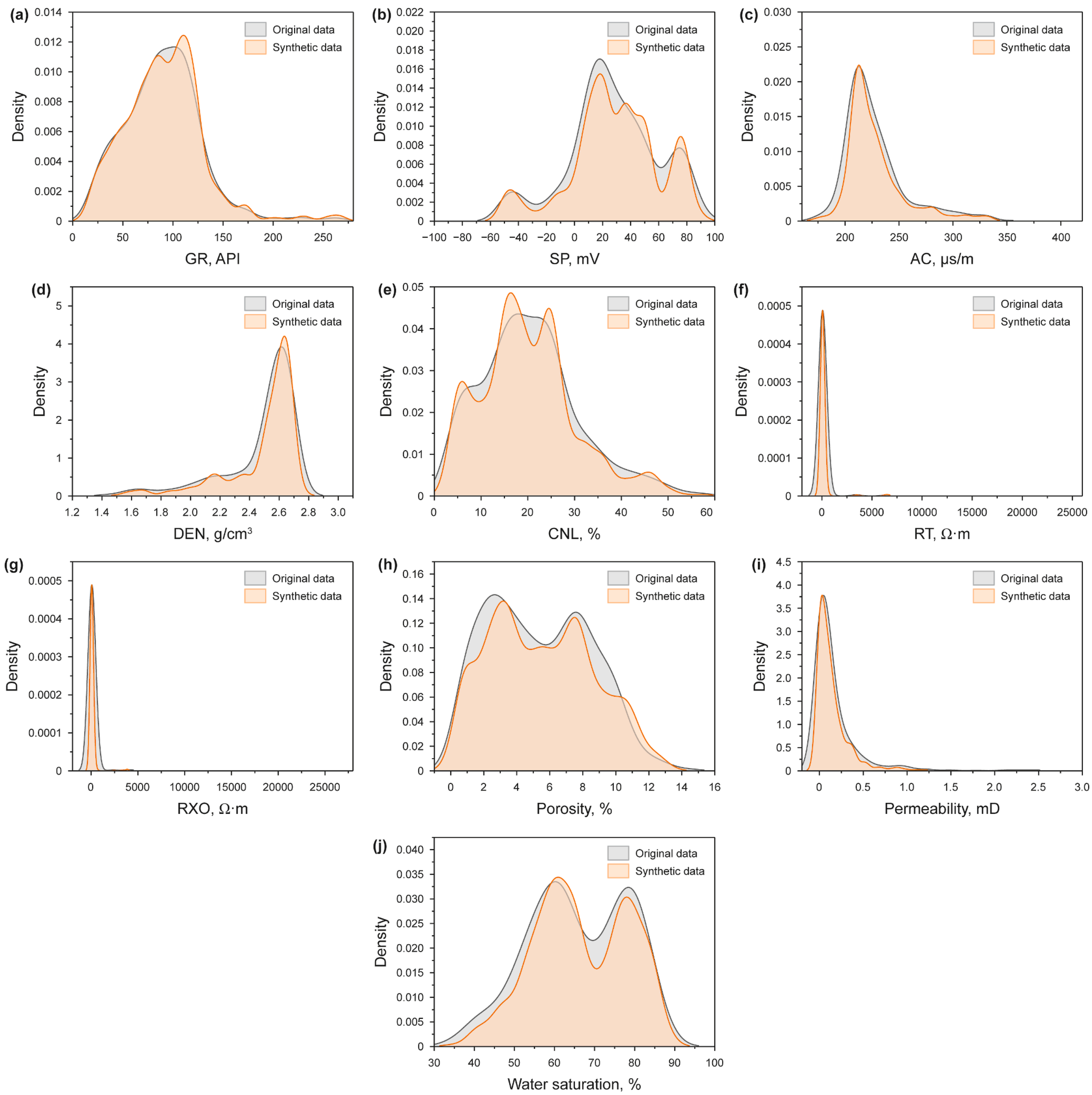

Fig. 5. Probability density distribution for original and synthetic well logging data and their predicted parameters. **(a)** GR. **(b)** SP. **(c)** AC. **(d)** DEN. **(e)** CNL. **(f)** RT. **(g)** RXO. **(h)** Porosity. **(i)** Permeability. **(j)** Water saturation.

4.2. Workflow for the CAF^2 model

During the training, the goal of the CAF^2 model is twofold: a) to minimize the supervised error between the predicted output $\mathbf{Y}^{\mathrm{W}}$ of the reservoir parameter and the actual core value; b) to minimize the sequence alignment error between the predicted reservoir parameters sequence $\mathbf{D}^{\mathrm{W}}$ (from a cross-depth perspective) and the predicted text sequence $\mathbf{D}^{\mathrm{t}}$. Therefore, the total loss function $\mathcal{L}^{\text{total}}$ during model training is defined as the weighted sum of the supervision loss $\mathcal{L}^{\text{sup}}$ and the modal consistency loss $\mathcal{L}^{\text{seq}}$:

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{sup}} + \lambda \mathcal{L}^{\text{seq}} \tag{13}$$

where λ is a hyperparameter. After the model training is completed, the final predicted output sequence $\mathbf{Y}^{\mathrm{W}}$ is compared with the actual reservoir parameters to evaluate the predictive efficacy of the proposed CAF^2 model.

Furthermore, during the optimization process, we use the L_1 loss function, which measures the average absolute difference between the predicted and true values. This loss function is particularly robust to outliers and provides a more accurate measure of prediction errors for reservoir parameters. To minimize the loss function, the gradients of the loss function with respect to the model parameters, such as the projection matrices $\mathbf{W}_\text{q}$, $\mathbf{W}_\text{k}$, and $\mathbf{W}_\text{v}$, are computed. The Adam optimizer then uses these gradients to adjust the model's parameters, aiming to reduce the loss. For our task, we set the learning rate to 0.0005 and the hyperparameter λ to 0.8 (for porosity and water saturation prediction) and 0.7 (for permeability prediction) to ensure stable and efficient convergence. Moreover, we employ an early stopping mechanism to halt the training process once the model's performance on a validation set stops improving. This helps prevent overfitting and ensures that the model retains its generalization capabilities. The algorithm for predicting parameters of tight sandstone reservoirs based on the CAF^2 model is outlined in Algorithm 1.

Algorithm 1: LLMs-guided cross-modal alignment with feature fusion

Input: Original well logging data $\mathbf{X}^\text{org}$

Output: Predicted results of the reservoir parameter $\mathbf{Y}^\text{W}$

 # data augmentation #

1. $\mathbf{X}^\text{aug} \leftarrow \text{RealTabFormer}(\mathbf{X}^\text{org})$

 # knowledge distillation #

2. $\mathcal{E} \leftarrow \text{PCA}(E)$

 # cross-modal alignment #

3. Calculate the attention output representation of the depth tokens sequence for the reservoir parameter:

$$\mathbf{X}^\text{W} \leftarrow \text{MHSA}(\text{Embedding}(\mathbf{X}^\text{aug}))$$

4. Calculate the projection matrixes:

$$\mathbf{Q} \leftarrow \mathbf{X}^\text{W}\mathbf{W}_\text{q}\,,\ \ \mathbf{K} \leftarrow \mathcal{E}\,\mathbf{W}_\text{k}\,,\ \ \mathbf{V} \leftarrow \mathcal{E}\,\mathbf{W}_\text{v}$$

5. Calculate the aligned text sequence representation:

$$\mathbf{X}^\text{t} \leftarrow \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V}$$

6. Calculate the hidden vector features:

$$\mathbf{H}_l^\text{W} \leftarrow \text{Transformer}(\mathbf{X}^\text{W} + \mathbf{X}_\text{PE}^\text{W})\,;\ \ \mathbf{H}_l^\text{t} \leftarrow \text{Transformer}(\mathbf{X}^\text{t} + \mathbf{X}_\text{PE}^\text{t})$$

7. Predict reservoir parameters sequence from a cross-depth perspective:

$$\mathbf{D}^\text{W} \leftarrow \text{Linear}(\mathbf{H}_L^\text{W} + \mathbf{X}^\text{W})\,;\,\mathbf{D}^\text{t} \leftarrow \text{Linear}(\mathbf{H}_L^\text{t} + \mathbf{X}^\text{t})$$

 # feature fusion #

8. Extract cross-variable and cross-depth features:

$$\mathbf{F}^\text{aug} \leftarrow MLP(\text{Embedding}(\mathbf{X}^\text{aug}))\,;\,\mathbf{F}^\text{W} \leftarrow MLP(\text{Embedding}(\mathbf{D}^\text{W}))$$

9. Predict the final reservoir parameters sequence:

$$\mathbf{Y}^\text{W} \leftarrow \text{Linear}(Concat(\alpha\bullet\mathbf{F}_t^\text{aug}, \beta\bullet\mathbf{F}_t^\text{W}))$$

4.3. Evaluation metrics

The performance of the proposed CAF^2 model is evaluated using mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R). Both MAE and RMSE quantify the difference between the predicted and actual values, with lower values indicating less prediction error from the model. Assume that there are n groups of well logging samples, where the actual values are denoted as $y_1, y_2, ..., y_n$, and the predicted values are represented as $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n$. The calculation of each evaluation metric is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{15}$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{16}$$

4.4. Experimental results

In this study, we initially conducted data augmentation on the original well logging data, ensuring that the synthesized data does not duplicate the original well logging data. Then, the synthesized data is selected as the training set, while the original well logging data is utilized as the test set, aiming to validate the performance of the model in predicting the parameters of tight sandstone reservoirs.

Specifically, the CAF^2 model is applied to predict porosity, permeability, and water saturation on the test set. Fig. 6 presents a comparison of the actual versus the predicted values. An intuitive observation from each sub-figure reveals that the CAF^2 model has demonstrated commendable predictive results across these three prediction tasks. The porosity predominantly ranges between 3% and 10%, indicating generally low porosity characteristics. Similarly, the permeability of the reservoir is primarily low, with values between 0.001 mD and 0.5 mD constituting the majority. In contrast, the water saturation levels in the reservoir are typically high, mostly falling between 50% and 80%. More concretely, from the perspective of curve fitting, the prediction of water saturation is the best. The primary prediction errors for water saturation stem from the presence of extrema, a phenomenon also observed in the prediction of porosity and permeability. The reason behind this may be attributed to the fluctuating values of reservoir parameters at adjacent depths, which the model struggles to capture perfectly. Additionally, the actual values of permeability contain some zero values, potentially impacting the model's predictive performance for permeability. However, this prediction outcome is understandable to a certain extent, as it reflects the inherent characteristics of the data and the limitations of the model in addressing these characteristics. Despite this, the proposed CAF^2 model still succeeds in yielding satisfactory results for reservoir parameter prediction.

Furthermore, Fig. 7 illustrates the histogram and the distribution of the probability density function (PDF) for prediction errors. The accuracy of the prediction results is evaluated by the distribution of the PDF curve concentrated around zero and the frequency of counts around zero. From Fig. 7, it can be observed that the distribution of the PDF curves for prediction errors associated with porosity, permeability, and water saturation are all concentrated around zero. Moreover, the histogram of prediction error exhibits a high frequency around zero, indicating that the range of prediction errors is relatively narrow. To quantify the predictive performance of the proposed CAF^2 model, the evaluation metrics of the CAF^2 model are also depicted in each sub-figure. In the prediction of porosity, the MAE, RMSE, and R are 0.4828, 0.9133,

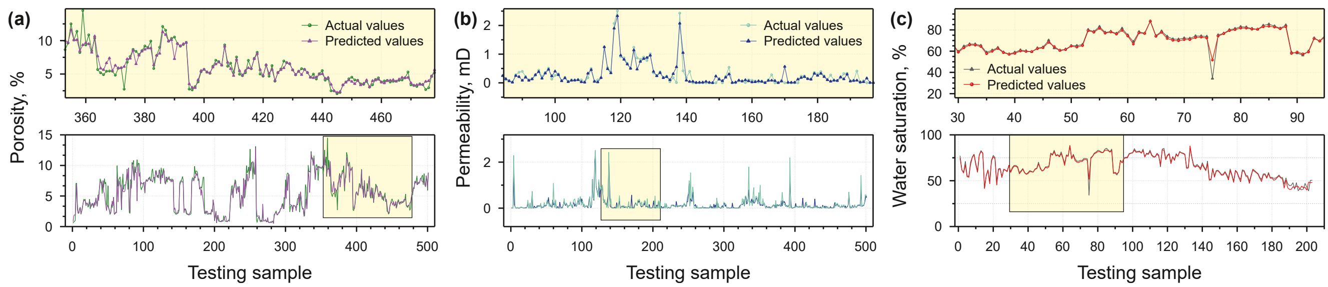

Fig. 6. Prediction results by our CAF^2 model on the test set. **(a)** Porosity, **(b)** permeability, **(c)** water saturation.

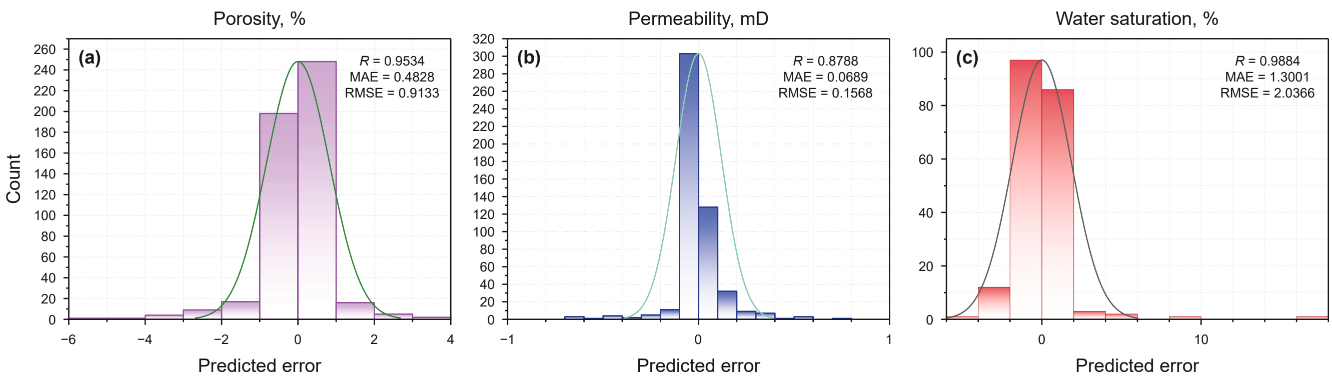

Fig. 7. The histogram of prediction errors. **(a)** Porosity, **(b)** permeability, **(c)** water saturation.

and 0.9534, respectively. For permeability prediction, these metrics are 0.0689, 0.1568, and 0.8788, respectively. And for water saturation prediction, the MAE, RMSE, and R are 1.3001, 2.0366, and 0.9884, respectively. These results collectively validate the effectiveness and feasibility of the proposed CAF^2 model.

5. Performance analysis

In this section, we conducted a detailed analysis of the CAF^2 model's performance. The necessity and significance of each designed module were mainly discussed. Furthermore, we highlight the advantages of the proposed CAF^2 model compared to other benchmarks. Finally, the visualization of the prediction results was demonstrated in the practical field gas wells of the research area.

5.1. Ablation analysis

To evaluate the contributions of the individually designed modules within CAF^2 model, we implemented ablation studies across various modules on the same test set (comprising 140 samples each for porosity and permeability, and 48 samples for water saturation). Table 1 presents the evaluation metrics derived from these ablation studies for each module. From the table, we observe that for the porosity prediction task, the MAE, RMSE, and R values of the model without the cross-variable and cross-depth fusion modules are 0.5030, 0.7586, 0.9382, and 0.6751, 1.0290, 0.8776, respectively. The MAE, RMSE, and R values without the loss function module are 0.9417, 1.3734, and 0.7816. Without the data augmentation module, the MAE, RMSE, and R values are 0.8324, 1.1801, and 0.8389. When these modules are used together, the model achieves optimal performance, with the lowest MAE and RMSE values of 0.3832 and 0.6665, and an R value of 0.9515. Similar results were obtained for the permeability and water saturation prediction tasks. Analysis of these results indicates that the data augmentation and loss function modules have a significant impact on the model's performance, followed by the feature fusion module. More importantly, the overall performance of the model is enhanced through the implementation of feature fusion technology.

Fig. 8 shows the cross-plot of the prediction results from the ablation studies conducted on the feature fusion module. By comparison, we can see that removing any cross-variable or cross-depth feature extraction module will lead to model performance

Table 1
The ablation studies of the individually designed modules with CAF^2 module.

Models	Porosity			Permeability			Water saturation		
	MAE	RMSE	R	MAE	RMSE	R	MAE	RMSE	R
CAF^2	0.3832	0.6665	0.9515	0.0641	0.1526	0.8378	1.9889	2.6175	0.9353
-w/o cross-variable FE	0.5030	0.7586	0.9382	0.0728	0.2156	0.5894	2.0503	2.8123	0.9243
-w/o cross-depth FE	0.6751	1.0290	0.8776	0.0747	0.2314	0.5749	2.4012	4.0077	0.8173
-w/o modality consistency loss	0.9417	1.3734	0.7816	0.0821	0.2245	0.5388	2.8210	3.4143	0.9243
-w/o data augmentation	0.8324	1.1801	0.8389	0.1115	0.2312	0.4826	3.0680	3.9012	0.8268

Note: "-w/o" denotes "without". "FE" denotes feature extraction.

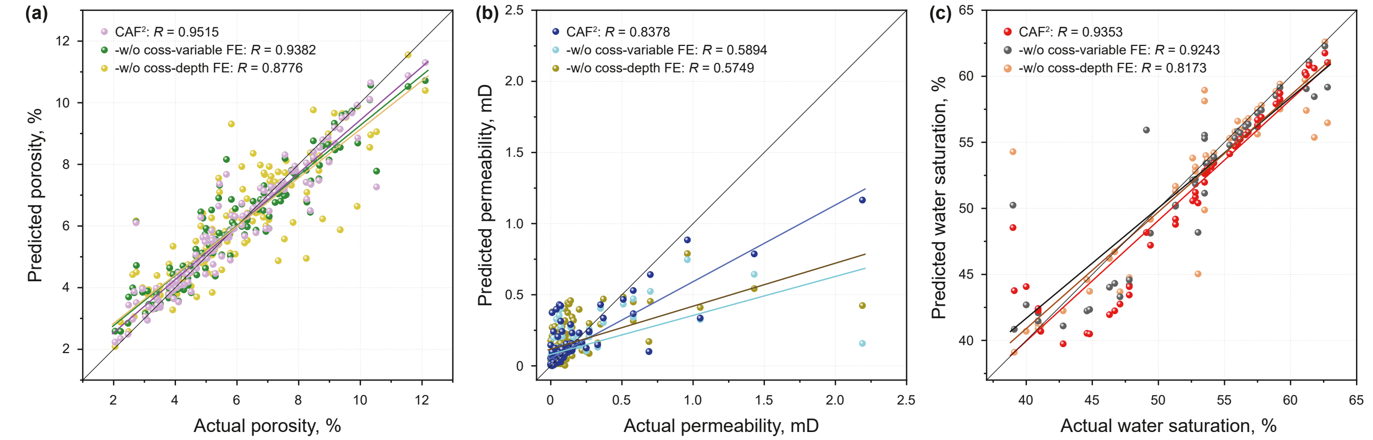

Fig. 8. The visualization results of ablation studies on feature fusion. FE denotes feature extraction. **(a)** Porosity, **(b)** permeability, **(c)** water saturation.

degradation. This is because CAF^2 model integrates the correlation of depth sequence well logging data. Such a structural design optimizes the extraction of key feature information from a cross-depth perspective, while eliminating interference factors during model learning. In addition, our findings indicate that predictions based on cross-depth contribute more significantly than those based on cross-variable (refer to Table 1). The R values of the models without the cross-depth model and the cross-variable model in the tasks of porosity, permeability, and water saturation are 0.9382, 0.5894, 0.9243, and 0.8776, 0.5749, 0.8173, respectively. The model's performance in the permeability task is suboptimal due to the strong heterogeneity of the reservoir in the study area. At the same porosity, the permeability varies greatly. However, the feature fusion technology effectively addresses the issue of insufficient model learning characteristics caused by strong heterogeneity. Consequently, following feature fusion, the CAF^2 model exhibits enhanced accuracy and reliability in predicting reservoir parameters.

Fig. 9 depicts the cross-plot of the prediction results from the ablation studies performed on the modal consistency loss. From the figure, we can find that with the introduction of modal consistency loss, the R values for the predicted results of porosity, permeability, and water saturation increased by 0.1699, 0.2990, and 0.0110, respectively. This demonstrates that the modality gap between well logging data and text knowledge can be significantly reduced by cross-modal alignment, where the modality consistency loss ensures the output consistency across different modalities.

Fig. 10 displays the scatter plot of prediction residuals from the ablation studies conducted on the data augmentation module. The figure reveals that the incorporation of the data augmentation module enhances the prediction accuracy across all three reservoir parameter prediction studies. Specifically, the MAE values for predicting porosity, permeability, and water saturation decreased by 0.4492, 0.0474, and 1.0791, respectively. Notably, a significant reduction in error is observed when permeability is high. This further illustrates the applicability of the data augmentation module, which effectively addresses the challenge of data scarcity in well logging data.

5.2. Comparative analysis

To further analyze the predictive performance of the CAF^2 model, a comparative analysis of various predictive models is listed in Table 2. In the prediction of porosity, the CAF^2 model's MAE, RMSE, and R values during the testing phase are 0.4828, 0.9133, and 0.9534, respectively. These R values are higher by 0.1895, 0.1337, 0.2605, 0.0396, and 0.0474 compared to the Support Vector Machine (SVM) (Moosavi et al., 2022), Convolutional Neural Networks (CNN) (Sun et al., 2024), LSTM, Random Forest (RF) (Li et al., 2023), and eXtreme Gradient Boosting (XGBoost) (Pan et al., 2022) models, respectively. In predicting permeability, the CAF^2 model's MAE, RMSE, and R values are 0.0690, 0.1568, and 0.8788,

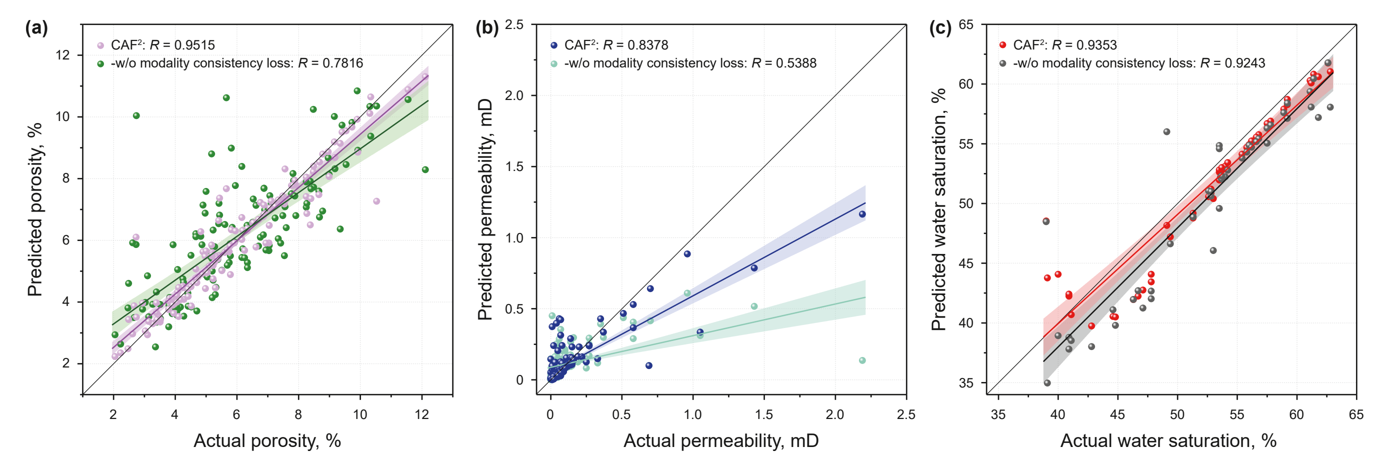

Fig. 9. The visualization results of ablation studies on cross-modal alignment. **(a)** Porosity, **(b)** permeability, **(c)** water saturation.

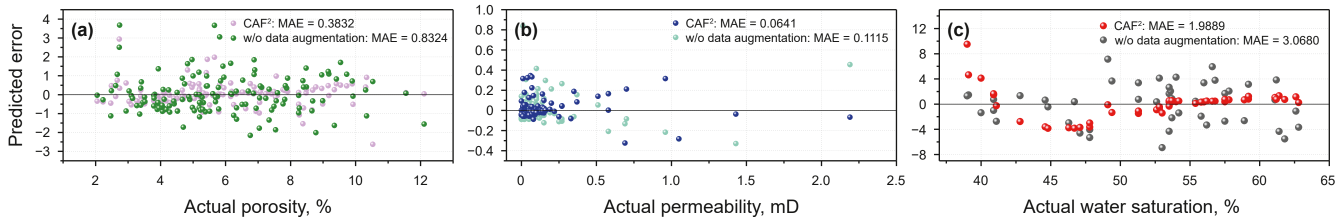

Fig. 10. The visualization results of ablation studies on data augmentation. **(a)** Porosity, **(b)** permeability, **(c)** water saturation.

Table 2
Performance comparison of reservoir parameter prediction with various competitive models. DA denotes data augmentation.

Models	Porosity			Permeability			Water saturation		
	MAE	RMSE	*R*	MAE	RMSE	*R*	MAE	RMSE	*R*
SVM-w/o DA	1.7474	2.2366	0.3075	0.1695	0.3057	0.1858	19.7185	20.8103	0.3865
CNN-w/o DA	2.1752	2.5176	0.4312	0.1587	0.2832	0.1783	20.8910	21.7833	0.3818
LSTM-w/o DA	1.7085	2.1629	0.3180	0.1847	0.2921	0.1862	16.3033	17.6034	0.0667
RF-w/o DA	1.4951	1.8531	0.6954	0.1842	0.3319	0.2098	7.3861	8.6466	0.6027
XGBoost-w/o DA	1.6445	2.0462	0.4157	0.2152	0.4644	0.1601	4.6424	5.3806	0.6707
CAF^2-w/o DA	0.8324	1.1801	0.8389	0.1115	0.2312	0.4826	3.0680	3.9012	0.8268
SVM	1.4989	1.9363	0.7639	0.1501	0.2996	0.3165	5.9298	8.5845	0.7250
CNN	1.2892	1.7747	0.8197	0.1451	0.2748	0.5149	5.9797	8.2645	0.7471
LSTM	1.6702	2.1589	0.6929	0.1575	0.2700	0.4341	5.6307	7.8955	0.7728
RF	0.7374	1.2076	0.9138	0.1040	0.2311	0.6953	2.8088	5.2949	0.9035
XGBoost	0.8032	1.2588	0.9060	0.1050	0.2151	0.7518	3.1496	5.3243	0.9022
CAF^2	0.4828	0.9133	0.9534	0.0690	0.1568	0.8788	1.3001	2.0366	0.9884

respectively, showcasing it as the superior model. For water saturation prediction, the CAF^2 model's *R* value is 0.9884, which surpasses the SVM, CNN, LSTM, RF, and XGBoost models by 0.2634, 0.2413, 0.2156, 0.0849, and 0.0862, respectively. These findings demonstrate that the proposed CAF^2 model exhibits superior predictive performance compared to other classical networks.

Despite the fact that the prediction results of RF and XGBoost do not surpass those of the CAF^2 model, their performance is nonetheless considerable due to their powerful ensemble learning computing methods. However, CNN and LSTM exhibit larger prediction errors. Compared with the CAF^2 model, the MAE for porosity prediction is higher by 0.8064 and 1.1874, respectively; the MAE for permeability prediction is higher by 0.0761 and 0.1157; and the MAE for water saturation prediction is higher by 4.6796 and 4.3306. The poor predictive performance of CNN and LSTM is primarily because they struggle to capture the nonlinear complex relationships within the data. The evaluation metrics delineated in the table reveal that the CAF^2 model consistently exhibits the lowest prediction error, regardless of the application of data augmentation. Note that without data augmentation, other methods falter in data scarcity scenarios, whereas our CAF^2 model significantly outperforms other benchmarks. Experimental results show that the CAF^2 model maintains its competitive results in reservoir parameter prediction, even in data scarcity logging scenarios. This demonstrates the significance and irreplaceability of the designed cross-modal alignment and feature fusion module, which effectively alleviates the issue of data scarcity in reservoir evaluation of well logging data.

Figs. 11–13 show the cross-plot between the actual core and the predicted values of various models following data augmentation. The data points are primarily distributed near the diagonal in Fig. 11, indicating that benchmarks are capable of predicting porosity, but our CAF^2 model performs the best. Notably, when the porosity value is greater than 7, the model's prediction results are generally underestimated. Fig. 12 shows the permeability prediction results, where it is evident that the prediction results of benchmarks deviate significantly from the actual values. The fitting line of our CAF^2 model aligns most closely with the diagonal. From Fig. 13, it can be observed that the prediction results of the RF and XGBoost are generally overestimated. The fitting line of our CAF^2 model essentially coincides with the diagonal, with slight error due to the presence of extremum. Overall, the proposed CAF^2 model outperforms other benchmarks in reservoir parameter prediction following data augmentation.

To further demonstrate the superiority of our CAF^2 method, we have compared it with advanced LLMs such as PromptCast (Xue and Salim, 2024) and TIME-LLM (Jin et al., 2024), as shown in Table 3. PromptCast transforms numerical inputs and outputs into prompts, framing the forecasting task as a sentence-to-sentence translation. It leverages language models, inspired by the success of pre-trained language models, to incorporate context and semantics around historical data. TIME-LLM reprograms input time series with text prototypes and utilizes the frozen LLM to generate outputs. It employs a prompt-as-prefix approach to enhance the input context and guide the transformation of reprogrammed input patches.

From Table 3, we observe that our method outperforms the advanced PromptCast and TIME-LLM methods across all three evaluation metrics: MAE, RMSE, and *R*. This superiority may be attributed to the fact that PromptCast uses explicit natural language prompts to characterize logging curves. However, it is challenging to describe a logging curve accurately and comprehensively through limited natural language. In contrast, our CAF^2 model adopts an implicit approach to learn knowledge from the LLM, effectively capturing the complex patterns and nonlinear relationships within the data. This results in superior performance in predicting reservoir parameters, as evidenced by the lower MAE and RMSE values and higher *R* values. Different from PromptCast, TIME-LLM employs a linear weight to implicitly learn prompts describing logging curves. However, it does not extract the primary knowledge from LLMs nor align this knowledge with the logging curves before inputting them into the LLMs, resulting in inferior

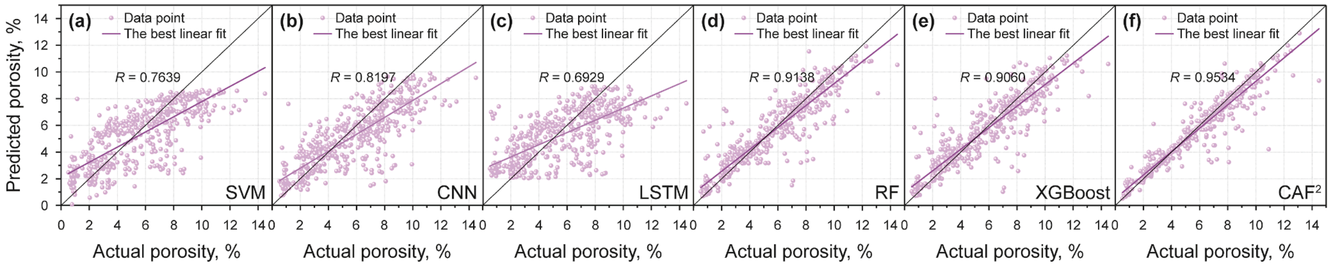

Fig. 11. Cross-plot of actual and predicted porosity with various models. **(a)** SVM. **(b)** CNN. **(c)** LSTM. **(d)** RF. **(e)** XGBoost. **(f)** CAF^2.

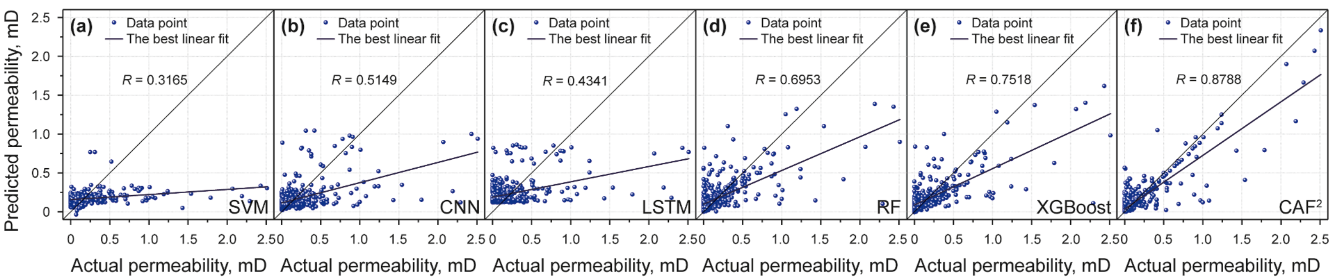

Fig. 12. Cross-plot of actual and predicted permeability with various models. **(a)** SVM. **(b)** CNN. **(c)** LSTM. **(d)** RF. **(e)** XGBoost. **(f)** CAF^2.

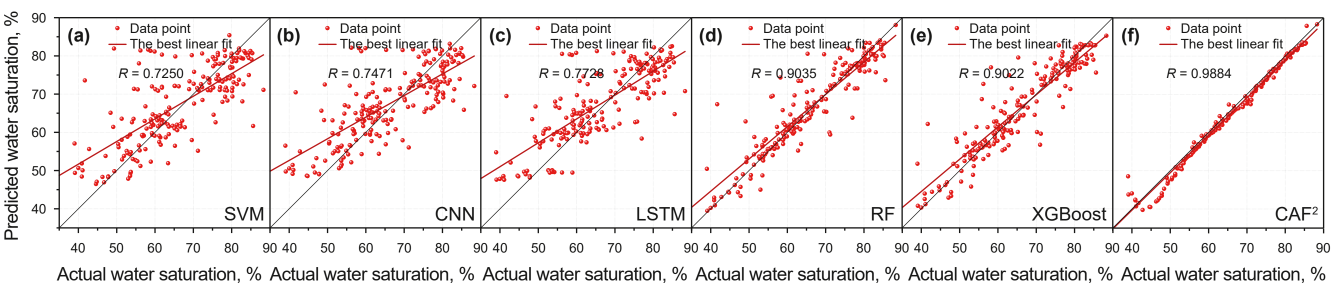

Fig. 13. Cross-plot of actual and predicted water saturation with various models. **(a)** SVM. **(b)** CNN. **(c)** LSTM. **(d)** RF. **(e)** XGBoost. **(f)** CAF^2.

Table 3
Performance comparison of reservoir parameter prediction with other advanced LLMs.

Models	Porosity			Permeability			Water saturation		
	MAE	RMSE	*R*	MAE	RMSE	*R*	MAE	RMSE	*R*
PromptCast	1.1145	1.5940	0.8482	0.0998	0.2414	0.6565	3.6497	5.9036	0.8329
TIME-LLM	0.6968	1.2121	0.9127	0.0776	0.1741	0.8189	2.5071	4.3349	0.9236
CAF^2	0.4828	0.9133	0.9534	0.0690	0.1568	0.8788	1.3001	2.0366	0.9884

outcomes due to the strong heterogeneity of the reservoir. This further demonstrates our CAF^2 model's ability to more effectively handle strong heterogeneity and data scarcity, leading to more accurate predictions in reservoir parameter evaluation tasks.

5.3. Case demonstration

The S2 well, located in the study area, is a water-producing well within the Sulige gas field, primarily composed of tight gas sandstone. Fig. 14 illustrates the prediction results of various models for the S2 well following data augmentation. The reservoir physical properties improve as rock grain size increases, with grain size exerting a significant control on these properties. GR is sensitive to lithology, decreasing as lithological quality improves. The experimental results indicate that the reservoir porosity values predominantly range from 3% to 10%, and the permeability values primarily range from 0.001 mD to 0.5 mD, characteristic of typical tight sandstone reservoirs with low porosity and low permeability.

As can be inferred from the figure, the reservoir parameter predictions made by several models are generally consistent with core analysis results. However, the reservoir heterogeneity in the study area is relatively pronounced, and the permeability range corresponding to the same porosity is quite large. Predictions of reservoir parameters using the CNN model exhibit significant errors in actual data interpretation. The RF model demonstrates an improvement in predictive results; however, its accuracy is constrained by data scarcity and the strong heterogeneity issues prevalent in this study area. The proposed CAF^2 model effectively addresses these challenges, thereby enhancing the model's predictive performance for reservoir parameters. Compared to

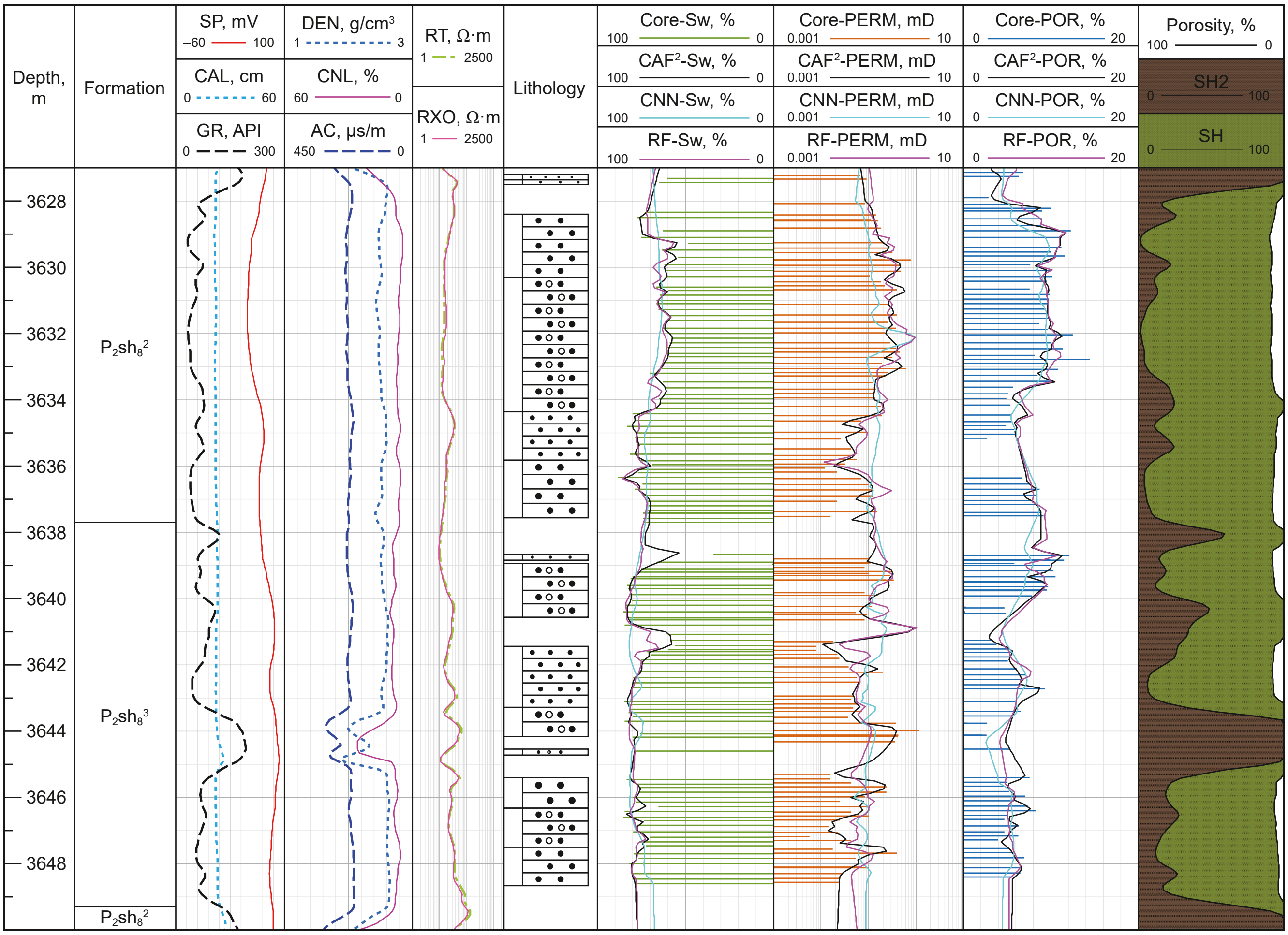

Fig. 14. Visualization results of reservoir parameters predicted by various models for the S2 well in the Sulige gas field, where SP, CAL, GR, DEN, CNL, AC, RT, RXO, Sw, PERM, POR, and SH denote spontaneous potential, caliper log, gamma ray log, density log, compensated neutron log, acoustic log, true formation resistivity, flushed zone formation resistivity, water saturation, permeability, porosity, and shale volume, respectively.

other models, the prediction results of the CAF^2 model align most closely with the actual results, which also confirms the significant applicability of the proposed CAF^2 model for reservoir studies in this area.

5.4. Parameters optimization analysis

The weight functions α and β are learnable parameters within our feature fusion network. These parameters are optimized through the process of calculating the loss and performing back-propagation. Specifically, the final optimized values obtained after training for the three prediction tasks are as follows: for porosity, the optimized weight parameters are $\alpha = 0.4664$ and $\beta = 0.4902$; for permeability, the values are $\alpha = 0.4375$ and $\beta = 0.4626$; and for water saturation, the optimized parameters are $\alpha = 0.4586$ and $\beta = 0.4937$.

In our study, the hyperparameter λ is primarily used to control and balance the supervision loss $\mathcal{L}^{\text{sup}}$ and the modal consistency loss $\mathcal{L}^{\text{seq}}$. To determine the optimal λ, we employed the grid search method within the range [0.1, 1], with an interval of 0.1, independently searching for the optimal parameter for the three prediction tasks. Given that we utilize the L$_1$ loss function to optimize the model, we use MAE to evaluate the optimal hyperparameter λ. The λ value that yields the lowest MAE is selected as the hyperparameter value. The experimental results are presented in Fig. 15. From the results, we observe that for the porosity and water saturation prediction tasks, the lowest MAE corresponds to $\lambda = 0.8$. For the permeability prediction task, the lowest MAE is achieved when $\lambda = 0.7$. Therefore, it is suggested to select $\lambda = 0.8$ for the

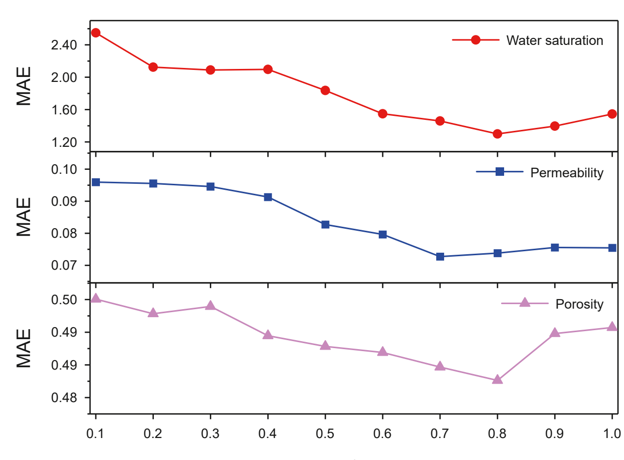

Fig. 15. Impact of λ on MAE for different prediction tasks.

porosity and water saturation prediction and $\lambda = 0.7$ for the permeability prediction.

5.5. Run time analysis

To evaluate the computational costs, we compare the run time of each method during training and testing, as presented in Table 4. Note that all results are obtained using a training data population of 2500 and a test data population of 510, ensuring that all methods are evaluated under the same conditions for a fair comparison.

From the table, it can be observed that among all methods, the training and testing times for the SVM model are the shortest. This efficiency is attributed to SVM's relatively straightforward network structure, where the training process involves finding the optimal hyperplane using convex optimization techniques. Additionally, the testing phase in SVM only requires calculating the hyperplane, making it computationally inexpensive. For RF and XGBoost, the training and testing times are slightly higher compared to SVM due to the ensemble nature of these methods. RF involves creating and averaging multiple decision trees, and XGBoost sequentially builds decision trees to minimize prediction errors, adding complexity to the training process. CNN and LSTM networks exhibit significantly higher training times due to their complex architectures. CNNs involve extensive computations such as convolutions and pooling operations, while LSTMs are designed to capture long-term dependencies in sequential data, requiring backpropagation through time during training. Lastly, our CAF^2 method, while consuming considerable time, remains within acceptable limits. The increased computational cost is due to the method's capability to thoroughly learn data features from both cross-variable and cross-depth perspectives. This comprehensive learning process enhances the prediction of reservoir parameters, significantly outperforming other methods.

6. Discussion

6.1. Advantages and limits of LLMs

LLMs, such as GPT-2, leverage extensive pre-training on diverse datasets, enabling them to capture complex contextual relationships and provide nuanced predictions. This gives them an edge over methods like RF and XGBoost, which may struggle with intricate patterns. While SVMs are effective for linear and nonlinear regression and are relatively easier to interpret with lower computational resource requirements than LLMs, they may also struggle with complex patterns. Additionally, the pre-training phase allows LLMs to learn general feature representations, making them more robust when dealing with limited labeled data during fine-tuning. This is particularly advantageous compared to CNN and LSTM models, which often require large amounts of labeled data for effective training. Overall, LLMs provide rich contextual understanding and offer flexibility and robustness in the face of data scarcity, particularly enhancing the prediction of reservoir parameters by capturing complex contextual dependencies. However, they require high computational resources and are more complex to interpret compared to other vanilla machine learning methods.

Table 4
Run time of each method in training and testing datasets.

Parameter	SVM	CNN	LSTM	RF	XGBoost	CAF^2
Train time, s	0.71	18.82	20.89	2.78	1.75	53.76
Test time, s	0.12	0.40	0.39	0.15	0.42	3.16

6.2. Insights and future work

In our study, we employed well logging data and core data from the Sulige gas field to predict key reservoir petrophysical parameters using machine/deep learning models such as SVR, CNN, RF, and CAF^2. Various evaluation criteria were employed to compare the performance of these models. The results demonstrate that our proposed CAF^2 model achieved the highest prediction accuracy for three petrophysical parameters, significantly outperforming other classical models. The established CAF^2 model significantly conserved manpower and material resources, overcoming the limitations of classical machine learning models that struggle with accurate prediction due to limited data and complex reservoir conditions, and providing reliable petrophysical parameters for reservoir well logging interpretation.

While the model developed in our study represents significant advancements over previous research methodologies, some limitations remain to be addressed. Generally, physical properties of rocks exhibit homogeneity within the same sedimentary facies, while significant disparities may be observed across different facies. In future research, we plan to further explore and design reservoir parameter prediction models that are versatile enough to accommodate various sedimentary facies, thereby addressing the influence of compaction on reservoir parameters. This endeavor, however, will involve the collection and rigorous analysis of a large amount of data on various types of sedimentary facies and corresponding well logging data, to ensure the robustness and efficacy of model training.

7. Conclusion

In this paper, we introduce a data-driven and LLMs-guided cross-modal alignment intelligent reservoir parameter prediction model (CAF^2) using well logging data. This model is designed to predict porosity, permeability, and water saturation parameters of tight sandstone reservoirs in the Sulige gas field, achieving high-precision predictions. The conclusions are as follows:

(1) By applying our proposed CAF^2 method to the tasks of predicting porosity, permeability, and water saturation in the Sulige gas field, we achieved MAE of 0.48 for porosity, 0.07 for permeability, and 1.3 for water saturation. The prediction errors are significantly lower than those of other models. These findings indicate that introducing LLMs into the field of geophysical exploration can greatly reduce exploration costs and provide a promising approach to enhance the accuracy of reservoir evaluations.
(2) To address the challenge of scarce core sample data, we have introduced a data augmentation method based on the RealTabFormer model. This approach effectively expands the number of training samples, laying a foundation for accurate prediction of reservoir parameters. Ablation experiments have shown that the introduction of this module significantly enhances the prediction performance, highlighting its necessity in scenarios with limited data samples.
(3) The designed cross-modal alignment method effectively mitigates the modality gap between well logging data and textual knowledge, facilitating the model's learning of depth features in well logging data. Additionally, the feature fusion technique fully utilizes the lateral and vertical information of well logging data, effectively overcoming the challenges posed by reservoir heterogeneity.

(4) The analysis of results indicates that the reservoir in this area is characterized as a typical tight sandstone reservoir with low porosity and low permeability. Porosity mainly ranges from 3% to 10%, and permeability is primarily distributed between 0.001 mD and 0.5 mD, demonstrating features of low porosity and low permeability. There is a significant positive correlation between permeability and porosity, identifying it as a typical pore-type reservoir. Our CAF^2 method is highly effective for such complex reservoir structures and achieves satisfactory parameter prediction results.

CRediT authorship contribution statement

Juan Wu: Writing – original draft, Conceptualization, Writing – review & editing, Methodology. **Ren-Ze Luo:** Funding acquisition, Supervision. **Lei Luo:** Writing – review & editing, Software, Validation. **Can-Ru Lei:** Writing – review & editing, Visualization. **Xing-Ting Chen:** Writing – review & editing, Data curation.

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.

Acknowledgement

We thank the editors and anonymous reviewers for their insightful comments and helpful suggestions. This research was supported by the Special Funds for Science and Technology Planning Projects of Sichuan Province (No. 2024YFHZ0158), and we would like express our heartfelt gratitude.

References

Bagheri, M., Rezaei, H., 2019. Reservoir rock permeability prediction using SVR based on radial basis function kernel. Carbonates Evaporites 34, 699–707. https://doi.org/10.1007/s13146-019-00493-4.

Bagheri, M., Riahi, M.A., 2015. Seismic facies analysis from well logs based on supervised classification scheme with different machine learning techniques. Arab. J. Geosci. 8, 7153–7161. https://doi.org/10.1007/s12517-014-1691-5.

Bagheri, M., Riahi, M.A., 2017. Modeling the facies of reservoir using seismic data with missing attributes by dissimilarity based classification. J. Earth Sci. 28, 703–708. https://doi.org/10.1007/s12583-017-0797-6.

Bai, Y., Tan, M., Shi, Y., et al., 2022. Regression committee machine and petrophysical model jointly driven parameters prediction from wireline logs in tight sandstone reservoirs. IEEE Trans. Geosci. Remote Sens. 60, 1–9. https://doi.org/10.1109/TGRS.2020.3041366.

Chen, Y., Tang, S., Xi, Z., et al., 2024. Estimation of shale adsorption gas content based on machine learning algorithms. Gas Sci. Eng. 127, 205349. https://doi.org/10.1016/j.jgsce.2024.205349.

Dai, R., Das, S., Bremond, F., 2021. Learning an augmented RGB representation with cross-modal knowledge distillation for action detection. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV, pp. 13033–13044. https://doi.org/10.1109/ICCV48922.2021.01281.

Feng, S., Xie, R., Radwan, A.E., et al., 2023. Accurate determination of water saturation in tight sandstone gas reservoirs based on optimized Gaussian process regression. Mar. Petrol. Geol. 150, 106149. https://doi.org/10.1016/j.marpetgeo.2023.106149.

Garza, A., Challu, C., Mergenthaler-Canseco, M., 2023. TimeGPT-1. https://doi.org/10.48550/arXiv.2310.03589.

Goliatt, L., Saporetti, C., Pereira, E., 2023. Super learner approach to predict total organic carbon using stacking machine learning models based on well logs. Fuel 353, 128682. https://doi.org/10.1016/j.fuel.2023.128682.

Huang, L., Qi, Y., Chen, W., et al., 2023. A geomechanical modeling method for shale oil reservoir cluster well area based on GridSearchCV. Well Log. Tech 47, 421–431. https://doi.org/10.16489/j.issn.1004-1338.2023.04.005.

Huo, F., Chen, Y., Ren, W., et al., 2022. Prediction of reservoir key parameters in 'sweet spot' on the basis of particle swarm optimization to TCN-LSTM network. J. Petrol. Sci. Eng. 214, 110544. https://doi.org/10.1016/j.petrol.2022.110544.

Jin, M., Wang, S., Ma, L., et al., 2024. Time-LLM: time series forecasting by reprogramming large language models. In: The Twelfth International Conference on Learning Representations. https://openreview.net/forum?id=Unb5CVPtae.

Jin, Y., Hu, G., Chen, H., et al., 2023. Cross-modal distillation for speaker recognition. Proc. AAAI Conf. Artif. Intell. 37, 12977–12985. https://doi.org/10.1609/aaai.v37i11.26525.

Li, Y., Li, Z., Shao, L., et al., 2023. A new physics-informed method for the fracability evaluation of shale oil reservoirs. Coal Geol. Explor. 51, 37–51. https://doi.org/10.12363/issn.1001-1986.23.02.0106.

Lin, L., Huang, H., Zhang, P., et al., 2024. A deep-learning framework for borehole formation properties prediction using heterogeneous well-logging data: A case study of a carbonate reservoir in the Gaoshiti-Moxi area, Sichuan Basin, China. Geophysics 89, WA295–WA308. https://doi.org/10.1190/geo2023-0151.1.

Lin, Z., Deng, C., Zhou, L., et al., 2023. GeoGalactica: A scientific large language model in geoscience. arXiv:2401.00434. https://doi.org/10.48550/arXiv.2401.00434.

Meng, M., Ge, H., Shen, Y., et al., 2023. Rock fabric of tight sandstone and its influence on irreducible water saturation in eastern Ordos Basin. Energy Fuel. 37, 3685–3696. https://doi.org/10.1021/acs.energyfuels.2c04011.

Moosavi, N., Bagheri, M., Nabi-Bidhendi, et al., 2022. Fuzzy support vector regression for permeability estimation of petroleum reservoir using well logs. Acta Geophys. 70, 161–172. https://doi.org/10.1007/s11600-021-00700-8.

Moosavi, N., Bagheri, M., Nabi-Bidhendi, M., et al., 2023. Porosity prediction using Fuzzy SVR and FCM SVR from well logs of an oil field in south of Iran. Acta Geophys. 71, 769–782. https://doi.org/10.1007/s11600-022-00944-y.

Noh, K., Kim, D., Byun, J., 2023. Explainable deep learning for supervised seismic facies classification using intrinsic method. IEEE Trans. Geosci. Remote Sens. 61, 1–11. https://doi.org/10.1109/TGRS.2023.3236500.

Pan, S., Zheng, Z., Guo, Z., et al., 2022. An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. J. Petrol. Sci. Eng. 208, 109520. https://doi.org/10.1016/j.petrol.2021.109520.

Sang, W., Yuan, S., Han, H., et al., 2022. Porosity prediction using semi-supervised learning with biased well log data for improving estimation accuracy and reducing prediction uncertainty. Geophys. J. Int. 232, 940–957. https://doi.org/10.1093/gji/ggac371.

Shao, H., Hu, Y., Wang, L., et al., 2024. LMDrive: closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15120–15130. https://doi.org/10.48550/arXiv.2312.07488.

Shao, R., Xiao, L., Liao, G., et al., 2022. A reservoir parameters prediction method for geophysical logs based on transfer learning. Chinese J. Geophys.-Chinese Ed. 65, 796–808. https://doi.org/10.6038/cjg2022P0057.

Solatorio, A.V., Dupriez, O., 2023. REaLTabFormer: Generating Realistic Relational and Tabular Data Using Transformers. https://arxiv.org/abs/2302.02041.

Sun, Y., Pang, S., Zhang, J., et al., 2024. Porosity prediction through well logging data: a combined approach of convolutional neural network and transformer model (CNN-transformer). Phys. Fluids 36, 026604. https://doi.org/10.1063/5.0190078.

Tang, J., Fan, B., Xiao, L., et al., 2021. A new ensemble machine-learning framework for searching sweet spots in shale reservoirs. SPE J. 26, 482–497. https://doi.org/10.2118/204224-PA.

Tavakolizadeh, N., Bagheri, M., 2021. Multi-attribute selection for salt dome detection based on SVM and MLP machine learning techniques. Nat. Resour. Res. 31, 353–370. https://doi.org/10.1007/s11053-021-09973-8.

Tian, Y., Stovas, A., Gao, J., et al., 2023. Frequency-dependent AVO inversion and application on tight sandstone gas reservoir prediction using deep neural network. IEEE Trans. Geosci. Remote Sens. 61, 1–13. https://doi.org/10.1109/TGRS.2023.3328183.

Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

Vobecky, A., Hurych, D., Simeoni, O., et al., 2022. Drive&Segment: unsupervised semantic segmentation of urban scenes via cross-modal distillation. Comput. Vis. – ECCV 2022, 478–495. https://doi.org/10.1007/978-3-031-19839-7_28.

Wang, M., Guo, X., Tang, H., et al., 2023. The prediction of shale gas reservoir parameters through a multilayer transfer learning network. Geophys. J. Int. 234, 1463–1475. https://doi.org/10.1093/gji/ggad145.

Xu, B., Xiao, L., 2024. Comparison of well logging formation evaluation using serial and parallel multi-task learning networks. Chinese J. Geophys.-Chinese Ed. 67, 1613–1626. https://doi.org/10.6038/cjg2023R0584.

Xue, H., Salim, F.D., 2024. Promptcast: a new prompt-based learning paradigm for time series forecasting. IEEE Trans. Knowl. Data Eng. 36, 6851–6864. https://doi.org/10.1109/TKDE.2023.3342137.

Yang, J., Lin, N., Zhang, K., et al., 2023a. A data-driven workflow based on multisource transfer machine learning for gas-bearing probability distribution prediction: A case study. Geophysics 88, B163–B177. https://doi.org/10.1190/geo2022-0726.1.

Yang, L., Fomel, S., Wang, S., et al., 2023b. Porosity and permeability prediction using a transformer and periodic long short-term network. Geophysics 88, WA293–WA308. https://doi.org/10.1190/geo2022-0150.1.

Zare, A., Bagheri, M., Ebadi, M., 2020. Reservoir facies and porosity modeling using seismic data and well logs by geostatistical simulation in an oil field. Carbonates Evaporites 35, 65. https://doi.org/10.1007/s13146-020-00605-5.

Zhang, H., Wu, W., 2023. DBN with IQPSO algorithm for permeability prediction: a case study of the Lizhai geothermal field, Zhangye Basin (Northern China). Nat. Resour. Res. 32, 1941–1957. https://doi.org/10.1007/s11053-023-10240-1.
Zhang, Q., Yan, Y., Li, W., et al., 2024. A mathematical model for predicting the productivity of fractured horizontal wells of tight sandstone gas: a case study in the Sulige gas field. Nat. Gas. Ind. B 11, 170–184. https://doi.org/10.1016/j.ngib.2024.04.003.
Zhao, X., Jin, F., Liu, X., et al., 2022. Numerical study of fracture dynamics in different shale fabric facies by integrating machine learning and 3-D lattice method: A case from Cangdong Sag, Bohai Bay basin, China. J. Pet. Sci. Eng. 218, 110861. https://doi.org/10.1016/j.petrol.2022.110861.
Zhou, X., Zhang, C., Zhang, Z., et al., 2019. A saturation evaluation method in tight gas sandstones based on diagenetic facies. Mar. Petrol. Geol. 107, 310–325. https://doi.org/10.1016/j.marpetgeo.2019.05.022.