Multimodal Large Language Models for Cross-Domain Knowledge Transfer and Integration in Intelligent Engineering Systems

Diansheng Yang
South China University of Technology

ABSTRACT

Multimode Wide Language Model (MLLM) is changing the intelligent engineering system. They allow unified reasoning of different data streams in a complex network physical environment. Based on system engineering theory, network physical system (CPS) principle and information theory, this review analyzes how MLLM can help interdisciplinary knowledge transfer. They use attention-based alignment, adjustment instructions and mixed integration with digital twins, reinforcement learning, computer vision and edge computing. Practical cases from semiconductor manufacturing, independent logistics, smart grid and infrastructure inspection show the progress of predictive maintenance, dynamic scheduling and real-time anomaly detection. Although MLLM reduces the separation between domains and accelerates adaptability, it still has important limitations. For example, there is a risk of illusion in the safety-critical control loop, and the computational load is sometimes too heavy for edge computing, so it is difficult to understand the physical world well. The initial contribution includes a multimodal information bottleneck framework for efficient transfer and elasticity index, and for measuring redundant knowledge paths in CPS. The review concludes that the mixed design of neural symbols and physical information, coupled with strict verification, is essential for reliable deployment. Future progress will require interdisciplinary progress in artificial intelligence, control theory and dependency engineering.

Keywords: Multimodal Large Language Models, Cross-Domain Knowledge Transfer, Cyber-Physical Systems, Intelligent Engineering Systems, Digital Twins, Knowledge Integration, Reinforcement Learning, Edge Computing.

1. INTRODUCTION

Intelligent engineering systems, such as intelligent factories, autonomous transportation networks and elastic energy infrastructure, are highly connected network physical systems (CPS); In these systems, physical dynamics, computer control and human monitoring always interact. These systems will generate heterogeneous multi-modal data, such as high-dimensional visual inspection, time series sensor readings, text maintenance logs and operator instructions, but it is difficult for traditional domain-specific models to integrate these data globally. System engineering emphasizes the consistency of the whole life cycle and the management of emerging behaviors, while information theory measures the effectiveness of knowledge flow between different modes. This makes cross-domain transfer and semantic integration easy without lengthy manual feature engineering work.

Recent models, such as LLaVA, show us how to connect pre-trained visual encoders to the backbone of large language models for general visual reasoning and follow instructions after specific adjustments [1]. These advances are directly dependent on contrastive pretraining, which aligns visual and textual representations on a large scale [2]. These functions are very useful in engineering environments, where operators need to interact with complex system models in natural language and must quickly transfer the knowledge of past simulations or operations to new configurations or product variants.

From the perspective of information theory, effective transmission needs to maximize the mutual information between mission-important modes and reduce the irrelevant noise from sensors. This principle extends the classic information bottleneck to multi-mode CPS environment. In addition, system engineering also requires the integration mechanism to support verification, traceability and flexibility, which cannot be guaranteed by pure statistical model. MLLM can be used as approximate solvers of these integration problems, when they are strengthened by appropriate constraints and verification layers.

Two representative examples illustrate practical problems. In semiconductor wafer manufacturing, MLLM is a classic CPS application, which combines optical and electron microscope images analyzed by computer vision, time series of multivariable process parameters and text report of equipment status to diagnose performance degradation and suggest corrective measures. The knowledge transferred from previous generations of processes speeds up the root cause identification of new equipment nodes. In urban traffic management, MLLM integrates video streams processed by real-time computer vision for event detection, text reports of vehicles and historical pattern data to optimize fire time and improve throughput and safety margin.

These benefits are accompanied by material risks. Illusion suggestion may exceed the actuator limit or thermodynamic constraint, and the computational intensity of large-scale model conflicts with the delay and energy budget of edge nodes distributed in CPS. Therefore, this review maintains a critical balance, reviews the theoretical basis, licensing technology, application fields and open challenges, and puts forward targeted extensions to the existing framework.

2. THEORETICAL FOUNDATIONS: MLLMS AS ENABLERS OF MULTIMODAL INFORMATION INTEGRATION IN CYBER-PHYSICAL SYSTEMS

CPS theory models the engineering system as a combination of physical plants, computing controllers and communication networks. Their interaction has produced new behaviors. The traditional information fusion in CPS uses the integration strategy sooner or later. When there are many different patterns or heterogeneous semantics, these strategies can't work well. MLLM overcomes these limitations through the transformational attention mechanism. These mechanisms dynamically align visual, text and time series feature in the shared token space, which is helpful for higher-order perception and reasoning.

Information theory provides the basis of quantification. The adaptive ability of a small number of shots allows the rapid embedding of knowledge in specific fields with little additional data [3]. The comprehensive study of MLLM development and architecture also explains how these mechanisms match the system engineering goals, such as reusing models in V model and reusing knowledge in the whole life cycle [4].

A new idea we put forward here is the concept of cross-domain knowledge reduction. In CPS, the apparent uncertainty comes from the incomplete causal model of the interaction between subsystems (for example, mechanical degradation that affects the stability of electrical control). MLLM, which is pre-trained on various multimodal corpora, implicitly reduces this uncertainty by discovering common physical principles expressed through visual patterns, sensor signatures and text descriptions. During the fine adjustment, the resulting reduction can be monitored to measure the transfer efficiency, thus providing a systematic supplement to the classical Shannon measurement.

A concrete example is the deployment of an autonomous fleet of aircraft for inspection of linear infrastructure. MLLM integrates visual telemetry of transformation visual processing, text history of tasks and maintenance, and time series of vibration or temperature to detect and detect cultural anomalies, such as corrosion. The transition from simulation to reality is realized by matching the output of digital twins with physical observations, which can be extended to different types of assets, such as bridges, pipelines and railways. Computer vision provides accurate spatial positioning, while language components link discovery with intervention history.

There is a parallel case in smart grid fault isolation. MLLM integrates the time series of phase measurement unit, operator's text annotation and infrared image of inspection drone to locate interference and provide recovery switching sequence. During extreme weather events, cross-domain migration from one network topology to another can improve resilience without completely retraining.

However, the statistical representation of MLLM learning does not naturally apply physical consistency or traceability. Excessive reliance on the results generated without redundant verification may lead to cascading failures in closed-loop control; The black box characteristics of attention distribution make the formal verification and verification process required by system engineering standards more complicated. These reflections inspire hybrid architectures that integrate domain constraints and physical simulators, such as verifying Oracle Bone Inscriptions.

3. ARCHITECTURES AND KEY TECHNOLOGIES FOR CROSS-DOMAIN KNOWLEDGE TRANSFER

The actual inter-domain transfer in MLLM depends on modular architecture. These architectures include an encoder specific to each mode, an alignment module and a basic language model that adapts instructions. Visual deformers extract features from images or videos; Specialized sequence encoders process time series or text logs. Lightweight projection layer or query converter converts these features into LLM token vocabulary. Then, cross-attention allows the context to be exchanged in two directions; The adjustment statement matches the behavior of the model with the specifications of the engineering task. At the same time, efficient parameterization methods, such as low-level adaptation, allow targeted updates without forgetting general functions.

The integration with digital twins provides a concrete mechanism for real-time synchronization of virtual and physical states [5]. The improved tuning instruction technology further improves the alignment between the model output and the engineering objectives in a specific field. This supports more reliable transfer between production variants or operating environments [6].

For dynamic scheduling in flexible manufacturing, reinforcement learning strategy benefits from the world model or simulation reward generated by MLLM. MLLM absorbs multi-modal state descriptions, such as visual observation of the workshop through computer vision, text production orders and machine state from sensors, and generates structured representations. These representations provide information for policy optimization. This provides real-time reordering, which is better than the static solver in case of interruption.

For online quality control, embedded computer vision can not only classify defects, but also explain them in natural language, and associate them with possible reasons found in historical logs. Edge computing deployment, through evaporation and quantification, puts the lightweight multimode components on the industrial gateway, and infers in less than one second, so as to adjust the closed-loop robot through vision and force feedback.

The hierarchical multimodal transfer framework is frozen by the perceptual alignment layer trained by objective comparison, so as to keep the solid characteristics applicable to all fields. The higher reasoning layer is trained by the special instructions of engineering corpus (CAD renderer with tracking simulation and maintaining ontology). This separation preserves the representation including physics, while allowing easy adaptation.

Advantages include fewer bespoke templates and better dialogue interface for human-computer collaboration. But the risk is great. Hallucinating actuator commands may disrupt the secure docking. The energy demand of large-scale aircraft makes it difficult for edge CPS to achieve the sustainable development goal; The transfer distribution between general preprocessed data and proprietary operation distribution may destroy loyalty transfer. Therefore, before using them, we must use high-fidelity physical simulator to benchmark and formally check the generated plans.

4. APPLICATIONS IN INTELLIGENT ENGINEERING SYSTEMS

Digital Twins is the main integration support for MLLM engineering applications. In intelligent manufacturing, constantly updated twins reflect physical assets through multi-mode flow. MLLM allows you to ask natural language questions in Twin and create explanations for optimization operations from reinforcement learning strategies. Knowledge transfer between production lines or product lines accelerates the growth of processes and reduces the demand for scarce experts.

The advanced model extends these functions to complex engineering problems of mixed vision, text and program input [7]. In automatic logistics and warehouse operation, MLLM combines computer vision based on camera, which is used for object recognition and posture estimation, text instructions of warehouse management system and time series telemetry data of self-driving vehicles. The resulting unified state indicates that dynamic task allocation and collision-free routing are allowed. Edge deployment meets strict delay requirements, while analog-real transmission is realized by multi-modal domain randomization technology.

Infrastructure construction is another example. MLLM observes images of unmanned aerial vehicles or fixed cameras, written inspection reports and time series of sound or stress to see whether the structure is healthy. Cross-domain transfer allows the model first trained on bridge data to adapt to tunnels or dams, using common degradation signs in different types of information.

These deployments have brought about real improvements in prediction accuracy, action speed and collaboration between humans and artificial intelligence. However, there are also limitations: possible results require security filters, privacy risks of images and texts, and possible deviations, because training data cannot fully display certain contexts or fault types. Therefore, human beings must always be vigilant in important situations.

5. CHALLENGES, LIMITATIONS, AND FUTURE PROSPECTS

Many obstacles hinder the reliable use of MLLM in intelligent engineering systems. Illusion and physical grounding are the most important problems: seemingly correct results without respect for physics may lead to dangerous control behavior, and its influence will spread to the CPS feedback loop. In order to solve this problem, we need to use a physical simulator to verify the results, and create a mixture of neural symbols to mix neurogenesis with formal logic from engineering tools.

Scalability and real-time performance bring other problems. The complete model uses too much calculation and energy for distributed edge nodes in CPS. Dilution and quantization are helpful, but they can't eliminate the compromise of reasoning depth and cross-modal alignment accuracy. For closed-loop control, hierarchical routing is needed, simple sensing tasks are given to lightweight and dedicated models, and complex reasoning is reserved for more powerful but slower components.

Data scarcity, heterogeneity and privacy make transmission more difficult. Engineering data sets are usually private, with unbalanced classes and not many multimodal annotations. Joint learning and continuous learning provide a way to share knowledge among organizations, while protecting privacy and dealing with exchange and distribution.

Interpretability and authentication have not been resolved. Note that weight is not easy to achieve the traceability required by functional safety standards. The future progress will depend on the multi-modal architecture and benchmarks of natural interpretation, which add physical stress satisfaction scores to precision measurement.

One of the original contributions of system theory is multimode elastic index. This index quantifies the improvement of overall reliability by redundant knowledge path realized through MLLM, and measures the soft degradation when a single sensor fails or attacks reversely. This index extends the classical concept of reliability engineering, and can help design trade-offs and certification. As emphasized in the basic literature on CPS, these quantitative frameworks are crucial for the development from prototype demonstration to certified industrial deployment [8].

Future directions include closer integration with neural networks provided by physics and basic robot models, standardization of multi-model data patterns in the framework of digital twins, and regulatory guidelines for the use of artificial intelligence in safety-critical projects. Continuous interdisciplinary cooperation among artificial intelligence, control theory and reliability engineering are essential.

6. CONCLUSION

Large-scale multimodal language model provides a powerful mechanism for transferring and integrating knowledge between different fields of intelligent engineering system. Based on the principles of system engineering, network physical system theory and information theory, MLLM unifies perception and reasoning through vision, language and time series, thus making knowledge flow from simulation to physical equipment and between different engineering fields. Technologies such as digital twins, reinforcement learning for adaptive planning, computer vision for real-time detection and edge computing for fast computing are enhancing these capabilities, as we have seen in manufacturing, logistics and infrastructure.

Strict examination shows that great advantages in adaptability and efficiency are accompanied by great risks related to illusion, computational cost and physical consistency. The viewpoint of multimodal information bottleneck and elasticity index provided by us provides a mixed design and rigorous verification tool for solving these problems. In order to use MLLM reliably in industry, it is necessary to continue to improve the efficient architecture, standardized multimodal protocol and interdisciplinary verification framework. When designed responsibly, MLLM can be a good knowledge intermediary and help to create a more flexible, efficient and people-oriented engineering system.

REFERENCES

[1] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In Advances in Neural Information Processing Systems, 36, 34892-34916. doi:10.48550/arXiv.2304.08485

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139, pp. 8748-8763). PMLR. doi:10.48550/arXiv.2103.00020

[3] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., et al. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716-23736. doi:10.48550/arXiv.2204.14198

[4] Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023). Multimodal large language models: A survey. In 2023 IEEE International Conference on Big Data (Big Data) (pp. 2247-2256). IEEE. doi:10.1109/BigData59044.2023.10386743

[5] Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), 2405-2415. doi:10.1109/TII.2018.2873186

[6] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 26296-26306). doi:10.48550/arXiv.2310.03744

[7] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. doi:10.48550/arXiv.2303.08774

[8] Khaitan, S. K., & McCalley, J. D. (2015). Design techniques and applications of cyberphysical systems: A survey. IEEE Systems Journal, 9(2), 350-365. doi:10.1109/JSYST.2014.2322503