EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Power grid inspection based on multimodal foundation models

Jingbo Hao[1,2,*] and Yang Tao[3]

[1]School of Information and Artificial Intelligence, Nanchang Institute of Science & Technology, Nanchang 330108, China
[2]Hunan Chaoneng Robot, Changsha 410003, China
[3]School of Engineering and Technology, Nanchang Vocational University, Nanchang 330007, China

Abstract

INTRODUCTION: With the development of large foundation models, power grid inspection is transmitting from traditional deep learning to multimodal foundation models.
OBJECTIVES: This paper aims to boost the application of multimodal foundation models for power grid inspection.
METHODS: Current research on foundation models and multimodal large language models (LLMs) is introduced respectively. Three application forms of multimodal foundation models in power grid inspection are explored. The reliability of these models is discussed as well.
RESULTS: These techniques can significantly reduce the time and cost of inspection by automating the analysis of large amounts of sensor data. They can also improve the accuracy and reliability of inspection by leveraging the understanding and reasoning abilities of LLMs.
CONCLUSION: These advanced techniques have shown great application potential in power grid inspection. But it is important to note that they should not entirely replace human inspectors who can validate automatic findings and address possible issues not captured by these models alone.

Keywords: power grid inspection, foundation model, large language model, multimodal application.

Received on 18 August 2024, accepted on 20 September 2024, published on 14 April 2025

doi: 10.4108/ew.9087

1. Introduction

Power grid equipment plays a critical role in ensuring the reliable and efficient distribution of electricity. Due to the essential nature of power grids, the devices equipped in the infrastructure have high reliability requirements. Sudden failures of power grid equipment may have a severe impact on the society causing economic losses and posing threats to the safety of the people's lives and property [1]. In view of this, power grid inspection is indeed a meaningful task for the regular inspection of the power grid to ensure its normal operation and safety. Specifically, it involves the inspection of transmission lines, substations, switchgears, insulators, grounding systems, protection devices and more in order to identify potential problems and prevent power grid failures and accidents.

A power grid inspection system consists of three parts: platforms, sensors and methods. The platforms include climbing robots, helicopters, unmanned aerial vehicles (UAVs) and hybrid platforms, while the sensors include thermal sensors, vision sensors, radar sensors and multi-sensors [2]. Currently inspection methods are primarily based on deep learning. However, deep learning-based techniques often struggle with overall perception and reasoning abilities, which leads to increased instances of wrong results. And recently there has been a growing tendency on integrating vision encoders with large language models (LLMs) to address the issue [3, 4]. LLMs have attracted significant

[*]Corresponding author. Email: jbhao@126.com

interest due to their impressive abilities in processing human languages and performing generic tasks. For the power industry, LLMs hold great potential since they can understand sophisticated prompts and reduce sensory overload [5]. Inspired by the excellent transferability of LLMs as language foundation models, much attention has also been paid to vision foundation models (VFMs) for generic vision tasks [6].

With the development of large foundation models, power grid inspection is at a juncture of transition from traditional deep learning to these models [7]. This paper aims to boost the application of multimodal foundation models for power grid inspection. Current research on foundation models and multimodal LLMs is introduced in Section 2 and 3 respectively. Three application forms of multimodal foundation models in power grid inspection are presented in Section 4. The reliability discussion is held in Section 5. Finally, the conclusion is drawn.

2. Foundation Models

The foundation model is an important paradigm shift from traditional feature engineering to base models that are pre-trained on massive data and subsequently fine-tuned for various downstream tasks [8].

2.1. Language foundation models

Language foundation models are revolutionizing the field of natural language processing (NLP) and becoming a part of the majority of NLP systems. The introduction of the Transformer architecture has brought about a significant transformation in language modeling. The Transformer's powerful strength in parallel computation and self-attention will make it possible to effectively integrate huge amounts of data. This capability has laid the groundwork for the advancement of LLMs.

Compared to traditional deep learning models, the advantages of LLMs are listed as follows: stronger learning ability adapting to complex and diverse scenarios; stronger generalization ability enabling good design reutilization; stronger creativity producing more intelligent behaviours; stronger effectiveness with higher accuracy.

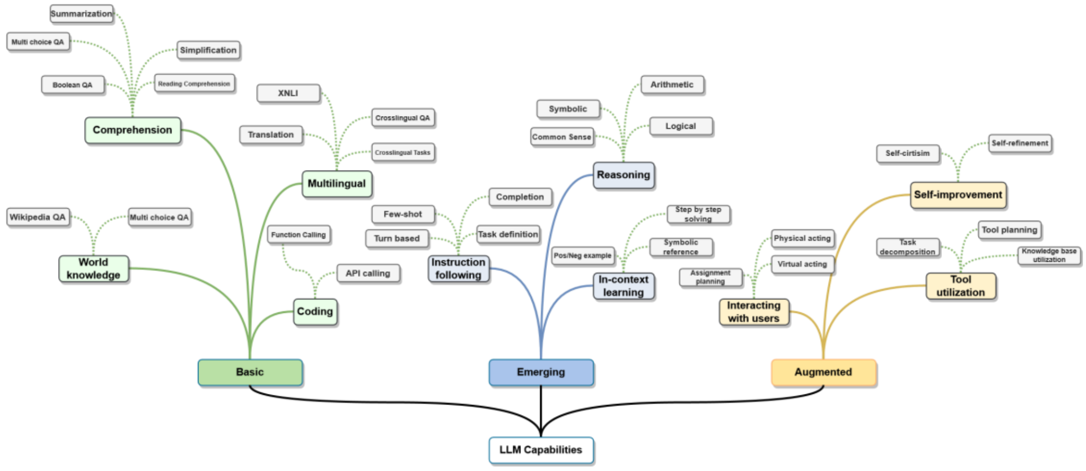


Figure 1. Capabilities of LLMs [9].

As shown in Figure 1, LLMs not only possess basic understanding and generation capabilities, but also exhibit emergent capabilities that traditional deep learning models lack. The emergent capabilities include in-context learning, Instruction following and step-by-step reasoning [9]. In addition, LLMs can be augmented through supplementary techniques, e.g. retrieval-augmented generation (RAG).

2.2. Vision foundation models

VFMs are used for converting images obtained from different sources into visual knowledge which is fit for multiple downstream tasks including image recognition, object detection and image segmentation.

The Recognize Anything Model (RAM) is a VFM presented for image recognition with Swin Transformer as its image encoder [10]. It showcases the remarkable zero-shot capability to accurately identify commonsensible objects and scenes.

GLEE is a VFM designed for object detection, visual grounding and image segmentation [11]. The model mainly consists of a text encoder, an image encoder, a visual prompter and an object decoder.

The Segment Anything Model (SAM) is proposed by Meta as a VFM for promptable image segmentation [12]. It actually defines a novel segmentation task and the model is primarily composed of an image encoder, a prompt encoder and a mask decoder.

The fusion of language models and vision models has led to the emergence of vision-language models (VLMs) which combine image encoders and language models to implement visual understanding and semantic reasoning simultaneously. These models have achieved significant success in many tasks, such as image captioning, visual question answering and visual reasoning. VLMs can be divided into four types: contrastive-based, masking-based, generative-based and LLM-based [13]. Within the last two years, LLM-based VLMs have shown high effectiveness by the aid of the reasoning and generalization abilities of open-source LLMs.

3. Multimodal LLMs

The multimodal LLM is defined as the LLM-based model which can input, process and output multimodal data that is a combination of diverse data types including images, text, audio, time series or composite data structures.

A multimodal LLM usually comprises of three layers: modality layer, alignment layer and LLM layer as shown in Figure 2. The modality layer is utilized to extract the features of specific modality data. The alignment layer is employed to align various modality features with the LLM encoding space. The LLM layer is used for the execution of a pre-trained LLM. At present prevailing multimodal LLMs are mainly LLM-based VLMs since visual tasks are in the highest demand among all kinds of tasks applicable to multimodal LLMs.

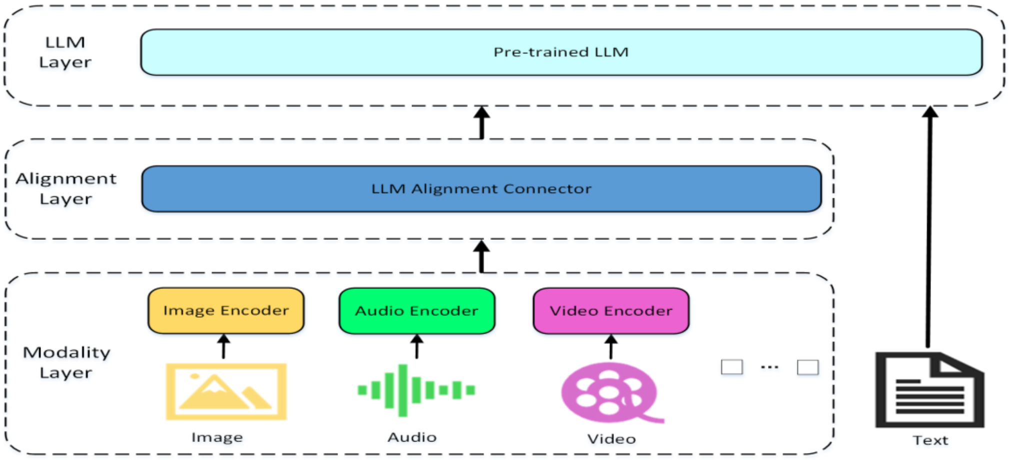


Figure 2. Typical architecture of a multimodal LLM.

3.1. Image encoders

The performance of a multimodal LLM is significantly influenced by the design of the image encoder as well as image resolution. The image encoder accepts image input and extracts visual feature vectors. These features will act as the input to the alignment connector.

Many recent studies have demonstrated empirically that employing higher image resolution can lead to impressive performance improvement. The methods used to scale up input resolution can be classified into two categories: direct scaling methods and patch-division methods [14]. The direct scaling method requires feeding higher resolution images into the encoder, which usually require additional adjustment to the encoder or the replacement of an encoder of a higher resolution version. The patch-division method divides a high-resolution image into small patches and then utilizes the low-resolution encoder for processing.

3.2. Alignment connectors

The alignment connector is responsible for the information alignment from both the textual and visual modalities and produces joint embeddings with shared semantics. The alignment connector is indispensable, but it possesses less importance compared to other system components.

For multimodal information fusion which is essential in alignment connectors, there are mainly two ways: token-level fusion and feature-level fusion [14]. Token-level fusion involves transforming the encoder's output vectors into tokens which are then joined together with text tokens before going into the LLM. To achieve this, a well-used solution is to employ a set of learnable queries to generate tokens in a query-based manner. On the other hand, feature-level fusion incorporates additional modules in order to facilitate deep feature fusion between textual and visual data.

3.3. Pre-trained LLMs

Pre-trained LLMs refer to large models that have been trained on a vast corpus of text data to learn the statistical patterns and linguistic structures of languages. They are typically trained using unsupervised learning techniques, such as masked language modeling or next sentence prediction, on tasks like predicting missing words in a sentence or determining if two sentences are coherent. Once pre-trained, LLMs can be fine-tuned on specific downstream tasks, such as text classification, sentiment analysis, or machine translation, by adding task-specific layers on top of the pre-trained model and training them on task-specific labelled data. This fine-tuning process helps the model adapt its knowledge to a specific task, improving its performance and generalization ability.

Pre-trained LLMs have gained significant popularity and success due to the ability to get contextual information and semantic representations. Nowadays they have become a crucial component in various applications including multimodal LLM systems.

4. Applications in Power Grid Inspection

Power grid inspection refers to the process of evaluating the condition and functionality of the electrical power grid infrastructure. Its purpose is to ensure the safe and reliable operation of the grid by identifying any potential issues or vulnerabilities. During the power grid inspection, various components, such as power lines, transformers, substations, and switchgears, should be examined visually.

4.1. SAM-based inspection

Manual annotation of the training data for infrared image segmentation is time-consuming and labour-intensive. Lin et al. [15] proposed to utilize the pre-trained foundation model SAM as shown in Figure 3 to control segmentation results by iteratively clicking on desired areas until ideal segmentation results are achieved. This approach facilitates the fast and accurate annotation of power equipment in infrared images and obviously reduces required manpower and time costs.

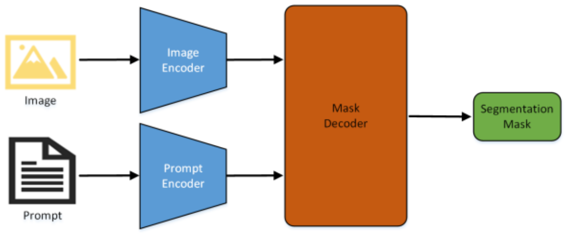


Figure 3. Segment Anything Model.

Current deep learning-based approaches for the image segmentation of power equipment are often struggling due to the poor generalization ability of feature extraction. Guo et al. [16] proposed SAMPE which leverages the excellent feature extraction capability of SAM through incorporating tailored adapters to facilitate effective feature transfer in power equipment image segmentation. In essence SAMPE utilizes adapters to fine-tune the learned features of SAM specifically for power scenarios. Specifically, it employs a convolutional neural network (CNN) encoder to enhance local features and an auto-prompting encoder to enable the seamless end-to-end functionality.

SAM can achieve accurate segmentation of multiple substations in an image without relying on any cues as shown in Figure 4 (a). However, compared to small deep learning models, large foundation models require massive data to leverage their advantages. The poor performance of large models in power scenarios is often due to a lack of high-quality training data rather than insufficient model designs. Due to the limited scale of data, the segmentation result of SAM on the insulator strings in an infrared image is not very ideal as shown in Figure 4 (b) [7].

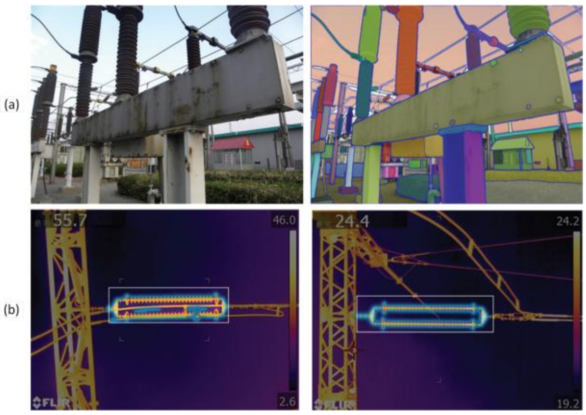


Figure 4. Segmentation examples of SAM [7].

4.2. VLM-based inspection

Existing object detection models usually struggle with holistic understanding and reasoning abilities, leading to increased occurrences of false and missed detection results. Gao et al. [3] designed a VLM named Grid-Blip focusing on the detection of wildfires in grid inspection. Grid-Blip adopts the BLIP architecture [17] with a high-resolution image encoder and a Baichuan-13B-Chat model fine-tuned with power industry data. On that basis, it can produce alarm notifications based on the detected target information as shown in Figure 5.

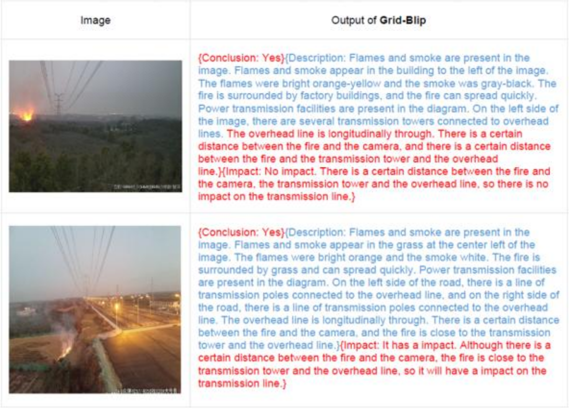
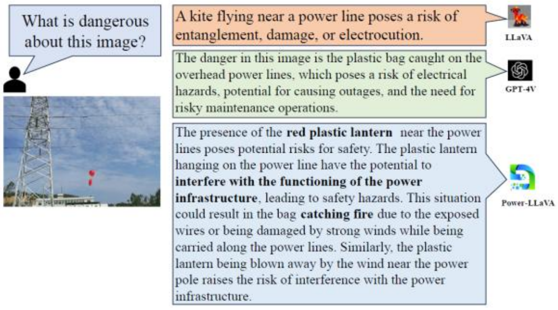


Image	Output of **Grid-Blip**
	{Conclusion: Yes}{Description: Flames and smoke are present in the image. Flames and smoke appear in the building to the left of the image. The flames were bright orange-yellow and the smoke was gray-black. The fire is surrounded by factory buildings, and the fire can spread quickly. Power transmission facilities are present in the diagram. On the left side of the image, there are several transmission towers connected to overhead lines. The overhead line is longitudinally through. There is a certain distance between the fire and the camera, and there is a certain distance between the fire and the transmission tower and the overhead line.}{Impact: No impact. There is a certain distance between the fire and the camera, the transmission tower and the overhead line, so there is no impact on the transmission line.}
	{Conclusion: Yes}{Description: Flames and smoke are present in the image. Flames and smoke appear in the grass at the center left of the image. The flames were bright orange and the smoke white. The fire is surrounded by grass and can spread quickly. Power transmission facilities are present in the diagram. On the left side of the road, there is a line of transmission poles connected to the overhead line, and on the right side of the road, there is a line of transmission poles connected to the overhead line. The overhead line is longitudinally through. There is a certain distance between the fire and the camera, and the fire is close to the transmission tower and the overhead line.}{Impact: It has a impact. Although there is a certain distance between the fire and the camera, the fire is close to the transmission tower and the overhead line, so it will have a impact on the transmission line.}

Figure 5. Detection examples of Grid-Blip [3].

Similarly, Wang et al. [4] presented Power-LLaVA that is well designed to provide a professional and dependable service for power transmission line inspection and capable of interacting naturally with humans as shown in Figure 6. It comprises a pre-trained ViT-L/14 model as the image encoder and a fine-tuned Vicuna-7B model as the LLM.



What is dangerous about this image?

A kite flying near a power line poses a risk of entanglement, damage, or electrocution.

The danger in this image is the plastic bag caught on the overhead power lines, which poses a risk of electrical hazards, potential for causing outages, and the need for risky maintenance operations.

The presence of the **red plastic lantern** near the power lines poses potential risks for safety. The plastic lantern hanging on the power line have the potential to **interfere with the functioning of the power infrastructure**, leading to safety hazards. This situation could result in the bag **catching fire** due to the exposed wires or being damaged by strong winds while being carried along the power lines. Similarly, the plastic lantern being blown away by the wind near the power pole raises the risk of interference with the power infrastructure.

Figure 6. Detection examples of Power-LLaVA [4].

In practise, Shenzhen Power Supply Bureau released the Zhurong 2.0 model by which the hidden dangers and risk levels to power grid equipment can be accurately described without the need for further manual investigation [18]. In addition, China Southern Power Grid Company released a power large model which is proficient in chatting, querying, diagramming and writing, and has been applied in multiple provinces across the country [19].

4.3. UAV-based inspection

UAV inspection of power grids is becoming an important and efficient maintenance approach, and integrating the abilities of artificial intelligence (AI) gradually. Through the empowerment of large foundation models, UAV inspection has made big progress thus far.

Large models enable automatic analysis of multimodal data, such as UAV flight trajectories, inspection targets and equipment status, improving the accuracy, efficiency, and safety of UAV inspection. The application of large models in UAV inspection primarily focuses on image recognition and data analysis. Utilizing the general reasoning ability of large models, high definition images captured by UAVs can be intelligently analysed to accurately identify minor damage, corrosion, foreign object suspensions and other abnormal conditions in key components. Leveraging the scene understanding and path planning abilities of large models, UAVs can skillfully make flight route planning, avoid obstacles and increase inspection frequency in key areas or historical fault points, which can thereby optimize inspection efficiency and coverage. Combining UAVs with AI models enables real-time monitoring of the power grid’s operational status. Once an anomaly is detected, the UAV will issue an alert and conduct focused investigation in the identified area. The powerful semantic processing ability of large models allows for efficient filtering of irrelevant information and focusing on key data, providing timely and effective information support for backend analysis. This will shorten the cycle from data collection to decision-making markedly.

5. Reliability Discussion

In critical application areas like medicine and electricity, there is usually a high demand for the reliability of computation models including adversarial robustness and model interpretability. Researchers strive to improve the accuracy of models all the time and these models even surpassed human performance in many cases. However, adversarial robustness and model interpretability have not yet received enough attention.

5.1. Adversarial robustness

Szegedy et al. [20] proposed the concept of adversarial examples referring to input samples crafted by artificially adding subtle perturbations to original data. Adversarial examples may lead to wrong predictions from the attacked model with high confidence and deep learning models are generally vulnerable to adversarial examples. In many cases, models with different structures trained on different datasets will still make incorrect predictions on the same adversarial example, meaning that adversarial examples have become a blind spot in the application of deep neural networks. Some research shows that the vulnerability to adversarial examples is not unique to deep learning models and exists in many machine learning models. Adversarial robustness represents a model's ability to resist adversarial examples, i.e., the ability to produce correct predictions on adversarial examples.

Adversarial attacks on power grids can be categorized into evasion attacks and poisoning attacks. The former manipulates input samples in order to evade the detection of a built model and the latter tries to pollute training data with malicious intent to corrupt the development of a new model. Furthermore, these attacks can also be categorized as either white-box attacks or black-box attacks. The former is applied with the understanding that the complete knowledge of the target model is available. On the other hand, the latter is conducted under the assumption that any information about the target model is unavailable.

The most common white-box evasion attack is generally defined as an optimization problem [21]:

$$\underset{x_e}{\operatorname{argmax}}\ \mathrm{L}(x_e, y, W) \quad s.t.\ \left\| x_e - x \right\|_d \le \varepsilon,\ x_1 \le x_e \le x_u \quad (1)$$

where x is the original sample with the label y and x_e is the generated adversarial example. For the attacked model, W is the parameter weights and L is its loss function. $\|.\|_d$ represents the L_d norm operation. ε is a small value used to limit the perturbation scale, and x_l and x_u fix the upper and lower bounds of an adversarial example.

Cui et al. [22] made an extensive study on the robustness of multimodal LLMs against adversarial attacks across various tasks. They found that on the whole multimodal LLMs are vulnerable to adversarial examples. But they also indicated that when prompts are given to the model, the context will contribute to alleviating the impact of those adversarial examples.

5.2. Model interpretability

Model interpretability is defined as the degree to which humans can consistently predict the output of a model. Deep learning models have shown excellent performance in many domains and the performance is largely dependent on the highly non-linear nature and massive number of parameters of the models. People cannot understand what knowledge a model has learned from mass data and how it makes final decisions as the end-to-end decision-making process leads to weak interpretability. The US Department of Defense put forward the idea of eXplainable Artificial Intelligence (XAI) firstly. It is essentially a set of concepts, methods, and processes that help the developer insert a transparency layer to a deep learning model to assess the reasonability of the model's output [23]. Cambria et al. [24] pointed out that so far there was only a scarcity of research dedicated to the XAI techniques specifically designed for LLMs. They suggested that developing XAI methods for LLMs is not only crucial from a technical point but also an important step to promote responsible computation.

5.3. Other consideration

In the context of LLMs, hallucination is defined as the generation of information that is either inconsistent with facts or meaningless. As for multimodal LLMs, model output is sometimes inconsistent with corresponding visual input, which will present significant hurdles to practical implementation and raises concerns about reliability [25]. Ensuring the safety of multimodal LLMs is a complex task due to several reasons which include the challenges posed by different data modalities, the access of black-box APIs, the reliance on unknown data sources and the emergence of unpredictable events like hallucination, data leakage and adversarial attacks [26].

6. Conclusion

Power grid inspection techniques are evolving into the age of multimodal foundation models. These techniques can significantly reduce the time and cost of inspection by automating the analysis of large amounts of sensor data. They can also improve the accuracy and reliability of inspection by leveraging the understanding and reasoning abilities of LLMs. However, it is important to note that these techniques should not replace physical inspection entirely. They should be used as complementary tools to assist human inspectors in decision-making. Physical inspection is still necessary to validate automatic findings and to address any issues that may not be captured by these models alone.

Acknowledgements.

The work is partially supported by the S&T Research Projects of the Jiangxi Provincial Department of Education (GJJ212521, GJJ2202908) and the Scientific Research Start-up Fund for Introduced Talents of Nanchang Institute of Science & Technology (NGRCZX-20-12).

References

[1] Wang L. A review of the application of machine vision in power safety monitoring. Zhejiang Electric Power. 2022; 41(10): 16-26.
[2] Yang L, Fan J, Liu Y, et al. A review on state-of-the-art power line inspection techniques. IEEE Transactions on Instrumentation and Measurement. 2020; 69(12): 9350-9365.
[3] Gao P, Rao Z, Gao S, et al. Research on grid inspection technology based on general knowledge enhanced multimodal large language models. Proceedings of the 12th International Symposium on Multispectral Image Processing and Pattern Recognition; 10-12 November 2023; Wuhan. Bellingham: SPIE; 2024. 130860B.
[4] Wang J, Li M, Luo H, et al. Power-LLaVA: large language and vision assistant for power transmission line inspection. arXiv Preprint. 2024; 2407.19178.
[5] Majumder S, Dong L, Doudi F, et al. Exploring the capabilities and limitations of large language models in the electric energy sector. Joule. 2024; 8(6): 1544-1549.
[6] Shen Y, Fu C, Chen P, et al. Aligning and prompting everything all at once for universal visual perception. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 17-21 June 2024; Seattle. Piscataway: IEEE; 2024. p. 13193-13203.
[7] Zhao Z, Feng S, Xi Y, et al. The era of large models: a new starting point for electric power vision technology. High Voltage Engineering. 2024; 50(5): 1813-1825.
[8] Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. arXiv Preprint. 2022; 2108.07258.
[9] Minaee S, Mikolov T, Nikzad N, et al. Large language models: a survey. arXiv Preprint. 2024; 2402.06196.
[10] Zhang Y, Huang X, Ma J, et al. Recognize anything: a strong image tagging model. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 17-21 June 2024; Seattle. Piscataway: IEEE; 2024. p. 1724-1732.
[11] Wu J, Jiang Y, Liu Q, et al. General object foundation model for images and videos at scale. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 17-21 June 2024; Seattle. Piscataway: IEEE; 2024. p. 3783-3795.
[12] Kirillov A, Mintun E, Ravi N, et al. Segment anything. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision; 1-6 October 2023; Paris. Piscataway: IEEE; 2023. p. 3992-4003.
[13] Bordes F, Pang RY, Ajay A, et al. An introduction to vision-language modeling. arXiv Preprint. 2024; 2405.17247.
[14] Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. arXiv Preprint. 2024; 2306.13549.
[15] Lin Y, Zhang F, Li Z, et al. Large model based interactive segmentation of infrared image for power equipment. Journal of Network New Media. 2024; 13(2): 53-60, 67.
[16] Guo Y, Gong R, Li D, et al. SAMPE: auto-prompting SAM for generalizable power equipment image segmentation. IEEE Access. 2024; 12: 104291-104299.

[17] Li J, Li D, Xiong C, et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Proceedings of the 39th International Conference on Machine Learning; 17-23 July 2022; Baltimore. Online: PMLR; 2022. p. 12888-12900.

[18] Ye Q, Yang J. “Power GPT” boosts the efficiency of safety hazard alarming by six times. Science and Technology Daily. 18 September 2023.

[19] Ye Q. Power large model: proficient in “chat, query, diagram and write.” Science and Technology Daily. 9 October 2023.

[20] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. Proceedings of the 2nd international conference on learning representations; 14-16 April 2014; Banff. Online: OpenReview; 2014.

[21] Hao J, Tao Y. Adversarial attacks on deep learning models in smart grids. Energy Reports. 2022; 8(S2): 123-129.

[22] Cui X, Aparcedo A, Jang YK, et al. On the robustness of large multimodal models against image adversarial attacks. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 17-21 June 2024; Seattle. Piscataway: IEEE; 2024. p. 24625-24634.

[23] Hao J. Towards reliable medical image analysis based on deep learning with XAI. Proceedings of the 2nd International Conference on Image, Signal Processing and Pattern Recognition; 24-26 February 2023; Changsha. Bellingham: SPIE; 2023. 1270754.

[24] Cambria E, Malandri L, Mercorio F, et al. XAI meets LLMs: a survey of the relation between explainable AI and large language models. arXiv Preprint. 2024; 2407.15248.

[25] Bai Z, Wang P, Xiao T, et al. Hallucination of multimodal large language models: a survey. arXiv Preprint. 2024; 2404.18930.

[26] Zhao T, Zhang L, Ma Y, et al. A survey on safe multi-modal learning systems. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 25-29 August 2024; Barcelona. New York: ACM; 2024. p. 6655-6665.