2025 NSF/ASME Student Design Essay Competition

Toward 2040: A Modular Vision-Language Framework for Adaptive, Multimodal Intelligence in Industrial Inspection

Zuxin Dai
University of Georgia
Athens, GA 30602
Email: zuxin.dai@uga.edu
Phone number: 706 380 2320

Mentor:
Dr.Beshoy Morkos
University of Georgia

Category: Graduate Level

Toward 2040: A Modular Vision-Language Framework for Adaptive, Multimodal Intelligence in Industrial Inspection

Abstract

The manufacturing industry is transitioning toward Industry 5.0, which emphasizes sustainability, human-centric design, and adaptability of systems. Conventional inspection methods are inadequate in addressing the growing complexity and customization challenges facing today's production lines. Thus, we propose a modular vision-language framework that consists of three main layers: (1) a perception layer that incorporates real-time CNN-based object detection, (2) a refinement layer that includes statistical, classical ML, temporal, and lightweight LLM methods to tune the post-detection outputs, and (3) an interaction layer driven by multimodal LLMs (MM-LLMs) for dynamic coordination and intuitive human-AI collaboration.

We note that our architecture will enable real-time perception and interpretable decision-making and allow for scalable implementation in various manufacturing contexts. We provide a phased implementation roadmap and discuss the system's capability for continual learning, feedback integration, and edge adaptation. Given that it aligns with Industry 5.0 principles, this framework presents a feasible pathway toward intelligent, modular, and human-aligned inspection systems for the factories of 2040.

Keywords: Industrial inspection, Industry 5.0, multimodal large language models, machine learning, continual learning, human-AI collaboration.

1. Introduction

The manufacturing paradigm of the 21st century is undergoing a significant change. We are now moving into Industry 5.0—a vision emphasizing sustainability, human-centric design, and resilience [1] —-and it has become evident that traditional approaches to quality inspection and automation are no longer adequate. The increasing complexity of products and demand for customization, together with the need for high-quality, productive manufacturing in distributed global factories, emphasizes the increased necessity for intelligent inspection systems that can not only detect defects but also adapt, learn, and collaborate seamlessly with human operators.

Recent advances in deep learning have transformed the domain of industrial inspection. Convolutional neural networks (CNNs)[2], particularly YOLO-style models [3, 4, 5, 6, 7], have facilitated real-time object detection and defect localization in a variety of manufacturing sectors. These models excel in high-speed visual recognition tasks but often lack adaptability, explainability, and support for user-specific criteria without retraining.

To close this gap, various lightweight techniques may serve practical utility in the post-detection fine-tuning task. For example, statistical techniques such as sliding-window smoothing[8] or majority voting offer some noise stabilization at negligible computational cost. Classical machine learning (ML) approaches, including Support Vector Machines (SVMs)[9] and Random Forests (RF)[10], may also be appropriate approaches to enhancing detection refinement, given that structured features can be extracted from the perception layer. Furthermore, inspection tasks are known to have strong temporal characteristics—for instance, conveyor-based production or continuous monitoring can produce downstream video frame-wise outputs, and there is an expectation of meaningful sequential patterns that exist within these process frames. In these cases, temporal models like Recurrent Neural Networks (RNNs)[11], Long Short-Term Memory networks (LSTMs)[12], and Transformers[13] are well-suited for modeling contextual dependencies and improving decision consistency over time.

Recently, the emergence of large language models (LLMs)[14] has opened up new possibilities for interpretability, reasoning, and multimodal interaction. LLMs can generate explanations, interpret human intent, and support complex decision-making by leveraging pre-trained semantic knowledge. Building on top of this, multimodal large language models (MM-LLMs) [15, 16, 17] have extended these capabilities by integrating inputs across vision[18, 19], language, audio/speech[20, 21], and video[22] to support more natural and flexible human-AI collaboration. Recently, these have been advanced to any-to-any MM-LLMs [23, 24], which support both multimodal inputs and outputs within a unified framework.

Despite these advances, current industrial applications often lack modularity, adaptability, and user-centered design. To address this challenge, we introduce a modular vision-language framework for future industrial inspection systems. Our framework consists of three integrated layers: (1) a perception layer that conducts real-time detections with CNN-based models; (2) a refinement layer supporting post-processing with statistical, ML-based, temporal, and lightweight LLM methods; and (3) an interaction layer, exploiting fine-tuned LLMs for multimodal reasoning and system coordination.

This modularity allows manufacturing companies to selectively deploy, scale, and adapt intelligent inspection systems to changing product lines, environmental conditions, and customer needs. Our framework also aligns state-of-the-art AI technologies with future-oriented industry principles associated with Industry 5.0 for the development of intelligent, explainable, and human-aligned manufacturing over the next decade.

2. Vision for the Design and Manufacturing Enterprise in 2040

As a company that operates globally in manufacturing and strives to be a technology leader, we recognize that success will no longer be based on the amount of production in 2040. Instead, companies that will ultimately gain a competitive advantage will demonstrate high degrees of intelligence, flexibility, and human-machine collaboration in their product and design systems and production processes.

More specifically, we anticipate that inspection—the foundation of product quality and trust—will change fundamentally. Static, discrete product quality control checks will transition into intelligent inspection systems capable of understanding context, reasoning across inputs, and continuous learning.

As we think about the future, our company should begin working on these fundamental capabilities:

- **Strategic Flexibility:** Leading organizations will adaptively reconfigure inspection workflows and quality criteria based on changed customer needs, regulations, or market conditions without lead time delays or system reengineering.

- **Empowered Workforce through Human-AI Collaboration:** All personnel will be able to engage with intelligent systems—not just experts in disciplines or skilled trades—in non-technical interfaces such as natural language, gestures, marked-up images, or sensor input. LLM-powered assistants will translate this input to actionable operations, improve speed, and reduce training requirements, thereby streamlining decision-making.

- **Modularity and Cost-Efficient Intelligence:** Next-generation systems will require a modular architecture for their AI components—especially the components responsible for refining models, getting feedback on implementation and performance, and getting feedback from users through interfaces—so that these components can be upgraded quickly and placed into use quickly—all, in order to minimize downtime and maximize reusability.

- **Institutionalized Knowledge and Reasoning Support:** With the increase in workforce transitions, long-term competitiveness will begin to rely on being able to embed expert logic into the AI models we create. LLMs are just another form of reasoning engine; they can capture and preserve (and scale) institutional reasoning across shifts, sites, teams, and generations.

- **Transparency and Explainability:** Regulatory compliance and customer trust will begin demanding more than simply whether the decision is ”accurate” but rather whether it is an ”explainable” decision. Organizations will need to design and adopt systems that are able to provide and explain—through interpretable and verifiable justification—a clear rationale for every one of their quality ruling.
- **Continuous Learning and Operational Agility:** Organizations that are able to leverage operator feedback, rectify mistakes, and learn new defect types as they happen will sustain a compounding advantage—transforming reactive quality control into proactive quality control.
- **Scalable Deployment Across Global Operations:** Systems that are ready for the future will perform reliably in prescribed and dynamic environments—from centralized hubs to resource-constrained remote locations—while leveraging intelligence that works at the edge and a globally consistent standard.

To conclude, the successful high-tech enterprise of 2040 will not compete solely on automation but on its ability to create intelligent, modular, interpretable, and human-aligned inspection ecosystems. In this way, these enterprises will turn quality control into a strategic advantage, improved responsiveness, risk reduction, and scalable excellence across the enterprise, no matter where it operates globally.

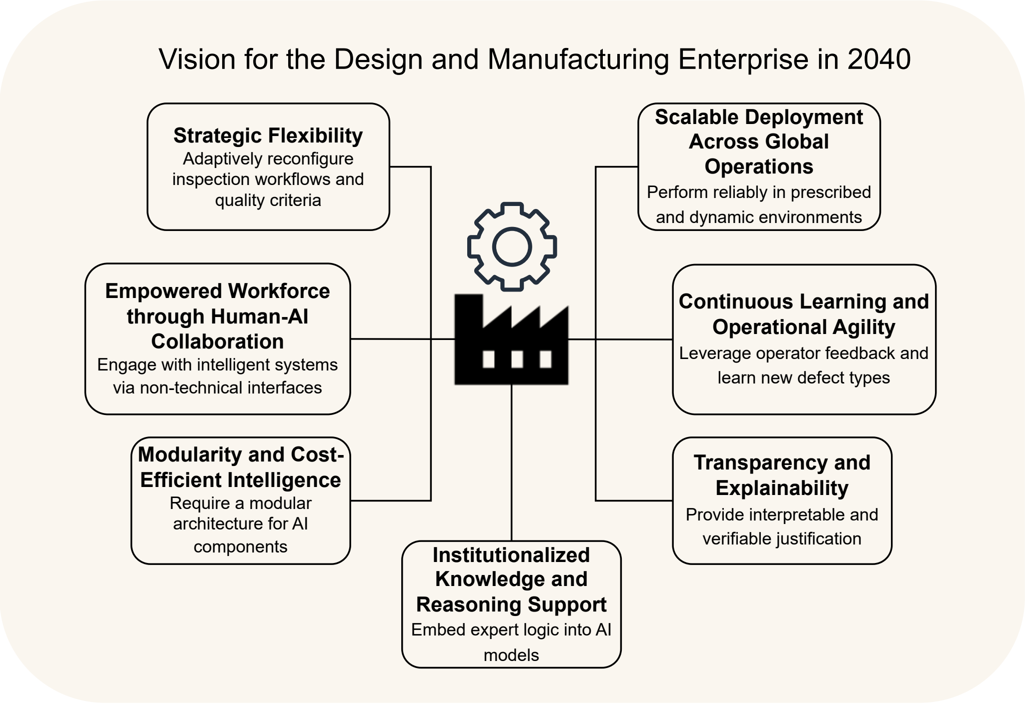
Vision for the Design and Manufacturing Enterprise in 2040
Strategic Flexibility
Adaptively reconfigure inspection workflows and quality criteria
Empowered Workforce through Human-AI Collaboration
Engage with intelligent systems via non-technical interfaces
Modularity and Cost-Efficient Intelligence
Require a modular architecture for AI components
Institutionalized Knowledge and Reasoning Support
Embed expert logic into AI models
Scalable Deployment Across Global Operations
Perform reliably in prescribed and dynamic environments
Continuous Learning and Operational Agility
Leverage operator feedback and learn new defect types
Transparency and Explainability
Provide interpretable and verifiable justification

Figure 1: Vision for the Design and Manufacturing Enterprise in 2040.

3. Core Challenges

Aligned with the compelling competitive vision for manufacturing companies in 2040, we must rethink “beyond” a set of fundamental challenges that have existed well before algorithmic-based advancements. These challenges are associated with system integration, organizational readiness, and human-machine cooperation and must be addressed to re-conceptualize today’s factories as adaptable, intelligent, multimodal enterprises.

- **Translating Human Intent into Actionable System Behavior:** Natural language interactions hold the potential of providing greater accessibility, but current systems have difficulty translating vague or context-rich instructions (e.g., “focus more on surface blemishes” or “use stricter criteria on this batch”) into actionable operations. Once enterprises begin to adopt multimodal LLMs, the challenge is not only mapping language but also translating visual cues, sketches, or sensor feedback into a coherent response from the system.
- **Fragmented Multimodal Inputs without Unified Understanding:** Factories generate a plethora of data, including defect images, sensor streams, maintenance logs, and spoken comments. These types of data are typically processed by AI systems independently of each other. Synthesizing multiple types of data into a single contextually relevant understanding of the factory requires more advanced multimodal reasoning and time-synchronized data pipelines, both of which are hardly commonplace in industrial settings.
- **Lack of Modularity in Current AI Infrastructures:** Many industrial AI implementations are developed as monolithic, single-function pipelines. In the absence of a modular architecture—where vision models, refinement logic, LLM-based interactions, and feedback loops can be easily added to or upgraded—companies will suffer from high integration costs, as well as limited adaptability to new products or client needs.
- **Limited Support for Fine-Grained Post-Deployment Adaptation:** While the vast majority of industrial AI systems are trained on data that is proprietary or specific to a product, they are often not highly flexible once they are implemented and deployed. Adapting to changes in customer tolerance, threshold levels for defects, or even different production conditions, especially in high-mix environments, often necessitates supporting returning or retraining the model from scratch. Supporting ongoing customization either through lightweight layers of refinement or user-influenced adjustment (potentially through natural language or example feedback) is an inadequately addressed function for almost all current systems.
- **Poor Interpretability and Trustworthiness:** As AI decisions become more autonomous, their explainability becomes essential for adoption. Yet most AI outputs remain opaque to factory personnel. Multimodal LLMs, while promising in generating natural-language justifications, still struggle to ground those explanations in verifiable visual or process-based reasoning.
- **Underdeveloped Continuous Learning Pipelines:** As artificial intelligence autonomy increases in decision-making, the explainable nature of AI decisions becomes paramount for many to be accepted and adopted. Yet, the vast majority of AI outcome results remain unclear to factory workers. Multimodal large language models—although promising in generating natural-language justifications—continue to struggle with providing justifications that are grounded in visual or process-based reasoning.
- **Integration Barriers Across Legacy Systems and Resource Constraints:** Organizations work in infrastructure that is extremely heterogeneous—from new cloud systems to old embedded controllers.

Deploying complex multimodal systems like LLMs will require lightweight edge-compatible models and solid APIs for cross-system communication. This represents an integration problem, and most tools available today are not equipped to solve it.

- **Organizational Resistance and Workforce Skill Gaps:** Even when faced with technically advanced solutions, adoption will continue to depend on acceptance by the workforce. LLMs and AI copilots may seem abstract or non-credible to those with no prior experience. Organizations will need to overcome inertia from cultural skepticism, as well as the absence of pathways to train non-experts to interact meaningfully with intelligent capabilities.

In combination, these challenges demonstrate that realizing the vision of human-centric, adaptive, and multimodal manufacturing is not simply a technical problem; it is a shift in the modes of structuring, surfacing, and integrating intelligence throughout the industrial enterprise.

4. Proposed Framework

In support of the strategic priorities of adaptability, interpretability, and competitive agility in industrial inspection, we propose a vision-language modular framework for human-aligned quality control that allows for dynamic adaptability and reliability. This architecture can benefit a broad assortment of manufacturing sectors, including electronics, automotive, pharmaceuticals, and consumer goods.

The framework is organized into three interconnected layers, with each addressing a set of respective operational challenges:

- **Perception Layer: Real-Time Visual Inspection**
  The core of this layer is a high-speed object detection module, such as a YOLO-style convolutional neural network, which detects product characteristics, defects, or anomalies reasonably in real-time using visual data. These models are trained on domain-specific datasets and output the raw predictions required for subsequent reasoning and decision-making.

- **Refinement Layer: Customizable and Modular Adaptation**
  Situated at the intersection of perception and final decision-making, the refinement layer provides the mechanism for finely tuning model outputs to meet product-specific tolerances, operational limits, and shifting customer-based specifications. This layer functions as a transitional logic layer to balance, modify, or contextualize predictions made by the primary detection model.

  This removable layer uses a variety of refinement approaches that cater to various data types, deployment conditions, and customization requirements:

  - **Statistical methods**, such as sliding-window smoothing and majority voting yield rapid, interpretable and low-complexity answers to noise reduction and temporal stabilization—appropriate for elementary tasks and resource-constrained environments.
  - **Classical machine learning models**, including Support Vector Machines and Random Forests, are typically well suited to environments where features are manually designed or structured as an output of the perception layer. They provide low-latency inference with satisfactory control over tuning parameters.
  - **Temporal sequence models**, such as Recurrent Neural Networks (RNNs) and Transformers, are effective in cases where outputs that were generated framewise have temporal correlations. Video-based inspection systems or fixed conveyor-line data streams are prime examples of technologies that lend themselves to sequence modeling since the context of prior frames will improve the accuracy of predictions and/or classifications.

  - **Lightweight large language models (LLMs)**, which were refined for tasks unrelated to language, may present some distinct alternatives for augmenting scenarios with structural or rule-based cases of defect detection, such as finding vertical cracks, parallel patterns, or patch-specific failure. While their use for the role of post-processing remains an early area of research, there is potential for these models to encode complex spatial or semantic logic that might be difficult to extract from using a standard model.

  Due to the diversity of the industrial context, some workflows are based on real-time sequential data, while others are based on static images, metadata, or sensor fusion data. The refinement layer is intended to support both paradigms, enabling flexible integration of multiple models or switching between multiple models, depending on the production context and performance requirements of the process.

  All elements in the refinement layer are modular and independent, meaning they can be updated, partially implemented quickly, and continually improved without retraining the core perception engine.

- **Interaction Layer: Multimodal Reasoning and Adaptive Coordination**
  At the apex of the framework resides a fine-tuned large language model (LLM), which serves as a multimodal reasoning and coordination layer that facilitates intuitive, high-level interaction with the system. In contrast to a traditional command interface, this layer supports a variety of input modalities; not only does it accept natural language prompts from operators (e.g., "inspect more thoroughly around the edges"), but it also accepts visual examples (e.g., annotated images of defects), sensory data (e.g., trends in vibration or temperature), and tabled production records.

  It assimilates these operator inputs with additional inputs generated internally from the detection layer and refinement layer, enabling the LLM to reason holistically about system behavior, production context, and inspection goals. Using this unified interface, users can issue vague or context-dependent commands, provide visual or quantitative input, ask for explanations, or ask about performance patterns.

  The LLM processes and analyzes these multimodal inputs to alter parameters based on the situation, including inspection thresholds, machine settings, or enhancement strategies for inspections. It can also consult experts and domain knowledge databases to suggest a process, provide assistance to less-experienced staff, and substantiate quality control decisions.

  In addition, the interaction layer has the ability to generate automated quality reports, summarize inspection results, highlight possible deviations, or recommend process changes upstream- thereby completing the loop between human intent, model behavior, and decision support at the enterprise level.

Continual Adaptation Loop: The system includes capturing operator feedback and identifying uncertain or edge-case predictions. These could be used to retrain refinement models (or even prompt updating the instruction mappings of LLMs)—creating a self-improving loop that continuously improves robustness over time while not requiring frequent manual reconfigurations.

Deployment Scalability: Modular components enable organizations to deploy only those layers appropriate for their use case—e.g., a lightweight single-edge variant for rapid classification or a full-stack version that integrates human interaction and continual learning for high-value inspections.

Overall, this framework provides a scalable and extensible approach to modern industrial inspection. It creates a loop between perception, reasoning, and human interaction while remaining light, interpretable, and adaptable to changing production.

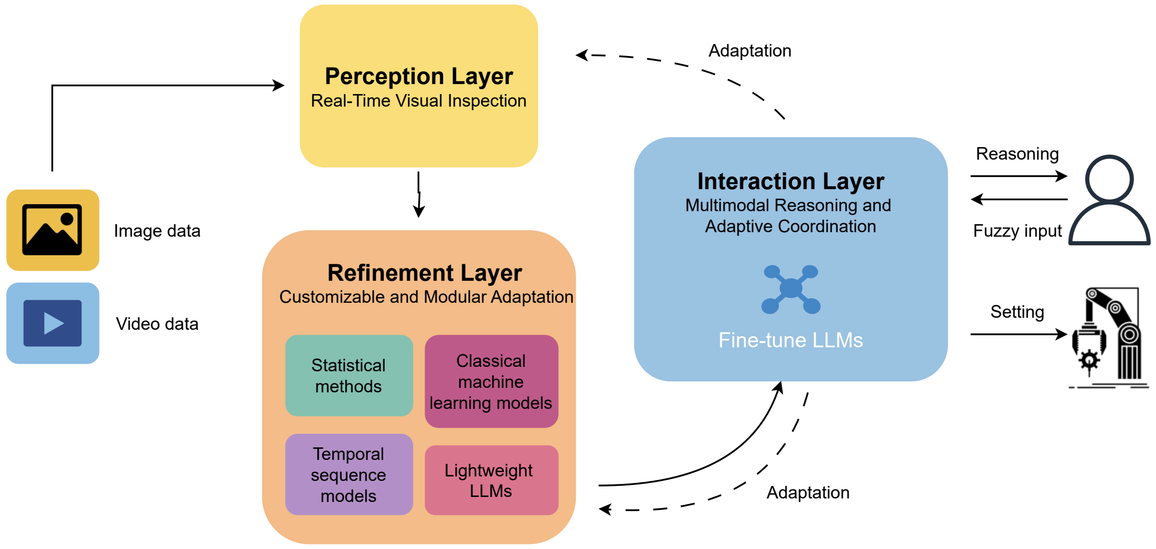

Figure 2: Proposed modular vision-language framework for adaptive, multimodal industrial inspection.

5. Implementation Roadmap

In order to actualize the intended framework in a manufacturing setting, we delineate a phased implementation roadmap related to incremental value delivery, scalable modularity, and integration with the existing manufacturing infrastructure.

1. Establishing Core Perception Infrastructure:
The first stage is the implementation of domain-specific object detection models to enable visual inspection in real time. This requires curated datasets that exemplify target defects, sufficient calibration of cameras to allow for consistent image collection across physical environmental deployments, and initial installation of YOLO-style architectures or similar object detection models. This first stage results in the perception of target classes that can be used for downstream refinement and interaction.

2. Integrating the Refinement Layer:
Once baseline perception has been established, refinement modules are built. These consist of simple statistical filters, classical classifiers (e.g., SVM, RF), and temporal models that are lightweight, such as RNNs or Transformers. The modules can be deployed in an iterative, scenario-driven approach. This allows for the refinement strategy to be applied to performance across different product lines or for differing customer quality specifications. Some Lightweight LLM-based refinement agents may be deployed in a selective manner where rule-based or spatially structured judgments are required.

3. Developing the Interaction Layer with Multimodal LLMs:
The next step is to build the interaction layer using fine-tuned LLMs that are able to reason over multiple modalities. This means combining language prompts, sensor streams, visual annotations, and structured logs into a unified command layer. Training or fine-tuning LLMs for intent classification, parameter tuning, and explanation generation in manufacturing is a key aspect of this work. This is also where we start integrating with knowledge bases.

4. Closing the Loop with Feedback and Continual Learning:
After initial implementation, feedback capture mechanisms are incorporated into the workflow. Operators have the flexibility to flag incorrect detections, signal a correction, or change their inspection behavior

through interaction. The operator has options to cancel or indicate that an inspection observation has been correlated with the detection observation and then log that information. These feedback marks are logged, verified, and used to update both the refinement models and LLM decisions in small incremental steps. This sets up a cycle of continuous improvement based on real-world operational activity.

5. Scalable Deployment and Edge Adaptation:
Finally, the architecture is modularized for scale. Organizations can select lighter-weight variants for edge deployment where bandwidth is limited or latencies are sensitive, or install full-architecture to specialized hubs for centralized quality control. Standard APIs and hardware abstraction layers are wholly portable across heterogeneous factory systems. Evaluation metrics include detection accuracy, false positive reduction, user satisfaction, and mean time to adaptation.

This staged roadmap will ensure the theoretical framework is developed in conjunction with operational maturity whereby a practical detection system is developed initially, followed by increasingly intelligent, adaptive, and interpretable industrial quality systems.

6. Conclusion

The successful development of the manufacturing sector into the future and towards 2040 will rely on an overall organization's capacity to adapt, learn, and collaborate at the boundaries of human and machine abilities. This paper addressed a proactive direction for intelligent, human-centered industrial systems that will focus not only on speed and precision but also on explainability, adaptability, and continuous development.

To achieve this vision, we proposed a modular vision-language framework to support tasks in real-time perception, customizable refinement, and multimodal interaction. This system combines state-of-the-art visual detectors with flexible finalization strategies and language-enabled coordination elements into a scalable framework for inspection tasks in manufacturing contexts.

The architecture also enables development with operator-in-the-loop control, domain-specific teams or adaptations, and integration with knowledge bases that connect the frontline and system-level intelligence bases. Including lightweight refiners and reasoning agents using fine-tuned LLMs expands the framework's ability to incorporate adaptations and changes in standard products, along with uncertainty and variation.

The proposed approach will advance a new generation of intelligent manufacturing systems that are automated, adaptive, interpretable, and user-centered. Future work should expand considerations for feedback and uncertainty-driven learning effects, examine real-world deployments in heterogeneous environments, and explore wider capabilities for multimodal interactions.

References

[1] M. Breque, L. De Nul, A. Petridis *et al.*, “Industry 5.0: towards a sustainable, human-centric and resilient european industry,” *Luxembourg, LU: European Commission, Directorate-General for Research and Innovation*, vol. 46, 2021.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 7464–7475. [Online]. Available: https://ieeexplore.ieee.org/document/10204762/

[5] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information,” Feb. 2024, arXiv:2402.13616. [Online]. Available: http://arxiv.org/abs/2402.13616

[6] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, “YOLOv10: Real-Time End-to-End Object Detection,” May 2024, arXiv:2405.14458. [Online]. Available: http://arxiv.org/abs/2405.14458

[7] Y. Tian, Q. Ye, and D. Doermann, “YOLOv12: Attention-Centric Real-Time Object Detectors,” Feb. 2025, arXiv:2502.12524 [cs]. [Online]. Available: http://arxiv.org/abs/2502.12524

[8] C.-S. J. Chu, “Time series segmentation: A sliding window approach,” *Information Sciences*, vol. 85, no. 1-3, pp. 147–173, 1995.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, pp. 273–297, 1995.

[10] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

[11] J. L. Elman, “Finding structure in time,” *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[15] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, “Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” in *International conference on machine learning*. PMLR, 2022, pp. 23 318–23 340.

[16] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[17] J. OpenAI Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report. arxiv," *arXiv preprint arXiv:2303.08774*, 2023.

[18] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu *et al.*, "Llava-plus: Learning to use tools for creating multimodal agents," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–142.

[19] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[20] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[21] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[22] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[23] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "NExT-GPT: Any-to-Any Multimodal LLM," Jun. 2024. [Online]. Available: https://openreview.net/forum?id=NZQkumsNlf

[24] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal, "CoDi-2: In-Context Interleaved and Interactive Any-to-Any Generation," 2024, pp. 27 425–27 434. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Tang$_{C}oDi - 2_{I}n - Context_{I}nterleaved_{a}nd_{I}nteractive_{A}ny - to - Any_{G}eneration_{C}VPR_{2}024_{p}aper.html$

Appendix: Language Editing Tools

To improve the clarity and readability of this essay, I used language assistance tools during the writing process. Specifically:

- **ChatGPT-4o (OpenAI):** Assisted in refining sentence structure, enhancing clarity, and ensuring coherence in technical descriptions.
- **Grammarly:** Used for checking grammar, punctuation, and stylistic consistency.

These tools were employed solely for language editing and did not influence the originality, technical content, or analytical contributions of the work.