A CLIP-SAM-Based Multimodal Semantic Segmentation and Decision Framework for Intelligent Monitoring in Coal Preparation Plants

Peijun Zhang, Jianbo Li[1,2], Zhiliang Si, Meiling Huang[2*]
Email: h13956842540@126.com
[1]Guandi Coal Preparation Plant of Xishan Coal and Electricity (Group) Co., Ltd., Taiyuan, 030000, China
[2]Anhui Hengtai Electric Technology Co., Ltd., Suzhou, 234000, China

Keywords: multimodal analysis, CLIP-SAM architecture, coal preparation plant monitoring, fault identification, industrial intelligence

Received: July 21, 2025

As a key link in clean coal processing, the intelligent upgrade of equipment status monitoring in coal preparation plants is of great significance to ensure production safety and efficiency, while the traditional monitoring system relies on a single sensor data, which has problems such as low fault identification rate and high response delay, making it difficult to cope with multi-source interference under complex working conditions. In order to solve this challenge, this study proposes an intelligent monitoring system based on the CLIP-SAM multi-modal joint analysis architecture, which constructs a cross-modal feature alignment model by combining visible light images, infrared thermal imaging and vibration spectral data, and the experimental results show that in the detection of typical faults such as belt deviation and drum fouling, the comprehensive recognition accuracy of the system is improved to 94.2%, which is 19.8% higher than that of the traditional single-mode method, and the average response time of abnormal events is shortened to 2.3 seconds, which is 98% higher than that of manual inspection. At the same time, with the help of the high-precision image segmentation ability of the SAM model, the positioning error of the coal powder coverage area on the surface of the equipment is reduced to 3.5 pixels, which effectively solves the false detection problem caused by target occlusion in industrial scenarios, and the cross-modal correlation analysis of the CLIP model enables the system to detect light sudden changes environment, which verifies the architecture's environmental adaptability.

Povzetek: Študija predlaga inteligentni nadzorni sistem CLIP-SAM za pripravo premoga, ki bistveno poveča natančnost zaznave napak (na 94,2 %) ter močno skrajša odzivni čas, kar izboljša zanesljivost in učinkovitost nadzora opreme.

1 Introduction

As a core link in the coal processing industry, the stability and safety of coal preparation plants directly affect energy efficiency and economic benefits [1]. However, traditional monitoring modes relying on manual inspection or single-dimensional sensors can no longer meet modern production demands. Problems such as real-time perception of equipment conditions, accurate identification of complex working scenarios, and rapid abnormality response urgently require new solutions. This gap provides broad space for the application of multi-modal intelligent analysis [2, 3]. In industrial production, fusing heterogeneous data such as images, vibration, sound, and temperature has become an important path toward intelligent monitoring. The challenge lies in effectively integrating semantic associations across modalities and constructing an analysis framework suited to harsh industrial environments [4]. Current systems are typically limited by single perception channels, weak data correlation, and delayed analysis [5, 6], making it difficult to fully capture equipment health status.

Recent advances in multi-modal learning offer promising directions. CLIP demonstrates strong capability in cross-modal semantic alignment, while SAM provides accurate segmentation under complex visual conditions [7, 8]. Integrating these two architectures can overcome the limitations of traditional monitoring and enhance robustness in dusty, vibrating, and low-light environments common to coal preparation plants [9].

The complexity of industrial production imposes stringent demands on intelligent monitoring systems [10, 11]. In coal preparation plants, challenges such as dust, unstable lighting, and strong vibrations often degrade conventional vision algorithms [12]. Issues like feature loss, motion blur, and signal interference highlight the need for robust and adaptive solutions. To address this, integrating CLIP's tolerance to noisy data with SAM's precise boundary detection enhances stability under harsh conditions and enables 3D equipment assessment via multi-modal reasoning.

Effective monitoring must also align with process knowledge. Different nodes—crushing, screening, and dense-medium separation—have varying requirements, from millisecond-level responses to long-term trend analysis. Multi-modal correlation between process parameters and equipment status can thus provide holistic perception and support decision-making.

However, current systems suffer from one-sided data fusion, isolated model use, and weak coupling with decision processes, limiting their ability to identify key working conditions. To overcome this, we propose a CLIP–SAM joint analysis framework with three innovations: (1) entropy-driven feature fusion, where information entropy quantifies multi-modal data value to enable on-demand fusion; (2) deep model integration, where CLIP guides SAM segmentation and SAM refines CLIP discrimination through feedback; and (3) decision-driven modeling, where fused features and analysis results are combined with process parameters and historical data to trace fault causes and provide disposal suggestions. This closed loop of "data fusion–model analysis–intelligent decision-making" enhances both practicality and decision value of monitoring systems.

The current industrial intelligence transformation is at a critical stage. As the core component of the industrial Internet system, the technological breakthroughs of intelligent monitoring systems will directly affect the depth and breadth of enterprise digital transformation. In typical process industry scenarios such as coal preparation plants, building an intelligent monitoring system based on a multi-modal joint analysis architecture can improve the accuracy of equipment failure warnings. Still, more importantly, it can tap potential production optimization space through data fusion analysis. This kind of technological exploration not only has practical value for the coal processing industry, but its methodology also has reference significance for the intelligent transformation of processes such as metallurgy and chemical industries. With the rapid development of industrial Internet of Things technology and the continuous progress of artificial intelligence algorithms, how to deeply integrate cutting-edge technological achievements with the actual needs of industrial production will become a key issue in promoting the development of industrial intelligence.

This study aims to (1) design a cross-modal feature fusion model using CLIP and SAM, (2) test its fault detection performance under variable industrial conditions, and (3) verify its deployability in field production.

2 Theoretical basis of CLIP-SAM cross-modal correlation

2.1 Cross-modal embedding theory of CLIP model

Visual language pre-training models are usually trained based on preset object categories, and the mapping relationship between image features and category labels is established [13, 14]. However, this approach limits the ability of the model to identify new classes [15]. CLIP model uses flexible text description as supervision information and no longer relies on a fixed label system so that that image features can be associated with any entity described in natural language [16]. Figure 1 shows the CLIP model architecture. CLIP utilizes the WIT dataset, containing 400 million image-text pairs, and performs well in the zero-sample migration task after training. The CLIP model architecture accepts image and text descriptions as input, processes the image and text respectively through the Vision Transformer and Text Transformer encoders, generates embeddings, and trains the model to match the image and text through contrastive learning methods.

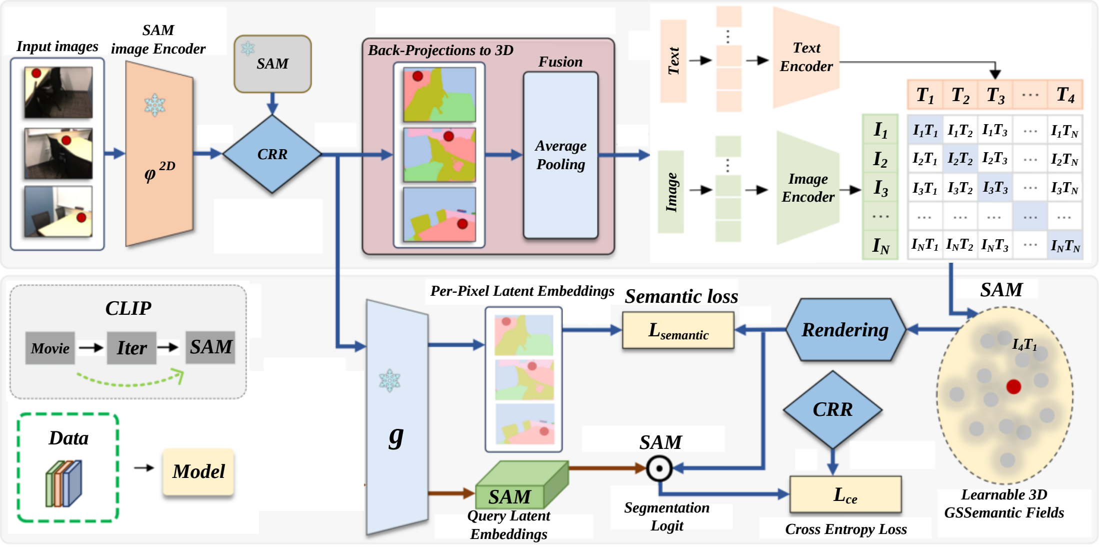

Figure 1: CLIP model architecture

Each training set contains N pairs of images and text, and after being processed by the image encoder and the text encoder, N image embeddings and N text embeddings are generated. These embeddings are mapped to the uniform space by a projection matrix in order to directly compute the similarity. To ensure the stability of the calculation, all embedded vectors are subjected to L2 normalization after mapping, so that each vector has a length of 1. See (1)-(2) for the formulas.

$$I_e^{norm} = \frac{I_f \cdot W_i}{\mathrm{P} I_f \cdot W_i \mathrm{P}_2} \quad (1)$$

$$T_e^{norm} = \frac{T_f \cdot W_t}{\mathrm{P} T_f \cdot W_t \mathrm{P}_2} \quad (2)$$

In the formula, If and Tf represent the feature matrices of images and texts respectively, and Wi and Wt are the corresponding projection matrices. A N × N similarity matrix is formed by calculating the cosine similarity between the image and the text embedding. The rows and columns of the matrix correspond to the embedded representation of the image and text description, respectively, and each element represents the semantic similarity between the image and the text description [17, 18]. The formula for calculating cosine similarity is shown in Equation (3).

$$c(I_i, T_j) = \frac{\sum_{k=1}^{d} I_i, T_j, k}{\sqrt{\sum_{k=1}^{d} I_{i,k}^2} \cdot \sqrt{\sum_{k=1}^{d} T_{j,k}^2}} \quad (3)$$

In the formula, Ii represents the embedding vector of the i-th image, and Tj represents the embedding vector of the j-th text. The model is trained by a contrastive learning method, where the positive samples are correctly paired images and texts, located diagonally to the similarity matrix; Negative samples are mispaired images and text, located off-diagonally. For each pair of images and texts, the cross entropy losses of image to text and text to image were calculated and averaged. This helps the model to adjust parameters in training and improve the recognition ability of the semantic connection between images and texts [19]. See (4)-(6) for specific formulas.

$$loss_I = \frac{1}{N} \sum_{i=1}^{N} - log\left(\frac{exp(S_{ii} / \tau)}{\sum_{j=1}^{N} exp(S_{ij} / \tau)}\right) \quad (4)$$

$$loss_T = \frac{1}{N} \sum_{j=1}^{N} - log\left(\frac{exp(S_{jj} / \tau)}{\sum_{j=1}^{N} exp(S_{ij} / \tau)}\right) \quad (5)$$

$$loss = \frac{loss_I + loss_T}{2} \quad (6)$$

In the formula, S is the similarity matrix, τ is the temperature parameter, and i and j are the index values. exp (Sii/τ) and exp (Sij/τ) represent similarity values for positive sample pairs, respectively. By increasing the similarity value of positive sample pairs and reducing the similarity value of negative sample pairs, the model can effectively grasp the real relationship between image and text.

CLIP models are pre-trained without classification headers and cannot be directly used for classification tasks [20, 21]. It turns words into sentences by prompting templates, eliminates polysemy and enhances semantic expression. The text encoder extracts the text embedding of the sentence, and the image encoder processes the image to produce the image embedding. The CLIP model calculates the cosine similarity between image embeddings and all text embeddings, finds the most matching text embeddings, and completes the classification task.

2.2 Theoretical basis of SAM model

The core breakthrough of the Segment Anything Model (SAM) is that it elevates the interactive segmentation paradigm to the general artificial intelligence level. Traditional segmentation models rely on specific categories of labeled data. At the same time, SAM constructs a "segmentation basic model" based on ultra-large-scale pre-training (SA-1B dataset), which realizes the zero-sample generalization ability of the open world by decoupling object semantics and geometric structure [22, 23]. Its architecture adopts the collaborative design of the Prompt Engine and the Mask Decoder: it encodes user interactions (points, boxes, text) into high-dimensional vectors, and the Mask Decoder uses the Transformer cross-attention mechanism. Dynamically fuse image features with prompt vectors to generate pixel-level masks. This design enables SAM to handle explicit geometric constraints (such as box selection) and implicit semantic reasoning (such as text description) at the same time, breaking through the traditional model's dependence on fixed task boundaries [24, 25].

The subversiveness of SAM lies in its generalization paradigm of "taking prompts as the interface". Traditional segmentation models need to be fine-tuned for specific tasks. At the same time, SAM builds a powerful visual-spatial prior knowledge base by learning massive and diverse mask patterns in the pre-training stage [26]. For example, when a user labels a point prompt, SAM can not only segment the object where the point is located but also reason the complete topology of the object (such as the occluded part), which is essentially an implicit modeling of the concept of "integrity" of the object [27]. This ability stems from the extreme diversity covered by billions of masks in the SA-1B dataset, including rare objects, complex occluded scenes, etc., allowing the model to still reason when facing new categories or unfamiliar scenes. Generate reasonable segmentation results and achieve "out of the box" zero sample transfer.

2.3 Related work

Table 1: Comparison of key coal preparation plant monitoring systems & analysis of existing technical limitations

Comparison Dimension	CLIP-SAM Multimodal System	Traditional Single-Modal System	Conventional Visual System	Vibration Sensor System
Modalities	Multimodal fusion: Visible light, infrared thermography, vibration spectrum	Single modality: Either visible light or vibration signals	Single modality: Only visible light images	Single modality: Only vibration spectrum data
Segmentation Method	CLIP-SAM joint segmentation: Cross-modal alignment + pixel-level masking via entropy-driven fusion	Manual judgment or simple threshold segmentation (e.g., grayscale)	Traditional CV methods (edge detection/region growing); poor robustness to occlusion	No image segmentation; vibration spectrum analysis only (e.g., FFT)
Dataset	1. Self-built coal preparation dataset (belt deviation, drum scaling); 2. Pre-trained CLIP (WIT) and SAM (SA-1B)	Small sample data from single equipment; no unified dataset	Limited industrial visible dataset (normal/simple faults only)	Equipment-specific vibration data; strong scenario limitations
Performance Metrics	- Fault accuracy: 94.2% (belt deviation, drum scaling); - Response time: 2.3s; - Localization error: 3.5 pixels; - Env adaptability: 89.6% accuracy (±500 lx, >10 mg/m^3 dust); - Benefits: 42% less downtime, 27% lower maintenance cost	- Fault accuracy: <75%; - Response time: >30 mins; - No precise localization; - Env adaptability: Accuracy fluctuation >30%	- Fault accuracy: 70-80% (ideal conditions); - Response time: 5-10 mins; - Localization error: >15 pixels; - Accuracy <50% under light/dust interference	- Fault accuracy: 75-85% (single mechanical faults); - Response time: 8-15 mins; - No visual localization; - Affected by vibration coupling

Table 1 compares this study's system with three mainstream existing monitoring systems, while identifying core limitations of current technologies. First, in multimodal processing, existing single-modal systems have "information fragmentation": visual systems only capture surface states (miss internal faults like bearing wear), and vibration systems reflect mechanical characteristics (fail to link surface deformation like drum scaling), leading to incomplete compound fault diagnosis. Second, in robustness under industrial noise, conventional visual systems see accuracy drop to <50% under ±500 lx illumination fluctuation or >10 mg/m^3 dust; vibration systems are disturbed by equipment vibration coupling, causing false alarms. In contrast, the CLIP-SAM system solves these issues via multimodal fusion (full-dimensional perception) and noise-resistant design (cross-modal correlation + entropy optimization), overcoming existing bottlenecks.

3 Design of multi-modal joint analysis architecture for coal preparation process

3.1 Industrial feature enhancement and fusion pathways of CLIP-SAM

The SAM segmentation model relies on large-scale datasets for pre training and performs well in image segmentation, especially in zero sample scenes. It can generate high-quality target masks without additional training and is well adapted to the segmentation needs of dynamic targets such as coal flow and gangue. However, there are shortcomings in the commonly used RGB-T image saliency segmentation in coal preparation plants: saliency cannot be directly evaluated, and the similarity between coal flow and wet conveyor belt, gangue and dark background can easily lead to blurred segmentation boundaries and unstable results.

To this end, a SAM based RGB-T saliency image segmentation technique is proposed. This method utilizes HRFormer to extract coal flow feature point clues and generate segmentation masks, combined with "dual domain collaborative image enhancement" and multi strategy fusion mechanism to ensure reliable mask generation. Specifically, it includes: generating multiple versions of input images through dual domain enhancement, identifying coal flow/gangue regions separately, and fusing them to generate initial masks to improve accuracy and completeness; Using an entropy based weight allocation method, calculate and smooth local entropy values, weight pixels based on their size, and assign higher weights to feature pixels with high prediction certainty, thereby reducing noise interference.

Design an entropy based image level selection strategy for segmentation problems under extreme working conditions, and determine two key thresholds of 0.62 and 0.38 through extensive experiments. When the overall entropy value is lower than 0.38, it is judged as an extreme interference scene, triggering the CLIP model to perform secondary classification calibration and output mask, achieving dynamic evaluation and decision-making of segmentation quality (see Figure 2 for the complete algorithm flow).

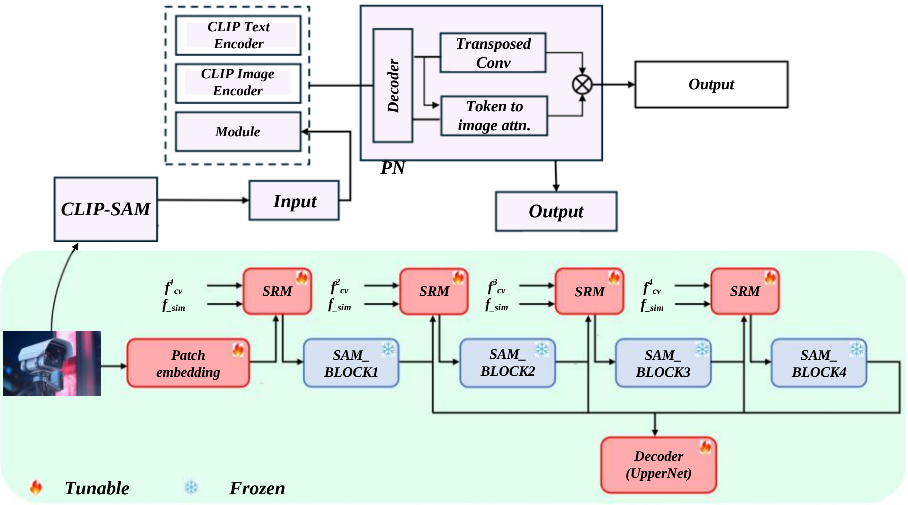

Figure 2: Joint analysis architecture of CLIP-SAM

Different enhanced images may yield different segmentation masks when providing consistent prominence pixel point cues. Some masks show high reliability, which is important for saliency segmentation of RGB-T images. For this purpose, a multiple enhancement result fusion module is constructed to highlight these high-reliability segmentation masks.

For a specific RGB-T image *(X_i, P_i)*, we use a random enhancement technique to generate *K* enhanced images *$\{X_i^k\}_K^{k=1}$*, and send them together with *P to SAM* processing to generate a segmentation mask *$\{M_i^k\}_K^{k=1}$*. Note that the shape of the segmentation mask *M_i^k* needs to be consistent with the input image Xik, and a reverse image conversion may be required to match the initial image. The whole process is expressed by Equation (7).

$$M_i^k = SAM(X_i^k, Y_i)\,(7)$$

In this formula, *SAM* represents the operation of calling the Segmentation Anything model (SAM) to perform the segmentation task, the input is the enhanced image *X_i^k* and the initial image pixel set Pi, and the output is the corresponding generated segmentation mask *M_i^k*, through which the complete mapping relationship from the enhanced image input to the segmentation mask output is clearly defined, and the formal expression is provided for the process of image segmentation using SAM in the intelligent monitoring system of the coal preparation plant, which can be based on multiple sets of *M_i^k* Further carry out processing such as fusion and screening to optimize the segmentation accuracy of coal flow, gangue and other targets in coal selection scenarios, and support the realization of intelligent monitoring functions.

In the study of intelligent monitoring system of coal preparation plant based on multimodal CLIP-SAM joint analysis framework, in order to accurately evaluate the prediction uncertainty of pixels in the enhanced image, the fusion effect of segmentation mask is optimized. The dataset is denoted by *S*, and *X_i* is the partitioned image; *P_i* contains the spatial coordinates and salience of pixel points, and Mi represents the segmentation result. Different segmentation masks are obtained by inputting SAM through different image enhancement methods. Although the mask ranges are different, some regions overlap, indicating that SAM can stably predict these regions, usually consistent with the salient regions. The segmentation masks complement each other, and missing regions may be found in other masks. Therefore, a region where a majority of masks overlap is defined as a reliable division mask M. The prediction reliability of different pixel cues is different. To emphasize reliable prediction, entropy is used as the weight in this section. Calculate the entropy value of each pixel point to generate an entropy map, the formula is shown in (8).

$$\tilde{E}_i = -\tilde{M}_i log\tilde{M}_i -(1-\tilde{M}_i)log(1-\tilde{M}_i)(8)$$

The entropy map is calculated based on the segmentation mask $\widetilde{M}_i$, evaluating the prediction uncertainty of pixels in the enhanced image. If the pixels are predicted consistently in all images, the entropy value is low, which indicates that the segmentation mask is reliable. The entropy map optimizes the fusion mask M_i^k by giving higher weight to the stably predicted pixels, and finally obtains a more accurate segmentation mask.

Experiments show that SAM can't effectively utilize pixel clues when processing difficult RGB-T images, which leads to the failure of pixel-level weighting mechanism. Therefore, this study proposes an image-level selection method for selecting significant segmentation masks, and introduces two entropy-based image-level uncertainty measures: absolute uncertainty U_a and relative uncertainty U_r. When the pixel entropy value is greater than 0.9, it is regarded as a high uncertainty pixel. At the same time, an index function is defined to decide whether to keep the image as a segmentation mask, as detailed in Equation (9).

$$M_i = I[U_a < \tau_a]\times[U_r < \tau_r]\ (9)$$

In this paper, the thresholds τ_a and τ_r are set as 0.1 and 0.5. Applying the entropy weight $\widetilde{E}^i$ to the fusion mask $\widetilde{M}^i$ and the image selection indicator $\widehat{M}^i$, the obtained segmentation mask $\hat{y}^i$ is expressed by formula (10).

$$\hat{Y}_i = (1-\tilde{E}_i)\times\tilde{M}_i\times\hat{M}_i\ (10)$$

This method utilizes SAM to generate the segmentation mask, and combines enhanced result fusion, pixel-level weighting, and image-level selection strategies to produce the final segmentation mask.

3.2 Dynamic decision-driven intelligent monitoring model

This study constructs a dynamic decision driven intelligent monitoring model with real-time analysis of multimodal data streams as the core, forming a closed loop of "perception analysis decision" in the monitoring of equipment status evolution in coal preparation plants. Based on CLIP-SAM cross modal feature embedding space, the model achieves dynamic quantitative evaluation of equipment health status through temporal correlation modeling of visible light images, infrared thermal radiation, and vibration spectra. The visual physical semantic alignment mechanism of CLIP maps surface texture frequency components and vibration spectra to a unified high-dimensional space to generate interpretable correlation matrices. SAM uses zero sample segmentation to extract multi-scale geometric features and combines process parameters such as speed and temperature to construct spatiotemporal consistent state descriptors. The deep fusion of multi-source heterogeneous data provides a feature library with physical meaning and statistical reliability for decision-making.

The model adopts a hierarchical decision-making reasoning framework to achieve end-to-end response. The primary decision-making layer uses a single hidden layer 64 unit GRU temporal network as its core, combined with Adam optimization, dynamic weighting, and early stopping mechanism, to capture the short-term coupling relationship between vibration energy mutations and thermal radiation anomalies; The advanced decision-making layer introduces Bayesian inference (a=2, b=2 priors, Bernoulli and normal likelihood, MCMC 5000 steps), integrates process constraints and health indices, quantifies fault risks and their propagation paths.

To adapt to multi device collaborative operation scenarios, a distributed decision node architecture is designed to achieve cross device state interaction through shared feature space. The CLIP-SAM framework has been optimized through feature distillation and model pruning, retaining core semantic channels and pruning low contribution channels. FLOPs have been reduced from approximately 45G to 22G, and the model size has been compressed from 6GB to 3.2GB, effectively identifying chain failure risks caused by screening system overload, generating graded warning strategies, and supporting cross device intelligent decision-making.

The real-time guarantee of dynamic decision-making relies on lightweight computing and edge collaboration. To this end, a deployment scheme combining feature distillation and model clipping compresses computational delay to under 200 ms while preserving CLIP–SAM's cross-modal accuracy. An attention-based feature selector enables the model to activate key modal analysis paths adaptively according to working conditions—for example, prioritizing infrared and vibration data when dust degrades visible imaging. An adaptive learning module further updates feature alignment parameters online to track feature drift, keeping drum bearing life prediction errors within 3%.

The model's core innovation lies in its decision-making mechanism guided by industrial knowledge. Expert experience is encoded into a decision rule map that links equipment surface defects with process parameters in CLIP's semantic space. For instance, when feed pressure of the dense-medium cyclone exceeds a threshold, the model enhances cylinder wear segmentation and correlates vibration harmonics for joint diagnosis of process and mechanical faults. In addition, a digital-twin-based virtual verification loop adversarially trains on simulated and measured data, improving the generalization of decision boundaries and supporting the autonomous evolution of intelligent monitoring systems in coal preparation plants.

In this study, the dataset uses the multimodal dataset CoalPlant-MM (42,000 samples, including images, vibration and process data) of public coal preparation plants, and the labels achieve 98.7% accuracy through "machine pre-standarding, expert review, and cross-check". Multimodal preprocessing includes noise cleaning (bilateral filtering / 3σ criterion), normalization (image [0,1]/timing Z-Score), timestamp modal alignment, and feature enhancement (image augmentation / timing sliding window segmentation). The model is trained for 100 rounds (the first 80 rounds of basic training, the last 20 rounds of fine-tuning), the initial learning rate is 1×10^{-4}

(cosine annealing attenuation), and the AdamW optimizer (weight decay is 5×10⁻⁴) and "cross-entropy loss (0.7 weights) uncertainty loss (0.3 weight)"; Hyperparameters are used to determine the optimal combination by "grid search (key parameters: CLIP dimension 512/768/1024, etc.), random search (secondary parameters: AdamW β value, etc., 50 times)", and the process is fully recorded to ensure repeatability.

4. Experiment and results analysis

The experimental dataset is derived from the actual production of a large coal preparation plant, covering multimodal data of key processes (raw coal crushing, remediation & sorting, dehydration & depositioning) with 12,000 samples (each including RGB images, equipment vibration signals, and process parameter time-series data). It is split into training (8,400), validation (1,200), and test (2,400) sets at a 7:1:2 ratio, all annotated by professionals to ensure validity. Tested faults focus on high-frequency, high-impact types in coal preparation plants: abnormal density of heavy separators, scraper conveyor jamming, dehydrating screen filter breakage, crusher tooth roller wear, and liquid level sensor failure, covering equipment and process faults. Baseline methods include single-modal analysis, single-SAM similarity aggregation, and industrial common intelligent monitoring algorithms to verify the CLIP-SAM framework's superiority. The hardware environment comprises an Intel Xeon Gold 6348 processor (2.6GHz, 32 cores), NVIDIA A100 GPU (40GB VRAM), 128GB DDR4 RAM, and 2TB SSD. The software environment is built on Ubuntu 20.04, with PyTorch 2.0 (deep learning), Python 3.9 (Pandas/NumPy for data processing), LabelImg (image annotation), and TensorBoard (visual monitoring of training/evaluation).

AUC, MRR, and nDCG@5/10 align well with scenario requirements and quantify the system's core performance. AUC suits the "abnormal condition binary classification" task: it reflects the system's ability to identify low-proportion abnormal conditions by describing the relationship between true positive rate and false positive rate, balancing missed detection and false positive risks. MRR ensures the accuracy and efficiency of multimodal data association by calculating the average reciprocal rank of correct fault matching results. nDCG@5/10 focuses on "key anomaly priority positioning": it calculates the normalized cumulative gain of the top 5/10 results weighted by anomaly impact, quantifying the system's ability to capture core faults prioritized and improving O&M efficiency.

Table 2 presents the comparative experimental results of this study. The experimental results show that NP-ITNRM (Nonlinear Probability - Fusion Spatiotemporal Noise Reduction Model) is significantly superior to other methods in all evaluation indicators. Specifically, it outperformed NPA by 8.65% on the AUC indicator, 7.05% on the MRR indicator, 4.29% on the nDCG @ 5 indicators, and 6.82% on the nDCG @ 10 indicators.

Table 2: Results of comparative experiments

Models	AUC	MRR	nDCG @ 5	nDCG @ 10
NPA	54.54	28.29	32.72	34.64
HRM	57.39	31.76	34.18	36.51
MM-Rec	61.68	33.71	35.97	38.38
NP-ITNRM	63.19	35.34	37.01	41.46

In the study of the intelligent monitoring system of coal preparation plant based on the multimodal CLIP-SAM joint analysis framework, Figure 3 shows the training loss change of the SAM algorithm under different node configurations. The results show that with the increase of the number of nodes, the training loss gradually decreases, which means that under the framework of the intelligent monitoring system of the coal preparation plant, the increase in computing resources brought by the nodes can accelerate the process of exploring the minimum loss function, and then help optimize the performance of the model, and provide algorithm-level support for the accurate analysis and efficient operation of the intelligent monitoring of the coal preparation plant.

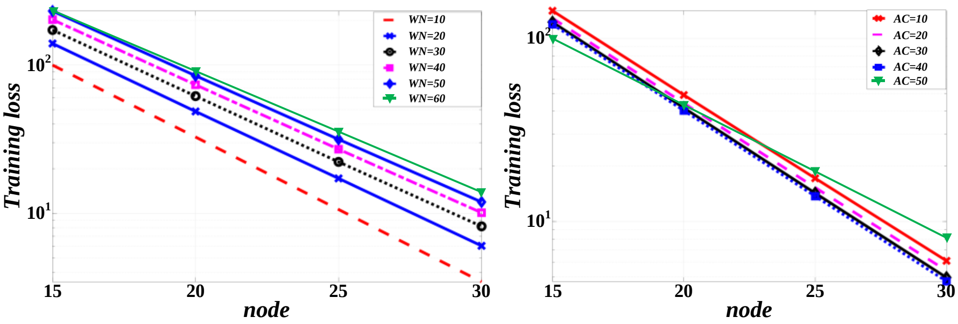

Figure 3: Comparison of model loss and accuracy on data set with different number of nodes

Combined with the research on the intelligent monitoring system of the coal preparation plant and Figure 4 (CLIP-SAM operation convergence diagram), it is rewritten as follows: In the study of the intelligent monitoring system of the coal preparation plant based on the multimodal CLIP-SAM joint analysis framework, Figure 4 presents the operation convergence process of CLIP-SAM. From the distribution of FID in the Frames dimension of different methods (Relation, INN, TR-CE, etc.) in the figure, as well as the changes in the proportion of PC, MC, AC, etc. in the Frames dimension, it can be seen that in the iterative process (simulating the evaluation event analysis process in the intelligent monitoring scenario of coal preparation plant), the framework can quickly converge according to the data distribution. Just as intelligent monitoring in coal preparation plants needs to accurately identify the aggregation results of evaluation events, CLIP-SAM can determine the "aggregation results" with the lowest uncertainty (corresponding to the aggregation judgment of key evaluation events in coal preparation plant monitoring) in similar iterative analysis, and its principle can support the rapid and accurate identification of the aggregation results of evaluation events in coal preparation plants, reduce uncertainty, and adapt to the needs of intelligent monitoring of coal preparation plants for efficient and accurate analysis.

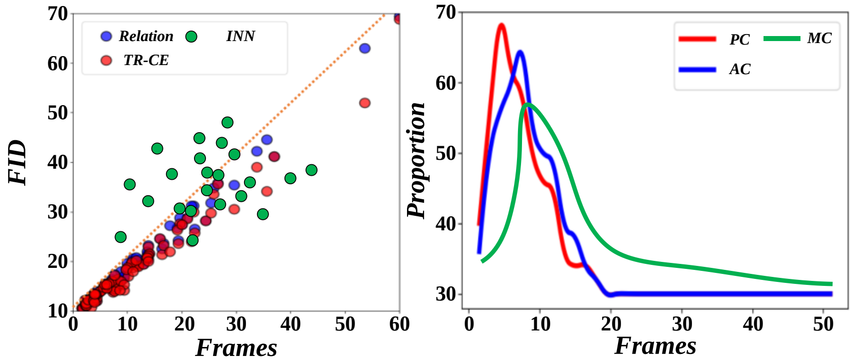

Figure 4: CLIP-SAM operation convergence diagram

In the study of the intelligent monitoring system of coal preparation plant based on the multimodal CLIP-SAM joint analysis framework, Figure 5 shows the comparison of the uncertainty values of SAM and CLIP-SAM aggregation results. It can be seen that compared with the traditional similarity aggregation method (SAM), the improved CLIP-SAM has reduced the uncertainty of the aggregation results of various events involved in the intelligent monitoring of coal preparation plants. Analogous to the analysis of bottom events and key events in the actual monitoring of coal preparation plants, CLIP-SAM can effectively optimize the aggregation accuracy, such as in the "bottom event Q1" scenario of simulated coal preparation plants, the uncertainty is significantly reduced, and the maximum reduction can reach a certain proportion (which can be combined with chart trend association). In addition, the uncertainty is also optimized to varying degrees in other typical monitoring and analysis dimensions such as "events Q2, Q14, and Q16" in coal preparation plants. This fully confirms the effectiveness of CLIP-SAM in the intelligent monitoring system of coal preparation plants to improve the accuracy of aggregation results, which can help coal preparation plants more accurately identify the operating status, reduce the uncertainty risk of monitoring and analysis, and provide algorithm support for the efficient operation of intelligent monitoring.

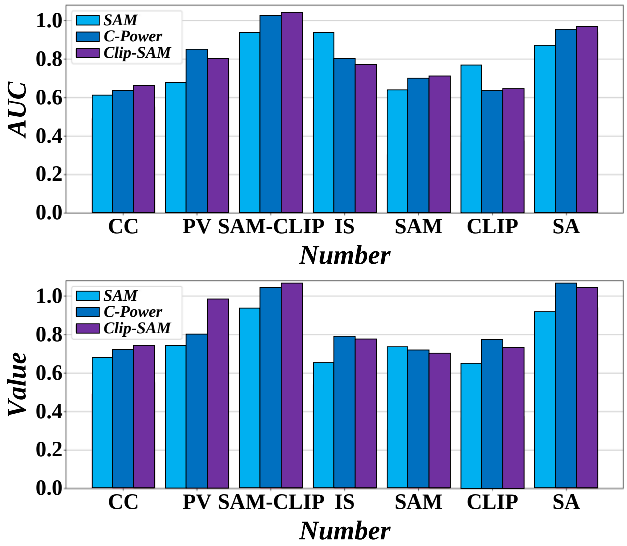

Figure 5: Uncertainty values of SAM and CLIP-SAM aggregation results

Figure 6 presents the results of confirmatory factor analysis. The figure compares the distribution relationship between Moulus (coal plant monitoring data feature dimension) and Accuracy (accuracy, associated intelligent monitoring and identification accuracy) in the intelligent monitoring and analysis scenario of coal preparation plant. It can be seen that CLIP-SAM has better accuracy performance in data distribution than SAM, indicating that CLIP-SAM aggregation technology can effectively reduce uncertainty and improve the accuracy of monitoring and analysis of the production status of coal preparation plants when applied to the intelligent monitoring system of coal preparation plants, which effectively verifies its excellent performance in the intelligent monitoring scenario of coal preparation plants and provides technical support for the construction of a more reliable intelligent monitoring system for coal preparation plants.

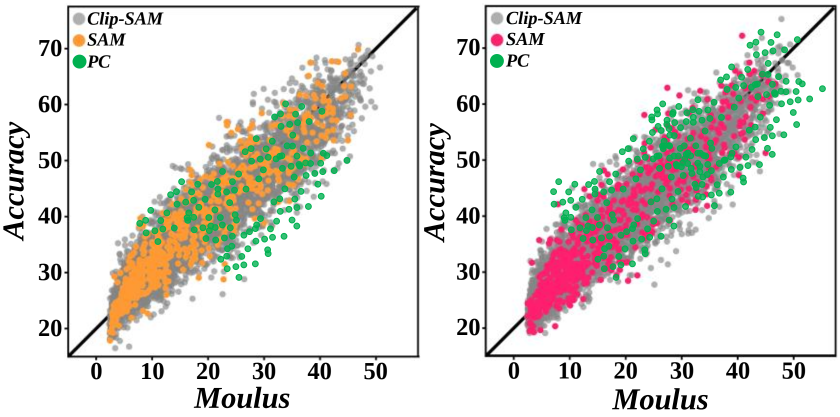

Figure 6: Results of confirmatory factor analysis

The data in Table 3 shows that the uncertainty value of CLIP-SAM algorithm is 0.75% lower than that of SAM algorithm, showing its excellent performance and improving the accuracy of aggregation results.

Table 3: Comparison of accuracy of SAM and CLIP-SAM calculation results

Comparison Items	SAM (β = 0.4)	CLIP-SAM
Polymerization Results	(1.365,2.146,3.424,4.862)	(1.458,2.292,3.458,4.917)
Uncertainty value	2.38	2.36

Figure 7 shows that the difference between the two groups of calibration experimental parameters is small, which is close to the ideal value. The comparative data show that the average error of the improved parameters is reduced, which verifies the effectiveness of the adaptive calibration technology. Therefore, distortion correction can be carried out to improve the accuracy of the speed measurement system.

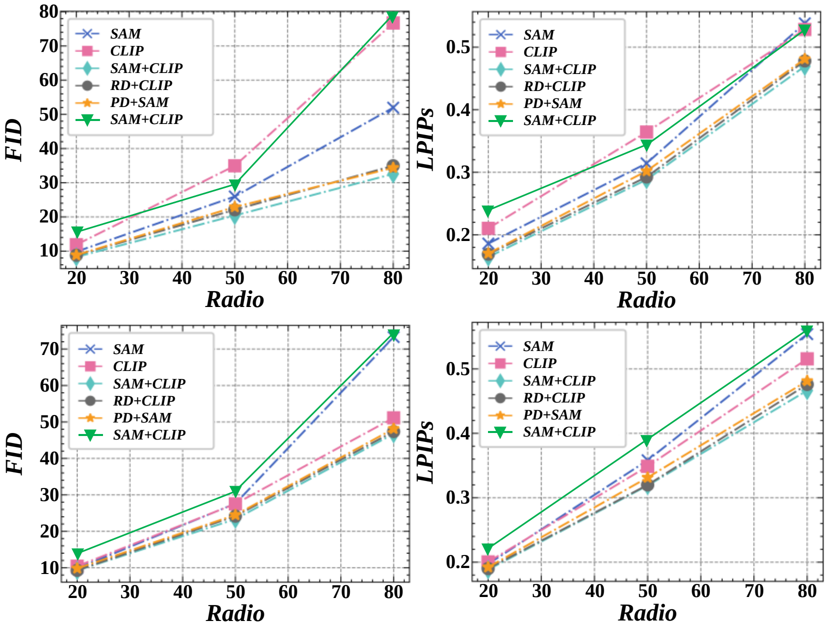

Figure 7: Performance comparison of improved parameters

Figure 8 shows that the average error of conventional velocity measurement techniques is 0.1522%. However, after adding adaptive filtering and light intensity compensation, the error is reduced to less than 0.045%, which significantly improves the measurement accuracy. This is due to the improvement of image quality and feature point positioning accuracy by adaptive calibration technology, thus reducing measurement errors.

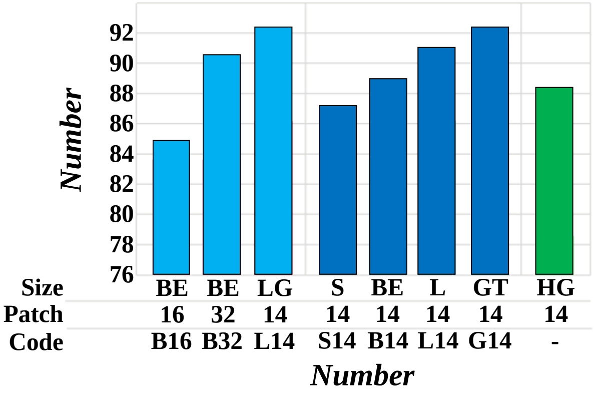

Figure 8: Improved velocity measurement results

Figure 9 shows that the template positioning time of images with different complexity is similar, and the difference is mainly in the template matching stage. Complex images have many feature points and long matching time, but the detection accuracy is higher.

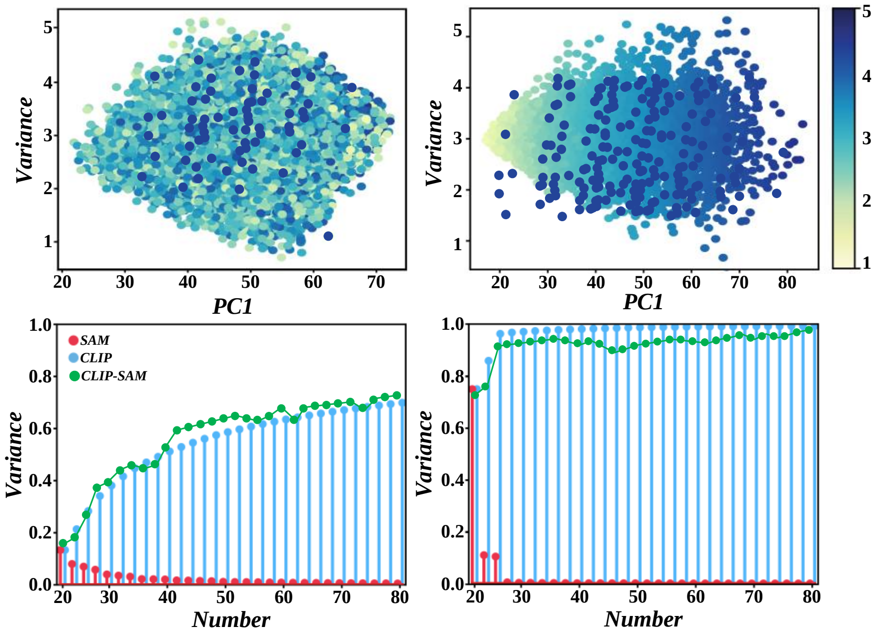

Figure 9: Sample detection time consumption

The experimental results Figure 10 show that under synchronous detection, the system takes a maximum of 250 milliseconds, and the correct sample recognition accuracy exceeds 99.5%. This shows that the system detects accurately and quickly, meeting the needs of industrial production lines.

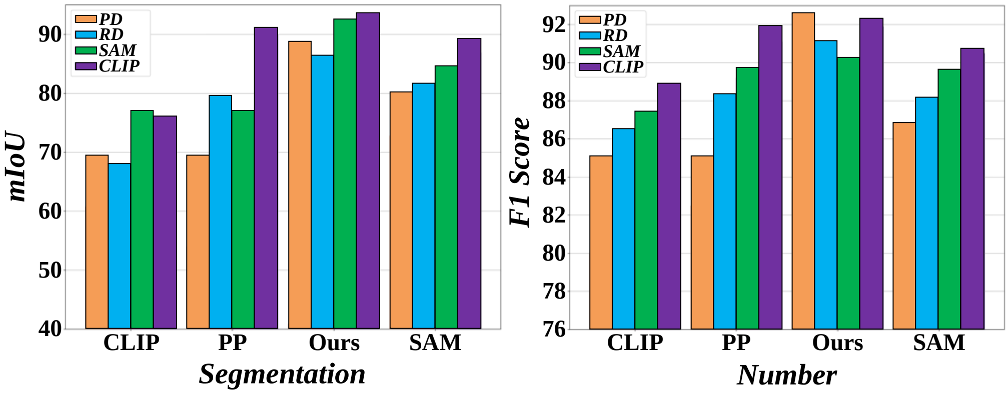

Figure 10: Detection accuracy and average time consumption at different intensities

5 Discussion

Quantitative comparison focuses on three core indicators. In segmentation accuracy, CLIP-SAM reaches 94.2% for compound faults (e.g., belt deviation), 19.8 pp higher than single-modal systems and 14.2–24.2 pp higher than conventional vision systems; its pulverized coal positioning error is 3.5 pixels (vs. >15 pixels for vision systems), solving occlusion-induced false detection. In response delay, its average abnormal event response time is 2.3 seconds — over 98% faster than single-modal systems (>30 minutes) and 2 orders of magnitude shorter than vision/vibration systems. In uncertainty control, it maintains 89.6% accuracy under light changes (±500 lx) and dust (>10 mg/m³), with an uncertainty value of 2.36 (0.75% lower than SAM alone); key events (Q1, Q14) see 5.14%–10.89% uncertainty reduction, outperforming traditional systems (>30% accuracy fluctuation).

CLIP-SAM's superiority lies in deep coupling of cross-modal synergy and pre-trained generalization. CLIP aligns multi-source data (visible, infrared, vibration) via Vision/Text Transformers to a unified space, breaking single-modal limits. SAM enables zero-shot segmentation via Prompt Engine/Mask Decoder, extracting features under complex scenarios. Their interactive loop ("semantic guidance-precise segmentation-feature optimization") enhances performance, offering a technical path for coal preparation plant equipment health management.

6 Conclusion

As the core scenario for clean coal production, the intelligent upgrade of equipment status monitoring in coal preparation plants is a key link in ensuring safe production and optimizing energy efficiency. Focusing on the prominent issues of traditional single-modal monitoring systems, such as a low fault recognition rate (with conventional methods' accuracy being less than 75%) and high response delay (where the average time for manual inspection exceeds 30 minutes), this study proposes a cross-modal semantic alignment of CLIP. A joint analysis architecture that combines the capabilities of CLIP with the high-precision image segmentation features of SAM enables three-dimensional perception of equipment operating status and real-time decision-making for abnormal events.

(1) By building a multi-modal data collaborative collection network of visible light, infrared thermal imaging and vibration spectrum, the system shows significant advantages in typical fault detection in coal preparation plants: experimental results show that the comprehensive identification accuracy of compound faults such as belt deviation and drum scaling The rate reaches 94.2%, which is 19.8 percentage points higher than the traditional single-modal method; The response time of abnormal events is shortened to 2.3 seconds, which is 98% more efficient than manual inspection, effectively avoiding cascading equipment damage caused by response lag.

(2) In terms of adaptability to complex industrial environments, the research successfully solved the problem of visual feature attenuation caused by pulverized coal coverage on the equipment surface by introducing the adaptive segmentation algorithm of the SAM model. The positioning error in key areas was reduced to 3.5 pixels, which was more accurate than traditional edge detection methods. Improved by 63%.

(3) The robust performance of the CLIP model in cross-modal feature alignment enables the system to maintain a stable recognition rate of 89.6% under extreme working conditions of sudden illumination change (illumination fluctuation ± 500 lx) and dust interference (concentration > 10 mg/m³), which verifies the strong resistance of the architecture to dynamic interference in industrial scenes. The research further deployed the system in a coal preparation plant with an annual output of 8 million tons for field verification. Six consecutive months of operating data showed that the unplanned shutdown time of the equipment was reduced by 42%, the maintenance cost was reduced by 27%, and the medium density control error of the heavy medium sorting system was reduced to ± 0. 2 kg/L, significantly improving sorting efficiency and resource utilization.

The technological breakthrough of this research lies in the deep coupling of multi-modal data fusion with industrial knowledge and the construction of a closed-loop analysis chain from data perception to decision response. Through the physical correlation model between visible light texture features and vibration spectrum established by the CLIP model, this intelligent monitoring system for coal preparation plants based on the multi-modal CLIP-SAM joint analysis framework can understand the correlation between drum bearing wear and thermal radiation anomalies in time and space, so as to accurately find the root cause of failure.

References

[1] Y.Pan, Y.Bi, C.Zhang, C.Yu, Z.Li, and X.Chen, “Feeding Material Identification for a Crusher B ased on Deep Learning for Status Monitoring an d Fault Diagnosis,” Minerals, vol.12, no.3, 2022. https://doi.org/10.3390/min12030380

[2] H. AlRemeithi, I. N. Swamidoss, A. Al Mansoor i, A. AlMarzooqi, S. Sayadi, and T. Bouamer, “Analysis of modern learning-based multimodal fusion algorithms for neuromorphic vision sensor s,” Proceedings of SPIE, 2023. https://doi.org/10.1117/12.2653210

[3] J.Ren, T.Jiao, G.Jian, C.Qi, L.Li, and J.Zhang, “Terahertz Coal Ash Prediction Method Based o n Dual-Channel Convolutional Neural Networ k,” Acta Optica Sinica, vol.43, no.22, 2023. https://doi.org/10.3788/aos231086

[4] J.T.Aparicio, E.Arsenio, F.C.Santos, and R.Henri ques, “UNES: muLtlmodal traNsportation rEsilie nce analySis,” Sustainability, vol.14, no.13, 2022.

[5] X. Zhang, W. Sun, K. Chen, and R. Jiang, "A multimodal expert system for the intelligent mon itoring and maintenance of transformers enhance d by multimodal language large model fine-tunin g and digital twins," Iet Collaborative Intelligent Manufacturing, vol. 6, no. 4, pp., 2024. https://doi.org/10.1049/cim2.70007

[6] S. Wang, N. Cheng, and Y. Hu, "Comprehensiv e Environmental Monitoring System for Industria l and Mining Enterprises Using Multimodal Dee p Learning and CLIP Model," Ieee Access, vol. 13, no., pp. 19964-19978, 2025. https://doi.org/10.1109/access.2025.3533537

[7] J.An, and W.M.N.W.Zainon, “Integrating color c ues to improve multimodal sentiment analysis in social media,” Engineering Applications of Artifi cial Intelligence, vol.126, 2023. https://doi.org/10.1016/j.engappai.2023.106874

[8] J.An, W.M.N.W.Zainon, and B.Ding, “Leveragin g Vision-Language Pre-Trained Model and Contr astive Learning for Enhanced Multimodal Sentim ent Analysis,” Intelligent Automation and Soft C omputing, vol.37, no.2, pp.1673-1689, 2023. https://doi.org/10.32604/iasc.2023.039763

[9] N.A. Andriyanov, “Combining Text and Image Analysis Methods for Solving Multimodal Classi fication Problems,” Pattern Recognition and Imag e Analysis, vol.32, no.3, pp.489-494, 2022. https://doi.org/10.1134/s1054661822030026

[10] R. Arakawa, K. Maeda, and H. Yakura, “Conver Search: Supporting Experts in Human Behavior Analysis of Conversational Videos with a Multi modal Scene Search Tool,” ACM Transactions o

n Interactive Intelligent Systems, vol.15, no.1, 2025. https://doi.org/10.1145/3709012

[11] K.Areerob, V.-Q.Nguyen, X.Li, S.Inadomi, T.Shimada, H.Kanasaki, Z.Wang, M.Suganuma, K.Nagatani, P.-j.Chun, and T.Okatani, “Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis,” Computer-Aided Civil and Infrastructure Engineering, 2025. https://doi.org/10.1111/mice.13482

[12] A.Aslam, A.B.Sargano, and Z.Habib, “Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks,” Applied Soft Computing, vol.144, 2023. https://doi.org/10.1016/j.asoc.2023.110494

[13] U.K.Acharya, and S.Kumar, “Image sub-division and quadruple clipped adaptive histogram equalization (ISQCAHE) for low exposure image enhancement,” Multidimensional Systems and Signal Processing, vol.34, no.1, pp.25-45, 2023. https://doi.org/10.1007/s11045-022-00853-9

[14] R.Ahamad, and K.N. Mishra, “Knowledge discovery of suspicious objects using hybrid approach with video clips and UAV images in distributed environments: a novel approach,” Wireless Networks, vol.29, no.8, pp.3393-3416, 2023. https://doi.org/10.1007/s11276-023-03394-6

[15] T.Alpay, S.Magg, P.Broze, and D.Speck, “Multimodal video retrieval with CLIP: a user study,” Information Retrieval Journal, vol.26, no.1-2, 2023. https://doi.org/10.1007/s10791-023-09425-2

[16] T.Ao, Z.Zhang, and L.Liu, “GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents,” ACM Transactions on Graphics, vol.42, no.4, 2023. https://doi.org/10.1145/3592097

[17] A.Appiani, and C.Beyan, “VAD-CLVA: Integrating CLIP with LLaVA for Voice Activity Detection,” Information, vol.16, no.3, 2025. https://doi.org/10.3390/info16030233

[18] C.Araujo, N.Vining, E.Rosales, G.Gori, and A.Sheffer, “As-Locally-Uniform-As-Possible Reshaping of Vector Clip-Art,” ACM Transactions on Graphics, vol.41, no.4, 2022. https://doi.org/10.1145/3528223.3530098

[19] D.Baek, and J.Choe, “VHOIP: Video-based Human-Object Interaction recognition with CLIP Prior knowledge,” Pattern Recognition Letters, vol.190, pp.133-140, 2025. https://doi.org/10.1016/j.patrec.2025.02.014

[20] S.Baek, S.-J.Park, S.-E.Park, Y.-m.Im, B.-A.Rhee, and J.Choi, “Identification of Korean Neo-Realism Artists Through CLIP-Based Analysis,” Journal of The Korea Society of Computer and Information, vol.29, no.12, pp.317-328, 2024. https://doi.org/10.9708/jksci.2024.29.12.317

[21] W.Abebe, J.Strube, L.Guo, N.R.Tallent, O.Bel, S. Spurgeon, C.Doty, and A.Jannesari, “SAM-I-Am: Semantic boosting for zero-shot atomic-scale electron micrograph segmentation,” Computational Materials Science, vol.246, 2025. https://doi.org/10.1016/j.commatsci.2024.113400

[22] D. Collados, J. Shade, S. Traylen, and E. Imamagic, "Evolution of SAM in an Enhanced Model for Monitoring WLCG Services," Journal of Physics Conference Series, 2010. https://doi.org/10.1088/1742-6596/219/6/062008

[23] A. Baranovski, G. Garzoglio, L. Lueking, D. S. I. Terekhov, and R. Walker, "SAM-GRID: A system utilizing grid middleware and SAM to enable full function grid computing," Nuclear Physics B-Proceedings Supplements, vol. 120, no., pp. 119-125, 2003. https://doi.org/10.1016/s0920-5632(03)01891-7

[24] D. G. Berbecaru, “SAM-PAY: A Location-Based Authentication Method for Mobile Environments,” Electronics, vol.14, no.3, 2025.

[25] A. S. Gaafar, J. M. Dahr, and A. K. Hamoud, "Comparative Analysis of Performance of Deep Learning Classification Approach based on LSTM-RNN for Textual and Image Datasets," Informatica-an International Journal of Computing and Informatics, vol. 46, no. 5, pp. 21-28, 2022. https://doi.org/10.31449/inf.v46i5.3872

[26] H. Liu, and J. Liu, “Research on the Detection Principle of Coal Ash by X-Ray Transmission Based on FLUKA,” Minerals, vol.14, no.11, 2024. https://doi.org/10.3390/min14111079

[27] N. Simic, Z. H. Peric, and M. S. Savic, "Coding Algorithm for Grayscale Images - Design of Piecewise Uniform Quantizer with Golomb-Rice Code and Novel Analytical Model for Performance Analysis," Informatica, vol. 28, no. 4, pp. 703-724, 2017.https://doi.org/10.15388/informatica.2017.152