International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(5s) 2026 pp. 832-845
https://doi.org/10.70917/ijcisim-2026-2533
Article

Optimising Domain-Specific Neuron Activation for Efficient Multimodal Language Understanding In Cloud AI Systems

Austin Olom Ogar[1], Joshua Abah[2], Muhammad Aliyu Suleiman[3], Oluwatobi Noah Akande[4], Faruk Obansa Muhammed[5]

[1] Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.
Email: austinolomogar@gmail.com
[2] Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.
Email: joshua.abah@nileuniversity.edu.ng
[3] Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.
Email: muhammad.suleiman@nileuniversity.edu.ng
[4] Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.
Email: akande.oluwatobi@gmail.com
[5] Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.
Email: faruklincoln@gmail.com
Corresponding Author: Austin Olom Ogar, austinolomogar@gmail.com

Abstract: —Multimodal large language models (MLLMs) have made it possible for artificial-intelligence systems to reason jointly across vision and language, supporting tasks ranging from image captioning and visual question answering to clinical decision support and autonomous perception. As MLLM scale grows, however, deploying these models in cost- and energy-bounded cloud environments has become a defining engineering challenge. This mini-review consolidates recent literature at the intersection of four research strands: (i) multimodal architecture design and fusion strategies, (ii) neural-activation patterns and mechanistic specialization, (iii) selective and conditional computation including mixture-of-experts, and (iv) cloud-deployment optimisation. We propose a unified four-quadrant taxonomy of efficiency strategies, trace the field's evolution through a decade-scale timeline, and synthesise sixteen primary studies in cross-cutting comparison tables. Particular attention is given to the under-explored intersection of domain adaptation and efficiency, where evidence is converging that neuron-level domain awareness can simultaneously reduce inference cost and improve interpretability. We identify five persistent limitations of current methods and five concrete research gaps that follow from them. The review closes with an integrated future-research agenda built on three converging innovations: adaptive cross-modal attention re-weighting, knowledge-injection pathways, and sparse domain-conditioned neuron gating, and outlines the cloud-aware evaluation framework that would validate them. The article is intended as a single-source reference for researchers and practitioners designing efficient, trustworthy multimodal AI for cloud deployment.
Keywords: Multimodal large language models, domain adaptation, sparse activation, conditional computation, cloud inference optimisation, mixture-of-experts, mechanistic interpretability

1. Introduction

Multimodal language understanding represents a major advance in artificial intelligence, enabling models to ground linguistic reasoning in visual context by jointly processing images and text. Recent MLLMs such as InstructBLIP (Dai et al., 2023), Qwen2-VL (Wang et al., 2024) and the LLaVA family (Zhang et al., 2024) have demonstrated remarkable capabilities across image captioning, visual question answering, document understanding and embodied reasoning (Liang et al., 2024). The architectural and engineering arc that produced these capabilities

Copyright: © 2026 by the authors

has been characterised by a steady move from task-specific dual-encoder pipelines to unified transformer-based models trained on broad multimodal corpora (Figure 2).

Utilising MLLMs on a scale, however, raises three intertwined challenges. The first is computational efficiency: MLLM training and inference involve very large parameter counts and long input sequences (especially when image patches and text tokens are concatenated), giving rise to substantial latency, energy and cost burdens (Pope et al., 2025). The second is domain adaptation: a single generalist MLLM rarely transfers without loss across specialised domains such as biomedical imaging, legal documents or remote sensing (Zhai et al., 2023; Wang et al., 2024). The third is interpretability: as MLLMs move into safety-critical applications, stakeholders increasingly require transparent reasoning that supports auditing and regulatory approval (Martin et al., 2025).

Several recent reviews have surveyed multimodal architectures (Liang et al., 2024), parameter-efficient fine-tuning (Chen et al., 2024), model compression (Li et al., 2023) and mixture-of-experts (Zhu et al., 2024). These reviews are valuable but tend to optimise one axis in isolation. Most efficiency techniques are evaluated independently of domain adaptation; most domain-adaptation techniques are evaluated independently of cloud deployment; and few existing reviews integrate insights from mechanistic interpretability with deployment engineering. The reader is left without a single resource that maps these strands onto one another or that motivates the coming generation of cloud-aware, domain-specific efficiency frameworks.

This mini review consolidates the literature at the intersection of four strands. The contributions are: (i) a unified four-quadrant taxonomy of efficiency strategies for multimodal LLMs (Figure 1); (ii) a decade-scale timeline of architectural milestones (Figure 2); (iii) nine illustrative figures and four comparative tables synthesising primary evidence across domain adaptation, conditional computation, efficiency-oriented techniques and evaluation frameworks; (iv) a critical analysis of five persistent limitations and five concrete research gaps; and (v) an integrated research agenda built on three converging innovations.

Section 2 describes the review methodology. Section 3 surveys multimodal architectures and fusion strategies. Section 4 examines neural-activation patterns and the evidence for neuron specialisation. Section 5 covers domain-adaptation techniques. Section 6 introduces selective and conditional computation. Section 7 surveys efficiency-oriented compression and adaptive inference. Section 8 focuses on domain-specific neuron activation. Section 9 covers cloud deployment. Section 10 presents the evaluation framework. Section 11 offers a critical analysis of limitations and research gaps. Section 12 articulates a future-research agenda, and Section 13 concludes.

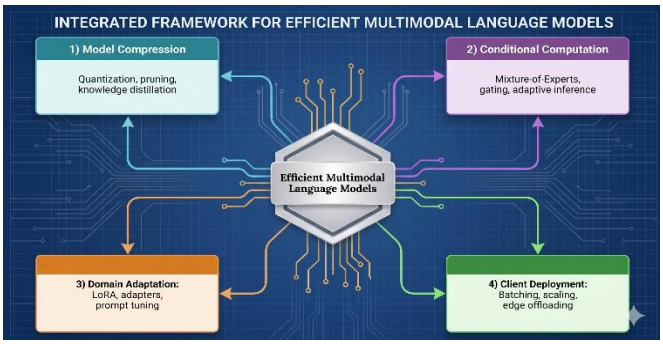

Fig. 1. A unified four-quadrant taxonomy of efficiency strategies for multimodal language models, mapped onto four cross-cutting evaluation axes.

2. REVIEW METHODOLOGY

A. Search strategy

The literature search covered the period 2015–2025 across IEEE Xplore, ACM Digital Library, arXiv and the proceedings of NeurIPS, ICML, ICLR, CVPR, ECCV and EMNLP. The search string was "(multimodal OR vision-language) AND (efficient OR sparse OR mixture-of-experts OR domain adaptation OR cloud deployment)". After de-duplication and relevance screening, 54 primary studies were retained for detailed review. We supplemented the database search with backward citation chasing on the eight most-cited papers in the set.

B. *Inclusion criteria*

Articles were included if they reported a multimodal language model with at least one explicit efficiency, domain-adaptation or interpretability evaluation. Single-modality NLP and single-modality vision papers were excluded except where they introduced foundational techniques (e.g. LoRA, MoE gating) that subsequently transferred to multimodal settings.

C. *Analytical framework*

We organised the literature along four axes: multimodal architecture, neural activation, domain adaptation and cloud deployment and overlaid four cross-cutting evaluation axes: accuracy, latency, energy and interpretability. The four-quadrant taxonomy (Figure 1) and the architectural timeline (Figure 2) are the conceptual anchors of the review.

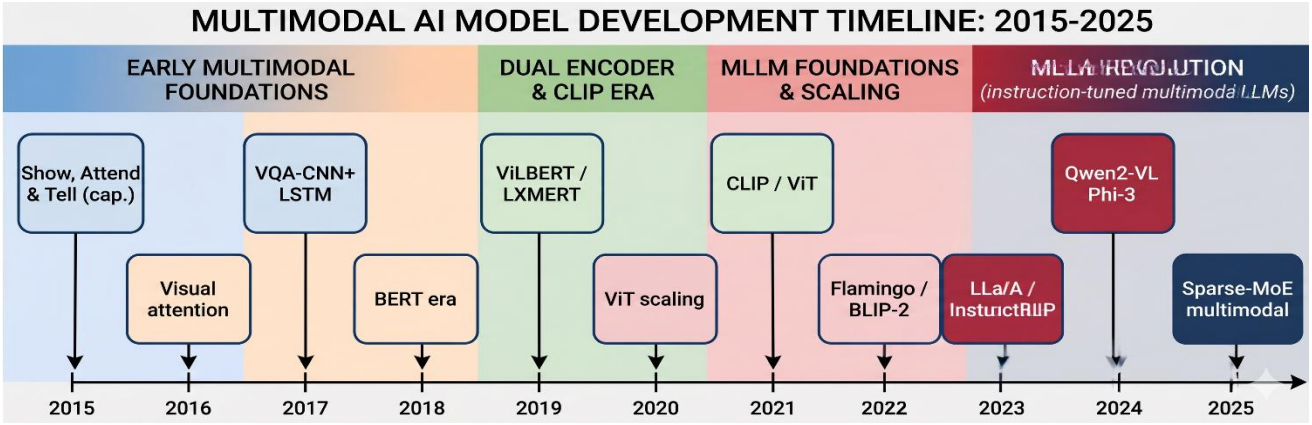

Fig. 2. Evolution of multimodal language model architecture from 2015 to 2025, highlighting the transition from CNN+LSTM captioners to instruction-tuned MLLMs.

3. MULTIMODAL LANGUAGE UNDERSTANDING IN MODERN AI SYSTEMS

A. *Architectural evolution*

The transition to unified MLLMs has evolved through identifiable stages. Early vision-language models (circa 2015–2020) used separate convolutional vision encoders and recurrent or transformer language encoders with a shallow fusion layer at the output (Jain et al., 2024). The Transformer revolution and the rise of contrastive image-text pre-training (CLIP, ViT) gave way, between 2021 and 2023, to instruction-tuned MLLMs that share a single language backbone and conditioning on visual tokens (Dai et al., 2023). The current generation of models (2024–2025) is characterised by mixture-of-experts variants, resolution-adaptive vision encoders and on-device sparse variants (Wang et al., 2024).

B. *Fusion strategies*

MLLMs differ in how and when visual and textual representations are combined. Three strategies dominate the literature (Figure 3). Early fusion concatenates visual embeddings with text tokens at the input layer; this maximises cross-modal interaction depth at the cost of attention quadratic in sequence length. Late fusion processes each modality through a dedicated encoder and merges representations near the output; this is computationally cheaper but limits multimodal interaction depth. Hybrid fusion adopted by InstructBLIP and a growing number of contemporary MLLMs uses modality-specific encoders for early layers and selective cross-attention in middle and later layers, providing a favourable accuracy-cost trade-off.

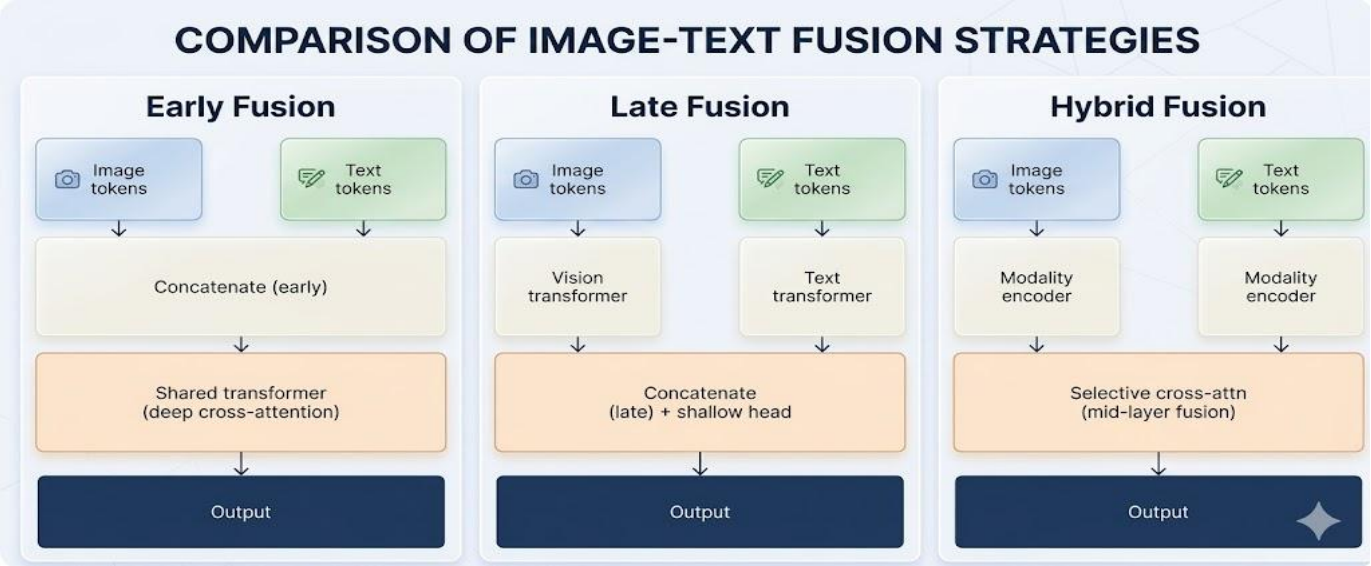

Fig. 3. Three fusion strategies for multimodal models. Early fusion maximises interaction depth; late fusion minimises cost; hybrid fusion compromises between the two.

C. Cross-attention mechanisms

A key innovation in MLLMs is cross-attention, where tokens from one modality attend to tokens of another. Cross-attention provides a principled mechanism for grounding language in visual context, but its compute and memory cost scales as O(NM) in the number of visual N and textual M tokens, a non-trivial overhead at production resolution.

D. Scaling challenges

As model size increases to billions of parameters and visual-input resolution increases, new bottlenecks emerge. High-resolution images produce many visual tokens which inflate the context length and slow inference (Pope et al., 2025). Domain-specific datasets are often small relative to the broader pre-training corpus, producing the well-documented catastrophic-forgetting effect when MLLMs are fine-tuned on narrow distributions (Zhai et al., 2023).

4. NEURAL ACTIVATION PATTERNS AND SPECIALISATION

A. Dense versus sparse activation

Biological brains use neurons sparsely: only a small group of neurons fires in response to a given stimulus. Artificial neural networks, by contrast, activate every neuron on every input. This dense regime is computationally wasteful when the input does not require the full representational capacity of the network. Recent work has accumulated evidence that artificial networks tolerate, and even benefit from, much sparser activation patterns than the dense default (Figure 4).

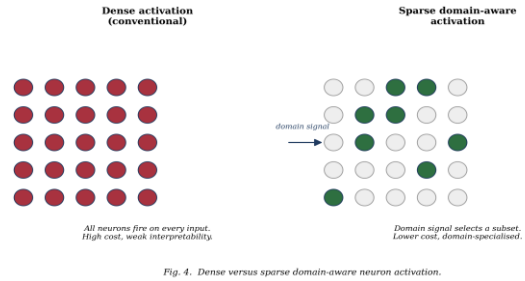

Fig. 4. Dense versus sparse domain-aware neuron activation.

Fig. 4. Dense versus sparse domain-aware neuron activation. The right panel illustrates how a domain signal can select a subset of neurons relevant to the input.

B. *Neuron specialisation and modularity*

Mechanistic interpretability studies have demonstrated that neurons in deep networks specialise functionally; some respond preferentially to specific concepts, syntactic structures or visual primitives (Martin et al., 2025). Multilingual large language models have been shown to contain language-specific neurons whose activation pattern correlates strongly with the language of the input (Zhao et al., 2024). Whether the analogous phenomenon, domain-specific neurons that activate preferentially for biomedical, legal or automotive content, exists in MLLMs is an open and tractable empirical question.

C. *Mechanistic interpretability methods*

Methods for identifying functional neurons include activation patching, feature attribution, the logit lens and probing classifiers. These techniques have made significant progress in text-only LLMs and are beginning to be applied to MLLMs (Liang et al., 2024). Their adoption in MLLM efficiency research is, however, still nascent.

5. DOMAIN ADAPTATION TECHNIQUES

A. *Full fine-tuning and catastrophic forgetting*

Supervised fine-tuning of all model parameters on a labelled domain dataset is the classical adaptation strategy. While effective for narrow benchmarks, it suffers from catastrophic forgetting: the model's pre-training knowledge is overwritten by the small domain dataset, degrading general capability (Zhai et al., 2023). Full fine-tuning is also expensive, requiring the full memory footprint of the model.

B. *Parameter-efficient fine-tuning*

To mitigate forgetting and reduce compute, researchers have introduced parameter-efficient methods including LoRA (Hu et al., 2022), adapter layers (Houlsby et al., 2019), prompt tuning (Lester et al., 2021) and Conv-Adapter (Chen et al., 2024). These approaches train only 0.1–5 % of model parameters, preserve the pre-trained weights, and have been shown to achieve task accuracy within 1–3 percentage points of full fine-tuning on a wide range of benchmarks (Figure 5).

Fig. 5. Trade-offs across parameter-efficient adaptation methods.

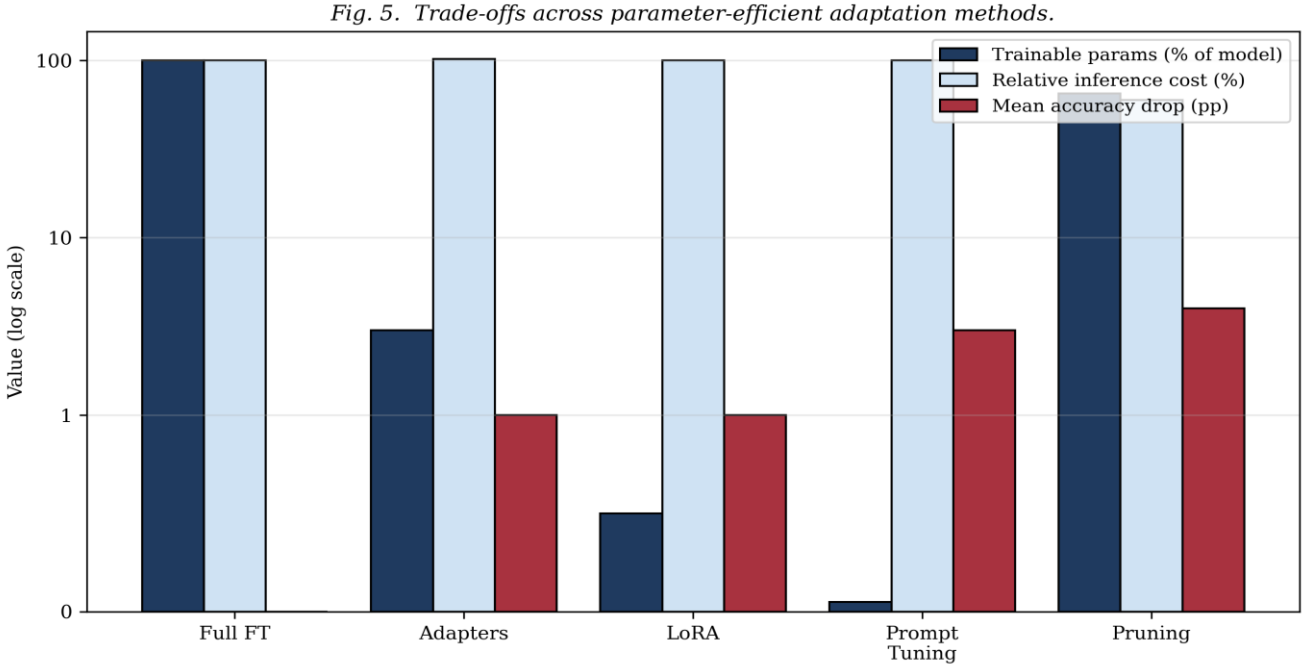

Fig. 5. Trade-offs across parameter-efficient adaptation methods on representative multimodal benchmarks.

C. *Domain-specific representation and modularity*

Evidence is converging that networks organise knowledge into modular, specialised subnetworks. Wang et al. (2024) introduce EfficientXpert, in which a domain-specific expert subnetwork is identified inside a generalist MLLM and selectively activated at inference. Soliman et al. (2025) propose GNN-MoE, a graph-neural-network-based context-aware router. These works strengthen the case that domain awareness can be integrated at the neuron and sub-network level rather than only at the loss function.

Table 1. Summary of parameter-efficient domain-adaptation techniques in multimodal large language models.

Method	Core idea	Trainable params	Inference cost	Strengths	Limitations	Refs
Full Fine-Tuning	Update all model parameters on domain data	≈100% of model	Very high	Strong adaptation	Catastrophic forgetting	Kirkpatrick et., al(2017)
Adapter Layers	Insert small bottleneck layers	≈1-5%	No reduction	Preserves base model	No neuron-level control	Houlsby (2019)
LoRA	Low-rank updates to attention	≈0.1-2%	No reduction	Memory-efficient training	Dense inference remains	Hu (2022)
Prompt Tuning	Learn soft prompts at input	<0.1%	No reduction	Simple; flexible	Limited expressiveness	Lester (2021)
Conv-Adapter	2D convolutional adapter blocks	≈1-3%	Minor overhead	Strong on vision tasks	Vision-only	Chen (2024)
Domain-Aware Pruning	Remove neurons irrelevant to domain	Reduced	Reduced	Improves efficiency	Often static	Wang (2024)
EfficientXpert	Selective expert sub-network	Reduced	Reduced	Domain-aware	Routing overhead	Wang (2024)

6. SELECTIVE AND CONDITIONAL COMPUTATION

A. *The mixture-of-experts paradigm*

Instead of activating all neurons for every input, conditional computation architectures route each input to a sparse subset of model components. The mixture-of-experts (MoE) paradigm is the dominant instantiation: a learned

gating network assigns each input token to k of N experts, where typically k « N (Figure 6). Switch Transformer (Fedus et al., 2022) and GShard (Lepikhin et al., 2021) demonstrated that MoE models can match or exceed dense counterparts while activating 2–5× fewer FLOPs per token.

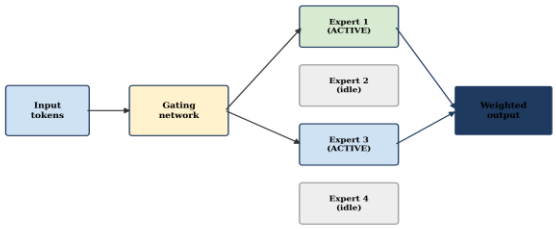

Fig. 6. Mixture-of-Experts: a gating network routes each token to a sparse subset of experts.

Fig. 6. Mixture-of-Experts: a gating network routes each token to a sparse subset of experts; only active experts contribute to the output, reducing per-token compute.

B. Gating networks and routing

Gating networks make per-token routing decisions and are typically implemented as small learned classifiers over the token representation. Routing decisions are made at every transformer layer, allowing different tokens to follow different computational paths through the network. Recent multimodal extensions include MoE Jetpack (Zhu et al., 2024) which converts dense vision-transformer checkpoints into adaptive MoEs, and GNN-MoE (Soliman et al., 2025) which uses graph neural networks to inform routing decisions in patch space.

C. Limitations of token-level routing

Despite their promise, MoE systems have well-documented limitations. The gating network itself adds 10–20 % overhead, partially offsetting expert sparsity savings (Zhu et al., 2024). Load imbalance during training causes some experts to dominate while others stagnate. Crucially, conventional MoE routing is domain-agnostic: routing decisions are based on token features alone, with no explicit signal about which domain the input belongs to. This is the central limitation that the next generation of conditionally computed MLLMs must address.

7. EFFICIENCY-ORIENTED TECHNIQUES

A. Model compression

Compression techniques reduce model size and inference cost without architectural change. Quantisation lowers numerical precision (Krishnamoorthi, 2019; Li & Gu, 2023) and can deliver 2–4× memory and energy savings with modest accuracy loss. Pruning removes redundant weights or neurons (Han et al., 2016). Knowledge distillation trains a smaller student model to mimic a larger teacher. All three techniques are typically applied uniformly across domains, ignoring the possibility that different domains require different compression strategies.

B. Sparse attention and low-rank approximations

Attention itself can be made more efficient. Sparse attention restricts each token to attend to a fixed-size local window or learned sparse pattern, reducing the quadratic cost. Low-rank approximations factorise the attention matrix into the product of two thin matrices, again reducing computation. Pope et al. (2025) survey contemporary sparse-attention variants for long-context multimodal inference.

C. Adaptive inference and early exit

Adaptive methods adjust computation per input. Early-exit networks add intermediate prediction heads and exit when an early head's prediction confidence exceeds a threshold. AdaLLaVA (Xu et al., 2025) adapts compute per multimodal request, exiting earlier on simple queries and continuing through deeper layers on complex ones. These methods are powerful but introduce control-flow overhead and require careful threshold calibration.

Table 2. Comparison of efficiency-oriented optimisation techniques for multimodal models.

Technique	Optimisation level	Primary benefit	Effect on accuracy	Domain awareness	Key limitation	Refs
Quantisation	Weight precision	Lower memory & energy	Minor drop	None	Hardware dependence	Krishnamoorthi (2019)
Structured Pruning	Neuron/layer removal	FLOPs reduction	Domain-dependent	Weak	May remove critical neurons	Han (2016)
Knowledge Distillation	Model-level	Smaller student	Reduced vs teacher	None	Requires retraining	Hinton (2015)
Sparse Attention	Attention matrix	Faster long-seq	Minimal	None	Token-level only	Pope (2025)
Mixture-of-Experts	Subnetwork routing	Decouples size and cost	Stable if balanced	Weak	Routing overhead	Muennighoff et., al (2025)
Adaptive Inference	Input-conditional	Saves compute	Minimal	Implicit	Control overhead	Xu (2025)
Low-Rank Approx.	Attention factorisation	FLOPs reduction	Minor	None	Less expressive	Pope (2025)

8. DOMAIN-SPECIFIC NEURON ACTIVATION STRATEGIES

A. Neuron masking and gating

Recognising the potential for explicit neuron-level selection, recent work has explored gating mechanisms that learn binary masks over neurons. Each neuron is associated with a learnable gate that turns it on or off based on the input or a domain signal. This is finer-grained than MoE routing, the gating is per-neuron rather than per-expert-block, and is in principle more expressive but training stability becomes a concern.

B. Expert subnetworks for domain specialisation

EfficientXpert (Wang et al., 2024) identifies domain-specific expert subnetworks within a generalist MLLM. At inference time, only the relevant subnetwork is activated, providing both efficiency and domain-specialisation gains. Conceptually related work in language-specific neurons (Zhao et al., 2024) demonstrates that this neuron-level specialisation emerges naturally during multilingual training and can be exploited at inference.

C. Contrast with parameter-efficient tuning

Methods like LoRA and adapters accelerate training and reduce trainable-parameter counts but do not, by themselves, reduce inference computation: at inference time the full set of frozen and adapter weights is still activated. Domain-aware neuron gating directly addresses inference cost in addition to adaptation cost.

Table 3. Domain-specific neuron-activation strategies reported in the literature.

Approach	Granularity	Activation mechanism	Adaptability	Interpretability	Efficiency gain	Refs
Static Pruning	Neuron	Permanent removal	Low	Medium	Moderate	Han (2016)

Task-Specific Subnetworks	Layer/block	Fixed subnet	Low	Medium	Moderate	Rusu (2016)
MoE Routing	Expert	Token-to-expert gate	Medium	Low	High	Soliman (2025)
Propagation-Aware Pruning	Path-level	Activation-flow pruning	Medium	Medium	High	Wang (2024)
Binary Neuron Masking	Neuron	Learned masks	High	High	High	Emerging
Language-Specific Neurons	Neuron	Naturally emerging	High	High	Variable	Zhao (2024)

9. CLOUD DEPLOYMENT OPTIMISATION

A. Cloud-tier constraints and service-level agreements

Cloud providers pack inference workloads onto large GPU clusters while meeting service-level agreements (SLAs) for latency, throughput and availability. State-of-the-art MLLM deployment uses several engineering techniques to make this packing efficient (Figure 7).

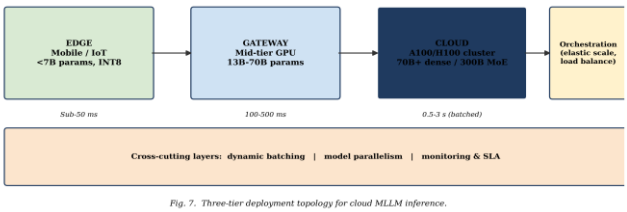

Fig. 7. Three-tier deployment topology for cloud MLLM inference.

Fig. 7. Three-tier deployment topology for cloud MLLM inference, spanning edge, gateway and cloud tiers with cross-cutting orchestration layers.

B. Dynamic batching and elastic scaling

Dynamic batching groups simultaneously arriving requests into batches to maximise GPU utilisation and amortise overhead across examples. Elastic scaling spins GPU instances up or down based on demand (Agarwal et al., 2024). Specialised hardware allocation assigns workloads to GPUs tuned for their profile; some GPUs are optimised for vision, others for language.

C. Edge and offloading paradigms

Serverless and edge paradigms offload lighter computations to the edge (closer to users). A model fragment may run on a smartphone or in a regional edge cache, with heavier computation forwarded to the cloud. Edge–cloud orchestration introduces new bandwidth, latency and privacy considerations that are now active research areas.

D. Runtime adaptation and load balancing

Real-time systems must adapt to fluctuating load and meet strict latency targets. Some frameworks tune model size, batch size or precision at runtime to maintain SLA conformance. Content-aware load balancing routes requests to the most appropriate model variant based on the request content rather than simple round-robin. Continuous monitoring of latency, accuracy and energy underpins these adaptive systems and feeds them with the signals needed for autonomic control.

10. EVALUATION FRAMEWORK FOR EFFICIENCY AND MULTIMODAL PERFORMANCE

A. Task-specific accuracy metrics

Evaluating MLLMs requires a mix of accuracy metrics reflecting the underlying tasks. For image captioning, BLEU and CIDEr (Rivera-Trigueros, 2022) are standard. For visual question answering, accuracy on the held-out test set is the primary metric. For document understanding, ANLS is dominant. These metrics measure functional correctness but say nothing about computational cost.

B. Efficiency metrics

Common efficiency metrics include latency (wall-clock inference time per request), throughput (requests served per second), energy consumption per inference (joules), and peak memory footprint. Computational cost is often reported as FLOPs, although FLOPs map imperfectly to wall-clock latency because of memory-bandwidth, cache and parallelism effects.

C. Cost-awareness and Pareto optimality

In cloud settings, minimising operational cost is usually the ultimate goal. Reducing FLOPs or peak memory is valuable only insofar as it translates into lower dollar cost per prediction. Pareto-optimal evaluation (Figure 8) plots accuracy against cost and identifies methods that are not dominated by any alternative.

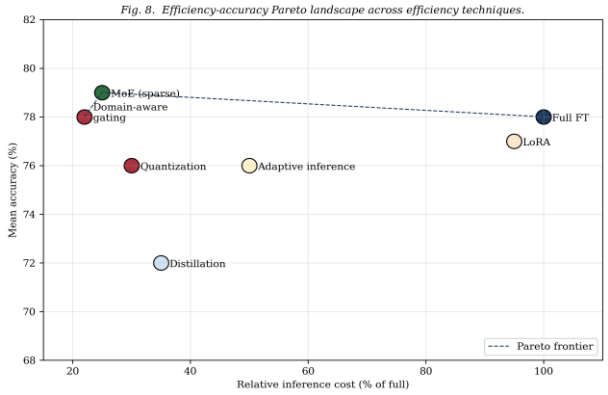

Fig. 8. Efficiency–accuracy Pareto landscape across efficiency techniques. Methods on the Pareto frontier are not dominated by any alternative in joint accuracy and cost.

D. Robustness and cross-domain generalisation

A crucial but sometimes overlooked dimension is robustness under domain shift. An efficiency-optimised model that drops 10 percentage points of accuracy on out-of-distribution data is not deployable in safety-critical settings. The Retention Ratio (Hendrycks et al., 2021) and related cross-domain transfer metrics quantify this dimension and should be standard alongside accuracy and efficiency.

Table 4. Recommended evaluation metrics for multimodal efficiency in cloud AI systems.

Category	Metric	Description	Relevance to cloud
Task accuracy	VQA accuracy	Correct answers in visual QA	Functional performance
Task accuracy	BLEU / CIDEr	Caption quality metrics	Text generation quality
Computational	Latency (ms)	Time per inference	SLA compliance
Computational	Throughput (req/s)	Requests served per second	Scalability

Energy	Energy per inference (J)	Power consumed per request	Sustainability
Cost	$ per 1k inferences	Operational cost	Total cost of ownership
Robustness	OOD accuracy / Retention Ratio	Accuracy under domain shift	Trustworthy deployment
Interpretability	Faithfulness / probing score	Quality of explanations	Auditability

11. CRITICAL ANALYSIS: LIMITATIONS AND RESEARCH GAPS

A. Five persistent limitations

Our review of the literature reveals five persistent limitations across the four research strands surveyed.

• Over-reliance on dense and global methods. Most efficiency techniques (quantisation, pruning, sparse attention) are applied uniformly across domains, ignoring the possibility that different domains require different compression strategies.

• Disconnect between adaptation and efficiency. Domain-adaptation and efficiency research have largely proceeded in parallel. Adaptation techniques rarely report inference-time efficiency; efficiency techniques rarely report domain transfer accuracy.

• Scalability and overhead of dynamic routing. MoE and adaptive-inference systems introduce deployment challenges including gating overhead, load imbalance and control-flow complexity that mitigate their theoretical efficiency advantage.

• Academic versus production gap. Most studies validate efficiency on static benchmarks with fixed compute budgets. Real cloud environments have mixed domains, variable batch sizes, and SLA pressure that academic benchmarks rarely model.

• Lack of interpretability focus. Efficiency papers typically optimise accuracy–cost trade-offs. In safety-critical domains, however, stakeholders also require transparent explanations of model behaviour, and few efficiency techniques are evaluated with respect to interpretability.

Five concrete research gaps

These limitations crystallise into five concrete research gaps that the next generation of work should address (Figure 9).

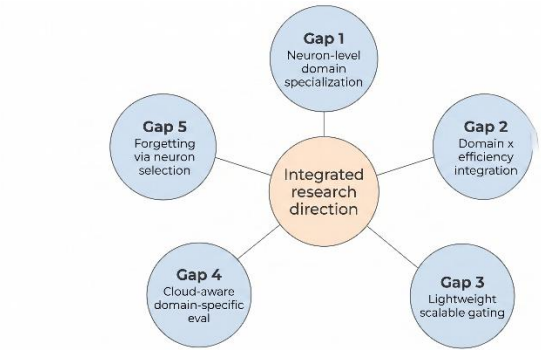

Fig. 9. Five research gaps identified by the review converge on a single integrated research direction.

• Gap 1: Neuron-level domain specialisation is unexplored. Although mechanistic interpretability has demonstrated functional specialisation of neurons, explicit domain-signalled neuron-level activation for MLLMs has not been systematically studied.

• Gap 2: Domain knowledge is not integrated into efficiency mechanisms. Current efficiency techniques (quantisation, sparse attention, MoE routing) are domain-agnostic; there is an opportunity to integrate domain signals into the routing function itself.

• Gap 3: Lightweight, scalable activation control is lacking. MoE gating overhead shows that lightweight per-neuron activation control is required for production scale.

• Gap 4: Cloud-aware evaluation of domain-specific efficiency is missing. Most evaluations focus on single-server benchmarks; we lack systematic evaluation of domain-specific efficiency under realistic cloud conditions (mixed domains, variable batches).

• Gap 5: Catastrophic forgetting via neuron selection is poorly understood. Parameter-efficient methods preserve weights to reduce forgetting, but the role of neuron-level selective activation in mitigating forgetting is largely uncharted territory.

12. FUTURE RESEARCH DIRECTIONS

An integrated research agenda

The literature points toward a converging research direction that fills these gaps by combining three complementary innovations:

• Adaptive cross-modal attention re-weighting. Domain-specific signals modulate cross-modal attention weights, letting the model focus on visual-textual interactions relevant to the domain while down-weighting irrelevant ones.

• Knowledge-injection pathways. Low-rank or sparse adapter modules inject new domain knowledge without overwriting pre-trained weights, reducing catastrophic forgetting while keeping the trainable parameter count low.

• Sparse domain-conditioned neuron gating. Learned gating mechanisms (e.g. binary masks) selectively activate only the neurons identified as domain-relevant, directly cutting inference compute and improving interpretability.

This integrated direction brings together four key scientific research areas: multimodal architecture design, mechanistic interpretability, cloud systems and domain adaptation. It takes the best from each area to develop a new generation of multimodal large language models capable of operating with limited cloud resources, geared toward specific domains such as healthcare and law, transparent enough for safety-critical use, and easily adaptable across deployment scenarios.

A. Open challenges

Implementing this agenda will require advances on several fronts. First, designing ultra-lightweight neuron-gating modules that add negligible overhead while supporting sparse binary masks at fine granularity. Second, methods to identify domain-relevant neurons on the fly, perhaps through mechanistic interpretability techniques applied at inference. Third, evaluation of the approach on multimodal benchmarks across diverse domains, including medical, legal, autonomous and remote-sensing applications. Fourth, testing under realistic cloud conditions with variable batches, mixed domains and multi-GPU setups. Fifth, analysing the interpretability and safety implications of neuron-level control, which become critical in safety-regulated deployment contexts.

13. CONCLUSION

Multimodal large language models are a major technological achievement, enabling machines to reason across vision and language at unprecedented scale. However, as MLLMs grow ever larger, their deployment faces critical challenges. The energy and dollar cost of huge models threaten sustainability and accessibility. Domain specialisation remains computationally expensive and fragile, with standard fine-tuning causing catastrophic forgetting and requiring heavy resources. Cloud deployment introduces SLA, scaling and orchestration concerns that academic efficiency benchmarks rarely model. And the move into safety-critical applications demands a level of interpretability that current efficiency techniques do not provide.

This review has synthesised evidence that selective, domain-aware neuron activation could address these challenges together. Networks naturally organise knowledge into specialised neurons; different domains shape which neurons are important; sparse activation reduces cost without hurting accuracy; and adaptive inference shows that computation can be tailored to input difficulty. Yet no existing framework systematically combines these insights.

We argue that future multimodal systems should be designed so that different domains use different computational strategies; different neurons carry different domain-specific knowledge; and selective activation, guided by domain signals and interpretability, enhances efficiency, specialisation and transparency. Research directions that unite adaptive attention re-weighting, knowledge injection and sparse gating represent a promising path forward. Such systems could be efficient enough for cloud inference, specialised for domains like medicine and law, and interpretable enough for safe deployment. Future research should implement and test domain-specific neuron-activation frameworks, evaluate their impact on cloud metrics and interpretability, and ultimately bring multimodal AI closer to both high efficiency and high versatility.

References

1. Agarwal, S., Phanishayee, A., & Venkataraman, S. (2024, April). Blox: A modular toolkit for deep learning schedulers. In Proceedings of the European Conference on Computer Systems (EuroSys).
2. Chen, H., Tao, R., Zhang, H., Wang, Y., Li, X., Ye, W., et al. (2024). Conv-adapter: Exploring parameter-efficient transfer learning for ConvNets. International Journal of Computer Vision.
3. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., et al. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Proceedings of NeurIPS. https://doi.org/10.48550/arXiv.2305.06500
4. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., et al. (2023, July). Scaling vision transformers to 22 billion parameters. Proceedings of ICML.
5. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformer: Scaling to trillion-parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1-39.
6. Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., ... & Hajishirzi, H. (2025, May). Olmoe: Open mixture-of-experts language models. In International Conference on Learning Representations (Vol. 2025, pp. 62061-62121)..
7. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Proceedings of ICLR.
8. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., et al. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. Proceedings of ICCV.
9. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint 1503.02531.
10. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., et al. (2019). Parameter-efficient transfer learning for NLP. Proceedings of ICML.
11. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. Proceedings of ICLR.
12. Jain, S., Zawar, S., Rupchandani, Y., & Chimanna, M. A. (2024, June). Image description generation using deep learning: a comprehensive overview. In 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS) (pp. 1-9). IEEE..
13. Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal, 6, 100050.
14. Krishnamoorthi, R. (2019). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of CVPR.
15. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., et al. (2021). GShard: Scaling giant models with conditional computation and automatic sharding. Proceedings of ICLR.
16. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. Proceedings of EMNLP.
17. Li, Z., & Gu, Q. (2023). I-ViT: Integer-only quantization for efficient vision transformer inference. Proceedings of ICCV.
18. Li, Z., Li, H., & Meng, L. (2023). Model compression for deep neural networks: A survey. Computers, 12(3), 60.
19. Liang, C. X., Tian, P., Yin, C. H., Yu, Y., An-Hou, W., Ming, L., et al. (2024). A comprehensive survey and guide to multimodal large language models. arXiv preprint.
20. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.
21. Martin, M. R., Chan, G., & Ma, K. L. (2025). Visualizing the Impact of Data Perturbation on Text-to-Image Models using Explanative AI (XAI).
22. Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., et al. (2025). Efficiently scaling transformer inference. Proceedings of MLSys.
23. Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: A systematic review. Language Resources and Evaluation, 56(2), 593-619.
24. Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2016). Progressive neural networks. arXiv preprint 1606.04671.

25. Soliman, M., Abdelaziz, O., Radwan, A., & Shehata, M. (2025). GNN-MoE: Context-aware patch routing using GNNs for parameter-efficient domain adaptation. Proceedings of AAAI.
26. Wang, J., Zhang, H., Wang, J., Yang, J., Cheng, L., & Wang, X. (2024). EfficientXpert: Efficient domain adaptation for large language models. Proceedings of EMNLP.
27. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., et al. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint. https://doi.org/10.48550/arXiv.2409.12191
28. Wen, M., Lin, R., Wang, H., Yang, Y., Wen, Y., Mai, L., et al. (2023). Large sequence models for sequential decision-making: A survey. Frontiers of Computer Science.
29. Xiao, Y., Liu, A., Zhang, T., Qin, H., Guo, J., & Liu, X. (2023). RobustMQ: Benchmarking robustness of quantized models. Visual Intelligence.
30. Xu, Z., Xu, Y., Li, H., & Sun, T. (2025). Learning to inference adaptively for multimodal large language models (AdaLLaVA). Proceedings of ICML.
31. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., et al. (2023). Investigating the catastrophic forgetting in multimodal large language models. Proceedings of NeurIPS.
32. Zhang, R., Zhou, Y., Chen, J., Gu, J., Chen, C., & Sun, T. (2024). LLaVA-Read: Enhancing reading ability of multimodal language models. arXiv preprint.
33. Zhao, Y., Zhang, W., Chen, G., Kawaguchi, K., & Bing, L. (2024). How do large language models handle multilingualism?. Advances in Neural Information Processing Systems, 37, 15296-15319..
34. Zhu, X., Guan, X., Liang, D., Chen, Y., Liu, Y., & Bai, X. (2024). MoE Jetpack: From dense checkpoints to adaptive mixture-of-experts for vision tasks. Proceedings of NeurIPS.