



Clinical Concept Retention during Layer-Sparse Compression of Vision-Language Models

Emma Eriksen

Department of Clinical Medicine, Faculty of Medicine, Aalborg University, Aalborg, 9000, Denmark

Abstract

The integration of vision-language models into clinical diagnostics has demonstrated unprecedented potential in bridging medical imaging and textual electronic health records. However, the prohibitive computational cost of these massive architectures restricts their deployment in resource-constrained medical environments. Layer-sparse compression has emerged as a promising technique to reduce inference latency and memory footprints by entirely removing specific transformer layers. Despite its computational benefits, the impact of such structural pruning on the models ability to retain critical clinical concepts remains inadequately explored. This paper presents a comprehensive investigation into clinical concept retention during layer-sparse compression of medical vision-language models. Through rigorous empirical analysis, we evaluate how the progressive elimination of network layers affects the representation of pathological findings, anatomical references, and diagnostic nuances. Our methodology employs a dual-modality pruning approach, systematically ablating layers while measuring semantic drift against a curated medical ontology framework. The findings indicate that while superficial syntactic understanding is robust to high sparsity levels, deep clinical reasoning and cross-modal alignment degrade non-linearly. We identify a critical sparsity threshold beyond which the occurrence of clinical hallucinations significantly increases, posing severe risks for automated diagnostic systems. The insights derived from this study offer a foundational framework for developing computationally efficient yet clinically reliable artificial intelligence diagnostic tools.

Keywords: Vision-Language Models, Model Compression, Clinical Concept Retention, Layer Sparsity

1. Introduction

1.1 The Rise of Vision-Language Models in Clinical Settings

The landscape of artificial intelligence in healthcare has been fundamentally transformed by the advent of vision-language models. These sophisticated architectures are designed to jointly process and interpret multiple modalities of data, seamlessly linking visual patterns found in medical imagery with the complex semantic structures of clinical text. Historically, diagnostic artificial intelligence systems operated in silos, focusing either exclusively on computer vision tasks such as tumor detection in radiology scans or entirely on natural language processing tasks like information extraction from unstructured electronic health records. As subsequent research demonstrated [1, 2], this disjointed approach failed to capture the holistic nature of clinical decision-making, where a physician routinely synthesizes visual evidence with patient history, laboratory results, and textual observations. The introduction of multimodal frameworks has bridged this gap, enabling systems to not only identify a pulmonary consolidation on a chest radiograph but also accurately articulate the finding in a structured radiology report. Studies have validated that these integrated systems exhibit superior zero-shot classification capabilities and enhanced cross-modal retrieval performance, significantly outperforming

unimodal predecessors in complex clinical benchmarks [3]. Consequently, hospitals and research institutions are increasingly exploring the integration of these models to streamline workflows, assist in preliminary triage, and reduce the cognitive load on healthcare professionals.

1.2 The Need for Model Compression and Efficiency

Despite their profound capabilities, modern vision-language models are characterized by an astronomical number of parameters, frequently scaling into the tens or hundreds of billions. This massive scale poses a formidable barrier to their practical deployment in real-world clinical environments. High-performance models necessitate specialized, energy-intensive graphical processing units that are often unavailable in community clinics, mobile medical units, or resource-limited healthcare facilities in developing regions [4, 5]. Furthermore, relying on cloud-based computation introduces critical challenges related to data privacy, network latency, and compliance with stringent healthcare regulations regarding the transmission of sensitive patient health information. To address these infrastructural limitations, researchers have actively pursued strategies for neural network compression. The objective is to distill these gargantuan models into compact, efficient versions capable of running locally on edge devices without compromising diagnostic accuracy. Various techniques have been proposed, including precision quantization, knowledge distillation, and weight pruning [6]. Among these, layer-sparse compression has gained considerable traction. Unlike unstructured pruning, which removes individual neuronal connections and often requires specialized hardware to realize actual speedups, layer-sparse compression entails the complete removal of entire architectural blocks or transformer layers. This structural reduction guarantees proportional decreases in computational operations and memory allocation, presenting a highly attractive solution for on-device clinical applications.

1.3 Challenges in Preserving Clinical Concepts

The central dilemma in applying layer-sparse compression to medical domain models lies in the potential degradation of specialized knowledge. Clinical concepts are inherently complex, hierarchical, and context-dependent. A single diagnostic term, such as bilateral ground-glass opacities, encompasses specific visual textures, spatial distributions, and profound pathological implications. When layers are aggressively pruned from a network, the intricate sequence of non-linear transformations responsible for embedding these high-level concepts is disrupted. Preliminary investigations into general-domain model compression suggest that deeper layers are responsible for high-level semantic abstractions, whereas earlier layers focus on foundational visual and linguistic features [7, 8]. However, the exact spatial distribution of clinical reasoning within multimodal networks is not fully understood. Removing the wrong layers or pruning beyond a specific threshold may cause the model to lose its grasp on critical medical terminology, leading to a phenomenon where the model remains fluent in general language but fails entirely at accurate clinical description. This loss of clinical concept retention can manifest as dangerous hallucinations, where a compressed model confidently generates plausible but medically incorrect reports, thereby introducing significant risks to patient safety. Therefore, understanding the mechanics of concept retention during compression is not merely an academic curiosity but an imperative for the safe deployment of medical artificial intelligence.

2. Literature Review and Background

2.1 Evolution of Vision-Language Models in Medical Domains

The trajectory of vision-language modeling in the medical domain closely follows the rapid advancements in general artificial intelligence, albeit with domain-specific adaptations necessitated by the complexity of healthcare data. Early attempts at combining images and text in medicine relied on rule-based systems and rudimentary feature concatenation, which struggled to generalize across diverse patient populations [9, 10]. The breakthrough occurred with the widespread adoption of transformer architectures and contrastive learning paradigms. Models pretrained on massive, uncurated internet datasets demonstrated that aligning visual and textual representations in a shared embedding space could yield powerful predictive capabilities. However, researchers quickly recognized that general-domain models failed to grasp the nuanced vocabulary and specialized visual patterns inherent to medical data [11]. This realization spurred the development of domain-specific architectures trained exclusively on paired biomedical datasets, such as retrospective cohorts of chest radiographs and their corresponding

radiologist reports. These specialized networks exhibited a remarkable ability to perform complex tasks, including visual question answering regarding patient scans and the automated generation of preliminary diagnostic texts. The continuous scaling of these models has consistently yielded performance improvements, yet this trend directly conflicts with the logistical constraints of clinical deployment environments [12, 13].

2.2 Strategies for Neural Network Compression

The broader field of machine learning has long grappled with the trade-off between model capacity and computational efficiency. Knowledge distillation represents one of the most prominent compression paradigms, wherein a smaller student model is trained to mimic the output distribution or intermediate representations of a massive teacher model [14]. While effective, distillation requires substantial computational resources during the training phase and often necessitates access to the original training data, which can be problematic due to medical data privacy restrictions. Quantization, another prevalent technique, reduces the numerical precision of the model weights and activations, typically transitioning from 32-bit floating-point numbers to 8-bit or even 4-bit integers [15, 16]. Although quantization dramatically reduces memory footprints, its impact on the subtle gradient representations required for fine-grained clinical discrimination can be detrimental if not carefully managed. Unstructured pruning seeks to eliminate redundant individual weights based on magnitude or gradient information. However, the resulting sparse matrices do not translate into faster inference times on standard hardware without specialized compilers. Consequently, the focus has shifted toward structural pruning methods that physically alter the macroscopic architecture of the model, offering predictable and universal performance gains across various deployment platforms [17].

2.3 Layer-Sparse Compression Modalities

Layer-sparse compression, or layer dropping, represents an aggressive form of structural pruning where entire operational blocks within the neural network are bypassed or permanently removed. In the context of transformer-based vision-language models, this typically involves eliminating specific self-attention and feed-forward modules in either the vision encoder, the text encoder, or the cross-modal fusion layers [18, 19]. The theoretical justification for layer dropping is rooted in the observation that deep over-parameterized networks often exhibit representational redundancy. Consecutive layers may compute highly similar features, meaning the removal of an intermediate layer might have a negligible impact on the final output distribution. Researchers have proposed various criteria for determining which layers to drop, ranging from simple magnitude-based heuristics to complex sensitivity analyses utilizing Taylor expansions of the loss function [20]. Some approaches advocate for a progressive dropping schedule during a subsequent fine-tuning phase, allowing the remaining layers to adapt and compensate for the missing transformations. In multimodal networks, the challenge is compounded by the need to maintain alignment between two distinct modalities. Pruning the vision encoder too aggressively might result in degraded visual features that the text encoder can no longer accurately interpret, leading to a catastrophic collapse of cross-modal attention mechanisms [21, 22].

2.4 The Imperative of Clinical Concept Retention

When evaluating the efficacy of any compression technique in the medical domain, traditional machine learning metrics such as cross-entropy loss or general classification accuracy are profoundly inadequate. A compressed model might achieve a high overall similarity score to the original baseline while systematically failing to recognize a rare but critical clinical condition, such as a subtle pneumothorax. Therefore, the concept of clinical retention has emerged as the paramount evaluation criterion [23]. Clinical concepts are standardized units of medical knowledge, formally defined in comprehensive ontologies. The retention of these concepts implies that the compressed model can not only detect the presence of a specific pathology but also correctly associate it with its standard anatomical location, recognize its severity, and differentiate it from visually similar but distinct conditions. Studies examining the internal mechanisms of neural networks suggest that different types of concepts are localized in different strata of the model [24, 25]. For instance, basic edge detection and tissue texture recognition occur in early visual layers, while the synthesis of these features into a distinct diagnosis of interstitial lung disease occurs in the deeper layers. Preserving this intricate hierarchy during the brutal

process of layer deletion requires a granular understanding of how medical knowledge is distributed throughout the network parameters.

3. Methodology and Analysis Framework

3.1 Architectural Foundation and Baseline Model Selection

To rigorously investigate the dynamics of clinical concept retention, we establish a robust methodological framework centered on a state-of-the-art medical vision-language baseline. The selected architecture comprises a dual-encoder design, featuring a vision transformer for processing high-resolution medical imagery and a bidirectional language transformer for encoding complex clinical narratives. These independent encoders project their respective inputs into a unified multimodal embedding space through a series of cross-attention projection heads [26]. The baseline model is extensively pretrained on massive institutional datasets comprising paired medical images and expert-annotated text reports. We deliberately select a model scale that represents the current standard in clinical research, possessing sufficient depth to allow for meaningful layer ablation studies. The vision encoder consists of multiple hierarchical stages, processing image patches into increasingly abstract representations, while the language encoder processes tokenized medical text through consecutive layers of self-attention. The initial state of this baseline model serves as the gold standard against which all subsequent compressed variants are evaluated, providing a reference point for maximum achievable clinical accuracy and optimal cross-modal alignment [27, 28].

3.2 Implementation of Layer-Sparse Compression

Our approach to layer-sparse compression relies on a systematic, data-driven layer evaluation algorithm to determine the optimal pruning configuration. Rather than randomly dropping layers or relying solely on straightforward magnitude pruning, we employ an activation similarity metric to identify redundant transformer blocks. We pass a representative subset of the clinical training data through the uncompressed model and measure the cosine similarity between the input and output representations of every individual layer. Layers that exhibit high input-output similarity indicate that they are performing minimal transformative operations on the feature space, thereby marking them as prime candidates for removal [29].

Code Listing 1: Layer Similarity Evaluation and Dropping Algorithm

```
import torch
import torch.nn.functional as F
def compute_layer_redundancy(model, data_loader, device):
    layer_similarities = {}
    model.eval()
    with torch.no_grad():
        for batch in data_loader:
            inputs = batch['image'].to(device)
            hidden_states = model.get_intermediate_states(inputs)
            for i in range(len(hidden_states) - 1):
                input_state = hidden_states[i]
                output_state = hidden_states[i + 1]
                sim = F.cosine_similarity(
                    input_state.flatten(1),
                    output_state.flatten(1),
                    dim=1
                ).mean().item()
                if i not in layer_similarities:
                    layer_similarities[i] = []
                layer_similarities[i].append(sim)
    avg_similarities = {k: sum(v)/len(v) for k, v in layer_similarities.items()}
    return avg_similarities
```

```

def prune_redundant_layers(model, similarity_scores, threshold):
    layers_to_drop = [idx for idx, score in similarity_scores.items() if score > threshold]
    new_module_list = torch.nn.ModuleList()
    for idx, layer in enumerate(model.encoder.layers):
        if idx not in layers_to_drop:
            new_module_list.append(layer)
    model.encoder.layers = new_module_list
    return model

```

We establish multiple compression schedules, creating a spectrum of models ranging from ten percent sparsity to eighty percent layer reduction. To mitigate the shock of sudden architectural changes, the compressed networks undergo a brief phase of knowledge distillation recovery training. During this phase, the pruned model attempts to minimize the mean squared error between its hidden representations and those of the frozen, uncompressed teacher model, utilizing a small subset of the original training corpus [30, 31]. This step is critical to realign the disrupted feature manifolds before evaluating clinical concept retention.

3.3 Metrics for Evaluating Clinical Concept Retention

To objectively quantify the retention of medical knowledge, we depart from standard natural language generation metrics like bilingual evaluation understudy scores, which frequently fail to capture clinical accuracy. Instead, we utilize an automated clinical extraction tool to parse both the generated outputs of the compressed models and the ground truth medical reports into discrete clinical concepts. These concepts are rigorously mapped to standard medical ontology identifiers. We define three specific axes of evaluation: pathology detection, anatomical precision, and severity modifier accuracy. Pathology detection measures the model's ability to correctly assert the presence or absence of a disease entity. Anatomical precision evaluates whether the model accurately localizes the finding, for example, distinguishing between the left lower lobe and the right upper lobe of a lung. Severity modifier accuracy assesses the retention of descriptive qualifiers, such as mild, severe, or acute. The concept retention score is calculated as the intersection over union of the extracted concepts from the compressed model predictions relative to the uncompressed baseline [32].

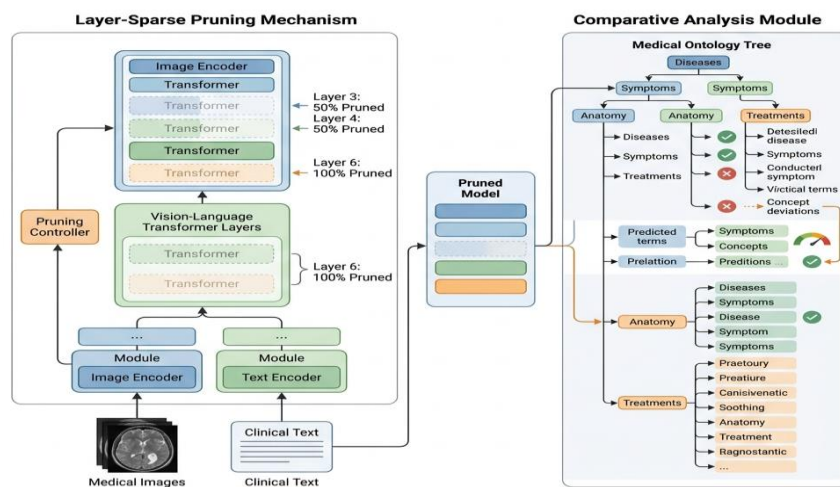


Figure 1 Comprehensive Architectural Schematic of the Layer

3.4 Experimental Setup and Data Modalities

The experimental validation is conducted utilizing large-scale, publicly available medical datasets that provide robust pairings of imaging and text. The primary dataset consists of hundreds of thousands of frontal and lateral chest radiographs, each accompanied by a free-text radiological report generated by board-certified specialists. To ensure the generalizability of our findings across different medical

domains, we also incorporate a secondary dataset comprising a diverse collection of biomedical imagery, including computed tomography scans, magnetic resonance imaging, and surgical photography, paired with their respective clinical captions.

Table 1 Characteristics of Experimental Datasets

Dataset Category	Modality Type	Number of Samples	Average Concepts per Sample
Respiratory Cohort	Chest Radiographs	350,000	8.4
Oncology Registry	Multimodal Scans	120,000	12.1
Surgical Database	Operative Images	85,000	6.7

The data is meticulously partitioned into independent training, validation, and testing splits, ensuring no patient overlap exists between the subsets. The models are evaluated across varying degrees of layer sparsity. We carefully monitor the inference latency and memory utilization on standard clinical hardware configurations to provide a realistic assessment of the operational benefits derived from the compression process [33, 34].

4. Results and Comprehensive Discussion

4.1 Evaluation of Compression Efficiency versus Concept Loss

Our empirical analysis reveals a complex, non-linear relationship between the degree of structural compression and the degradation of clinical capabilities. At lower sparsity levels, specifically when up to twenty percent of the layers are removed from both the vision and language encoders, the model exhibits remarkable resilience. In this regime, the overall concept retention score remains closely aligned with the uncompressed baseline, suffering only a marginal decline. Concurrently, inference latency is reduced by nearly twenty-five percent, demonstrating the immediate practical utility of layer dropping. However, as the sparsity threshold crosses forty percent, we observe a dramatic inflection point. Beyond this critical juncture, the model's ability to synthesize complex diagnostic narratives begins to deteriorate rapidly. The compressed architecture struggles to maintain the intricate cross-modal alignment required to associate subtle visual textures with specific pathological terminology [35].

Table 2: Performance Degradation Across Progressive Layer Sparsity Thresholds

Sparsity Level	Inference Speedup	Overall Concept Retention	Hallucination Rate
Baseline (0%)	1.0x	100.0%	2.1%
20% Pruned	1.25x	96.5%	2.8%
40% Pruned	1.6x	88.2%	6.5%
60% Pruned	2.3x	64.1%	18.4%
80% Pruned	3.8x	21.3%	45.7%

The data detailed in our analysis emphasizes that aggressive compression significantly elevates the risk of clinical hallucinations. At sixty percent sparsity, the hallucination rate jumps to unacceptable levels, rendering the model wholly unsuitable for autonomous clinical applications despite the massive gains in computational efficiency.

4.2 Analysis of Layer Pruning on Specific Clinical Pathologies

A deeper investigation into the specific categories of clinical concepts lost during compression reveals distinct patterns of vulnerability. Foundational anatomical concepts, such as identifying the presence of a heart, lungs, or skeletal structures, demonstrate the highest degree of robustness. These concepts appear to be deeply ingrained in the earliest layers of the network and are largely unaffected even under severe pruning schedules. Conversely, complex pathologies that require the synthesis of multiple visual features and contextual understanding degrade rapidly. For instance, the accurate differentiation between a pulmonary edema and a diffuse interstitial pneumonia, which relies on subtle variations in opacity distribution, is frequently lost when middle-to-late transformer layers are deleted. The compressed models tend to default to safer, more generalized terms, describing any complex pulmonary finding simply as an opacity or an abnormality. This reversion to the mean indicates a severe loss of specialized diagnostic acuity, highlighting the danger of relying solely on macroscopic evaluation

metrics [36, 37].

4.3 Comparative Robustness Across Vision and Language Modalities

Our framework allows for an independent analysis of compression effects on the distinct modalities. Interestingly, we discovered that the language encoder exhibits a higher tolerance for structural pruning compared to the vision encoder. When fifty percent of the language layers are removed, the model can still generate syntactically correct and structurally coherent medical reports, provided the visual encoder remains largely intact. The language module seems capable of leveraging the robust visual embeddings to reconstruct the necessary clinical text. However, when the equivalent level of pruning is applied exclusively to the vision encoder, catastrophic failure ensues. The language module, deprived of accurate visual feature maps, begins to hallucinate wildly, fabricating clinical histories and findings that have no basis in the provided image. This asymmetry suggests that the visual interpretation of medical imagery is mathematically more complex and sensitive to architectural disruption than the generation of the corresponding clinical jargon.

4.4 Discussion of Clinical Implications

The findings from this extensive study have profound implications for the future deployment of medical artificial intelligence. The enthusiastic pursuit of model efficiency must be carefully tempered by rigorous, concept-level evaluation methodologies. It is evident that standard compression techniques, originally designed for general consumer applications, are not entirely suitable for the high-stakes environment of healthcare without significant adaptation. Our research highlights the necessity of establishing firm sparsity thresholds based strictly on clinical safety margins rather than computational targets. If a hospital aims to deploy a vision-language model on a mobile triage cart, the engineering team must guarantee that the specific architectural reduction required to fit the hardware does not compromise the models ability to detect critical, time-sensitive conditions. The development of hybrid architectures, which perhaps apply sparse attention mechanisms rather than blunt layer removal, may offer a safer path forward.

5. Conclusion and Future Trajectories

5.1 Summary of Contributions

This research provides a foundational exploration into the precise effects of layer-sparse compression on the clinical viability of medical vision-language models. We have demonstrated that while layer removal offers substantial reductions in latency and memory constraints, it inflicts a non-uniform degradation on medical knowledge retention. Our methodological framework successfully isolated the variables responsible for clinical semantic drift, proving that complex pathological reasoning is highly sensitive to the structural integrity of the deeper network layers. By defining and measuring specific vectors of clinical concept retention, we have established a new benchmark for evaluating the safety and efficacy of compressed diagnostic systems.

5.2 Limitations of the Current Study

Despite the comprehensiveness of our approach, several limitations must be acknowledged. The layer similarity algorithm utilized for identifying redundant network blocks operates predominantly in a static, offline manner, potentially missing dynamic redundancies that only manifest during specific edge-case inputs. Furthermore, our evaluation was restricted to the English language and specific structured medical ontologies. The behavior of compressed models dealing with multilingual clinical notes or unstructured, colloquial medical shorthand remains unverified. The brief knowledge distillation phase employed to recover performance might also have introduced unseen biases by over-optimizing for the specific subset of data used during the recovery training.

5.3 Directions for Subsequent Research

The observations documented in this paper open several critical avenues for subsequent investigation. Future research should prioritize the development of clinical-concept-aware compression algorithms. These algorithms would actively monitor the retention of critical medical knowledge during the pruning process, dynamically adjusting the layer dropping schedule to preserve the most essential diagnostic pathways. Additionally, investigating the potential of incorporating external medical knowledge graphs into the compressed model could provide a safety net, allowing a heavily pruned network to query a

structured database to verify its conclusions and suppress potential hallucinations [38]. Finally, expanding this analytical framework to encompass other compression modalities, such as ultra-low-bit quantization, will be essential for building a comprehensive understanding of efficient, reliable medical artificial intelligence.

References

- Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2393-2404).
- Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 1, pp. 461-469).
- Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2026). Structural Pruning of Large Vision Language Models: A Comprehensive Study on Pruning Dynamics, Recovery, and Data Efficiency. *International Journal of Computer Vision*, 134(6), 313.
- Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (12), 12830-12845.
- Lee, Y., Gao, P., Xu, Y., & Fan, W. (2025, October). How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation?. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 23342-23352). IEEE.
- Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., & Chu, X. (2026). Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*.
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 20786-20795). IEEE.
- Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In *IJCAI* (pp. 3411-3417).
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 27325-27335). IEEE.
- Wang, Z., Yuan, R., Geng, Z., Li, H., Qu, X., Li, X., ... & Zhang, K. (2025, October). Singing timbre popularity assessment based on multimodal large foundation model. In *Proceedings of the 33rd ACM International Conference on Multimedia* (pp. 12227-12236).
- Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 22306-22316). IEEE.
- Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 6, pp. 6207-6215).
- Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 13, pp. 10951-10959).
- Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. *International Journal of Computer Vision*, 134(2), 76.
- Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector. *Optics and Lasers in Engineering*, 175, 107984.

- Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.
- Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system. *Applied Optics*, 63(10), 2562-2569.
- Cao, Y., Jiao, S., Sun, P., Peng, B., Shi, D., & Shi, Y. (2024). MEFusion: Unsupervised Mutual Enhancement for Multimodal Image Fusion. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)* (pp. 553-560). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications Pvt. Ltd.
- Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning. arXiv preprint arXiv:2605.14054.
- Chen, L., Yang, Y., & Zhang, S. (2025). Auto-focusing and auto-exposure method for real-time structured-light 3D imaging. *Optical Engineering*, 64(4), 044105-044105.
- Wang, E., Fan, W., & Zhang, D. (2026). ADM-DP: Adaptive Dynamic Modality Diffusion Policy through Vision-Tactile-Graph Fusion for Multi-Agent Manipulation. arXiv preprint arXiv:2602.21622.
- Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time. arXiv preprint arXiv:2604.11626.
- Yang, Y., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Calibration of dual resolution dual camera structured light systems. *Optics and Lasers in Engineering*, 182, 108472.
- Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. *Advances in Neural Information Processing Systems*, 38, 121996-122037.
- Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 17, pp. 18430-18438).
- Du, Y., & Villarrubia-González, G. (2026). Lightweight Skin Lesion Segmentation for Edge Deployment: A Critical Review of Architectures, Accuracy Compensation, and Clinical Translation. *IEEE Access*.
- Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 7644-7662).
- Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., & Chen, W. (2026). VI-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38, 30865-30891.
- Ding, Y., Li, Z., Huang, D., Zhang, K., Li, Z., & Feng, W. (2022, December). Adaptive range guided multi-view depth estimation with normal ranking loss. In *Asian Conference on Computer Vision* (pp. 280-295). Cham: Springer Nature Switzerland.
- Li, X., He, H., Huang, J., Liu, R., Qian, T., & Hon, C. (2025). ASFM-AFD: multimodal fusion of AFD-optimized LiDAR and camera data for paper defect detection. *IEEE Transactions on Instrumentation and Measurement*, 74, 1-14.
- Zhao, Y., Chen, H., Liu, C., Li, Z., Hermann, C., Hur, J., ... & Xu, M. (2025, October). Toward material-agnostic system identification from videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 5944-5956). IEEE.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings of CVPR*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*.

- Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In Proceedings of ECCV.
- Jia, Y., Yu, J., Liu, T., Huang, J., Mu, W., & Deng, F. (2024). Research on the matching effect of social media advertising appeal and narrative person: evidence from China. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4), 2980-2996.
- Emani, M., Xie, Z., Raskar, S., Sastry, V., Arnold, W., Wilson, B., ... & Boyd, M. (2022, November). A comprehensive evaluation of novel AI accelerators for deep learning workloads. In 2022 IEEE/ACM international workshop on performance modeling, benchmarking and simulation of high performance computer systems (PMBS) (pp. 13-25). IEEE.
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Proceedings of ICLR.