ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 13 December 2025; Accepted: 16 March 2026

https://doi.org/10.17559/TV-20260203003363
Original scientific paper

Research on the Construction of Multimodal Large Models and Self-supervised Learning for Panoramic Perception of Power Equipment

Guoshun ZHENG

Abstract: The global information perception of power equipment is the key to supporting the efficient and stable operation of the new power system. This paper adopts the digital twin technology and constructs a new framework for panoramic perception of transformer vibration status. To address the difficulties in obtaining power defect samples, the dominance of normal samples, and the reliance on large-scale data of multi-modal large models, a power defect data enhancement method based on diffusion models is proposed. Under the premise of ensuring the rationality of the generated image structure, this method utilizes the trained multi-modal large model Qwen-VL-Max to extract the high-order semantic information of real power scene images and combines the prompt engineering technology to generate synthetic images with power defect features and high quality. Moreover, for the common problem of data sparsity in multi-modal recommendation systems, a multi-modal fusion recommendation algorithm based on collaborative self-supervised learning is proposed. This algorithm effectively enhances the representation ability of multi-modal data through the joint learning of the deep features of the data, thereby alleviating the problem of performance decline in recommendations caused by data sparsity. Experimental results show that, compared with the current mainstream multi-modal recommendation algorithms, this algorithm has significant improvements in multiple recommendation evaluation indicators.

Keywords: digital twin; electric power equipment; multimodal large model; panoramic perception of transformers; self-supervised learning

1 INTRODUCTION

According to the "14th Five-Year Plan", power enterprises need to take digitalization as the core focus to drive the transformation and high-quality development of the power grid, and the digital twin of equipment is the key foundation for building a digital technology support system for the new power system [1]. Digital twins precisely, comprehensively and dynamically map physical entities through digital means. Relying on data, they reproduce the actual operating status of equipment in a virtual space, providing a new path for achieving an overall perception of the production process. In terms of twin modeling of mechanical equipment, reference [2] established a simulation model for the curved driving behavior of trains, demonstrating the dynamic changes of various safety indicators when the train passes through. Reference [3] proposed a global virtual perception system for rotating blade discs, achieving real-time monitoring of the displacement field of the blade discs. This type of modeling focuses on the rapid characterization of the static and dynamic responses of the equipment. Its twins are mostly modeled with 3D Max and data-driven through Unity 3D [4], thereby achieving the effect of "controlling the virtual with the real and presenting with the virtual".

In contrast, the twin model of power equipment still needs to rely on multi-physical field coupling to endow the geometric model with "vitality" internally, and then move towards a higher stage of "controlling the real with the virtual and interacting between the virtual and the real". As a key device in the power grid, the operational reliability of transformers directly affects the overall safety of the system. In actual operation, transformers are in an environment where multiple physical fields such as electricity, magnetism, heat, force, sound and vibration interweave [5], and the parameters of each field quantity reflect their health conditions in different aspects. Most of the existing research based on these parameters focuses on state perception [6] and fault diagnosis [7], which is a local abstraction of the equipment status and makes it difficult to achieve a global perception of its overall operating status.

In addition, power scenarios and defect identification, as core links in inspection work, can also achieve cross-modal intelligent analysis with the help of multi-modal large models, thereby enhancing the efficiency of human-machine collaboration. However, there are still several challenges in applying general multimodal large models to the power field. On the one hand, the pre-training data of the existing models are mostly derived from natural scenes, lacking a deep understanding of the professional norms and defect characteristics of the power industry. This leads to insufficient accuracy in identifying professional defects such as insulator damage and rust on hardware, and the generated description texts often deviate from the power professional terminology system.

Therefore, although multimodal large models have significant value for intelligent power identification, their current limitations cannot be ignored. To address the above issues, based on the concept of digital twin, this paper proposes a panoramic perception method of transformer vibration state driven by digital twin. Centering on the performance analysis of transformers, a coupled modeling framework integrating multiple physical fields such as electricity, magnetism and vibration is constructed to establish a twin model of the transformer. To ensure the fidelity of the twin, this paper dynamically updates the model in combination with the aging parameter curve of insulating materials, thereby achieving an accurate panoramic mapping of the transformer's vibration state and providing effective support for the digital management and control of power equipment. Meanwhile, an intelligent power image recognition technology based on fine-tuning of multimodal large models is proposed. A multimodal data set of power defects is constructed with the help of target detection and classification algorithms to enhance the model's graphic and textual understanding ability of power professional scenarios. By combining efficient fine-tuning with cross-modal alignment methods, efficient parameter fine-tuning is implemented on large models to enhance their visual-semantic alignment performance in power defect recognition at a lower computational cost,

ultimately achieving intelligent power image recognition based on multi-modal large models.

2 RELATED WORK

At present, the exploration of the application of multimodal large models in the field of power defect detection still needs to be deepened. To address this task, this section systematically reviews the development trajectory of multimodal models, especially large visual language models, covering their phased improvements, existing deficiencies, and current progress. Based on this, algorithms and models suitable for power defect scenarios are screened and optimized. The proposal of the Transformer architecture has brought about a transformative impact in the field of natural language processing [8]. The self-attention mechanism it introduced for the first time has not only achieved remarkable results in text tasks but also demonstrated outstanding performance in the field of computer vision, becoming a key technology for connecting text and visual modalities. Subsequently, the ViT (Vision Transformer) model further promoted the development of large visual models [9], performing excellently in tasks such as image classification, but there is still room for improvement in capturing local image details. The CLIP (Contrastive language-Image Pre-training) model realizes the fusion of visual and Language modalities [10], can understand the content of images based on text descriptions, and promotes the development of unified representation of text and images. However, its training process relies on large-scale datasets and high computing resources. The BLIP (Bootstrapping Language-Image Pretraining) framework has cross-modal encoding and decoding capabilities [11] and can be used to generate image descriptions or prompt texts. Its improved version, BLIP-2 [12], integrates a pre-trained visual encoder, a large language model, and a learnable Q-Former module. By reusing the parameters of existing visual and language models, it provides high-quality visual representation and powerful language generation capabilities while reducing training costs. In the same year, GPT-4, built on natural language processing technology, attracted widespread attention due to its outstanding multimodal image-text understanding and question-answering capabilities, driving a related research boom. Similar models include MiniGPT-4 [13], InstructBLIP [14] (Miniature GPT-4), and LLaVA-v1.5-13B [15] (Large Language and Vision Assistant version 1.5 with 13 billion parameters), etc. However, the detailed training strategies and weight parameters of most models have not been made public, which brings certain difficulties to further in-depth research.

In recent years, self-supervised learning methods have demonstrated performance comparable to that of supervised learning in computer vision and natural language processing tasks [17], providing useful references for the modeling of recommendation systems. The core idea is to train the model by enhancing the consistency of the data under different disturbances, thereby introducing additional supervisory signals. However, in the field of recommendation, users and products are usually identified only by ID, and there is a lack of explicit association between feature dimensions. The perturbation strategies originally designed for visual or language tasks are not directly applicable. To explore the application of self-supervised learning in recommendation, three perturbation strategies, namely node discard, edge discard and random walk, have been designed on the user-product bipartite graph [18]. The multi-view representation of nodes is obtained by using the generated subgraph, and the learning quality of user and product embedding vectors is improved by constrelling the similarity of different views of the same node and the difference of views of different nodes. The AdaGCL framework proposes an adaptive graph contrastive learning paradigm [19]. It constructs contrastive views through learnable graph generation and denoising modules, and introduces high-quality self-supervised signals for the recommendation model. However, this paradigm still regards self-supervised learning as an auxiliary task. Its main learning objective still relies on the negative sampling strategy, and the process of generating subgraphs through perturbation on bipartite graphs requires high storage and computational overhead. The LATTICE model proposes a latent structure mining method [20]. By constructing a multimodal product-commodity relationship graph, it integrates the semantic connections between commodities into the feature learning process, thereby enhancing the recommendation performance. Although this method has made progress in utilizing the relationships among commodities, it mainly relies on the nearest neighbor method to capture semantic associations, which may focus more on local relationships and fail to fully reveal the complex deep semantic structures among commodities. The FREEDOM model has been improved on the basis of LATTICE [21]. By denoising the user-commodity interaction graph and fixing the product-commodity graph structure before training, its performance has been further enhanced. However, it still inherits the limitations of LATTICE to a certain extent.

Inspired by the research on self-supervised learning in data augmentation [22, 23], another approach to alleviating the problem of data sparsity in multimodal recommendation is to introduce a self-supervised learning mechanism. Recently, some studies have applied self-supervised learning to recommendation systems to enhance model performance. The BM3 method proposes a novel self-supervised learning framework [24], which generates contrastive views through random dropout techniques and optimizes the representation by combining three contrastive loss functions to alleviate the problems caused by computational costs and insufficiently supervised signals. MMSSL (Multi-Modal Self-Supervised Learning) designed a cross-modal contrastive learning task [25], aiming to maintain the commonality of cross-modal semantics and the diversity of user preferences. Although self-supervised learning has made certain progress in recommendation systems, existing research mostly focuses on self-supervised learning of single features (such as data features or user-product interaction embedages), which may fail to fully utilize multi-perspective information, thereby limiting the algorithm's ability to mine data consistency features from different levels and failing to fully unleash the potential of multimodal data.

3 PANORAMIC PERCEPTION FRAMEWORK

The panoramic perception architecture of transformer vibration state driven by digital twin mainly consists of four core parts: twin data, twin body modeling, model fidelity and panoramic perception. Its overall framework is shown in Fig. 1.

1) Twin data module

This module integrates the structural field data of the equipment, the production site data and the operating environment data, laying a necessary data foundation for the subsequent construction of a multi-physics field coupling model.

2) Twin modeling module

This module provides crucial model support for the panoramic perception of transformers and establishes a twin model precisely corresponding to physical entities in the information space through a multi-physics coupling process. Specifically, on the one hand, a three-dimensional geometric model of the transformer is constructed relying on the ANSYS platform, and a real-scene digital twin expression is achieved in combination with the on-site measured data. On the other hand, by integrating the electro-magnetic-vibration multi-physical field coupling parameters of the transformer entity, its operating status can be accurately characterized, meeting the requirements for model accuracy in the digital control of the entire life cycle.

3) Model fidelity module

This module is a key link to ensure the continuous synchronization of the transformer twin model with the physical entity. Given that the aging of insulating materials can cause additional vibrations and affect the perception accuracy, this architecture dynamically updates insulating aging characteristic parameters such as dielectric loss factor based on the real-time operating status of the transformer, thereby maintaining the fidelity performance of the twin model.

4) Panoramic perception module

This module refers to the visualization of the vibration state of the twin in the form of 3D cloud maps through functions such as dynamic rendering and auxiliary hiding. This function has expanded from the limited sensor information of physical entities to the full-domain perception of twins, enabling operation and maintenance personnel to intuitively monitor the physical field signals of each unit inside the transformer, and ultimately achieve virtual-real synchronization and comprehensive perception of the operating status.

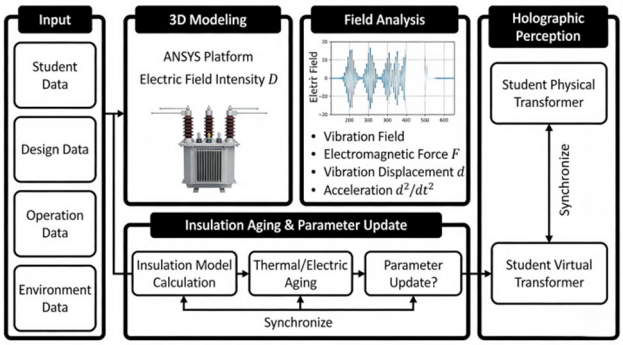


Figure 1 Panoramic perception architecture of transformer vibration state driven by digital twin

3.1 Twinconstruction

Taking the S11-M-2 000 kVA oil-immersed transformer as the research object, the size parameters of each component were extracted from its production drawings, and the geometric model of the transformer twin was established based on the ANSYS platform.

The internal structure of a transformer is complex, constructed based on physical entities, and involves a huge amount of calculation. Given that the influence of the transformer's clamping, fixing and supporting structures on the electromagnetic field is relatively small [22], in order to improve the calculation efficiency of the model, the main structural model of the transformer, namely the core, windings, insulating oil paper and oil tank, is established. Meanwhile, assume that insulating oil fills the space between the body and the box wall [23]. Using the RMxpry magnetic circuit module, a core model with a length of 1820 mm, a width of 760 mm and a height of 1740 mm was created. The thickness of the silicon steel sheet material was 0.35 mm and the thickness of the insulating oil paper was 0.18 mm. The winding is equivalent to a solid circular ring. The coil material adopts TBY1 type soft copper flat wire, wound in a spiral manner. The low-voltage winding is on the inner side with a radius of 441.68 mm and 424 turns per phase. The high-voltage winding is on the outer side with a radius of 605.62 mm and 980 turns per phase, with a height of 1380.48 mm. The volume of the oil tank is 2.44 × 1.26 × 1.92 m, with a wall thickness of 8 mm. x, y, and z are defined as the transverse, longitudinal, and axial directions respectively. A 3D transformer structure model is established, as shown in Fig. 2.

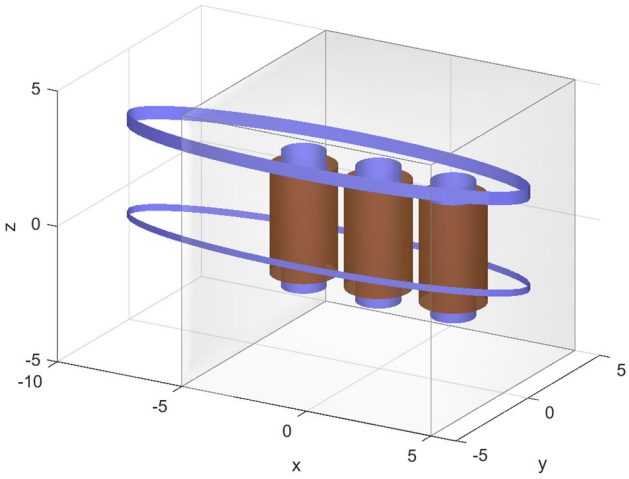


Figure 2 Transformer structure diagram

The transformer twin is based on the structural field and senses the vibration of the transformer through the coupling of multiple physical fields such as electricity, magnetism and vibration. The boundary conditions are set as follows:

Fixed support constraints are applied to the upper and lower end areas of the core to simulate the constraint effect of clamps on the vibration of the transformer core.

2) Set the transformer oil tank as the magnetic insulation boundary.

3) To ensure the stability of the model, the insulating oil is set as a non-slip boundary condition at the insulating boundary.

4) According to the actual operating season of the transformer, set the ambient temperature to 34 ℃ and apply the on-site current data to the low-voltage winding.

In engineering practice, to meet the demand for panoramic perception of the vibration characteristics of transformers, a three-axis vibration measurement method is adopted. Among them, the x, y, and z axes, which are three orthogonal directions, respectively represent the vibration conditions left and right, front and back, and up and down, to effectively guide the optimization and maintenance of transformers.

A section of the excitation during the operation of the transformer is extracted, as shown in Fig. 3. To achieve panoramic perception of the transformer, electromagnetic coupling is carried out in the finite element form. The 3D transient solver of ANSYS Maxwell is adopted to solve the internal elements and obtain the state changes of the electric field and magnetic field during the operation of the transformer.

The magnetic field parameters of the materials of each component of the transformer are shown in Tab. 1. According to the national standard GB/T7600-2008, the transformer oil is 45 # insulating oil with a relative magnetic permeability of 1.0.

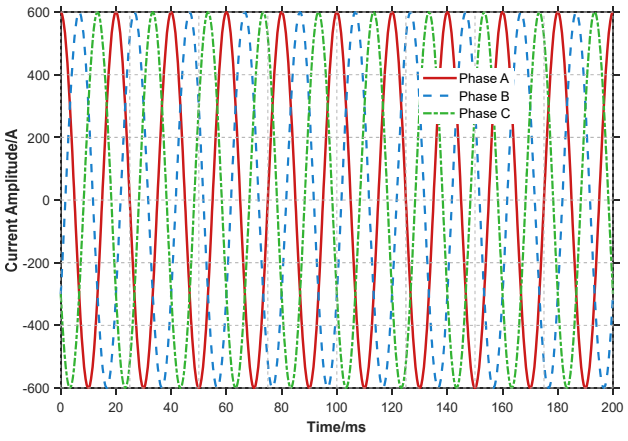


Figure 3 Current excitation

Table 1 Magnetic Field Parameters of Component Materials

Component	Material	Relative permeability	Electrical conductivity / S·m^{-1})
Iron core	Silicon steel sheet	Curve	1.96×10^6
Winding	copper	1	5.8×10^7

When a time-varying sinusoidal alternating current $I = I_m \sin(\omega_t + \varphi_0)$ with amplitude I_m, angular velocity ω and initial phase ϑ_0 passes through the winding, a time-varying magnetic field is generated. Electromagnetic fields are generated in two ways: conducting current and time-varying electric fields, satisfying Maxwell's Ampere's Law, that is:

$$\nabla \cdot H = J + \frac{\partial D}{\partial t} \tag{1}$$

In the formula: H represents the magnetic field intensity; J is the current density; D is the electric displacement vector; t represents time. The constitutive relationship between the electromagnetic field quantity and the medium characteristic quantity is:

$$D = \varepsilon E;\ B = \mu H \tag{2}$$

In the formula: E represents the electric field intensity; B represents the magnetic induction intensity; ε is the dielectric constant; μ represents magnetic permeability. The electric field intensity E can be adjusted by changing the driving voltage, the charge distribution, or the dielectric constant ε of the medium; the magnetic field intensity H can be controlled by altering the current in the coil, the magnetic permeability μ of the magnetic medium, or the external magnetic field.

From the perspective of the convenience of solving the internal magnetic induction intensity B of the transformer, the vector magnetic potential A is introduced as the independent variable. A is defined as:

$$\nabla \cdot A = B \tag{3}$$

Solve the vector magnetic potential of the grid nodes to obtain the magnetic induction intensity distribution of the transformer, as shown in Fig. 4.

Based on the calculation results of the electromagnetic field, the result data is coupled to the transformer vibration field as a load function.

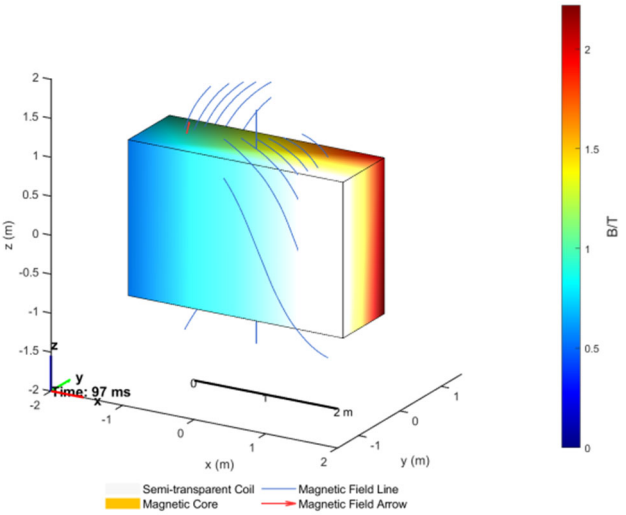


Figure 4 Distribution of magnetic induction intensity of transformer

3.2 Panoramic Perception Model Update

Since these electrolytes are not ideal insulators, on one hand, under the influence of an external electric field, charge carriers will move, and some protons will collide with the windings, thereby altering the stress and strain of the insulator. On the other hand, the electromagnetic field will cause resistance losses within the transformer, and the accumulation of heat will cause the temperature of the insulating material to rise, thereby generating thermal stress. Frequent thermal stress cycles cause thermal fatigue in insulating materials. Meanwhile, moisture and oxygen in the environment cause insulating materials to absorb moisture, accelerating the aging of insulation. The insulation performance is commonly characterized by the

relative dielectric constantεr and the dielectric loss factor tanδ. ε_r is used to measure the dielectric properties of insulating materials, and tanδ is used to describe the degree of energy dissipation of insulating materials in an electric field. The relationship between the two is:

$$\tan\delta = \frac{\varepsilon^*}{\varepsilon'} \tag{4}$$

Insulating oil paper and insulating oil are important components of transformer insulation. Compared with insulating oil, insulating oil paper is thinner and has a higher oil immersion rate. Therefore, it is assumed that it follows the aging law of insulating oil. High temperature is the direct cause of insulating oil aging. When the oil temperature exceeds 60 ℃, the aging rate increases exponentially. This paper selects the tanδ variation curves with time and temperature from 0 to 34 days, that is, from 0 to 12 years of natural aging, as shown in Fig. 5.

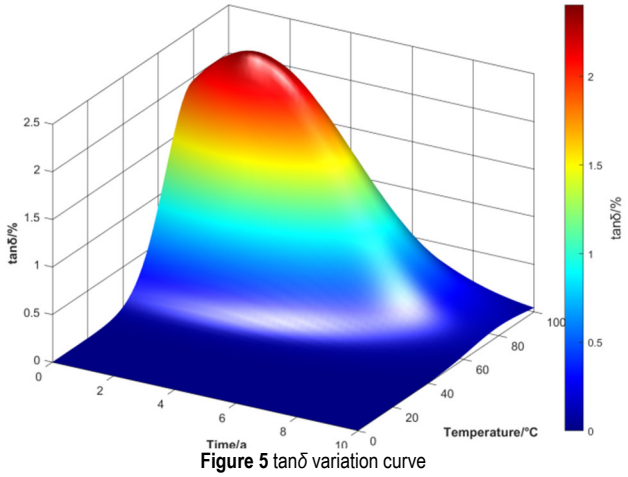


Figure 5 tanδ variation curve

As can be seen from Fig. 5, the aging characteristic parameter tanδ of insulating oil shows an upward trend with the increase of time and temperature.

Based on the transformer entity, sensor data is obtained. According to the operating duration, tanδ is updated in accordance with the insulating oil aging curve, and the virtual model is input to meet the fidelity performance. By comparing the virtual-real deviation, if tanδ is lower than 10%, the operation will continue; otherwise, a warning will be issued.

To analyze the accuracy of the digital twin model of transformer vibration, the mean absolute percentage error (e_{MAPE}) is adopted to quantify the model accuracy:

$$e_{\mathrm{MAPE}} = \frac{1}{N}\sum_{t=1}^{N}\frac{\hat{a}_t - a_t}{a_t} \tag{5}$$

In the formula: a_t is the vibration acceleration measured by the sensor at time t; N represents the number of sampling points. The smaller the e_{MAPE} value, the higher the accuracy of the transformer twin model.

3.3 Construction of a Multimodal Large Model Based on Panoramic Perception of Power Equipment

This study aims to construct a multimodal large model for the power field and apply it to the intelligent recognition task of power images. Through cross-modal graphic and text interaction, it realizes scene understanding, defect identification, and professional domain knowledge Q&A of power inspection images. The specific implementation path is as follows: Firstly, the YOLOv8 algorithm is adopted to automatically and efficiently construct a multimodal dataset of power defects, thereby enhancing the model's ability to recognize and express professional graphic and textual information in the power field. Secondly, the low-rank adaptation (LoRA) method and the Q-Former module are combined to conduct efficient parameter fine-tuning of the model. While reducing the consumption of computing resources, the visual and semantic alignment ability of the model in the power context is improved. Finally, the performance of the proposed model in the task of power image recognition was verified through experiments, and a comparative analysis was conducted with current mainstream multimodal large models with similar parameter numbers (such as InstructBLIP, LLaVa-Med, Mini-GPT4, LLaVa-v1.5-13B and VisualGLM-6B, etc.). The experimental results show that the method proposed in this paper performs better in the accuracy of intelligent recognition of power images.

Among the YOLO series of algorithms, YOLOv8 strikes a good balance between accuracy and efficiency. For this purpose, a total of five YOLOv8 models were trained in this study: one was used for power scene classification (denoted as YOLOv8-FenLei), and the other four were respectively used for defect detection in transmission, transformation, distribution, and safety monitoring (Ansupervision) scenarios. The number of defect categories that each model could identify was 13, 7, 13, and 2 respectively. These models play a crucial role in building a high-quality multimodal dataset of power defects.

Given the huge parameter scale of multimodal large models, it is usually difficult to fine-tune them to all parameters to adapt to tasks in the power field. Therefore, in this paper, low-rank adaptation (LoRA) technology is adopted to fine-tune the multimodal large model, thereby significantly reducing the number of trainable parameters, enhancing the computational feasibility and multimodal recognition effect of the model during downstream task fine-tuning, while maintaining its core performance. The implementation process of LoRA mainly includes the following steps:

(1) Introduce a bypass structure beside the weights of the pre-trained language model to approximately represent its inherent low-rank characteristics through dimensionality reduction and dimensionality increase operations.

(2) Initialize the dimensionality reduction matrix A using a random Gaussian distribution and initialize the dimensionality increase matrix B with a zero matrix; During the training process, the parameters of the original pre-trained model are frozen, and only the introduced matrices A and B are optimized. B and A: Two trainable low-rank matrices. During training, only B and A are updated, significantly reducing the number of trainable parameters while maintaining the fine-tuning effect. After the training is completed, the incremental weights obtained

by multiplying matrix B by A are combined with the parameters of the original pre-trained model to obtain the fine-tuned final model.

$$W_0 + \Delta W = W_0 + BA \tag{6}$$

Based on VisualGLM-6B, an open-source multimodal question-answering large model, it supports both Chinese and English, and has a large amount of basic knowledge as well as a good image description ability. This chapter selects VisualGLM-6B as the basic model. VisualGLM-6B was trained on the CogView dataset, which contains 30 million high-quality Chinese image-text pairs and 300 million filtered English image-text pairs. These data are used to pre-train the model, enabling it to understand image and text information in different languages. By combining the created multimodal dataset and integrating the LoRA method and the Q-Former method, the Power-VGLM was finally constructed. This model reduces the number of training parameters and simultaneously enhances its image-text dialogue capability in the field of electrical defects. The principle structure of Power-VGLM is shown in Fig. 6.

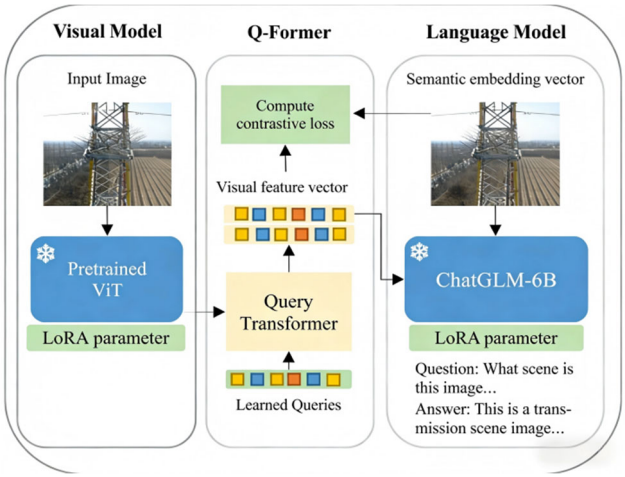


Figure 6 Schematic diagram of Power-VGLM

The overall framework is divided into three key parts, including the image model, Q-Former and language model. In terms of the image model, the pre-trained Vision Transformer is selected for image feature extraction. ViT captures image information with local and global contexts through a continuous multi-layer attention mechanism and generates a set of image embedding vectors. Q-Former extracts the most informative visual representation of the text content. A series of learnable parameters such as queries were set to learn the connection between visual features and semantic features. By training the contrastive loss function between visual feature vectors and semantic vectors, the goal of aligning visual features and semantic features was achieved. Then, the features output by the Q-Former are input into the language model to ensure that the model has dynamic adjustment adaptability and strong learning ability. This semantic vector re-enters the Q-Former part to calculate the contrastive loss between the visual feature vector and the semantic embedding vector, and further align the visual features and semantic features.

3.4 Joint Self-Supervised Learning

Self-supervised learning methods can effectively improve the data sparsity problem faced in recommendation systems by automatically learning universal features from the original data. To this end, this paper designs a joint self-supervised learning component. This component builds multiple enhanced views on the original data through a cross-modal alignment mechanism and integrates self-supervised learning tasks based on feature representation and graph representation. Then, consistent features are extracted from these views to improve the robustness of the recommendation results. This joint self-supervised learning component mainly includes three core tasks: multi-layer cross-modal feature alignment, multi-view feature enhancement, and graph structure perturbation.

During the modal fusion process, the feature distributions of different modalities vary greatly, which may lead to the generation of a large amount of noise and affect the effect of modal fusion. Therefore, it is necessary to align different modal features to reduce the influence of noise in modal fusion. To achieve this goal, this paper introduces a multi-level cross-modal alignment component. Inspired by MENTOR (Multi-level sElf-supervised learNing for mulTimOdal Recommendation) [22], an effective self-supervised learning method is proposed. The features of different modalities can be aligned from the perspective of data distribution.

The multi-layer cross-modal feature alignment component in this article includes four methods: alignment of ID modality and fusion modality, alignment of ID modality and visual text modality, alignment of fusion modality and visual text modality, and alignment of visual modal and text modality. Among them, the ID mode is used to guide the feature alignment between different modes. By leveraging the ID modality, the algorithm can better align the features of other modalities such as vision and text in self-supervised learning tasks. Firstly, Gaussian distribution parametric fusion modes, ID modes, visual modes and text modes are adopted. This parametric method enables the feature distribution of each mode to be clearly characterized and provides a foundation for the subsequent mode alignment process.

The multi-perspective feature enhancement task generates different views from different perspectives of features and enhances the algorithm's representation ability of data by introducing feature transformation to assist the task. Based on the existing two types of feature transformation auxiliary tasks, an auxiliary task is proposed, namely the feature coarse-grained learning task. This task projects the feature vector into a low-dimensional space and then maps it back to the original space. This process captures the coarse-grained structure of the data by reducing the dimension of features, thereby enriching the intrinsic expression of features. Perform three feature transformation-assisted tasks, namely feature discard, feature masking, and feature coarse-grained learning. Through these three auxiliary tasks ring the training process to prevent them from affecting the parameter update of the algorithm. Finally, the generalization ability of the algorithm is improved by calculating the loss function.

$$E_{ud} = E_u \cdot \text{Bernoulli}(p) \tag{7}$$

In this paper, a graph-based enhancement method is adopted to construct the structural perturbation of the contrast view based on the user-item diagram for the visual and text modalities. Generate two comparison views for each modal. Finally, the general structural meaning of user-item interaction is extracted through the InfoNCE loss function.

When constructing the user-project diagram, build two comparison views for each modality. Formally, the embedding of contrast views can be expressed as:

$$E_m = \sum_i \frac{1}{N_u N_i} \cdot E_m^{i-1} \tag{8}$$

A nonlinear feature projection layer is introduced in the algorithm of graph structure perturbation to enhance the performance of self-supervised learning. This improvement is based on the SimCLRv2 algorithm, which enhances the expressive power of features by adding a nonlinear transformation network. Therefore, a nonlinear feature projection layer is added to the graph structure perturbation. The feature projection layer consists of two fully connected layers and one ReLU (Rectified Linear Unit) activation function. The calculation formula is:

$$\overline{E}_m = \text{ProjectHead}\left(\cdot \overline{E}_m\right) \tag{9}$$

Finally, the loss of the joint self-supervised learning task is the sum of the loss of feature masking and the loss of graph structure perturbation, which is formally expressed as:

$$\gamma_{\text{enhance}} = \lambda_g \gamma_{\text{enhance}_g} + \lambda_f \gamma_{\text{enhance}_f} \tag{10}$$

Among them, λ_g and λ_f are equilibrium hyperparameters.

4 SIMULATION VERIFICATION

The superimposed vibration of the transformer core and windings is transmitted to the surface of the box through the connecting parts. Under the current excitation in Fig. 3, vibration analysis was conducted on test points ① to ⑨ respectively, and the acceleration signals in the *x*, *y*, and *z* axes were obtained. The results are shown in Fig. 7.

As can be seen from Fig. 7a, the vibration signal trends at positions ①, ② and ③ on the top of the box are close, and the acceleration signal in the *z* direction is the most intense. As can be seen from Fig. 7b, due to the fixed support set at the bottom of the transformer, the vibration acceleration signal at positions ④ and ⑦ on the front of the box shows a weakening trend from top to bottom, and the acceleration signal in the *y* direction is the most intense. This rule also applies to positions ⑤⑧ and ⑥⑨.

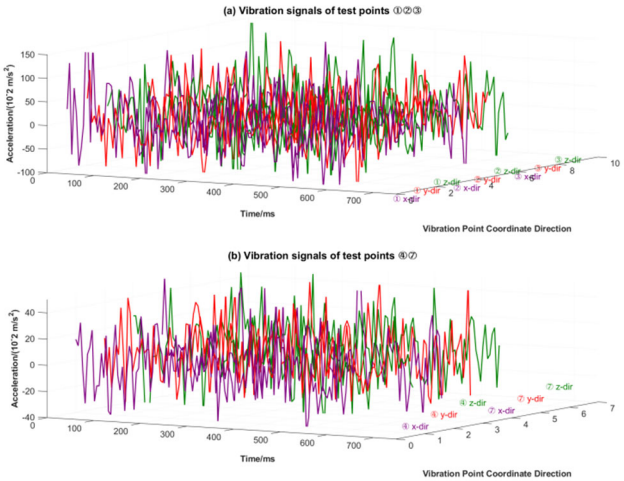


Figure 7 Three-direction vibration test of the box surface along the *xyz* axes

With the passage of time and changes in operating conditions, insulating oil gradually ages. 82 sets of experimental data were selected from the change curve, as shown in Tab. 2. The variation law of the characteristic parameter tanδ of transformer insulating oil aging was statistically obtained, as shown in Fig. 8.

Table 2 Experimental Data of tanδ, the aging parameter of insulating oil

Experiment Serial Number	Variable		tanδ
	t / a	θ	
1	1	50	0.11
2	1	70	0.24
3	4	75	0.54
...	...	...	...
48	5	80	0.75
49	6	75	0.86
50	7	85	1.38
...	..	...	...
82	10	100	2.46

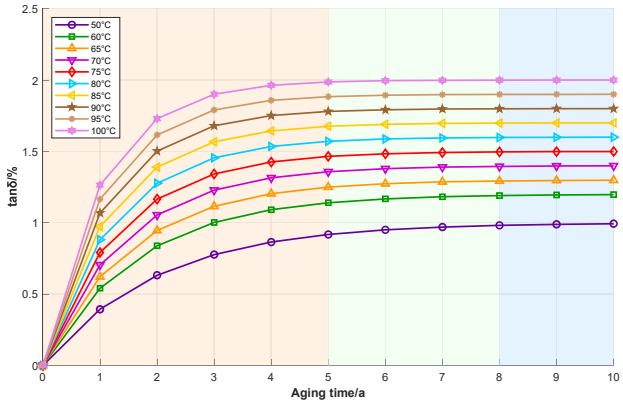


Figure 8 Change of tanδ when insulating oil ages

As can be seen from Fig. 8, the period from 0 to 5 a is the initial stage of insulating oil aging, with a relatively small slope and tanδ remaining relatively stable. During this period, the performance of the oil does not change much and the oil quality remains good. The period from 5 to 10 years is the mid-term stage, during which the slope begins to increase and tanδ shows a significant upward trend. As the decomposition products in the oil start to accumulate, the increase in moisture and contaminants leads to a gradual rise in tanδ, and the performance of the insulating oil begins to deteriorate. In the later stage of insulation aging, tanδ tends to stabilize, and the

transformer can still operate normally for 1 to 2 years. After that, the insulation oil deteriorates rapidly, and the insulation is prone to breakdown.

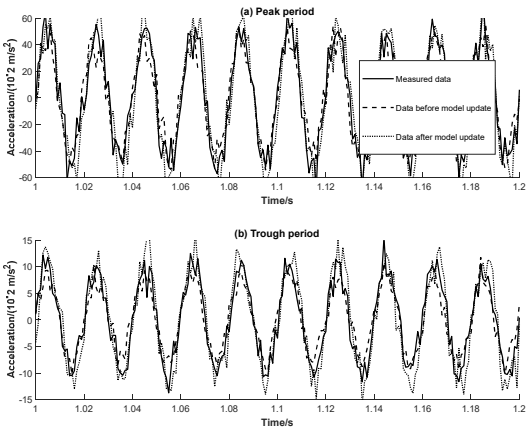


Figure 9 Comparison chart between the model and the actual data

At present, the operating temperature of transformers is generally 65 to 85 ℃. The tanδ variation within this temperature range is not significant. To improve the calculation efficiency, the variable is only updated based on time. The specific update principle of the aging parameters is as follows: Select the time-varying curve of the tanδ of the insulating oil at 75 ℃. Update the tanδ every 1 a between 0 and 5 a, and every six months between 5 and 10 a. At the same time, replace the new insulating oil every 10 a. Compare the output data of the transformer model before and after aging update with the measured operation data, as shown in Fig. 9.

The panoramic perception of the transformer creates a 3D scene object by importing the structural field model, combines the on-site operation data such as current and voltage, and uses a 3D cloud map to intuitively obtain the vibration information of the transformer, including the panoramic perception domain of the twin and the internal information perception, as shown in Fig. 10.

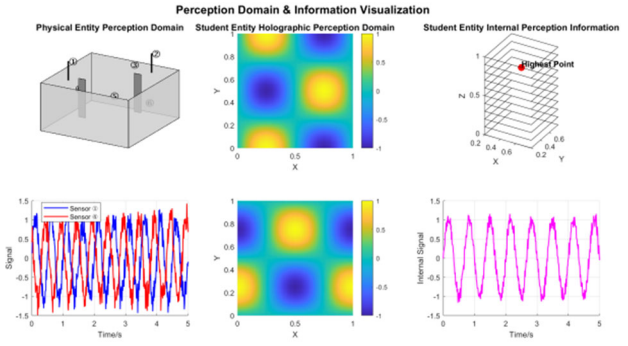


Figure 10 Panoramic perception map of the transformer status

As can be seen from Fig. 10, the 3D model of the twin, driven by external production data, uses dynamic rendering and auxiliary observation functions such as rotation, scaling, and translation to display the overall vibration state of the transformer from 360°, expanding the observation field of the physical entity. In particular, the twin body has a structural hiding function, which makes it convenient to observe the vibration state of the specified structure, such as the core and windings.

To further optimize the knowledge extraction effect of large models, the LoRA+ fine-tuning method was adopted for the best model to fine-tune it. Fig. 11 shows the F1 values of test sets at different scales.

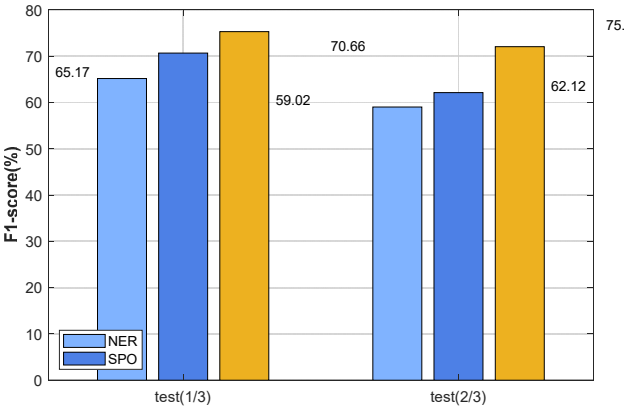
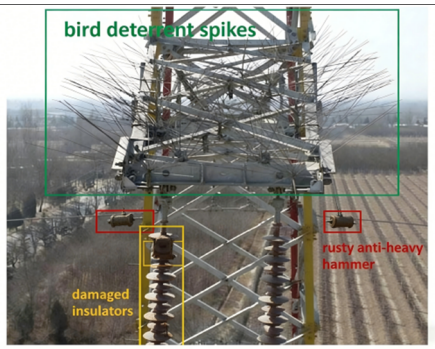


Figure 11 F1 value fine-tuned by LoRA

As can be seen from Fig. 11, in the NER task, the F1 value has increased by 0.8%. The F1 value increased by 2.39% in the SPO task. The F1 value represents the overall performance of the model and has improved in both tasks. These improvements demonstrate the effectiveness of the hybrid fine-tuning method in optimizing the model extraction effect.

Table 3 Model test results

Input image	
Input questions	What kind of scene is this picture of? Please describe the following picture.
Ground truth	This is a picture of a power transmission scene, depicting normal bird stinger protection, damaged insulators, and rusted shockproof hammers.
Power-VGLM(ours)	This is a picture of a power transmission scene. In the picture, it is depicted that the transmission tower is normal, the insulator is damaged, and the bird protection is normal.
VisualGLM-6B	This picture depicts a beautiful natural scene, with a towering transmission tower located between the fields. It is composed of many metal structures and glass, forming a unique landscape. Surrounded by thick trees and vast grasslands, it feels as if one is in the embrace of nature. People can feel an atmosphere of tranquility and harmony.

Tab. 3 shows that the fine-tuned model Power-VGLM in this work demonstrates higher accuracy in scene recognition and related description. It not only accurately identified the "power transmission scene", but also provided detailed descriptions of the power equipment and defects in the image, such as "insulator damage" and "normal bird puncture prevention". The fine-tuned

multimodal large model significantly improves the accuracy of scene recognition and related defect description, and greatly reduces irrelevant background information. Observe Tab. 3. Power-VGLM provides more standardized and accurate answers. Based on the comprehensive experimental results and analysis, Power-VGLM significantly outperforms baseline models and common multimodal large-scale models in Power scene classification, power defect detection, and defect knowledge question answering.

5 CONCLUSION

This paper designs a multimodal image-text question-answering technology for power defect recognition, which is based on a pre-trained large model for domain adaptation and fine-tuning. By adopting YOLOv8 to construct a multimodal dataset for describing power defects, the cost of data annotation has been significantly reduced, while the semantic understanding of the model for professional graphic and textual content in the power field has been enhanced. Furthermore, by means of LoRA and Q-Former methods, the language and visual models were co-fine-tuned, and a power-specific multimodal large model, Power-VGLM, was constructed. This model achieves efficient alignment of text and visual features and performs better than several existing mainstream large models in the task of power defect identification. In addition, this paper also proposes a panoramic perception framework for the vibration state of transformers based on digital twins. By dynamically correcting the time-varying aging parameter tanδ, the twin model is updated to achieve precise perception of the vibration acceleration signal of the transformer. Finally, by integrating the joint self-supervised learning component and conducting collaborative learning on multi-source data features, the common problem of data sparsity in recommendation systems has been significantly alleviated. This paper still has certain limitations. It lacks manual meticulous verification. Although the self-supervised learning part has achieved improvements in multimodal recommendation, in the design of graph structure perturbation and feature enhancement strategies, it still relies on heuristic methods and lacks theoretical optimality guarantees. Next, a multi-agent system is constructed to simulate the process of human collaborative learning. The reward mechanism is used to drive the agents to explore the multimodal feature space, and self-supervised learning is employed to optimize the classification strategy.

6 REFERENCES

[1] Hang, H. & Li, Z. (2025). Research on the construction of intelligent art design system based on multimodal perception and generative AI. *Discover Applied Sciences*, *7*(9),1-28. https://doi.org/10.1007/s42452-025-07513-0

[2] Timothée, D., Jabaian, B., & Fabrice, L. (2025). FlowAct: A Proactive Multimodal Human-Robot Interaction System with Continuous Flow of Perception and Modular Action Sub-Systems. *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods*, *1*, 771-779. https://doi.org/10.5220/0013265700003905

[3] Wang, R., Xie, Y., & Liu, H. (2025). Center-of-Mass-Based Object Regrasping: A Reinforcement Learning Approach and the Effects of Perception Modality. *IEEE/ASME Transactions on Mechatronics*, *30*(2), 1356-1365. https://doi.org/10.1109/TMECH.2024.3433435

[4] Fan, L., Wang, Y., & Zhang, H. (2024). Multimodal Perception and Decision-Making Systems for Complex Roads Based on Foundation Models. *IEEE transactions on systems, man, and cybernetics. Systems*, *11*(1), 54-67. https://doi.org/10.1109/TSMC.2024.3444277

[5] Cui, H., Feng, Z., & Tian, J. (2023). MAG: a smart gloves system based on multimodal fusion perception. *CCF Transactions on Pervasive Computing and Interaction*, *5*(4), 19-24. https://doi.org/10.1007/s42486-023-00138-5

[6] Liu, L. (2024). Optimization of a Business English Tutoring System Based on Intelligent Recommendation Algorithms and Multimodal Data Analysis. *2024 7th International Conference on Education, Network and Information Technology (ICENIT)*,187-193. https://doi.org/10.1109/ICENIT61951.2024.00041

[7] Liu, J., Luo, D., & Fu, X. (2023). Design Strategy of Multimodal Perception System for Smart Environment. *EAI/Springer Innovations in Communication and Computing*, 93-115. https://doi.org/10.1007/978-3-031-09729-4_6

[8] Tsiotras, P., Gombolay, M., & Foerster, J. (2024). Editorial: Decision-making and planning for multi-agent systems. *Frontiers in Robotics and AI*, *11*(11), 1422344-1422356. https://doi.org/10.3389/frobt.2024.1422344

[9] Li, H., Li, Q., & Yang, C. (2025). Task Knowledge Injection: Training-Free Adaptation of Multimodal Large Language Models for Remote Sensing Image Understanding. *IEEE Geoscience and Remote Sensing Letters*,*22*(3), 1-5. https://doi.org/10.1109/LGRS.2025.3581558

[10] Langlentombi, L, C. (2024). Development of Spectral Signature of Chir Pine Using Hyperion Data. *International Journal of Bio-resource and Stress Management*, *15*(Feb, 2), 1-06. https://doi.org/10.23910/1.2024.5026

[11] Stephen, S. & Kumar, V. (2023). Detection and Analysis of Weed Impact on Sugar Beet Crop Using Drone Imagery. *Journal of the Indian Society of Remote Sensing*, *51*(12), 2577-2597. https://doi.org/10.1007/s12524-023-01782-1

[12] Zhang, P., Zhang, Y., & Wu, H. (2025). Language-Guided Object Localization via Refined Spotting Enhancement in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *63*(3), 1-15. https://doi.org/10.1109/TGRS.2025.3562439

[13] Wang, Y., Ye, F., & Chen, Y. (2025). A multi-modal dental dataset for semi-supervised deep learning image segmentation. *Scientific Data*, *12*(1), 4306-4315. https://doi.org/10.1038/s41597-024-04306-9

[14] Wu, Y., Wu, G., & Lin, J. (2025). Role Exchange-Based Self-Training Semi-Supervision Framework for Complex Medical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(5), 8372-8386. https://doi.org/10.1109/TNNLS.2024.3432877

[15] Wang, K., Wang, Y., & Zhan, B. (2022). An Efficient Semi-Supervised Framework with Multi-Task and Curriculum Learning for Medical Image Segmentation. *International journal of neural systems*, *32*(9), 2250043-2250056. https://doi.org/10.1142/S0129065722500435

[16] Li, H., Yang, J., & Qu, M. (2025). A semi-supervised multi-task assisted method for ultrasound medical image segmentation. *Neurocomputing*, *639*, 130217. https://doi.org/10.1016/j.neucom.2025.130217

[17] Qiu, Y., Lu, H., & Xu, B. J. (2024). Towards semi-supervised multi-modal rectal cancer segmentation: A large-scale dataset and a multi-teacher uncertainty-aware network. *Expert Systems with Application*, *255*(Part C), 124734. https://doi.org/10.1016/j.eswa.2024.124734

[18] Yang. J., Li, H., & Wang, H. (2024). 3D medical image segmentation based on semi-supervised learning using deep co-training. *Applied Soft Computing*, *159*(3), 13-39. https://doi.org/10.1016/j.asoc.2024.111641

[19] Bashir, R. M. S., Qaiser, T., & Raza, S. E. A. (2024). Consistency regularisation in varying contexts and feature perturbations for semi-supervised semantic segmentation of histology images. *Medical Image Analysis*, *91*(2), 14-37. https://doi.org/10.1016/j.media.2023.102997

[20] Torbati, M. E., Minhas, D. S., & Laymon, C. M. (2023). MISPEL: A supervised deep learning harmonization method for multi-scanner neuroimaging data. *Medical Image Analysis*, *89*(2), 16-39. https://doi.org/10.1016/j.media.2023.102926

[21] Li, W., Yu, J., & Chen, D. (2025). Fine-grained building function recognition with street-view images and GIS map data via geometry-aware semi-supervised learning. *International Journal of Applied Earth Observation and Geoinformation*, *137*, 104386-104398. https://doi.org/10.1016/j.jag.2025.104386

[22] Hao, J., Wong, L. M., & Shan, Z. (2024). A Semi-Supervised Transformer-Based Deep Learning Framework for Automated Tooth Segmentation and Identification on Panoramic Radiographs. *Diagnostics*, *14*(17), 1948-1965. https://doi.org/10.3390/diagnostics14171948

[23] Anzabi, R. M., Badkoobeh, A., & Nabian, M. (2024). Machine Learning Approaches for Integrating Clinical and Radiographic Data in the Early Detection of Osteonecrosis of the Jaw. *Galen Medical Journal*, *13*(SP1), e3623-e3636. https://doi.org/10.31661/gmj.v13iSP1.3623

[24] Shi, J., Li, P., & Shen, C. S. (2023). You Only Label Once: 3D Box Adaptation From Point Cloud to Image With Semi-Supervised Learning. *IEEE robotics and automation letters*, *8*(10), 6811-6818. https://doi.org/10.1109/LRA.2023.3310433

[25] Tengfei, H. (2025). The Construction of Multimodal Metaphors in Animation: An Example from "Soul". *International Journal of Linguistics, Literature & Translation*, *8*(2), 13-26. https://doi.org/10.32996/ijllt.2025.8.2.13

Contact information:

Guoshun ZHENG
Extra-high Voltage Branch of State Grid Fujian Electric Power Co., Ltd.
E-mail: zhengguoshun_fj@163.com