Energy and AI 24 (2026) 100744

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

A multimodal feature fusion and large language model approach for the combustion stability diagnosis of 660 MWth coal-fired boilers

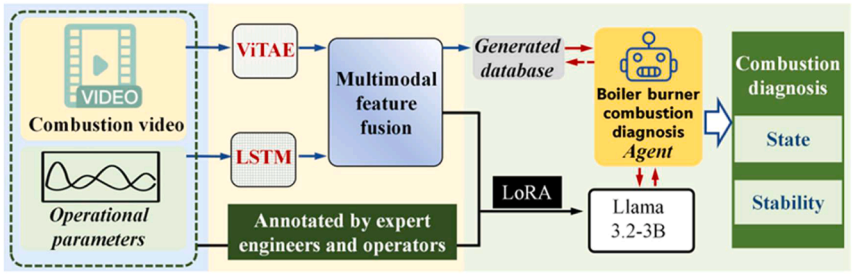
Sixu Pu [a], Yi Zhou [a], Md.Moinul Hossain [b], Xiaoyu Zhu [a], Guoqing Chen [c], Chuanlong Xu [a,*]

[a] National Engineering Research Center of Power Generation Control and Safety, School of Energy and Environment, Southeast University, Nanjing, 210096, China
[b] School of Engineering, University of Kent, Canterbury, Kent, CT27NT, UK
[c] State Key Laboratory of Low-Carbon Smart Coal-fired Power Generation and Ultral-clean Emission, China Energy Science and Technology Research Institute Co., Ltd, Nanjing, 210096, China

HIGHLIGHTS

- A domain-specific LLM for boiler burner combustion stability diagnosis is proposed.
- A multimodal fusion model is developed to integrate combustion features effectively.
- Expert annotations improve interpretability and reasoning capability of the model.
- A multimodal feature database and a combustion stability index are established.
- A quantitative stability assessment is demonstrated on a 660 MW boiler burner.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Coal-fired power plant
Burner
Combustion stability
Flame imaging
Multimodal feature
Large language model

ABSTRACT

Under low-load flexible operation conditions, where coal-fired boilers often need to respond quickly to deep load changes, the combustion state of boiler burners is vital for furnace efficiency and safety. Although image-based flame detection offers higher monitoring accuracy than traditional methods, it cannot provide quantitative evaluations and clear diagnoses of flame stability. To address these issues, this study introduces a large-language-based approach utilizing a pre-trained Llama 3.2–3B model with lightweight fine-tuning for flame stability diagnosis. Temporal features from operational data and spatial features from flame videos are extracted using a Long Short-Term Memory network and a Vision Transformers advanced by exploring intrinsic inductive bias model. A custom multimodal fusion network then blends these complementary features to create a unified representation of combustion characteristics. Expert annotations are integrated during fine-tuning to improve understanding and diagnostic reasoning specific to combustion. Additionally, a multimodal feature database of stable conditions is built to enable quantitative evaluation of flame stability. Tests conducted on a 660MW opposed-fired boiler demonstrate the effectiveness of the proposed model. The results show the model's ability to distinguish between stable and unstable flame states across different operational loads. It accurately detects transitions between safe and unsafe combustion states and offers interpretable recommendations for adjustments.

* Corresponding author.
E-mail address: chuanlongxu@seu.edu.cn (C. Xu).

https://doi.org/10.1016/j.egyai.2026.100744

Available online 2 April 2026

This provides a practical pathway for the safe, efficient, and intelligent operation of coal-fired boilers under variable load conditions.

1. Introduction

In the context of the green and low-carbon transition [1], coal-fired power plants must enhance peak-shaving capacity to accommodate large-scale renewable integration and ensure grid stability [2]. Among various influencing factors, the stability of flames plays a pivotal role in determining boiler combustion efficiency and operational safety. Flame instability can lead to local overheating and temperature field imbalance [3], thereby degrading combustion performance and increasing pollutant emissions [4]. Therefore, accurately identifying and diagnosing the stability of burner flames is of great significance for improving the safety [5], environmental performance, and economic efficiency of power generation units [6].

Digital imaging technologies have experienced continuous advancement in recent years [7], expanding their capability for high-quality visual data acquisition [8]. Image-based flame diagnosis methods have consequently been applied across a wide range of fuel types and complex combustion conditions [9], due to their non-invasiveness, high sensitivity, and rapid response [10]. These methods primarily identify combustion states by extracting flame features [11], such as geometric parameters [12], luminous intensity [13], oscillatory behavior [14], and texture information [15]. Chi et al. [16]. observed that the oscillation frequency, fluctuation ratio, and combustion kinetic energy near the flame root are susceptible to coal types, making them effective indicators of flame stability during ignition. Yang et al. [17]. proposed a method combining Gabor filters and gray-level co-occurrence matrix (GLCM) to extract texture features from industrial boiler flame images. This approach enhances image texture and converts it into quantifiable statistical descriptors to identify combustion conditions. Gaidhane et al. [18]. employed local binary patterns, dual-threshold classification, and Levenberg-Marquardt optimization to map flame contour and area features to identify combustion states. Sun et al. [19]. quantified the oil flames' stability using flame brightness and Hue-Saturation-Intensity (HSI) color space features. However, traditional image-based combustion diagnostics rely on handcrafted features and statistical methods, making them susceptible to noise and limiting their performance under varying operation conditions. Deep learning (DL)-based methods [20] can automatically extract multi-level and multi-scale feature information directly from raw images [21], thereby effectively enhancing feature representation capability and improving model adaptability to complex scenes [22]. For example, Lyu et al. [23]. used a deep belief network (DBN) to extract global features from flame images, effectively avoiding information loss caused by image segmentation. Zhou et al. [24]. developed a convolutional neural network (CNN) and named as BIM (Basis Image Monitor) to extract critical morphological features from flame images and detect thermoacoustic instability. Qiu et al. [25]. utilized a convolutional autoencoder (CAE) combined with unsupervised clustering to achieve automated classification of combustion states. Han et al. [26]. employed a stacked sparse autoencoder (SSAE) to extract flame features and introduced the squared prediction error (SPE) as a metric for combustion stability assessment. Although DL methods have made progress in flame recognition and combustion state classification, they are limited in understanding combustion processes under complex operation conditions. Moreover, current DL-based approaches often overlook the representation and utilization of diagnostic logic and expert knowledge, making it difficult to achieve accurate identification and intelligent diagnosis of combustion stability.

The development of large language models (LLMs) has demonstrated broad applications in areas such as multi-source information integration [27], complex fault analysis [28], and expert knowledge generation [29]. In recent years, LLMs have also been gradually introduced into the field of combustion research to enhance the understanding and representation of complex physical processes [30]. For instance, Sharma et al. [31]. integrated a large language model (e.g., GPT-3.5) with a combustion knowledge graph, using retrieval-augmented generation (RAG) and causal graphs to enable literature-based question answering, experimental recommendations, and logical reasoning. For instance, the framework can respond to domain-specific queries such as key mechanisms of NOx formation or factors influencing flame stabilization, thereby enhancing its adaptability and reliability in combustion applications. Xu et al. [32]. developed a GPT-based agent that generates fire dynamics simulations from natural language, utilizing chain-of-thought reasoning and memory to streamline design, reduce skill demands, and enhance efficiency. In the field of flame image recognition, Alkhammash et al. [33]. applied GPT-4 models to fire-incident image classification and demonstrated that LLMs can provide richer contextual information (e.g., fire type and risk level), thereby improving decision-support capability. To integrate control strategies and human behavior modeling, Martinez et al. [34]. proposed a multimodal flame-analysis agent that combines a large language-and-vision assistant (LLaVA) with reinforcement learning from human feedback (RLHF) to estimate combustion quality. The system analyzes real-time flame imagery and iteratively adjusts combustion-related parameters to improve accuracy. Xie et al. [35]. developed WildfireGPT, a domain-specific model for wildfire analysis, integrating knowledge-enhanced pretraining and instruction tuning with domain knowledge on fire evolution, vegetation, and meteorology, enabling wildfire identification, risk prediction, and emergency strategy generation. In terms of edge perception and spatial localization, Wu et al. [36]. adapted the Contrastive Language-Image Pre-training (CLIP) foundation model through multimodal prompt tuning to enhance fire-smoke recognition under pseudo-smoke interference and regional domain shift. Moreover, Seidel et al. [37]. embedded a multimodal vision-language model (VLM) into an Unmanned Aerial Vehicle (UAV) remote sensing system. They designed a bidirectional image-text interaction framework for early wildfire detection, enabling the model to automatically generate fire descriptions, spatial localization, and intervention suggestions from aerial imagery.

Although image-based and DL methods have advanced flame recognition and stability analysis, they remain limited by insufficient semantic understanding, weak interpretability, and inadequate use of expert knowledge, making it difficult to address causal reasoning and knowledge-driven diagnostics under complex operation conditions. LLMs show potential in multi-source information integration, knowledge reasoning, and interpretable output, but general-purpose models still face shortcomings in domain adaptation, data understanding, and diagnostic capability. To address these issues, this study constructs a domain-specific large language model dedicated to combustion diagnosis in coal-fired boiler burners. By integrating multimodal features from operation data and flame images with expert knowledge annotations, the model enables quantitative assessment of flame stability and interpretable adjustment recommendations, providing an innovative solution for safe and efficient operation under deep peak-shaving conditions. The principal contribution of this study is the construction of a multimodal feature fusion network utilising Long Short-Term Memory (LSTM) and Vision Transformers Advanced by Exploring Intrinsic Inductive Bias (ViTAE) networks. This network achieves the coordinated extraction and integration of temporal features from boiler operational data and spatial features from flame images.

- A domain-specific LLM for flame stability diagnosis in coal-fired boiler burners is developed by integrating fused features with expert knowledge annotations.
- A multimodal feature database covering stable combustion characteristics under different operational loads is constructed. Using the combustion stability index (CSI), derived from this database, combustion stability is quantitatively evaluated.
- An intelligent agent, built upon the proposed domain-specific LLM and deployed for real-time inference, is equipped with semantic understanding and interactive capabilities. Based on the diagnostic results of combustion, it offers suggestions for optimizing operational parameters.

2. Multimodal feature extraction and fusion

This study presents a large language model-based intelligent diagnostic approach for burner flame stability in coal-fired power plant boilers, as illustrated in Fig. 1. Synchronously acquired burner combustion videos and operational parameters are processed through a ViTAE network to extract spatial features and an LSTM network to extract temporal features. A custom multimodal feature fusion network integrates these features, which, together with time-aligned expert annotations, are fed into the Llama 3.2–3B model. Low-Rank Adaptation (LoRA) fine-tuning imparts domain-specific knowledge, understanding and diagnostic representation capability to the model. During deployment, the combustion agent extracts multimodal features from new operating conditions, matches them against a database of known conditions, infers the current flame stability, and outputs optimized operational parameter adjustment strategies.

2.1. Feature extraction of operational parameters via LSTM

The operational parameters of the boiler are defined as the key variables that reflect and regulate the combustion process. These mainly include secondary air volume, primary air velocity, coal feed rate, boiler load, and water wall temperature. Among them, secondary air volume, primary air speed, and coal feed rate are considered as primary control parameters, because they can be directly adjusted and monitored in real time, also playing a crucial role in the fuel-air mixing state and combustion intensity. In contrast, boiler load and water wall temperature are regarded as response parameters, reflecting the overall output level of the combustion process and the thermal distribution within the furnace, respectively. Before encoding these time-series feature parameters, it is essential to analyze the correlation and redundancy among them to optimize the feature space, reduce redundant input dimensions, and improve the model's generalization. Therefore, this study introduces the Time-Varying Mutual Information (TVMI) method to dynamically quantify the dependency between different time-series parameters by computing the mutual information within a sliding time window, thus capturing the temporal evolution of shared data. For example, parameters with low dynamic correlation are discarded to suppress redundancy and enhance the discriminative power of the extracted features.

Mutual information is a concept from information theory used to measure the degree of mutual dependence between two random variables. For instance, for two discrete random variables X and Y, the mutual information is defined as:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

TVMI is a temporal extension of mutual information, designed to capture the dynamic dependency between two time-series parameters $X(t)$ and $Y(t)$, such as time-varying coupling strength, lead-lag effects, and nonlinear interactions that evolve with operating conditions. The core idea of TVMI is to compute local mutual information within each time point or time window, thereby reflecting how the strength of the association between the parameters evolves.

$$TVMI(t)=I(X(t);Y(t))=\sum_{x\in X(t)}\sum_{y\in Y(t)}p_t(x,y)log\frac{p_t(x,y)}{p_t(x)p_t(y)} \tag{2}$$

After eliminating operational parameters with weak cross-correlation, a Long Short-Term Memory (LSTM) network is employed to extract features from the time-series data. LSTM is an improved variant of the Recurrent Neural Network (RNN), consisting of multiple basic units connected in a chain-like structure, as illustrated in Fig. 2. Compared to traditional RNNs, LSTM introduces memory cells and gating mechanisms that enable selective retention and updating state information, thereby effectively modeling long-term dependencies in

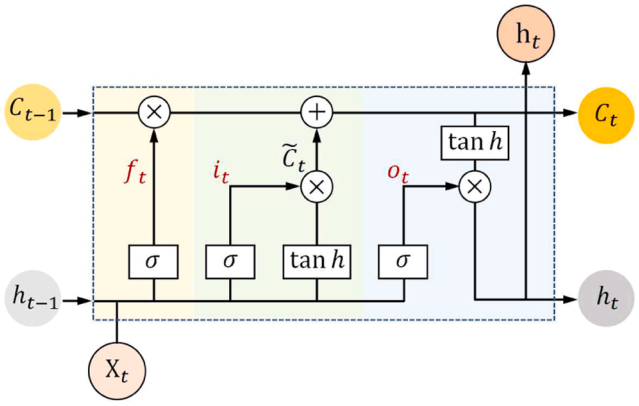


Fig. 2. The LSTM network unit structure for operational parameters feature extraction.

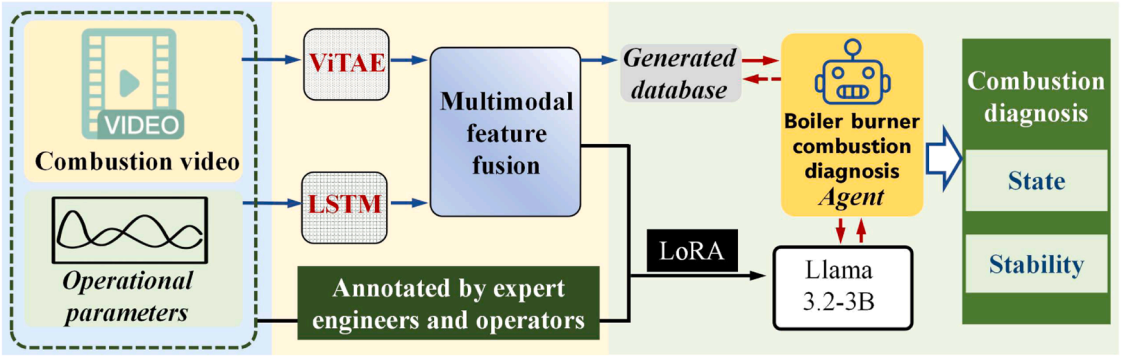


Fig. 1. Overview of the architecture for combustion stability diagnosis.

sequential data [38].

In the LSTM unit, the transmission of information and the updating of internal states primarily rely on three gating mechanisms: the forgetting, input and output gates. Specifically, the forget gate f_t determines how much of the previous cell state should be retained; the input gate i_t controls the extent to which new input information is written into the cell state; the candidate state $\widetilde{C}_t$ represents the potential update generated from the current input, and the output gate o_t determines how much information from the cell state is propagated to the hidden state output. These gating mechanisms jointly regulate the updating process of the cell state C_t and hidden state h_t, thereby enabling effective modeling of long-term dependencies in sequential data. The corresponding computation processes are as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_c\right) \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_0) \tag{6}$$

where x_t denotes the input vector at the current time step, and h_{t-1} represents the hidden state from the previous time step. W_f, W_i, W_c, W_o are the weight matrices corresponding to the forget, input and output gates, and candidate cell states, respectively. b_f, b_i, b_c, b_o are the associated bias terms. σ denotes the sigmoid activation function; *tanh* denotes the hyperbolic tangent activation function, used to generate the candidate cell state content.

Based on the above gating mechanisms, the current cell state is updated by the weighted summation of the forget gate output and the candidate state modulated by the input gate:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{7}$$

After the cell state is updated, the output gate regulates the generation of the current hidden state based on the updated cell state, as expressed by the following equation:

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{8}$$

In this study, the network is used for feature extraction from time-series parameters rather than for complex tasks such as prediction, allowing for a relatively simple structure with two stacked layers and 32 hidden units per layer.

2.2. Flame image features extraction

Flame images typically exhibit complex texture patterns and pronounced dynamic characteristics, which pose unique challenges to feature extraction models. Traditional methods based on Convolutional Neural Networks (CNNs) perform well in capturing local features. However, their inherently limited receptive fields make it difficult to model the global spatiotemporal dependencies inherent in combustion processes. In addition, to enable large models for better understanding and to integrate multimodal information, flame features need to be represented as structured sequences compatible with the input mechanisms and semantic modeling of such models. To address these issues, this study adopts the Vision Transformers Advanced by Exploring Intrinsic Inductive Bias (ViTAE) architecture [39], which overcomes the dual limitations of conventional approaches in global modeling and local structural perception.

As shown in Fig. 3, unlike the standard Vision Transformer that divides images into fixed-size patches and obtains embedding sequences through linear projection, ViTAE introduces a Reduction Cell (RC module) at the embedding stage. This module integrates multi-scale contextual information based on local structural encoding, thereby constructing sequence representations with spatial hierarchical awareness. The RC module consists of two parallel branches:

The multi-scale branch is responsible for capturing long-range dependencies. Specifically, it utilizes a Pyramid Reduction Module (PRM) to perform multi-scale dilated convolutions, enabling the extraction of rich contextual features across different receptive fields:

$$f_i^{ms} \underset{=}{\Delta} PRM_i(f_i) = Cat\left(\left[Conv_{ij}\left(f_i; s_{ij}; r_i\right)\right] \middle| s_{ij} \in S_i, r_i \in R\right) \tag{9}$$

where f_i^{ms} denotes the multi-scale output feature, f_i is the input feature of the ith RC module, S_i represents the predefined set of dilation rates, r_i is the stride, and $Conv_{ij}$ denotes the jth dilated convolution.

Subsequently, the multi-scale features f_i^{ms} are transformed into a sequence and fed into a Multi-Head Self-Attention (MHSA) module to model long-range dependencies.

$$f_i^g = MHSA_i\left(Img2Seq\left(f_i^{ms}\right)\right) \tag{10}$$

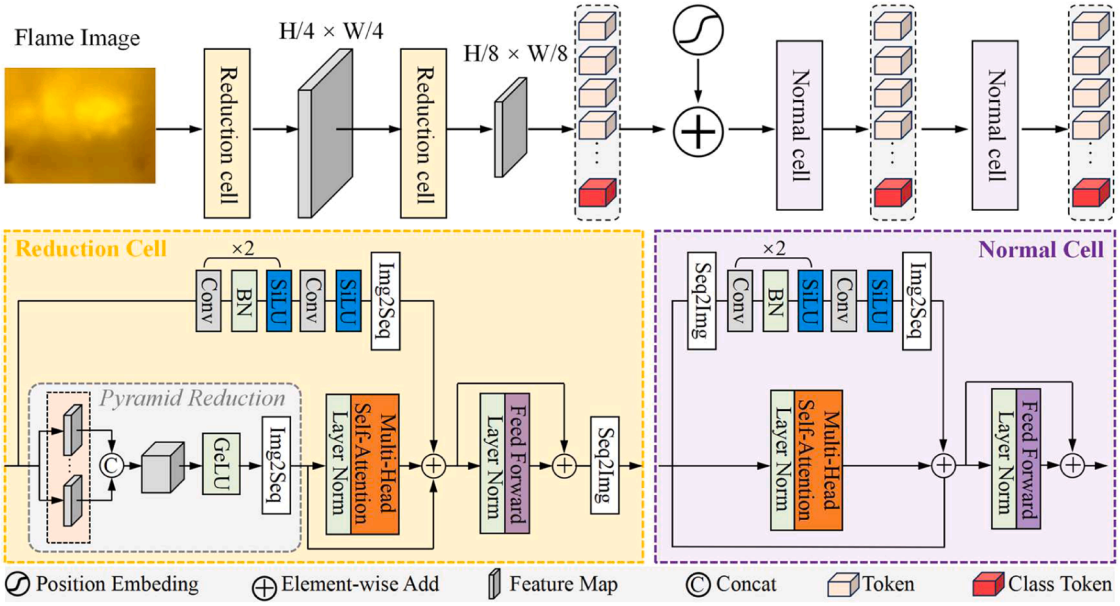


Fig. 3. Architecture of the ViTAE-based model for flame feature extraction.

where f_i^g denotes the global feature extracted from the ith RC module. *Img2Seq* denotes the transformation of a feature map into a one-dimensional sequential representation, and $MHSA_i$ represents the multi-head self-attention mechanism.

The local branch extracts fine-grained details through a Parallel Convolution Module (PCM). The output of PCM is fused with the previously obtained global feature f_i^g to generate f_i^l, which encodes both contextual information and local features, as formulated below.

$$f_i^{lg} = f_i^g + PCM_i(f_i) \tag{11}$$

Finally, the fused token representation is processed through a feed-forward network and reshaped into a feature map, which serves as the input to the subsequent RC module:

$$f_{i+1} = Seq2Img\left(FFN_i\left(f_i^{lg}\right) + f_i^{lg}\right) \tag{12}$$

where *Seq2Img* performs a reshape operation to convert the token sequence back into a feature map, and $FFN_i(\cdot)$ denotes the feed-forward network in the ith RC module.

Two cascaded RC modules are employed, which progressively reduce the spatial resolution of the input image by factors of 4 × and 2 ×, respectively, ultimately generating a multi-scale fused feature map of size [*H/8, W/8, D*]. Considering that the target types at the burner outlet are relatively simple, primarily flame regions, despite their complex morphology, the semantic categories are limited. Therefore, deep hierarchical abstraction is not necessary. The two RC modules are sufficient to extract dynamic texture and structural information while preserving relatively high spatial resolution.

The Normal Cell (NC module) shares a similar structure with the RC module, with the key difference that it does not include the multi-scale extraction module (PRM), as the spatial resolution of the input feature map has already been reduced to one-eighth of the original image size, making further downsampling unnecessary. The output of the NC serves as the final fused representation, which encodes both global contextual information and fine-grained local features, and is used as the input for subsequent tasks.

2.3. Multimodal feature fusion

In **LLM,** efficiently integrating heterogeneous modalities, such as text, images, audio, and sensor data, is one of the key influencing factors of overall model performance. Due to inherent differences in data distribution, semantic representation and statistical properties across modalities, direct fusion at the task decision level may lead to suboptimal information utilization and semantic misalignment. Therefore, a well-designed cross-modal feature fusion strategy is essential for enhancing the model's representational capacity and generalization ability, thereby enabling better performance in various multimodal tasks [40].

In this study, to address the heterogeneous multimodal information present in the combustion process, separate feature extraction modules were developed for operational parameters and flame images, upon which a multimodal feature fusion mechanism was designed. The overall framework for multimodal feature fusion extraction is illustrated in Fig. 4. Specifically, an LSTM network is employed to model the temporal patterns of boiler operational parameters. The input comprises four operational parameters, and a sequence of length five is constructed using a five-second sliding time window. This sequence is processed by two LSTM layers to extract a temporal feature representation of size 5 × 32, which is subsequently mapped to a 5 × 64 feature space via a linear projection to align with the spatial dimension of image features. Meanwhile, flame images with a resolution of 224 (H) × 224 (V) are fed into the ViTAE model to extract image token representations. The model produces a token sequence of 196 length, with each token having a dimension of 64, resulting in a 196 × 64 image feature matrix. Subsequently, the features extracted from the operational parameters and flame images are concatenated along the token dimension, resulting in a unified multimodal sequence of size 201 × 64, which integrates both temporal and spatial representations. This sequence is passed through a learnable fusion module and projected into a higher-dimensional semantic space of 201 × 768 via a multilayer perception (MLP), thereby enhancing the cross-modal representational capacity. The fused features are further processed by a weighted fusion mechanism and a cross-attention module to strengthen semantic alignment and complementary interactions between modalities. Finally, a class token is introduced as the aggregated representation, and the fused sequence is fed into a pretrained large model (Llama) for downstream task modeling and intelligent inference. The class token was selected instead of the full set of fused tokens mainly because it can effectively aggregate the global information from both the image modality and the temporal modality, thereby forming a compact and representative feature representation that is more suitable for subsequent combustion state discrimination. Compared with feeding all fused tokens into the large language model, the class token preserves the key global semantic information while effectively reducing input redundancy, lowering the computational complexity and resource consumption during both training and inference, and thus improving the overall efficiency of combustion diagnosis.

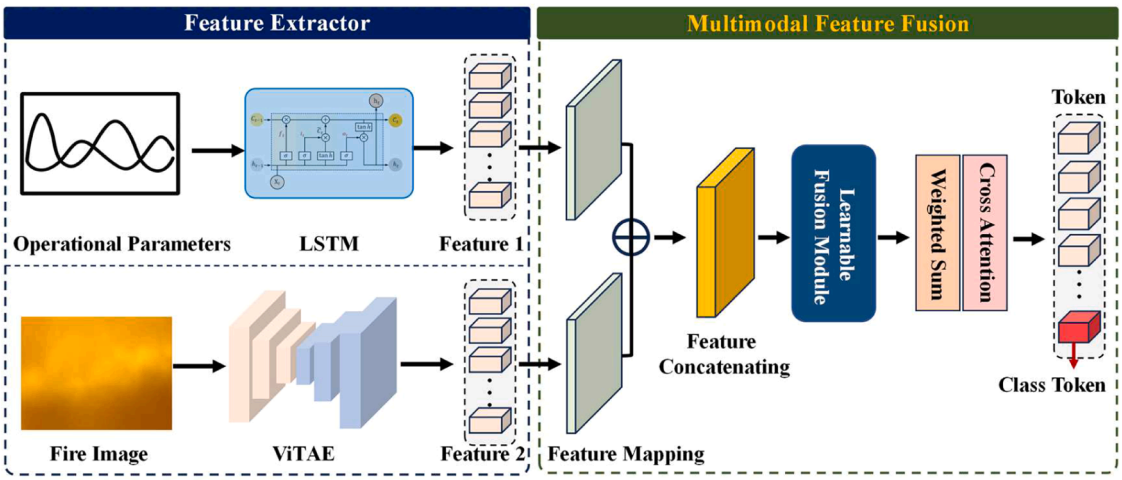


Fig. 4. Overview of a network architecture for multimodal feature fusion.

3. Model establishment

3.1. Data collection from a 660MWth coal-fired boiler

Experiments were conducted on a 660 MWth opposed-fired boiler equipped with a rich-lean fuel separation swirl burner. A flame image acquisition system was installed at the flame observation port. The system consists of an optical probe and a CCD camera. As shown in Fig. 5, the optical probe has a field of view of 110°, ensuring coverage of a sufficiently wide combustion region. The CCD camera captures flame images at a frame rate of 25 Hz. To prevent image distortion or saturation, the exposure time was set to 150 μs. Air cooling was employed to protect the imaging equipment and ensure its stable operation under high-temperature environments.

Fig. 6 presents the 30-hour load variation data. To ensure comprehensive coverage of the load reduction process, the dataset consists of two parts. Part A includes data collected from 15:00 to 20:00, corresponding to load conditions ranging from 600 MW to 330 MW. Due to nighttime safety restrictions, no flame images were recorded between 21:00 and 09:00. Part B covers the period from 15:00 to 18:00 on the following day, representing load conditions from 330 MW to 270 MW. For the generalization performance verification, data collected on the following day were used, including the high-load condition (560 MW, 11:00–11:30), the medium-load condition (400 MW, 12:30–13:00), and the low-load condition (280 MW, 19:30–20:00).

The training was conducted in an environment configured with Python 3.12, an NVIDIA Tesla V100 GPU (16 GB memory). After training, the model was deployed locally for inference and testing on a workstation equipped with an Intel i9-10940X CPU, 128 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU.

3.2. Model fine-tuning and implementation

A publicly released Llama3.2-3B [41] model is adopted as the base model for fine-tuning. With approximately 3 billion parameters, it offers a balance between lightweight architecture and expressive capacity, enabling efficient multimodal feature fusion under limited computational resources. Compared to larger-scale models (e.g., 13B, 65B), Llama3.2-3B significantly reduces memory consumption while maintaining strong representation and generalization capabilities, making it feasible to train and infer the model on a single GPU. In addition to the factor of parameter size, although general-purpose multimodal large models excel in cross-modal alignment, these models are primarily trained on large-scale natural image and text data and are not specifically designed for combustion systems. Therefore, while they possess strong visual understanding capabilities, the goal of this study is not general-purpose visual understanding, but rather the development of a feature-driven combustion stability diagnosis framework. In this framework, feature extraction is performed outside the model, and the model itself is mainly used for understanding and reasoning. This approach not only better ensures the interpretability within the research domain but also enhances the controllability in industrial applications.

When fine-tuning large-scale pre-trained models, the traditional Full Parameter Fine-Tuning approach requires updating all model weights, which causes significant computational and storage costs and may lead to catastrophic forgetting. To reduce training expenses and enhance task adaptability, this study uses the parameter-efficient fine-tuning method Low-Rank Adaptation (LoRA) [42] to adjust the Llama3.2-3B model. By altering internal weight structures, LoRA allows effective adaptation to multimodal tasks while maintaining the original model's knowledge. The main idea of LoRA is to freeze most of the pre-trained model parameters and insert low-rank trainable matrices within the attention layers. This method greatly reduces computational requirements while effectively adapting the model to new tasks. Unlike traditional full-parameter fine-tuning, which involves updating billions of parameters, LoRA updates only a small set of additional parameters, reducing memory use and computational complexity while preserving the model's ability to represent data.

In the Transformer architecture, the attention mechanism involves three weight matrices: Query (Q), Key (K), and Value (V). Taking the query matrix as an example, the standard linear projection can be expressed as:

$$X' = XW_0 \tag{13}$$

where X denotes the input features, W_0 is the original trainable weight matrix, and X' represents the transformed output. In traditional full-parameter fine-tuning, all parameters in W_0 must be updated, resulting in substantial computational overhead.

To improve parameter efficiency, the LoRA method introduces two low-rank matrices, A and B, which are used to model weight updates in an incremental and compact form:

$$W = W_0 + \Delta W \tag{14}$$

$$\Delta W = AB \tag{15}$$

where A and B are trainable low-rank matrices with a rank much smaller than that of W_0, significantly reducing the number of parameters and computational cost. Since only ΔW is updated while keeping the original weights W_0 frozen, LoRA effectively mitigates the risk of catastrophic forgetting and preserves the expressive power of the pretrained model.

To enable interpretable analysis and understanding of combustion states, expert annotations by senior operational engineers were incorporated during the data preprocessing stage to semantically label the flame states across different time periods. Given the pronounced non-stationary and transient characteristics of furnace flames, the choice of

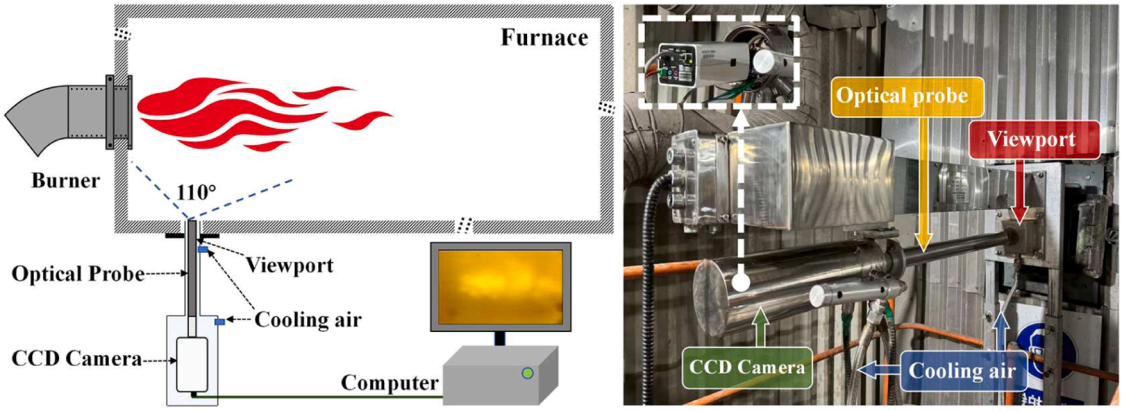


Fig. 5. Experimental setup of the flame imaging system.

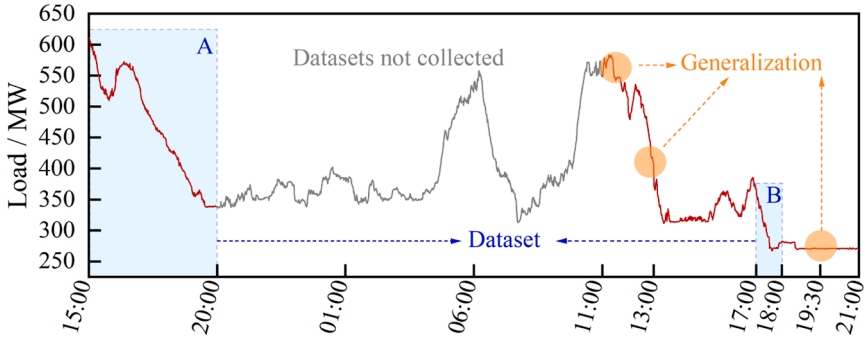


Fig. 6. Combustion operation under 30-hour boiler load variations.

data duration is critical for accurate state characterization. An excessively long-time window may overlook short-term dynamic variations in the flame, whereas an excessively short window may fail to capture the complete evolutionary trend. To strike a balance between local stability and global representativeness, each video segment and its corresponding operational parameters segment were uniformly set to 5 s. Specifically, each flame video was first segmented into 5-second clips, and the operational parameters were then selected from the same corresponding 5-second time interval based on the recorded time information. The temporal alignment accuracy mainly depends on the timestamp resolution of the acquisition system, and such alignment was sufficient for the combustion stability diagnosis task considered in this study. On this basis, expert assessments of the combustion state were further used to generate structured semantic annotations, providing the subsequent large model with precise and context-aware inputs. Annotations by expert engineers and operators consist of two levels of information. First, the current flame video is semantically described to characterize the combustion state of the flame. Second, targeted optimization suggestions are provided based on the features of the current state. After fine-tuning, the constructed model was endowed with specialized capabilities for combustion stability diagnosis, hereafter referred to as Combustion Stability Diagnosis-Llama (CSD-Llama), as shown in Fig. 7.

3.3. Combustion stability index

This study constructs a multimodal feature database covering various load conditions, systematically storing the stable combustion feature vectors for each load range. In this approach, the current operating condition is matched with the corresponding load-specific template features from the database, enabling similarity-based quantitative assessment of combustion stability. The cosine similarity [43] between the current condition and the stable combustion template under the same load level is calculated. For clarity, this similarity score is defined as the Combustion Stability Index (CSI), which quantifies how closely the current state resembles a stable combustion state. The calculation is provided in Eq. (16).

$$CSI = \frac{A \cdot B}{\| A \| \cdot \| B \|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{16}$$

where A denotes the fused feature vector extracted from the current operating condition, while B represents the reference feature vector corresponding to stable combustion states under different load levels in the database. Specifically, a_i and b_i denote the i th elements of the feature vectors A and B, respectively. The CSI quantifies the similarity between the current multimodal features and the reference stable condition. A higher CSI value indicates that the current combustion state is closer to the stable reference, suggesting more stable boiler operation. Conversely, a lower CSI value indicates potential combustion abnormalities or instabilities that warrant attention.

4. Results and discussion

4.1. Selection of effective operational parameters

Based on key combustion-related time-series parameters such as load, secondary air volume, primary air speed, coal feed rate, and temperature, the pairwise TVMI was calculated. As shown in Table 1, these operational parameters are indexed as Parameters 1–5. A higher TVMI value indicates stronger dynamic correlation and coupling between corresponding parameters. During the boiler load-reduction process, stagewise upward fluctuations in load occur. To improve the accuracy and representativeness of the correlation analysis, operational parameters from both the load-increasing and load-decreasing stages were selected for analysis. As shown in Fig. 8, (a) presents the operational parameter sequences, while (b) illustrates the TVMI matrix of the operational parameters. During the load-increasing phase, the TVMI between load and coal feed rate reaches 0.69, significantly higher than other off-diagonal elements. This indicates a strong interaction between load and fuel supply during ramp-up, reflecting a typical load, fuel

Table 1
Key parameters of the combustion operation.

Operational parameters	Specification
1	Load (MW)
2	Coal feed rate (T/H)
3	Primary air speed (m/s)
4	Secondary air volume (m3/min)
5	Water wall temperature (°C)

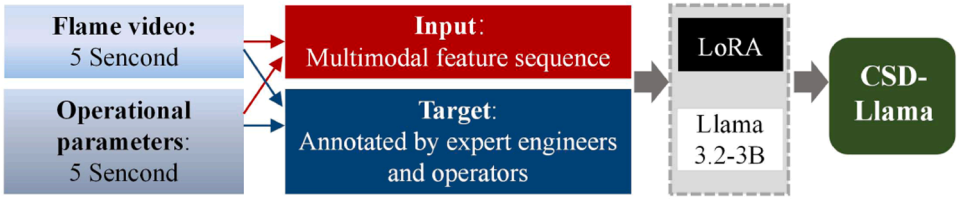


Fig. 7. Training and annotation framework for developing the CSD-Llama model.

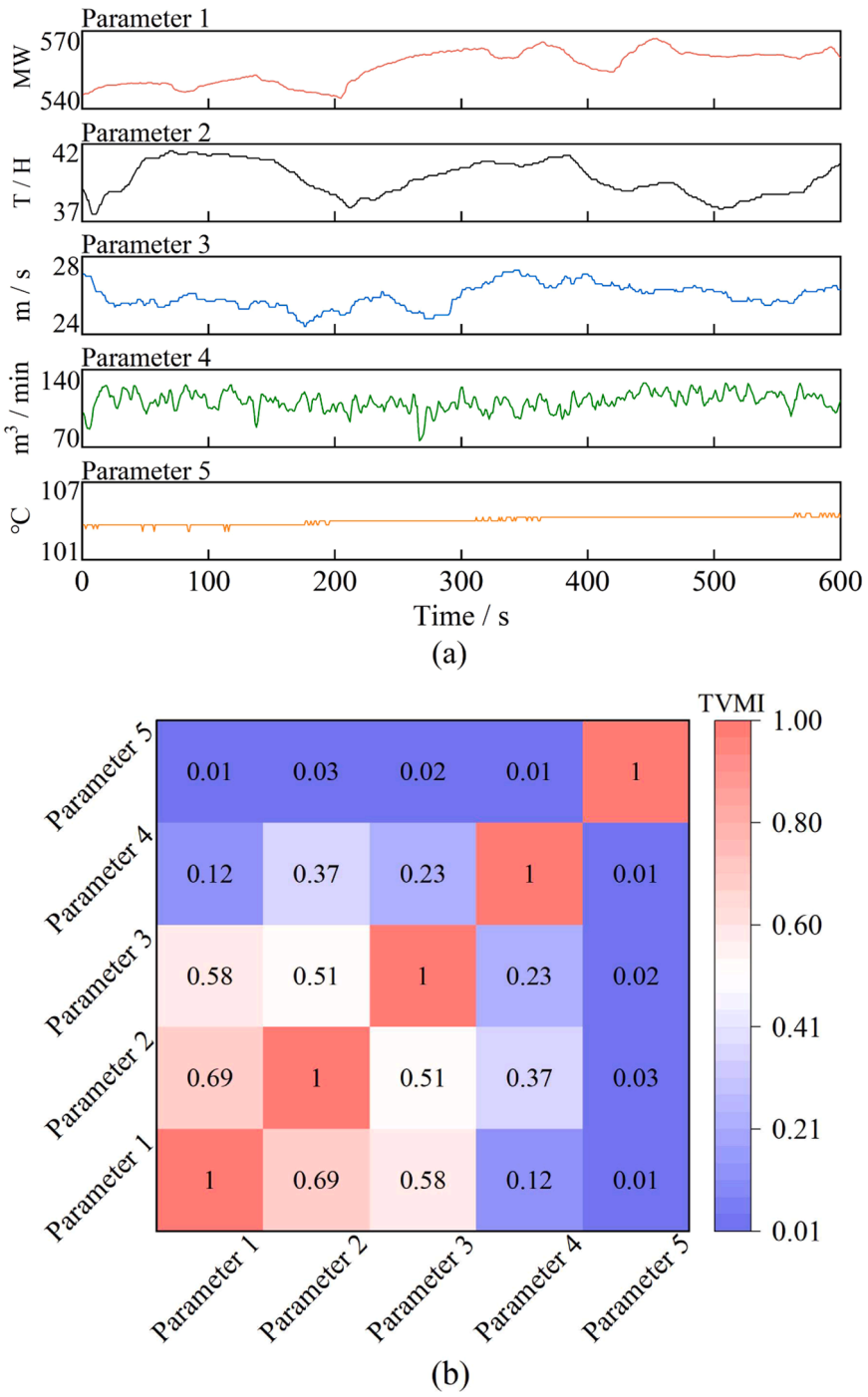


Fig. 8. (a) Overview of operational parameters [higher loads from 540 MW - 570 MW] and (b) Time-Varying Mutual Information.

dynamic ratio control feature. Additionally, the TVMI between load and primary air speed is 0.58, while that between coal feed rate and primary air speed is 0.51. These results suggest that primary air speed not only responds to changes in coal feed rate but also plays a compensatory role to ensure flame stability, making it a key control variable in air-fuel ratio regulation. In contrast, the secondary air volume exhibits relatively weak TVMI values (ranging from 0.12 to 0.37) with load, coal feed rate, and primary air speed, indicating its more independent regulation behavior. Its variation does not directly follow the main load trend and is primarily responsible for flame structuring and staged combustion control. The water wall temperature shows consistently low TVMI values with other variables, generally in the range of 0.01 to 0.03.

As shown in Fig. 9, (a) presents the operational signal sequences, while (b) illustrates the TVMI matrix of the operational parameters. The TVMI between load and coal feed rate reaches 0.89, which is higher than that observed during the load-decrease phase. This indicates a tighter coupling between coal adjustment and load variation during load reduction. This is because load-reducing operations typically adopt more precise and centralized control strategies to ensure flame stability and avoid risks such as thermal imbalance or flameout. The TVMI between load and primary air speed is 0.48, slightly lower than in the load-increase phase, while the TVMI between coal feed rate and primary air speed increases to 0.62, suggesting that primary air becomes more sensitive to fuel changes during load decrease. The secondary air volume still maintains low correlation in this stage, with TVMI values ranging from 0.21 to 0.32 when paired with major variables, indicating that it

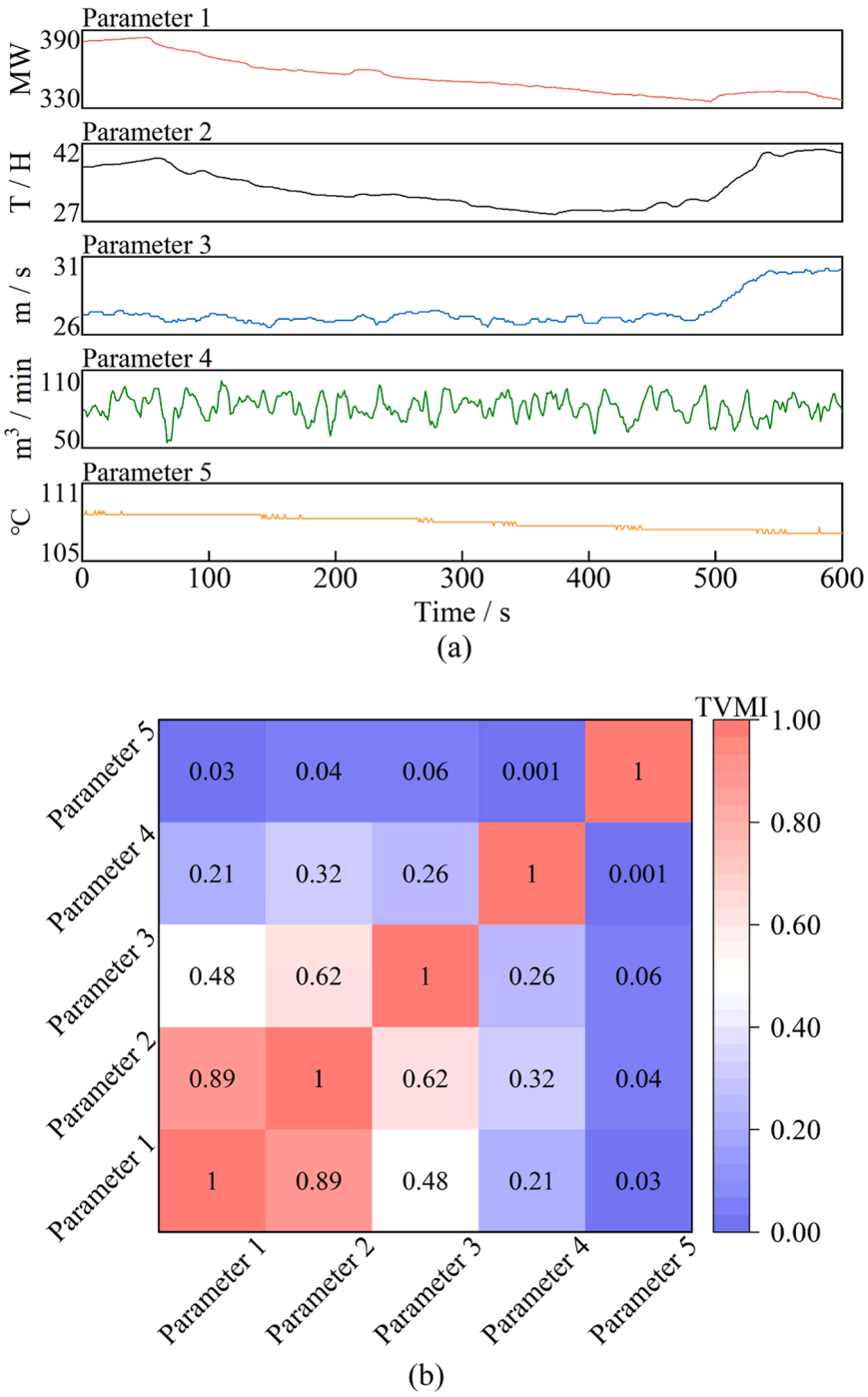


Fig. 9. (a) Over of operational parameters [lower loads from 330 MW - 390 MW] and (b) Time-Varying Mutual Information.

retains independent regulation characteristics. Its fluctuations primarily serve to support flame structure rather than participate in core combustion control. The water wall temperature continues to exhibit very low TVMI with other variables (0.01–0.06), reaffirming its slow variation under short-term dynamic conditions and suggesting it is not suitable as a fast-response indicator for combustion state perception.

In summary, during both load-increasing and load-decreasing conditions, parameter 5, the water-cooled wall temperature, consistently shows extremely low TVMI values (typically below 0.06) in relation to other key operating parameters, indicating a lack of significant dynamic coupling. This weak association suggests a delayed response that fails to capture rapid changes in combustion conditions. Consequently, to avoid introducing redundant or non-informative Parameters, this variable is excluded from the input features used in LSTM-based temporal feature extraction.

4.2. *Model performance evaluation*

Model performance was evaluated based on dataset A. As shown in Fig. 10, the training performance of the LoRA method with ranks 8 and 32 was compared against that of the full fine-tuning approach. The results indicate that increasing the rank enables LoRA to better approximate the performance of full-parameter fine-tuning. Specifically, when the rank is set to 8, LoRA exhibits noticeable gaps in training loss and validation performance compared to full fine-tuning, suggesting that the representational capacity of low-rank approximations is limited.

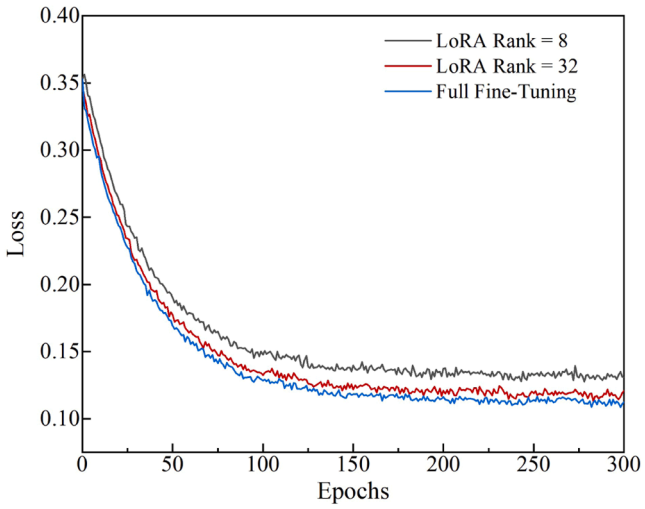


Fig. 10. Training loss under different LoRA ranks.

However, when the rank increases to 32, the training and validation curves of LoRA closely align with those of full fine-tuning, indicating that a higher rank can effectively capture the essential information required for model adaptation.

From a computational efficiency and resource consumption perspective, LoRA with rank 8 significantly reduces the number of trainable parameters, resulting in lower memory usage and computational cost, albeit with a slight performance compromise. When the rank is increased to 32, model performance improves substantially with training loss nearly identical to that of full fine-tuning while still requiring far fewer additional parameters. This demonstrates a favorable trade-off between performance and efficiency. Further increasing the rank (e.g., to 64) was not pursued, as the marginal gains in performance beyond rank 32 were limited and unlikely to offset the additional resource demands, making such an extension unnecessary.

To comprehensively evaluate the performance of the trained large model, multiple evaluation metrics, including Accuracy, Precision, Recall, and F1-Score [44], are employed to ensure a thorough assessment across various operating conditions.

Accuracy is defined as the proportion of correctly predicted samples to the total number of samples in a classification task, and is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

Precision indicates the proportion of true positive samples among all samples classified as positive by the model, which is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{18}$$

Recall represents the proportion of actual positive samples correctly identified by the model, calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{19}$$

F1-score is the harmonic mean of Precision and Recall, designed to balance the trade-off between the two. It is defined as:

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{20}$$

where TP (True Positive) represents the number of positive samples correctly predicted by the model, TN (True Negative) refers to the number of negative samples correctly predicted, FP (False Positive) indicates the number of negative samples incorrectly predicted as positive, and FN (False Negative) represents the number of positive samples incorrectly predicted as negative.

For the intelligent diagnosis of flame combustion states, the large model must first possess an accurate capability for recognizing and understanding the current flame image. As shown in Fig. 11, when the burner flame is in a fully developed combustion state, the model can easily learn and identify its visual features. However, three complex conditions pose significant challenges to the model's perception and interpretation: (1) flame-coal interweaving; (2) interference from adjacent flames; and (3) flame impingement on the water-cooled wall.

As illustrated in Fig. 12, this study utilizes 5-second flame video clips along with corresponding operational parameters as multimodal inputs, incorporating expert annotations to provide semantic descriptions of flame states. Flame images are grouped by load level: (a) S1 represents high load (560 MW), (b) S2 represents medium load (400 MW), and (c) S3 represents low load (280 MW); within each group, samples are labeled sequentially as Sx-1 to Sx-5 based on time. By examining samples under different load conditions, it can be observed that as the boiler load gradually decreases, the flame region tends to contract while the pulverized coal region expands, indicating a significant change in the spatial structure of combustion.

As shown in Fig. 13, two types of abnormal combustion conditions may occur under high and low load conditions: (a) flame impingement on the water-cooled wall at 560 MW and (b) interference from adjacent flames at 280 MW. First, during high-load conditions, improper adjustment of operational parameters may cause the flame to deviate from its intended direction and impinge on the water-cooled wall (S4-1 to S4-5), posing a significant risk of furnace heat load deviation. Second, during low-load conditions, a large accumulation of pulverized coal particles is visible within the combustion region (S5-1 to S5-5), but without a clearly defined flame structure, indicating a risk of incomplete combustion or flameout. In addition, flame interference from adjacent burners appears in the upper portion of the image, further increasing the likelihood of misdiagnosis for the current burner and disrupting accurate combustion state recognition.

The experimental results, as shown in Table 2, demonstrate that the proposed multimodal fusion model exhibits strong robustness and generalization capability when dealing with complex flame scenarios. For the task of identifying flame-coal interweaving, the model achieved an accuracy of 92.1% and an F1-score of 0.899, effectively capturing the coupled features of coal particle distribution and combustion morphology. In the task of detecting flame impingement on the water-cooled wall, the model similarly achieved 89.5% accuracy and an F1-score of 0.886, indicating strong interpretability of flame directional changes and temporal sequence patterns. These results further validate the effectiveness of the proposed multimodal feature fusion strategy in integrating visual, thermal, and flow-related information, significantly enhancing the model's ability to detect abnormal combustion states. For the task involving interference from adjacent flames, although the interactions and blurred boundaries between flames increase the difficulty of recognition, the model still reached an accuracy of 88.2% and an F1-score of 0.871, showing strong classification performance. The mutual interference between adjacent flames often causes traditional methods to fail in distinguishing flame regions. In contrast, the multimodal model, by leveraging diverse sources of information, effectively identifies such interference and maintains high recognition accuracy.

To comprehensively evaluate the contribution of different modalities to model performance, this study designs a series of ablation experiments that systematically assess the effectiveness of using only operational parameters, only the image modality, and both modalities in fusion as inputs. The experimental results, as shown in Table 3, demonstrate that the model exhibits limited discriminative capability when solely relying on operational parameters, achieving an accuracy of 70.3% and an F1 score of 0.694. This suggests that although operational parameters can capture certain trends in combustion fluctuations, they are insufficient to characterize the spatial structure of the flame, thereby limiting the model's ability to assess combustion stability. When using

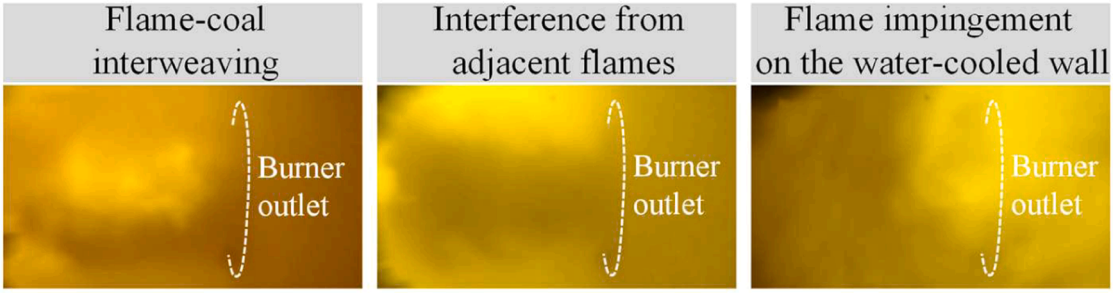


Fig. 11. Examples of flame recognition states.

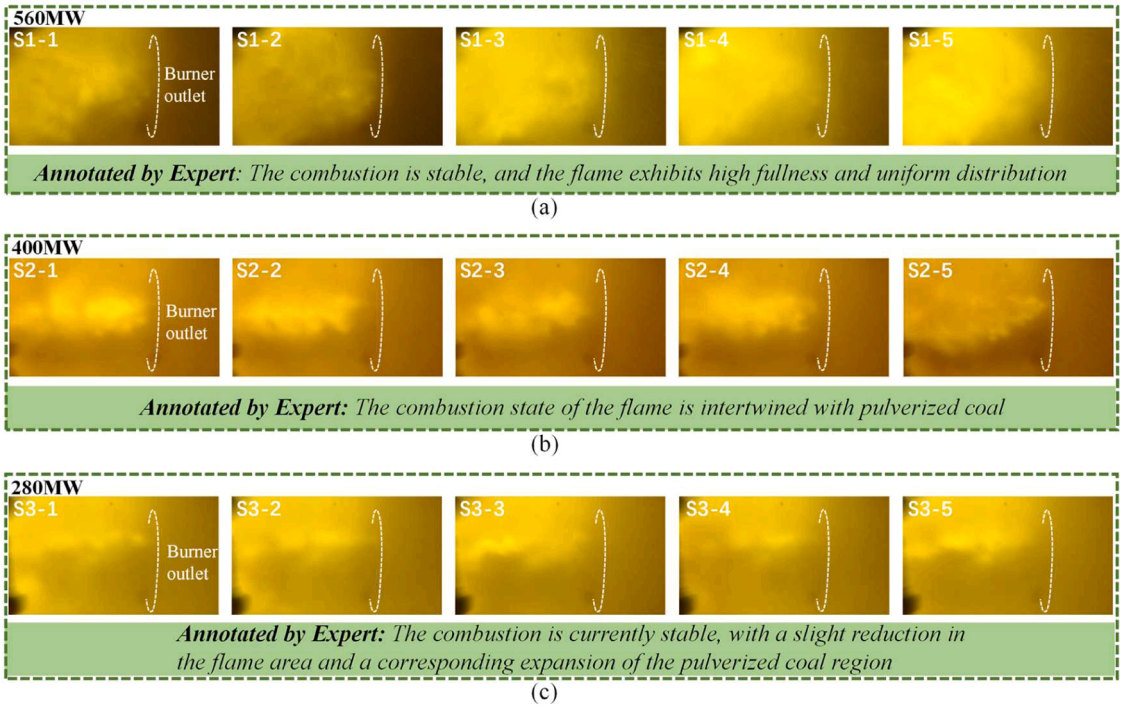


Fig. 12. Examples of flame recognition states under varying loads annotated by experts.

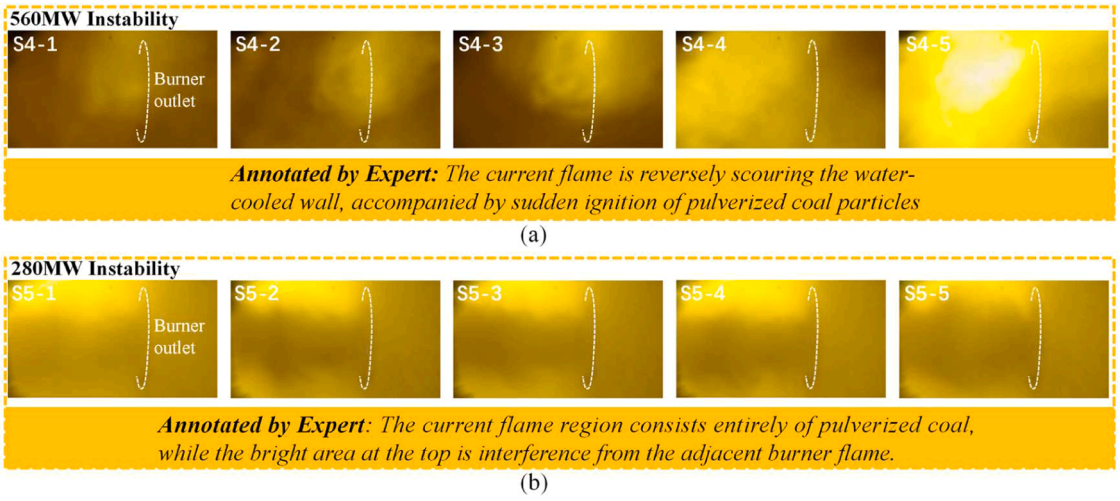


Fig. 13. Examples of flame recognition states annotated by experts at (a) 560MW and (b) 280MW unstable combustion conditions.

Table 2
Model evaluation across different tasks of recognition.

Task	Accuracy (%)	Recall (%)	F1
Interference from adjacent flames	88.2	85.8	0.862
Flame impingement on the water-cooled wall	89.5	87.8	0.886
Flame-coal interweaving	92.1	89.5	0.899

Table 3
Ablation study on different modality combinations.

Modality	Accuracy (%)	Recall (%)	F1	Inference time
Operational parameters	70.3	68.5	0.694	335ms
Single image	78.9	76.8	0.778	612ms
Image-operational parameters	89.5	87.8	0.886	759ms

only the image modality, the model achieves improved performance, with an accuracy of 78.9% and an F1 score of 0.778, indicating that image-based features provide valuable spatial and semantic information for identifying flame morphology. Further performance gains are observed when the image and operational parameters modalities are fused, yielding the best results with an accuracy of 89.5%, a recall of 87.8%, and an F1 score of 0.886. These findings verify the complementary strengths of spatial and operational parameters and highlight the effectiveness of multimodal fusion in enhancing the robustness and generalization capability of combustion stability assessment. In terms of inference time, the average processing times for Operational parameters, Single image, and Image-operational parameters were 335 ms, 612 ms, and 759 ms, respectively. Combined with the performance metrics, it can be seen that multimodal fusion significantly improves the diagnostic performance of the model while still maintaining high inference efficiency, thereby meeting the application requirements of combustion state recognition and online diagnosis.

In summary, the experimental results clearly demonstrate the complementarity between different modalities. The image modality provides spatial structure and semantic information, while the temporal parameters contribute dynamic evolution characteristics. Their combination not only enhances the model's robustness but also improves its generalization ability in recognizing complex combustion states, highlighting the significant advantages of multimodal fusion in flame diagnosis tasks.

4.3. Diagnosis of combustion stability

The trained CSD-Llama is deployed locally to enable real-time diagnosis of combustion stability. As illustrated in Fig. 14, the user issues a query to the agent, the CSD-Llama, which functions as the central processing unit. It interprets and processes the input, invokes the combustion stability diagnostic tool and returns the current combustion state to the user. The tools were not constructed as independent components from scratch; instead, they were further encapsulated from the functional modules established in the previous sections into callable tools. Specifically, the program for reading flame videos and operational parameters was encapsulated as the Data Acquisition Tool, the ViTAE and LSTM models were encapsulated as the Feature Extraction Tool, the multimodal feature fusion network was encapsulated as the Feature Fusion Tool, and the CSI computation module was encapsulated as the CSI Calculation Tool, thereby forming a toolchain for combustion stability diagnosis.

To demonstrate the diagnostic capability of CSD-Llama, the variation of the CSI under different load conditions is presented. As shown in Fig. 15, under high load conditions (560MW), the flame exhibits complete combustion, characterized by a higher proportion of saturated intensity regions in the image and a relatively stable flame morphology. Meanwhile, the control strategies and parameter configurations remain consistent. In this state, the CSI remains stable around 0.8 with minimal fluctuations, indicating good combustion stability. As the system enters the middle load conditions (400MW), although the boiler load remains steady, variations in parameter settings occur across different cycles. Additionally, the intertwining of flame and coal powder becomes more frequent, leading to more unstable flame shapes. As a result, the average CSI decreases to approximately 0.75, accompanied by increased fluctuations. When the boiler load is further reduced to the low load conditions (280MW), combustion conditions continue to deteriorate, and the average CSI drops further to around 0.7. In summary, under the combined influence of increasing flame variability and shifting operational parameters, the CSI exhibits a downward trend as boiler load decreases. This trend aligns with the practical evolution of combustion stability observed in engineering operations. Therefore, this study sets a CSI threshold for stability assessment, enabling effective evaluation of combustion states under variable load conditions.

Although combustion tends to be more stable under 560MW

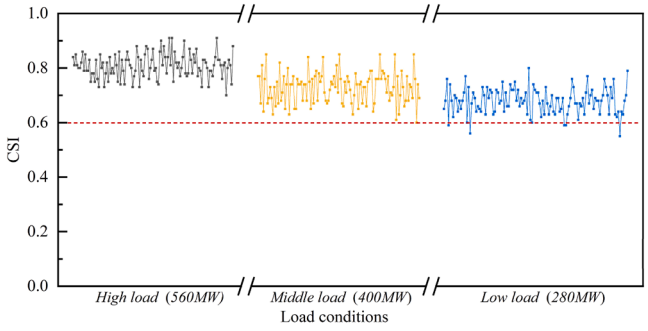


Fig. 15. Combustion stability index under different load conditions.

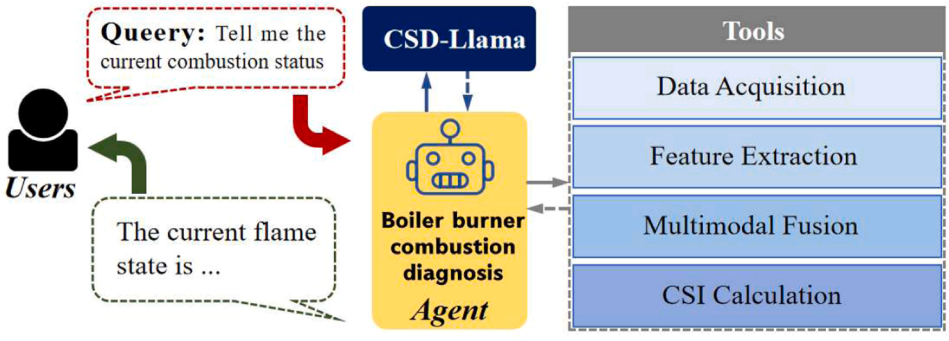


Fig. 14. Intelligent Agent-Based Combustion Diagnosis.

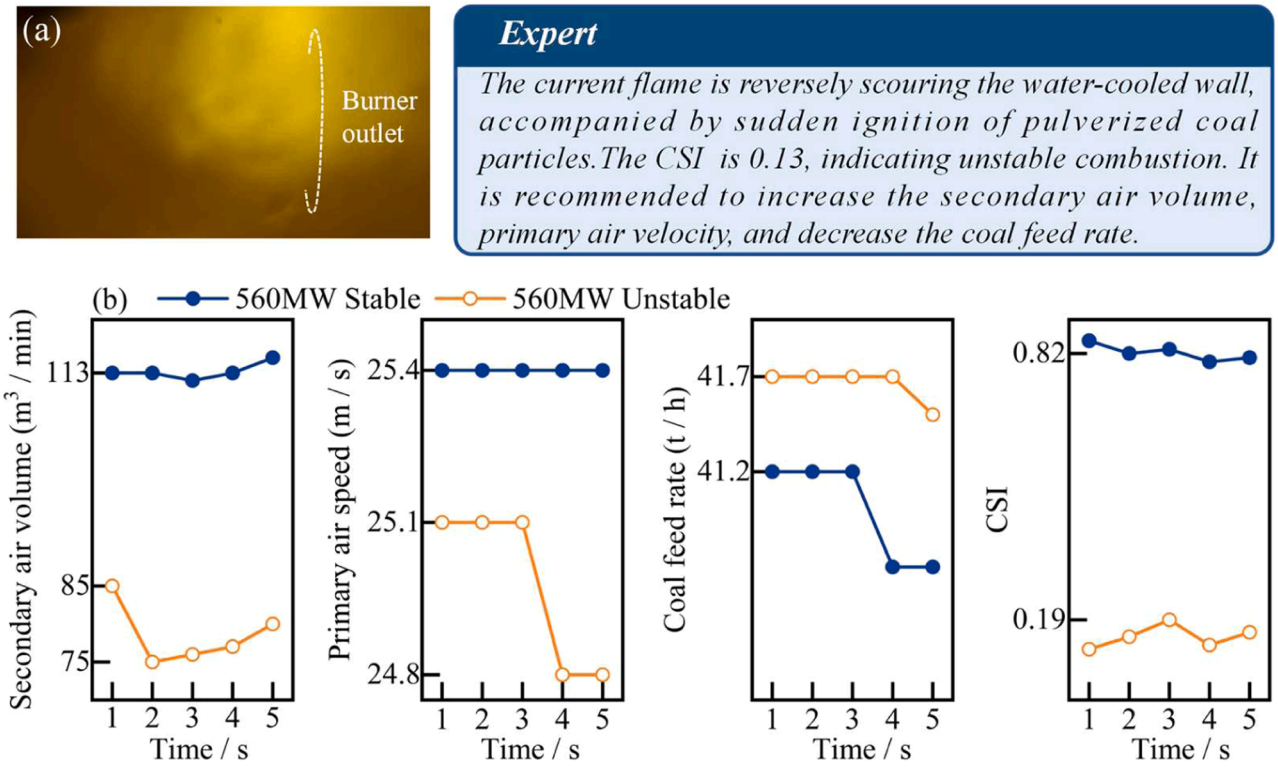


Fig. 16. Intelligent diagnosis of instability conditions under a high load.

conditions and operational parameters are generally consistent, adjustments are still required based on real-time operating demands. During this process, improper coordination among operational parameters can lead to flame impingement on the water-cooled walls, as illustrated in Fig. 16a. A comparison of the actual parameters under this abnormal condition with those of stable combustion at the same load (Fig. 16b)

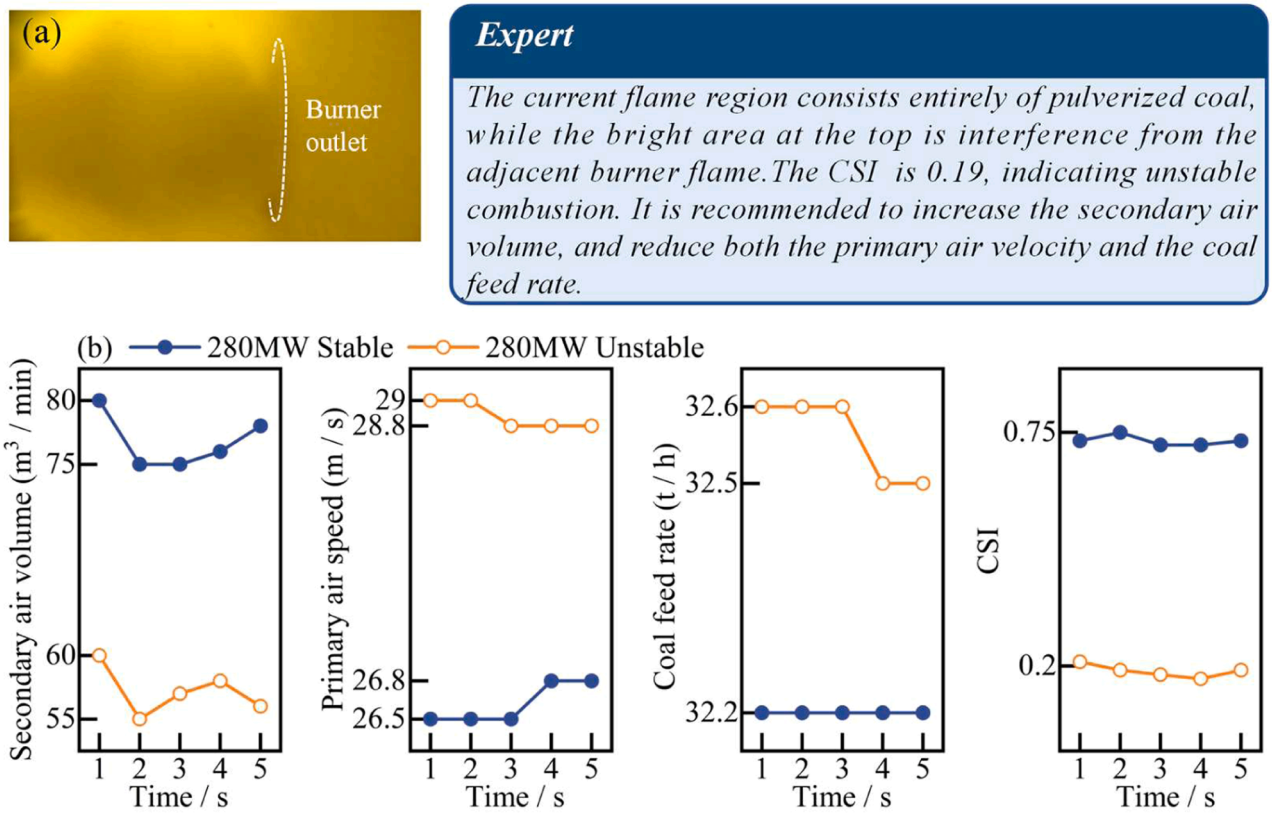


Fig. 17. Intelligent diagnosis of instability conditions under a low load.

reveals significant deviations. Specifically, the secondary air volume drops markedly to approximately 85 m^3/min, the primary air velocity decreases to 24.8 m/s, while the coal feed rate increases to 41.7 T/h. Under the same boiler load, this mismatch results in insufficient combustion air per unit mass of coal, causing a decline in oxygen concentration, which hinders complete combustion. The reduction in primary air velocity weakens the momentum of the carrier airflow, impairing the stability of coal particle injection. More critically, the decreased secondary air volume directly reduces the swirl intensity of the burner. Swirl burners rely on the rotational flow generated by secondary air to anchor the flame and ensure stable combustion. When the swirl weakens, the flame spreads over a wider angle and becomes more prone to deviating towards the furnace walls, potentially leading to contact with the water-cooled walls and forming an asymmetric flame structure. Based on expert knowledge, it is recommended to increase both the secondary air volume and the primary air velocity while simultaneously reducing the coal feed rate. This coordinated adjustment helps restore a stable combustion state and prevents wall impingement under high-load conditions.

As illustrated in Fig. 17, under 280MW conditions, the secondary air volume is reduced to 60 m^3/min. In contrast, both the primary air velocity and coal feed rate exceed those typically observed during normal low-load operations. Theoretically, a moderate increase in coal feed at low loads can help maintain adequate pulverized coal concentration in the furnace, enhancing flame intensity to prevent flameout. However, in this abnormal case, there is a clear mismatch among operational parameters: the insufficient secondary air supply fails to provide adequate oxygen for complete combustion, while the elevated coal feed further increases the combustion burden. Based on expert knowledge, it is determined that the secondary air volume should be increased, while the primary air velocity and coal feed rate should be reduced.

5. Conclusions

In this study, an intelligent diagnostic model for coal-fired boiler burners is developed by integrating multimodal features with a large language model. Temporal features of boiler operational parameters are extracted using an LSTM network, while spatial features of flame videos are extracted using a ViTAE network. A multimodal feature fusion framework integrates information from different modalities. Building upon this, an intelligent combustion diagnostic model based on the Llama 3.2–3B large language model incorporates expert-annotated diagnostic labels and enables interactive capabilities and parameter optimisation suggestions. Combustion diagnosis experiments and model validation were conducted on a 660 MW coal-fired boiler, achieving an intelligent assessment of combustion stability. The main findings are as follows:

- When a model relies on a single input modality, its recognition performance is limited. Using only operational parameters leads to lower diagnostic accuracy, while incorporating only image features provides a moderate improvement. In contrast, a multimodal model that integrates both image and operational data achieves superior performance across all evaluation metrics. This demonstrates that multimodal fusion effectively combines temporal information and spatial-semantic features, enhancing the model's ability to diagnose and interpret complex flame behaviours.
- The system introduces the combustion stability index, a metric based on multimodal feature similarity, to quantify the proximity of current operating conditions to stable reference states. Results show that combustion state identification decreases progressively as boiler load decreases, consistent with real-world patterns of combustion stability. A combustion state index threshold of 0.6 is proposed as an effective criterion for evaluating combustion stability under variable load conditions.
- By integrating visual information, operational parameters, and expert knowledge representation, the multimodal fusion approach significantly enhances diagnostic accuracy and adaptability. Furthermore, the Llama3.2-based language model architecture empowers the agent with semantic understanding and reasoning capabilities, enabling it to produce interpretable diagnostic insights.

Future research will further expand the multimodal framework by incorporating additional sensing modalities such as furnace temperature fields, flue gas composition and acoustic signals to achieve a more comprehensive representation of combustion dynamics. Furthermore, reinforcement learning and knowledge-graph-based reasoning will be explored to enable adaptive optimisation and causal inference of combustion processes.

CRediT authorship contribution statement

Sixu Pu: Writing – original draft, Visualization, Methodology, Conceptualization. **Yi Zhou:** Formal analysis, Data curation. **Md.Moinul Hossain:** Investigation. **Xiaoyu Zhu:** Software. **Guoqing Chen:** Project administration. **Chuanlong Xu:** Writing – review & editing, Resources, Funding acquisition.

Declaration of competing interest

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

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2023YFB4102904) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX22_0041).

Data availability

Data will be made available on request.

References

[1] Dong Z, Ye X, Jiang J, Li C. Life cycle assessment of coal-fired solar-assisted carbon capture power generation system integrated with organic Rankine cycle. J Clean Prod 2022;356:131888.
[2] Yang K, Li Z, Cao X, Du T, Liu L. Numerical simulation study on the stable combustion of a 660 MW supercritical unit boiler at ultra-low load. Processes 2024;12(11):2573.
[3] Wang H, Fu Z, Wang S, Zhang W. Analysis of CO_2 emissions in the whole production process of coal-fired power plant. Sustainability 2021;13(19):11084.
[4] Zheng H, Song M, Shen M. The evolution of renewable energy and its impact on carbon reduction in China. Energy 2021;237:121639.
[5] Zhang H, Shu Y, Wang X, Zhou X, Li W, Zheng H, et al. Improving the flexibility of coal-fired power plants via a pre-gasification burner with ultra-enhanced flame stability. Engineering 2025.
[6] Wei M, Wang J, Gao S, Li J, Pang X. An operating zone model for safety and efficiency monitoring of power generation units in thermal power plants. Control Eng Pract 2024;153:106101.
[7] Shan L, Huang H, Hong B, Zhao J, Wang D, Kong M. Temperature measurement method of flame image fusion with different exposures. Energies 2020;13:1487.
[8] Lin J, Zhang X, Liu K, Zhang W. Emissivity characteristics of hydrocarbon flame and temperature measurement by color image processing. Energies 2019;12:2185.
[9] Choi O, Choi J, Kim N, Lee M. Combustion instability monitoring through deep-learning classification of sequential high-speed flame images. Electronics 2020;9:848.
[10] Yan Y, Lu G, Colechin M. Monitoring and characterisation of pulverised coal flames using digital imaging techniques. Fuel 2002;81(5):647–55.
[11] Wang X, Li M, Liu Q, Chang Y, Zhang H. Multi-scale flame situation detection based on pixel-level segmentation of visual images. Appl Sci 2023;13:11088.
[12] Pu S, Li J, Han Z, Zhu X, Xu C. Flame region segmentation of a coal-fired power plant boiler burner through YOLOv8-MAH model. Fuel 2025;398:135518.

[13] Samantaray B, Mohanta C. Analysis of industrial flame characteristics and constancy study using image processing technique. J Mech Eng Sci 2015;9:1604–13.
[14] Lu G, Yan Y, Colechin M, Richard H. Monitoring of oscillatory characteristics of pulverized coal flames through image processing and spectral analysis. IEEE Trans Instrum Meas 2006;55:226–31.
[15] Cao Z, Lyu Y, Peng J, Qiu P, Liu L, Yang C, et al. Experimental study of flame evolution, frequency and oscillation characteristics of steam diluted micro-mixing hydrogen flame. Fuel 2021;301:121078.
[16] Chi T, Zhang H, Yan Y, Zhou H, Zheng H. Investigations into the ignition behaviors of pulverized coals and coal blends in a drop tube furnace using flame monitoring techniques. Fuel 2010;89:743–51.
[17] Yang G, He Y, Li X, Liu H, Lan T. Gabor-GLCM-Based Texture Feature Extraction Using Flame Image to Predict the O2 Content and Nox. ACS Omega 2022;7:3889–99.
[18] Gaidhane V, Hote Y. An efficient edge extraction approach for flame image analysis. Pattern Anal Appl 2018;21:1139–50.
[19] Sun D, Lu G, Zhou H, Li X, Yan Y. A simple index based quantitative assessment of flame stability. In: IEEE International Conference on Imaging Systems and Techniques; 2013. p. 190–3.
[20] Li D, Deng Q, Lee D, Jeon C. Prediction of attrition rate of coal ash for fluidized bed based on chemical composition with an artificial neural network model. Fuel Process Technol 2022;225:107024.
[21] Yin W, Xia H, Huang X, Zhang J, Miyombo M. A fault diagnosis method for nuclear power plant rotating machinery based on adaptive deep feature extraction and multiple support vector machines. Progr Nucl Energy 2023;164:104862.
[22] Liu S, Zhou X, Yu J, Wang Y, Xu T, Wang H. Graph attention Network-Based model for multiple fault detection and identification of sensors in nuclear power plant. Nucl Eng Des 2024;419:112949.
[23] Lyu Y, Chen J, Song Z. Image-based process monitoring using deep learning framework. Chemometr Intell Lab Syst 2019;189:8–17.
[24] Zhou Y, Zhang C, Han X, Lin Y. Monitoring combustion instabilities of stratified swirl flames by feature extractions of time-averaged flame images using deep learning method. Aerosp Sci Technol 2021;109:106443.
[25] Qiu T, Liu M, Zhou G, Wang L, Gao K. An unsupervised classification method for flame image of pulverized coal combustion based on convolutional auto-encoder and hidden Markov model. Energies 2019;12:2585.
[26] Han Z., Hossain M., Wang Y., Li J., Xu C. Combustion stability monitoring through flame imaging and stacked sparse autoencoder based deep neural network. Appl Energy, 2020, 259: 114159.
[27] Alsaif K, Albeshri A, Khemakhm M, Eassa F. Multimodal large language model-based fault detection and diagnosis in context of industry 4.0. Electronics 2024;13:4912.
[28] Khan A, Nahar R, Chen H, Constante F, Li C. FaultExplainer: leveraging large language models for interpretable fault detection and diagnosis. Comput Chem Eng 2025;199:109152.
[29] Qaid H., Zhang B., Li D., Kiong S., Li W.FD-LLM: Large language model for fault diagnosis of machines. 2024, https://doi.org/10.48550/arXiv.2412.01218.
[30] Lin L, Zhang S, Fu S, Liu Y. FD-LLM: Large language model for fault diagnosis of complex equipment. Adv Eng Inform 2025;65:103208.
[31] Sharm V, Raman V. A reliable knowledge processing framework for combustion science using foundation models. Energy AI 2024;16:100365.
[32] Xu L., Mohaddes D., Wang Y. LLM agent for fire dynamics simulations. 2024, https://doi.org/10.48550/arXiv.2412.17146.
[33] Alkhammash E. Leveraging Large language models for enhanced classification and analysis: fire incidents case study. Fire 2025;8:7.
[34] Martinez F, Rendon A, Penagos C. Flame analysis and combustion estimation using large language and vision assistant and reinforcement learning. Int J Artif Intell 2025;14:1853–62.
[35] Xie Y., Mallick T., Bergerson J., Hutchison J., Verner D., Branham J., et al. WildfireGPT: tailored large language model for wildfire analysis. 2024, https://doi.org/10.48550/arXiv.2402.07877.
[36] Wu S, Qiao Y, He S, Zhou J, Wang Z, Li X, et al. FireCLIP: enhancing forest fire detection with multimodal prompt tuning and vision-language understanding. Fire 2025;8:237.
[37] Seidel L, Gehringer S, Raczok T, Ivens S, Eckardt B. Advancing early wildfire detection: integration of vision language models with unmanned aerial vehicle remote sensing for enhanced situational awareness. Drones 2025;9:347.
[38] Yuntia A, Pratama M, Almuzakki M, Ramadhan H, Akashah E, Mansur A, et al. Performance analysis of neural network architectures for timeseries forecasting: a comparative study of RNN, LSTM, GRU, and hybrid models. MethodsX 2025;15:103462.
[39] Xu F., Zhang Q., Zhang J., Tao D. ViTAE: Vision transformer advanced by exploring intrinsic inductive bias. 2021, https://doi.org/10.48550/arXiv.2106.03348.
[40] Tang W, Qing L, Wang P, Li L, Peng Y. Cross-attention fusion of graph and MLLM for social relation recognition. Inf Fusion 2025;123:103259.
[41] Mondillo G, Colosimo S, Perrotta A, Frattolillo V, Masino M, Martino M, et al. Artificial intelligence for solving pediatric clinical cases: aRetrieval-Augmented approach utilizing Llama3.2 and structured references. Int J Med Inform 2025;203:106027.
[42] Luo D, Zheng K, Wu C, Wang X, Wang J. ERAT-DLoRA: parameter-efficient tuning with enhanced range adaptation in time and depth aware dynamic LoRA. Neurocoputing 2025;614:128778.
[43] Sohangir S, Wang D. Improved sqrt-cosine similarity measurement. J Big Data 2017;4:25.
[44] Avazov K, Mukhiddinov M, Makhmudov F, Cho Y. Fire detection method in smart city environments using a deep-learning-based approach. Electronics 2022;11:37.