Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29794

Integration Mechanisms of Multi-modal AI in Energy Systems

Jiangdong Wang

Swanburn Institute, Shandong University of Science and Technology, Qingdao, China
wangjiangdong.xp@gmail.com

Abstract. Under the backdrop of the "carbon neutrality" goal, the integration of the integrated energy system (IES) and MMAI is expected to become a key research direction.Firstly, IES plays a significant role in optimizing energy allocation, promoting the intelligent transformation of energy, and enhancing energy utilization efficiency. However, it still has issues such as complexity, multiplicity, and uncertainty. Secondly, the optimization methods of traditional integrated energy systems encounter numerous problems when dealing with high-dimensional, heterogeneous, and time-varying multimodal data, including excessively high computational complexity and difficulties in handling heterogeneous data.Therefore, Therefore, the introduction of MMAI technology is required the system's ability to handle the complexity and diversity of data. This paper adopts the research approach of "theoretical analysis - architecture construction - mechanism explanation - challenge outlook" to explore the integration mechanism of multimodal AI in IES. Research has shown that MMAI can achieve efficient processing and rapid adaptation of multimodal data through a closed loop of "perception - cognition - decision - control", thereby enhancing the intelligence level of the system. However, the technological development of MMAI integrated with IES still faces multiple challenges. To address these challenges, in the future, our research efforts should be focused on the data level, algorithm level, and system level. This technology has multiple research directions, such as developing lightweight and interpretable multimodal fusion models, constructing an IES multimodal open benchmark dataset and simulation platform, and exploring new paradigms for the integration of physical mechanisms and data-driven approaches.

Keywords: Integrated Energy System (IES), Multi-modal Artificial Intelligence Technology (MMAI), SCADA time series data

1. Introduction

The increase in global temperatures due to greenhouse gases emitted from human activities represents the most significant consequence [1]. In response, China has introduced the "Dual Carbon Strategy," which aims to reach a peak in carbon emissions by 2030 and achieve carbon neutrality by 2060 [2]. Accordingly, China intends to expedite the transition to a new energy framework, enhance energy efficiency, secure energy resources, and foster green, low-carbon growth. However, the conventional energy system is inadequate for fulfilling the requirements of low-carbon development.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29794

The integrated energy system (IES) offers complementary energy sources and coordinated supply, which can improve the efficiency and cleanliness of energy consumption at the end user level [3]. The Integrated Energy System (IES) has garnered significant interest from both researchers and industry professionals. IES integrates diverse energy forms such as electricity, heating, cooling, gas, and hydrogen, along with the cooperative use of three primary types of modal data: data-based, image-based, and text-based modalities [4]. It is widely accepted that this system can enhance overall energy efficiency [5] and improve the alignment of supply with demand [6]. Consequently, the Integrated Energy System is seen as a crucial focus for the future transformation and development of energy and power [7]. Nevertheless, it also involves the intricacies of multi-energy flow interactions, along with the conflicts arising from various entities and differing objectives. As a result, its capability to address complex issues still has limitations. Without the right system configuration and optimal operational strategies, this system cannot fully tap into its potential. Consequently, to enhance the advantages of Integrated Energy Systems (IES), optimization of the system is essential [8]. Employing multimodal artificial intelligence (MMAI) for system optimization is a practical and effective strategy. The functioning of conventional IES relies on optimization techniques derived from mathematical programming and intelligent algorithms. The foundation of these approaches involves the construction of precise multi-energy flow coupling models and efficient intelligent algorithms, among other components. However, it is important to note that this approach is not without its limitations, which include deficiencies in model accuracy and generalization. To illustrate this point, consider the application of machine learning in the domains of load forecasting and fault diagnosis. Machine learning algorithms are capable of extracting potential patterns from historical operation data, thereby constructing prediction and decision-making models. This, in turn, enhances the accuracy of predictions and the rationality of scheduling. Concurrently, it can detect the normal operation of equipment and achieve the function of fault diagnosis. Multimodal learning research has made significant progress. In the domain of computer vision, the rapid advancements in artificial intelligence (AI) technology, particularly the significant progress in deep learning algorithms, have led to noteworthy achievements in ophthalmic disease diagnosis by AI systems based on single-modal data, such as fundus photographs or optical coherence tomography (OCT) [9]. In the field of natural language processing, multimodal AI finds extensive application in areas such as machine translation and sentiment analysis. However, extant research continues to operate at the level of single-modal data or a single technical approach, impeding comprehensive elucidation of the intricate coupling relationships and dynamic mechanisms in multi-energy systems. A paucity of in-depth analyses exists from a system theoretical framework and mechanism level regarding the role mechanism of multimodal artificial intelligence (MMAI) in the optimized operation of integrated energy systems (IES). This has resulted in the failure to realize the full potential of MMAI, hindering support for the global optimization and intelligent development of IES. Consequently, this paper adopts "multimodal fusion" as the research starting point, constructs the MMAI-IES integrated theory framework, analyzes its core integration mechanism, deduces its application value, and systematically examines the challenges it faces.

2. Integrated energy system multimodal data characteristics and integration theoretical framework

2.1. Abbreviations and acronyms

During the optimization and operation of the integrated energy system, a variety of data sources are utilized. However, a taxonomy can be proposed based on the presentation forms of these

DOI: 10.54254/2755-2721/2026.TJ29794

phenomena. This taxonomy would include three main modalities: data modalities, image modalities, and text modalities. Consequently, the varied data modalities mirror the multifaceted dimensions of information concerning the operational mode, information sources, and management of operations of the IES. The following is a list of the primary characteristics that were identified during the analysis.

1) Numeric modal data： Representative sources include: The term "SCADA" (Supervisory Control and Data Acquisition system) refers to a system that acquires and processes data from sensors.

Characteristic analysis: Continuity and real-time nature: Information collection systems, such as sensors, sample data at a high frequency, thereby reflecting the dynamic characteristics changes of the system.

Multi-source integration: The integration of heterogeneous data from multiple sources, including sensors and SCADA systems, is a critical component of the data management process.

The present study demonstrates a high degree of structuring. Numerical data are characterized by a fixed format, a property that facilitates storage and calculation.

The presence of noise and missing values is a concern. Such issues may include mechanical accuracy, communication delay, or equipment failure, which can result in abnormal points or missing values.

2) Image-based modal data： Representative sources include: The utilization of infrared thermal imaging and video surveillance techniques constitutes a methodical approach to surveillance.

Characteristic analysis: The presence of a substantial amount of information: A single frame image can contain over ten thousand pixels.

The level of uncertainty is high. The acquisition of data is influenced by various factors, including weather conditions, lighting conditions, and the angle at which the image is captured. This results in a significant degree of uncertainty.

The following are notable spatial distribution characteristics: The image has the capacity to directly reflect the spatial state.

3) Textual Modal Data： Typical sources: Operation and Maintenance Reports, Weather Texts.

Characteristic analysis: Unstructured and highly semantic: The text data is mostly described in natural language, with flexible information expression but low degree of structuring.

High domain specificity: The operation and maintenance reports contain a large number of industry terms, abbreviations, and equipment codes.

Timeliness and context dependence: For example, weather forecast texts have a close relationship with operation scheduling and have strong constraints on timeliness.

Data acquisition relies on natural language processing technology: Semantic parsing and knowledge extraction need to be carried out by means of techniques such as word segmentation, named entity recognition, and knowledge graphs.

2.2. Theoretical foundation of multi-modal AI integration

The theoretical foundation of multimodal artificial intelligence (MMAI) mainly consists of three core aspects: multimodal representation learning, cross-modal alignment, and information fusion. The integrated energy system (IES) involves the acquisition of multi-source heterogeneous data and the coupling of multiple energy flows. Therefore, the rational application of multimodal artificial intelligence technology to the integrated energy system can enable the efficient processing and rapid adaptation of multi-modal data by the integrated energy system.

DOI: 10.54254/2755-2721/2026.TJ29794

1) Multimodal Representation Learning: Multimodal representation learning refers to the process of representing data from multiple different sources using machine learning techniques. Through translation, alignment, fusion, and collaborative learning, it maps these data to the same feature space to fully preserve semantic information and correlations [10]. Multimodal learning representation can be divided into joint representation and coordinated representation.

Joint representation: Joint representation is achieved by constructing a shared feature space, mapping the inputs of multiple modalities to a unified representation, thereby capturing the intrinsic connections among multiple modalities.

Coordinated representation: Coordinated representation ensures the comparability between modalities while preserving the independent features of each modality by using constraints and mapping functions.

In the IES system, through multimodal representation learning, data modalities such as data type, image type, and text type can be unifiedly represented, enabling cross-modal state analysis and prediction.

2) Cross-modal Alignment: Different modalities differ in terms of time scale, spatial distribution rate and semantic hierarchy, and the establishment of corresponding relationships requires the implementation of alignment mechanisms.

Temporal alignment: The cross-modal alignment system ensures the establishment of the relationship between multi-modal data by synchronizing SCADA time series data with video surveillance frames.

Spatial alignment: The cross-modal alignment system needs to align infrared images with equipment topology or sensor positions to achieve spatial consistency.

Semantic alignment: The cross-modal alignment system can correlate fault events described in operation and maintenance text with abnormal systems of images or numerical data through natural language processing.

Cross-modal alignment is an important prerequisite for achieving complementary and coordinated multi-modal information.

3) Information Fusion: The purpose of information fusion is to integrate multi-modal information, thereby enhancing the expression ability and decision-making accuracy of each modality. The fusion strategies can be divided into three parts: early fusion, mid-fusion, and late fusion.

Early fusion: Early fusion combines and maps multi-modal information at the feature level, but this is only applicable to data with high correlation between modalities.

Mid-fusion: Mid-fusion captures the correlations between modalities through sharing a multi-modal interaction module in a dynamic manner.

Late fusion: Late fusion integrates the outputs of independent models of each modality at the decision level to enhance the robustness of the system.

In the IES system, information fusion not only compensates for the deficiencies of single-modal data but also provides more comprehensive basis in fault diagnosis, assessment of operating status, and predictive scheduling.

2.3. "Perception-cognition-decision-control" integrated architecture design

In the intelligent journey of the integrated energy system (IES), traditional single-modal information processing is unable to fully depict the state of the system and is insufficient for supporting adaptive scheduling in complex environments. Therefore, we need to enhance and improve it with artificial intelligence technology. The common artificial intelligence algorithms in multi-modal systems

DOI: 10.54254/2755-2721/2026.TJ29794

include convolutional neural networks (CNN), deep learning, and ensemble methods [9]. This paper proposes a "perception - cognition - decision - control" four-layer integrated framework based on multi-modal artificial intelligence algorithms, which can achieve systematic design from data acquisition to closed-loop control.

1) Perception Layer: The perception layer forms the foundation of the architecture and is primarily responsible for the collection and preprocessing of multi-source heterogeneous data.

Data collection: Data collection encompasses various types of data such as data values (SCADA, sensors), image-based (infrared thermal imaging, video surveillance), and text-based (operation and maintenance reports, weather text).

Preprocessing: Preprocessing can perform tasks such as noise reduction, data completion, normalization, and structuring on the raw data to ensure the accuracy and usability of the data.

This layer ensures that the system possesses comprehensive and real-time environmental perception capabilities.

2) Cognitive Level: The cognitive layer is the core of the entire architecture and plays a role in understanding and unifying the representation of multimodal information. The main function of the cognitive layer in the IES system is to achieve feature extraction, cross-modal alignment, and multimodal fusion.

Feature extraction: Feature extraction uses deep learning methods to represent high-level features of different modal data.

Cross-modal alignment: Cross-modal alignment solves the differences in sampling frequency, spatial resolution, and semantic level of multimodal data through temporal synchronization and semantic mapping.

Multimodal fusion: Multimodal fusion refers to using methods such as attention mechanisms, graph neural networks, or contrastive learning to fuse multimodal features into a unified state representation, forming a comprehensive representation of the operating conditions of the IES.

3) Decision-making level: The decision-making layer performs optimization calculations and intelligent decisions based on the unified representation of the integrated state. The functional mechanism of the decision-making layer mainly consists of three parts: optimization scheduling, intelligent prediction and diagnosis, and decision-making mechanism.

Optimization scheduling: Utilizing multi-objective algorithms to achieve coordinated operation of energy flow, information flow, and material flow.

Intelligent prediction and diagnosis: Based on the fusion model, load forecasting, fault diagnosis, and risk assessment are conducted.

Decision-making mechanism: Combining methods such as reinforcement learning and game theory to enhance the system's adaptive ability in uncertain environments.

4) Control Layer: The control layer is responsible for converting the results of the decision-making layer into executable control instructions and applying them to the system through actuators. The functions of the control layer mainly include two parts: instruction issuance and system feedback.

Instruction issuance: Dispatching instructions to various energy devices (such as power grids, energy storage systems, and cold and heat source devices).

System feedback: Real-time collection of system operation results and feeding them back to the perception layer, thereby forming a closed loop of the "perception-cognition-decision-making-control" four-layer integrated framework.

This layer ensures the dynamic controllability and self-adaptive optimization of the IES operation.

DOI: 10.54254/2755-2721/2026.TJ29794

5) Overall framework analysis： This four-layer framework is based on data collection from the perception layer, centered on multi-modal fusion in the cognitive layer, driven by optimization computing in the decision-making layer, guaranteed by execution and feedback in the control layer, forming a complete and intelligent closed-loop system. This architecture provides a systematic theoretical framework support for the efficient, stable and intelligent operation of IES.

3. Explanation of core integration mechanism and application analysis

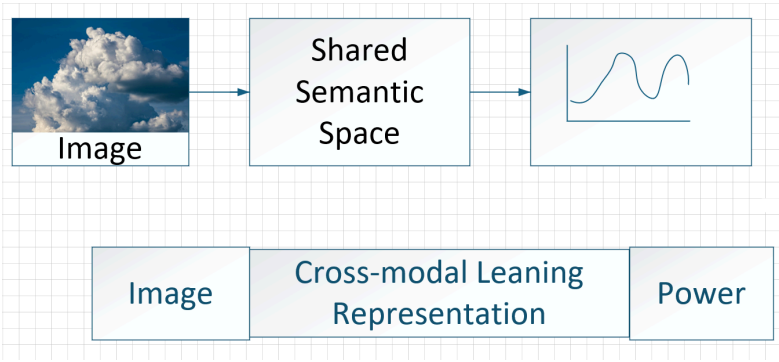

Figure 1. The process of mapping cloud charts and power curves into a shared space

The key to the application of multimodal artificial intelligence (MMAI) in integrated energy systems lies in achieving an enhancement of the system's perception, cognition, decision-making, and control capabilities by deeply integrating heterogeneous data such as images, time series, and text. Within the "perception-cognition-decision-control" four-layer framework, the cross-modal representation and semantic alignment mechanism completes the unified representation of multimodal data at the perception layer, realizes feature fusion and understanding of complex operating conditions at the cognition layer, and provides reliable input and semantic consistency verification at the decision-making and control layers. Figure 1 presents this process in an image format in detail. Here, "image" represents the sky cloud map, and the coordinate graph represents the output of photovoltaic power. The two share data through the shared representation space. The image data and power output data are represented in the same feature space through cross-modal representation learning.

3.1. Cross-modal representation and semantic alignment mechanism

In the integrated energy system, there are significant differences in heterogeneous data. For instance, sky cloud images and power drops. Sky cloud images represent weather boundaries in the form of images, while the output of photovoltaic power is expressed in a numerical time sequence to indicate the energy conversion results. Although both sets of data originate from meteorological factors, due to the different data forms, there exists a "semantic gap". The role of MMAI is to construct cross-modal mapping through shared representation space, enabling heterogeneous data to be characterized by consistent semantic vectors.

The implementation path for integrating MMAI with the integrated energy system mainly consists of two steps: Firstly, feature extraction for different modalities is carried out using deep neural networks. Convolutional Neural Network (CNN) can automatically learn image features, thereby significantly improving classification accuracy and efficiently processing complex image data [11]. Therefore, we can extract spatial texture and structural information as image data through

DOI: 10.54254/2755-2721/2026.TJ29794

CNN; time series data can be captured by Recurrent Neural Network (RNN), Long Short-Term Memory Network (LSTM), or self-attention mechanism. Subsequently, through projection networks or contrastive learning methods, the features of different modalities are embedded into a shared latent space. In this space, visual features such as cloud thickness and movement trajectories in the cloud image can be correlated with the power decline trend.

Secondly, we can enhance the correlation between modalities through the semantic alignment mechanism. Specifically, we employed maximum mutual information constraints, cross-modal attention mechanisms, or contrastive loss functions to enable the model to continuously adjust the functions during training, ensuring that "semantically equivalent" features converge in the shared space, thereby achieving deep correlations between the data. For instance, when the system detects rapidly moving thick clouds in the satellite image, the model can respond quickly and predict that the power curve is about to decline.

3.2. Mechanism of information complementation and redundancy verification

In addition to semantic alignment, the core of multimodal intelligence also includes a key mechanism of information complementarity and redundancy verification. The operation of integrated energy systems is complex, and a single information source often has issues such as insufficient coverage, excessive noise, and failure risks. However, MMAI can achieve more comprehensive and robust perception and decision-making through different forms of complementarity and redundancy.

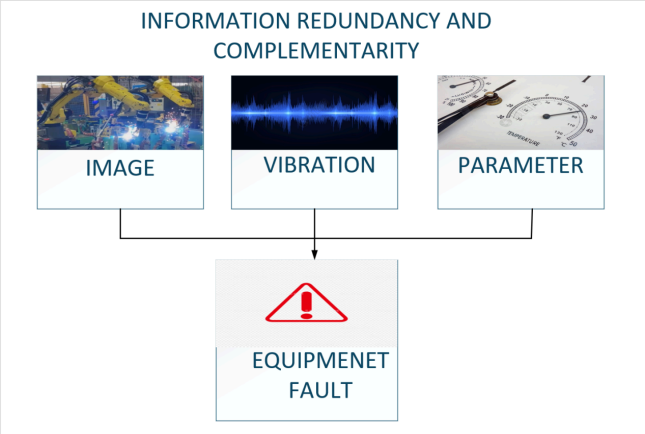

Figure 2. The process of jointly diagnosing equipment failures through images, vibrations and parameter data

Firstly, complementarity is manifested in the fact that different modalities can depict the same state from multiple perspectives. For instance, when a wind turbine is in the monitoring state, the image mode can identify the spatial patterns of ice formation on the blades or structural damage, while the vibration sensor can capture the resulting minor mechanical abnormalities; the operating parameters can supplement the load and environmental background information. The specific process is shown in figure 2. Here, "image" represents the spatial patterns, "vibration" represents minor mechanical abnormalities, and "parameter" represents the operating parameters. When these three are integrated, the system can form a complete state perception, thereby overcoming the limitations of a single mode.

DOI: 10.54254/2755-2721/2026.TJ29794

Secondly, redundancy manifests as repeated measurements of the same physical quantity by multiple modalities. For instance, the temperature of a device can be derived from an infrared thermal imaging image, or directly obtained using a thermocouple sensor. When one of the sensors fails for some reason, the data from the other modalities can provide verification and compensation, thereby preventing data loss. This redundancy mechanism significantly improves the fault tolerance and security of the system, ensuring that the system can still maintain stable monitoring of critical operating parameters under extreme conditions. The dual effects of information complementarity and redundancy verification enable multi-modal fusion not only to be a data overlay, but also to achieve an enhancement in the information hierarchy, thereby enhancing the robustness and fault-tolerant capability of the system's perception.

3.3. Performance analysis of typical application scenarios

After integrating MMAI with the comprehensive energy system through the above mechanism, it demonstrates strong application potential. From this, we infer that three classic scenarios can be used as a basis for performance analysis:

Scene 1: Ultra-short-term prediction of renewable energy

The difficulty of accurately predicting the output of wind and solar power due to rapid changes in weather conditions has always been a challenge for this technology. If relying solely on data-based weather forecasting, there will be a problem of insufficient time resolution, while relying only on sky images will lack quantitative characterization capabilities. Therefore, by integrating two types of modal data, namely "data weather + sky images", this system can capture the corresponding relationship between cloud movement trajectories and radiation intensity in the shared representation space, thereby achieving precise data prediction within seconds to minutes.

Scene 2: Predictive Maintenance of Equipment

The operational health of energy equipment is closely related to the reliability of the system. A single-modal energy equipment is difficult to detect the early signs of system failures, while modal fusion can achieve high-precision diagnosis of faults. For example, when monitoring the same operating system, infrared images can identify abnormal temperature increases, vibration signals can reveal the damage of mechanical components, and operating parameters can provide background information on load and environmental conditions. MMAI maps these three types of modalities to a shared space, enabling the capture of potential signs before the faults have expanded, thus achieving early and precise diagnosis. Utilizing artificial intelligence (AI), especially machine learning (ML) and deep learning (DL) technologies, is a promising research field for addressing the limitations of traditional predictive models [12].

Scene 3: Multi-scale Coordinated Scheduling of Time

The traditional scheduling of integrated energy systems needs to consider the physical coupling of multiple energy flows (such as electricity, heat, and gas), economic constraints from market signals, and load requirements from user behaviors. However, in a single mode condition, the scheduling model often has difficulty simultaneously taking into account multiple aspects such as economy and environmental protection. However， by integrating "multi-energy flows + market signals + user behavior" data, intelligent energy resource management and planning can be achieved [12]. For example, in the short term, rapid responses can be achieved based on user load characteristics, and in the medium and long term, reasonable strategies can be formulated in combination with market economic conditions. Thus, it is possible to achieve dual optimization of economic cost and carbon emissions, improve energy utilization efficiency and promote green development.

DOI: 10.54254/2755-2721/2026.TJ29794

4. Discussion

4.1. Theoretical and technical challenges

Although MMAI has significant application potential in the integrated energy system, there are still many difficult problems at the theoretical and technical levels that need to be solved. Among them, issues at the data level, algorithm level, and system level are the key challenges in the integration process.Data level: Due to the complex multi-modal data types involved in the integrated energy system, which cover various data such as images, time-series sensor data, market signals, and user behaviors, some unavoidable problems will arise. Among them, issues such as data heterogeneity, lack of annotations, low quality, and privacy security have always been the key challenges in research. Firstly, there is a significant heterogeneity in data. Different modalities have considerable differences in dimensions, temporal and spatial resolutions, and expression forms. Direct fusion would encounter the problem of "semantic gap". Secondly, annotation is often missing, especially in scenarios such as fault diagnosis of energy equipment and prediction of extreme weather, where limited data volume makes it difficult for the model to obtain sufficient monitoring signals. Thirdly, the quality of data varies greatly. For instance, issues such as sensor noise, communication interruptions, and missing measurements significantly affect the stability and robustness of the model. Fourthly, there are user privacy and security concerns. Data security in cloud systems also faces risks from advanced network threats, such as network-based attacks, like XSS vulnerabilities [13]. Therefore, how to strike a balance between data sharing and protection is a key challenge for the application to be implemented.

Algorithmic level: The current mainstream polymorphic models mostly rely on deep neural networks, and their "black box" nature leads to insufficient interpretability. Moreover, energy system decision-making often directly affects operational safety and economic effects, thus the current mainstream polymorphic models have difficulty gaining technical trust. Furthermore, the energy multimodal data is massive, and both the training of the model and its inference involve high computational complexity, which limits its application in scenarios with extremely high real-time requirements. On the other hand, the model's generalization ability is insufficient. The model performs well in a single scenario, but when it is transferred to different regions, different devices, or systems with different energy structures, its performance will significantly decline, indicating that the system has a weak adaptability to complex and changing environments and insufficient generalization ability.

System level: The operation of IES relies on the close coupling of software and hardware. Firstly, there is a high degree of difficulty in seamlessly integrating multimodal AI modules into the existing energy management system, especially the issue of insufficient compatibility, which limits the cross-platform and cross-protocol data exchange and functional collaboration to a certain extent. Secondly, energy systems typically demand extremely high reliability and practicality. Therefore, to meet the operational standards of energy systems, the fault-tolerant and stable design of the MMAI model needs to be utilized. However, current research mostly focuses on the algorithm level and lacks fault-tolerant and security mechanisms for the entire system.

4.2. Future research directions

In response to the above issues, future research can be conducted in the following aspects.

At the model level: We need to develop lightweight and interpretable multimodal fusion methods. On one hand, we can reduce computational costs through techniques such as pruning, knowledge

DOI: 10.54254/2755-2721/2026.TJ29794

distillation, and model compression, enabling adaptation for edge computing and real-time applications. On the other hand, we should introduce methods such as causal reasoning, visual attention mechanisms, and symbolic logical constraints to enhance model transparency and interpretability, thereby increasing the trust of industry decision-makers in AI.

At the data level: We need an open and shared IES multi-modal benchmark dataset and simulation platform. Firstly, by standardizing data formats, evaluation metrics, and scenario designs, it can effectively promote the horizontal comparability and vertical accumulation of research results. Secondly, the virtual data generated based on the bionic platform can make up for the shortcomings of real data, providing a broader training environment for the improvement of the robustness and generalization of the model.

At the paradigm level: We need to explore the deep integration of physical mechanisms and data-driven approaches. Firstly, we need to understand that models solely relying on data-driven methods are difficult to capture the complex physical constraints and operating laws in the energy system. However, the paradigm fusion based on methods such as physical information neural networks can embed physical equations and energy conservation laws during the training process, achieving better adaptation to small sample cases or abnormal working states, and providing a reliable guarantee for the safe and controllable optimization of the system.

System level: We need to enhance the cross-level collaborative design of "AI-software-chip". Through the collaborative optimization of software and hardware, we can significantly optimize the real-time inference efficiency and energy consumption performance of the model, thereby meeting the computing power requirements of the energy system at different levels. At the same time, we should promote cross-level open standards, covering model interfaces, data exchange, security fault tolerance, etc., to form a complete industrial ecosystem and a sustainable development path.

In summary, the application of MMAI in IES is still in the exploration stage. However, through continuous research and breakthroughs at the four levels of model, data, paradigm and system, it is expected to achieve a leap from experimental verification to continuous deployment, thereby providing strong support for the intelligent and low-carbon transformation of future energy systems.

5. Conclusion

This paper focuses on the integration of the integrated energy system (IES) and multimodal artificial intelligence (MMAI), and conducts research by constructing a systematic "MMAI-IES integration framework". Through logical analysis and mechanism exploration of each layer including perception, cognition, decision-making, and control, it reveals the unique advantages of MMAI in achieving multi-source data fusion, precise state representation, and intelligent optimization decision-making. The study shows that MMAI has the ability to enhance the operational efficiency and stability of IES, and also possesses great potential to drive the energy system towards intelligent and low-carbon transformation.

In terms of theory, this article innovatively presents two concepts: Firstly, it systematically proposes the overall framework of the MMAI-IES integration, which compensates for the shortcomings of existing research that mostly focuses on a single modality or a single technology; Secondly, through mechanism analysis, it clarifies the role mechanism of multi-modal learning in complex energy systems, providing theoretical support and methodological references for future research.

However, this study is mainly based on literature review and theoretical deduction, and lacks large-scale experimental verification. Future research should further combine real energy system operation data to conduct framework validation and optimization application. At the same time, it is

DOI: 10.54254/2755-2721/2026.TJ29794

necessary to explore the adaptability and scalability of MMAI in multi-objective optimization, real-time control, and cross-system collaboration, in order to promote the intelligent development of integrated energy systems.

References

[1] Kanna. I V, Roseline S, Balamurugan K, et al. The Effects of Greenhouse Gas Emissions on Global Warming [A]. In: Encyclopedia of Renewable Energy, Sustainability and the Environment (First Edition) (Rahimpour MR, ed). Oxford: Elsevier, 2024: 143–154.
[2] Zhang S, Wang K, Xu W, et al. Policy recommendations for the zero energy building promotion towards carbon neutral in Asia-Pacific Region [J]. Energy Policy, 2021, 159: 112661.
[3] Zhang Z, Fedorovich KS. Distributed robust cooperative optimization of multi-integrated energy systems based on variational inequality-driven non-cooperative game theory [J]. Applied Energy, 2025, 401: 126696.
[4] Hua H, Du C, Chen X, et al. Optimal dispatch of multiple interconnected-integrated energy systems considering multi-energy interaction and aggregated demand response for multiple stakeholders [J]. Applied Energy, 2024, 376: 124256.
[5] Dong X, Wu J, Hao J, et al. Stochastic optimal coordination of hydrogen-enabled zero-carbon integrated energy systems in buildings [J]. International Journal of Hydrogen Energy, 2025, 111: 1–11.
[6] Ma H, Sun Q, Chen L, et al. Cogeneration transition for energy system decarbonization: From basic to flexible and complementary multi-energy sources [J]. Renewable and Sustainable Energy Reviews, 2023, 187: 113709.
[7] Li J, Chen H, Qi Y, et al. Collaborative optimization for cross-regional integrated energy systems producing electricity-heat-hydrogen based on generalized Nash bargaining [J]. Energy, 2025, 333: 137444.
[8] Han G, You S, Ye T, et al. Analysis of combined cooling, heating, and power systems under a compromised electric–thermal load strategy [J]. Energy and Buildings, 2014, 84: 586–594.
[9] Jin K, Yu T, Grzybowski A. Multimodal artificial intelligence in ophthalmology: Applications, challenges, and future directions [J]. Survey of Ophthalmology, 2025.
[10] Baltrušaitis T, Ahuja C, Morency LP. Multimodal Machine Learning: A Survey and Taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423–443.
[11] Li D, Sun Y, Yuan Y, et al. A quantum-inspired medical scalable convolutional neural network for Intelligent pneumonia diagnosis [J]. Biomedical Signal Processing and Control, 2026, 112: 108440.
[12] Krishnamurthy S, Adewuyi OB, Luwaca E, et al. Artificial intelligence-based forecasting models for integrated energy system management planning: An exploration of the prospects for South Africa [J]. Energy Conversion and Management: X, 2024, 24: 100772.
[13] Hameed K, Maqsood F, Wang Z. Artificial intelligence-enhanced zero-knowledge proofs for privacy-preserving digital forensics in cloud environments [J]. Journal of Network and Computer Applications, 2025, 243: 104331.