From Models to Systems: A Comprehensive Survey of Efficient Multimodal Learning

Anonymous authors
Paper under double-blind review

Abstract

The rapid expansion of multimodal models has surfaced formidable bottlenecks in computation, memory, and deployment, catalyzing the rise of Efficient Multimodal Learning (EML) as a pivotal research frontier. Despite intensive progress, a cohesive understanding of *what*, *how*, and *where* efficiency is manifested across the learning stack remains fragmented. This survey systematizes the EML landscape by introducing the first structured, model-to-system taxonomy. We distill insights from over 300 seminal works into three hierarchical levels—*model*, *algorithm*, and *system*—addressing architectural parsimony, execution refinement, and hardware-aware orchestration, respectively. Moving beyond a purely categorical review, we offer a methodological synthesis of the vertical synergies between these layers, elucidating how cross-layer co-design resolves the fundamental "Efficiency-Utility-Privacy" trilemma. Through an integrative case study of Multimodal Large Language Models (MLLMs), we trace the field's evolutionary trajectory from initial structural adjustments to modern full-stack resource orchestration. Furthermore, we provide a holistic discussion and application-specific optimization blueprints for diverse domains and posit a paradigm shift toward self-regulating intelligence, where efficiency is an intrinsic, emergent property of the model's fundamental design rather than a post-hoc constraint. Finally, we present open challenges and future directions that will define the trajectory of EML research. This survey establishes a formal foundation for multimodal systems that are not only high-performing and generalizable but natively efficient and ready for ubiquitous deployment. We also maintain a Github repository to continuously update related work for research community.

1 Introduction

The paradigm shift toward multimodal learning has revolutionized artificial intelligence, enabling systems to jointly perceive, align, and reason over heterogeneous signals such as vision, language, audio, and sensor data (Baltrušaitis et al., 2018; Mo et al., 2024; 2023). This unification underpins critical advances in domains ranging from embodied robotics and autonomous driving to precision healthcare (Jin et al., 2025). However, this scaling success faces a formidable bottleneck: computational inefficiency. The quadratic complexity and massive parameter counts of modern multimodal transformers demand exorbitant memory and energy resources, often precluding deployment in real-time or resource-constrained environments. As the field moves to democratize these models beyond high-end clusters, establishing a systematic understanding of Efficient Multimodal Learning (EML) has become a critical academic and industrial frontier.

Unlike unimodal efficiency targeting homogeneous inputs, multimodal efficiency must orchestrate heterogeneous modalities with disparate semantics, resolutions, and temporal dynamics. Consequently, optimization cannot be isolated to a single layer; it must span the entire stack—from architecture and algorithms to hardware execution. To achieve this, a robust system must allocate computation adaptively, determining *what*, *how*, and *where/when* to process each modality under varying constraints. This evolution marks a paradigm shift from accuracy-centric learning to intelligence that is intrinsically efficient, resource-adaptive, and deployable at scale.

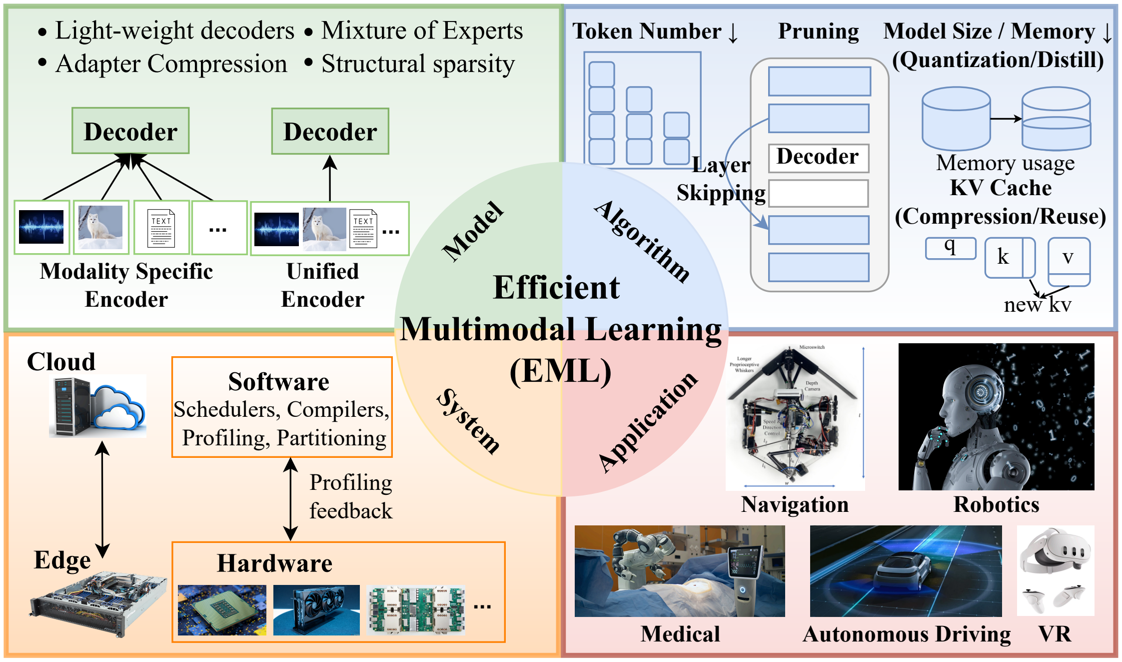

Figure 1: Overall landscape of Efficient Multimodal Learning (EML), organized across three interconnected levels—Model, Algorithm (Compression & Acceleration), and System—that jointly optimize architectural design, computation, and deployment, with representative applications illustrated.

Despite the explosion of research, the path to multimodal efficiency remains obscured by a fog of fragmented techniques. Current literature typically organizes methods by isolated stages or granular mechanisms (Jin et al., 2025; Shinde et al., 2025), which, while useful for cataloging, fails to answer a fundamental structural question: *Where exactly in the multimodal stack can efficiency be injected, and how do these injections interact?* Without a clear topological map, researchers risk optimizing specific modules in isolation, often neglecting the broader landscape of cross-layer opportunities. This lack of a full-stack perspective limits the community's ability to systematically exploit efficiency bottlenecks across the entire computing pipeline of EML development.

To bridge these isolated domains, this survey proposes the Model–Algorithm–System (MAS) taxonomy, a unified framework that synthesizes over 300 representative studies into a layered ecosystem. Rather than viewing efficiency as a collection of disparate tricks, we systematically map optimizations to their precise locus within the computing stack: ***(1) Model-level:*** reshaping architectural topology to define *what* to compute—encompassing modality-specific or unified encoders, sparse expert routing, and modular adapters. ***(2) Algorithm-level:*** modulating information flow to determine *how* to compute—via token compression, pruning, quantization, distillation, speculative decoding, and cache reuse. ***(3) System-level:*** orchestrating physical execution to decide *where* and *when* to compute—integrating cache management, edge–cloud collaboration, latency-aware scheduling, and hardware-software co-design.

Moving beyond a purely categorical review, this survey provides a methodological synthesis of the vertical synergies between these layers. We analyze Efficient Multimodal Large Language Models (MLLMs) as a primary proving ground, demonstrating how the convergence of perception, execution, and scheduling resolves the fundamental "Efficiency-Utility-Privacy" trilemma. Furthermore, we establish application-specific optimization blueprints for diverse domains—from affective computing to spatial understanding and reasoning—and posit a paradigm shift toward self-regulating intelligence. In this nascent regime, efficiency is reframed not as a post-hoc constraint, but as an intrinsic, emergent property of the model's fundamental design. By delineating these critical open challenges and future directions, we offer the community a coherent guide to realizing natively efficient and scalable multimodal intelligence.

This survey makes the following contributions:

- **Unified Taxonomy:** We present the first blueprint integrating model, algorithm, and system efficiency into a holistic MAS framework for EML.
- **Cross-Level Analysis:** We deconstruct the dependencies between architectural sparsity, algorithmic compression, and system orchestration, offering a principled view of resource-aware intelligence.
- **MLLMs Synthesis:** We synthesize recent breakthroughs and evolution in efficient MLLMs as a critical convergence within the MAS framework, where vertical integration empowers scalable, real-world multimodal intelligence.
- **Future Roadmap:** We identify emerging applications and open questions, showing future directions toward sustainable, adaptive, and deployable multimodal learning.

2 Scope and Taxonomy

This survey focuses on techniques that explicitly enhance the *efficiency* of multimodal learning systems under realistic resource constraints such as limited FLOPs, latency, memory, or energy budgets. We include methods that *reduce*, *reallocate*, or *reuse* computation while maintaining multimodal performance, spanning model, algorithm, and system-level optimization. In contrast, topics orthogonal to efficiency—such as representation learning, pretraining objectives, or interpretability—are discussed only when they directly integrate efficiency-driven mechanisms (e.g., sparsity-inducing alignment, compression-guided adaptation). Within multimodal large language models (MLLMs), we focus on how their efficiency mechanisms align with our framework rather than exhaustively reviewing all variants; readers seeking comprehensive MLLM surveys may refer to specialized reviews (Jin et al., 2025; Xu et al., 2024; Shinde et al., 2025).

Figure 2 illustrates our MAS taxonomy, structuring the EML landscape across three interdependent levels: **model-level**, which reshapes architectural topology via modality-specific and -unified encoders, structural sparsity, structural decoding, and lightweight modular adaptation; **algorithm-level**, which refines execution via token compression, pruning, quantization, knowledge distillation, Prompting and speculative decoding, caching and reuse, and runtime sparsity; and **system-level**, which enables deployable efficiency through cache management and serving, edge cloud collaboration, latency-aware scheduling and pipelining, hardware-software codesign, and federated learning. Although each level targets a distinct stage of the multimodal pipeline, their interactions are tightly coupled—for example, model-level sparsity often depends on algorithm-level quantization for accuracy retention, while system-level scheduling can further amplify the efficiency gains of dynamic inference.

The remainder of this survey: Sections 3–5 examine efficiency across model, algorithm, and system levels, followed by analyses of efficient MLLM (Sec. 6), applications (Sec. 7), holistic discussion (Sec. 8), and open research challenges and opportunities (Sec. 9). Each section reflects our unique discussion and insights. Besides, we also present a case study for the edge EML system and list a subset of recent efficient multimodal models under our MAS framework in the Appendix 10.

3 Model

As shown in Fig. 3, model-level efficiency fundamentally reshapes the architectural topology to define *what* to compute. Efficient architectures seek to minimize redundant processing while preserving alignment, interaction, and representational richness. They reshape computation through explicit structural choices—encoder specialization, unified encoders, sparsity-aware routing, structural decoding, and modular adaptation—each offering a pathway to achieve scalable and expressive multimodal learning under limited budgets.

3.1 Modality-specific Encoders

Vision. The evolution of vision encoders reflects a continuous search for efficient topologies that balance perceptual fidelity with computational constraints. Early efficiency-oriented designs, such as Mo-

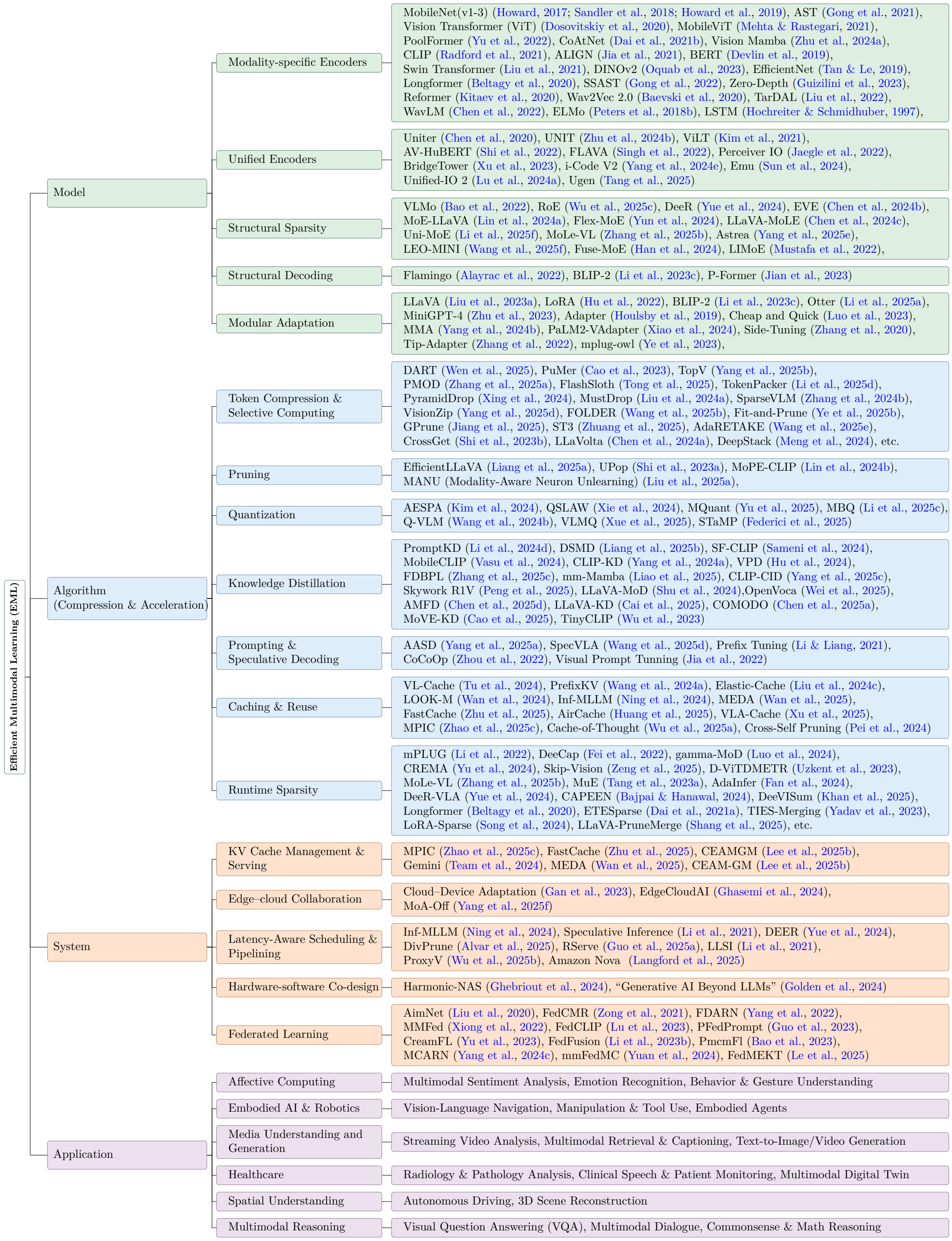

Figure 2: The MAS taxonomy for EML with representative works.

bileNet (Howard, 2017), ShuffleNet (Zhang et al., 2018), and EfficientNet (Tan & Le, 2019), mitigate computational redundancy via depthwise separable convolutions or compound scaling rules. While effective for local feature extraction, their lack of global context prompts the development of hybrid architectures like MobileViT (Mehta & Rastegari, 2021) and CoAtNet (Dai et al., 2021b), which synergize convolutional efficiency with Transformer expressivity. Pushing this structural evolution further, PoolFormer (Yu et al.,

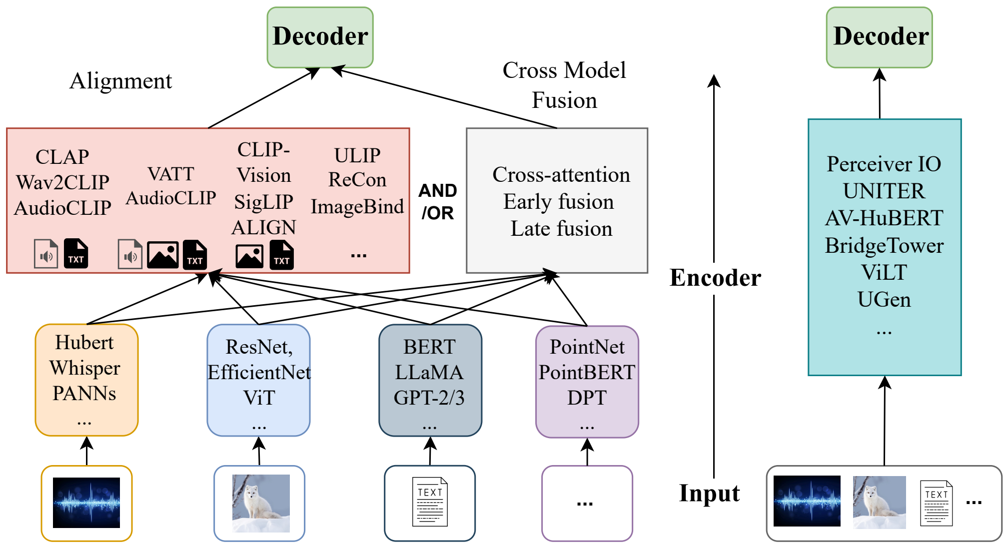

Figure 3: Structural paradigms of multimodal encoders. The taxonomy contrasts (left) decoupled modality-specific pipelines utilizing post-hoc alignment or fusion mechanisms with (right) natively unified encoders that collapse heterogeneous signals into a shared parameterized core. This architectural evolution reflects a shift toward functional consolidation, where unification acts as a structural prerequisite for efficiency.

2022) demonstrates that simple pooling operations can replace complex attention mechanisms within a "MetaFormer" architecture, achieving efficiency through pure topological simplification.

The subsequent shift to Vision Transformers (ViT) (Dosovitskiy et al., 2020) fully enables global reasoning but incurs quadratic complexity. To address this, hierarchical variants like Swin Transformer (Liu et al., 2021) reintroduce shifted-window attention to linearize complexity while preserving local priors. Simultaneously, Masked Autoencoders (MAE) (He et al., 2022) and BEiT v2 (Peng et al., 2022) resolve the scalability bottleneck by turning masked image modeling into an efficiency primitive, enabling large-scale pretraining with reduced overhead. More recently, alternatives emerge to challenge the dominance of attention entirely: Mamba-based backbones (Gu & Dao, 2024; Zhu et al., 2024a) and Kolmogorov–Arnold Networks (KANs) (Liu et al., 2024b) utilize state-space models or learnable splines to capture long-range dependencies with linear complexity $O(N)$, offering superior scaling laws.

Ultimately, modern encoder design moves beyond purely structural optimization to prioritize cross-modal alignability. Architectures like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) employ dual-stream encoders to project visual and textual features into a shared semantic space via the InfoNCE objective:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\big(\text{sim}(v_i,t_i)/\tau\big)}{\sum_{j=1}^{N}\exp\big(\text{sim}(v_i,t_j)/\tau\big)}, \tag{1}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ is a temperature parameter, and N is the number of samples. Building on this foundation, SigLIP (Zhai et al., 2023) identifies the softmax normalization as a scalability bottleneck and replaces it with a pairwise sigmoid loss, decoupling memory usage from batch size. Complementing these global alignment methods, dense self-supervised encoders like the DINO series (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) provide fine-grained visual features essential for multimodal understanding. This paradigm transforms visual encoders from static classifiers into flexible foundations for open-vocabulary multimodal systems.

Text. The evolution of text encoding traces a cyclical trajectory: moving from memory-efficient recurrence to parallelized attention, and finally converging on architectures that unify the strengths of both to support massive multimodal contexts. The foundational era relies on Recurrent Neural Networks (RNNs), where architectures like LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014) maintain hidden states to capture temporal dependencies. While deep contextualized representations like ELMo (Peters et al., 2018a) demonstrate the representational power of this paradigm, the inherent sequentiality of recurrence prohibits parallel training, creating a fundamental barrier to scaling on modern hardware. Efficiency-oriented variants, such as IndRNN (Li et al., 2018) and LightRNN (Li et al., 2016), attempt to mitigate this by decoupling matrix operations or compressing vocabularies, yet the underlying throughput bottleneck persists.

The introduction of the Transformer (Vaswani et al., 2017) and BERT (Devlin et al., 2019) breaks this serial constraint by enabling fully parallel context aggregation. However, this architectural shift exchanges sequential latency for quadratic computational complexity ($O(N^2)$), which becomes prohibitive when processing long sequences of interleaved text tokens. To reconcile global context with computational viability, structural variants dismantle the dense attention matrix: Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) introduce sparse window mechanisms to reduce complexity to linear time, while Reformer (Kitaev et al., 2020) employs locality-sensitive hashing (LSH) to approximate global interactions. Simultaneously, approaches like Linformer (Wang et al., 2020) demonstrate attention matrices are low-rank, allowing for projection-based approximations that further compress the computational footprint.

As decoder-only LLMs like LLaMA (Touvron et al., 2023) become the backbone for modern VLMs, the frontier shifts toward revisiting recurrence to handle the explosive growth of multimodal context windows. Recent architectures, including Linear Transformers (Katharopoulos et al., 2020) and state-space models (SSMs) like TextMamba (Zhao et al., 2025b), abandon the standard softmax attention mechanism entirely. By combining the parallelizable training of Transformers with the constant inference memory of RNNs ($O(1)$), these designs unlock a critical capability for multimodal learning: the ability to sustain effectively infinite context windows. This transforms text encoders from a computational bottleneck into a scalable semantic anchor, capable of maintaining long-term dialogue history and reasoning over extensive descriptive inputs essential for multimodal systems.

Audio. The evolution of audio encoders parallels that of vision, transitioning from rigid convolutional priors to unified, efficiency-aware tokenization that aligns seamlessly with broader multimodal architectures. Early efficiency-oriented designs rely on architectural inductive biases: by treating log-Mel spectrograms as 2D image-like signals, convolutional encoders such as VGGish (Hershey et al., 2017) and CNN14 (Kong et al., 2020) leverage local connectivity to extract harmonic patterns. While effective, this approach defines efficiency primarily through parameter sharing and locality, often overlooking the temporal continuity intrinsic to acoustic signals. To overcome this and leverage massive unlabeled data, the paradigm shifts toward self-supervised learning (SSL) within hybrid topologies. Foundational frameworks like Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) combine lightweight convolutional front-ends for local feature extraction with Transformer backbones for global context. By solving contrastive or masked unit prediction tasks, these methods redefine efficiency through the lens of data scalability, producing robust representations that generalize across tasks with minimal supervision.

A critical turning point toward architectural unification arrives with the adoption of patch-based masked modeling, inspired by Vision Transformers. The Audio Spectrogram Transformer (AST) (Gong et al., 2021) and SSAST (Gong et al., 2022) eliminate the convolutional hierarchy, treating spectrogram patches as discrete tokens for global self-attention. This shift not only simplifies the design space but also aligns audio encoding structurally with visual and textual modalities, facilitating cross-modal transfer and unified pretraining. However, the quadratic complexity of global attention poses a bottleneck for long-form audio processing. Addressing this, recent architectures like Audio Mamba (Yadav & Tan, 2024) adopt selective SSMs to bypass the attention mechanism. By capturing long-range dependencies with strict linear complexity ($O(N)$) and constant inference memory, these models naturally align with the continuous nature of sound. This trajectory culminates in a vital capability for multimodal learning: the creation of efficient, linear-time audio encoders that allow high-fidelity acoustic streams to be integrated into LLMs as native tokens, establishing a scalable foundation for real-time, omni-sensory understanding.

Beyond Canonical Modalities. Beyond vision, text, and audio, modalities such as thermal imagery, depth sensing, and time-series data extend multimodal learning into domains characterized by low resolution, sparsity, or temporal irregularity. Although heterogeneous, the evolution of their encoders reveals a convergent trajectory: moving from rigid priors to flexible, computation-aware abstractions.

Thermal imagery captures radiometric intensity rather than clear texture, often resulting in low-contrast, high-noise data. To process this efficiently, architectures must isolate informative features from clutter. Early designs such as TarDAL (Liu et al., 2022) employ dual-stream topologies to disentangle target semantics from noise via sub-networks. More recently, approaches like FW-SAT (Jiang & Chen, 2024) transition to ViT-based backbones but restrict computation through local window attention. This design mitigates the quadratic cost of global modeling while focusing resources on informative regions, preserving the structural details vital for interpreting low-quality thermal inputs without incurring the overhead of full self-attention.

Depth sensing provides explicit geometric cues but frequently suffers from sparse or missing measurements due to sensor limitations. To balance structural fidelity with computational cost, the field converges on hybrid architectures. Frameworks like Lite-Mono (Zhang et al., 2023a) and MonoDETR (Zhang et al., 2023b) integrate lightweight convolutions for high-frequency surface completion and Transformer blocks for global geometric reasoning. This hybrid topology effectively leverages the efficiency of convolutions for handling local sparsity (filling "holes" in the depth map) while reserving expensive attention operations solely for establishing long-range scale consistency.

Time-series data pose unique challenges regarding irregular sampling and extremely long-range dependencies. While RNNs capture local trends, their sequential nature prevents parallel training on massive datasets. To address this, efficient Transformers like Informer (Zhou et al., 2021) introduce sparse attention mechanisms, such as ProbSparse attention, to approximate global context with sub-quadratic cost ($O(N \log N)$). Most recently, the focus shifts to continuous-time architectures: State-space models like Mamba (Gu & Dao, 2024) and Liquid-S4 (Hasani et al., 2023) model temporal evolution via linear recurrence. By processing massive horizons with strict linear complexity ($O(N)$), these models resolve the memory bottleneck of Transformers, establishing a scalable paradigm for long-term forecasting and sequential reasoning.

3.2 Unified Multimodal Encoders

Unified multimodal encoders aim to collapse redundant, modality-specific pipelines into a shared computational backbone(Wang et al., 2022). Instead of maintaining parallel branches for each signal, these architectures introduce a centralized parameterized core—such as a joint Transformer trunk—that processes heterogeneous tokens within a single vector space. This paradigm shifts the definition of efficiency from simple resource reduction to functional consolidation, where cross-modal interactions occur deeply and repeatedly, turning unification itself into a mechanism for parameter and inference efficiency.

From Dual-stream Fusion to Shared Trunks. Early efforts focus on integrating visual and textual streams within a single Transformer. Models like UNITER (Chen et al., 2020), ViLT (Kim et al., 2021), and BridgeTower (Xu et al., 2023) discard heavy modality-specific extractors, instead encoding image patches and text tokens directly through a shared attention backbone. ViLT, in particular, demonstrates that a unified Transformer could replace complex convolutional front-ends, significantly reducing model footprint. These works establish the principle of parameter sharing as efficiency, proving that cross-modal understanding does not require independent feature hierarchies.

Multi-sensory Unification. Subsequent frameworks extend this unified paradigm to include audio and video. AV-HuBERT (Shi et al., 2022) generalizes SSL by jointly masking and predicting clustered audio–visual units, achieving strong recognition accuracy with orders of magnitude less labeled data. FLAVA (Singh et al., 2022) further scales this by employing a single Transformer to process image, text, and audio tokens, demonstrating that cross-modal co-training acts as an implicit regularizer. This holistic approach reduces the total pretraining compute compared to maintaining separate unimodal models, validating unification as a path to scalability.

Latent-Core and Autoregressive Unification. Recent designs unify capacity through latent bottlenecks or sequence-level modeling. Perceiver IO (Jaegle et al., 2022) and Perceiver-VL (Tang et al., 2023b) encode arbitrary modalities via a fixed-size latent array, decoupling computational cost from input size and resolution. Alternatively, UNIT (Zhu et al., 2024b) maintains lightweight modality-specific heads during pretraining but merges them at inference for a single shared encoder. More radically, autoregressive models such as Emu (Sun et al., 2024), Unified-IO 2 (Lu et al., 2024a), i-CodeV2 (Yang et al., 2024f), UGen (Tang et al., 2025), and Grok-1.5V (xAI / Grok team, 2024) embrace generative unification: they tokenize all modalities into a single discrete sequence and optimize a unified next-token prediction objective:

$$\mathcal{L}_{\mathrm{AR}} = -\sum_{t=1}^{T} \log p\big(y_t \,\big|\, y_{<t},\, \mathrm{ctx}\big), \tag{2}$$

where y_t denotes the target token at step t, $y_{<t}$ represents the multimodal history, and ctx is the context. This formulation enables seamless conditioning and generation across modalities, establishing sequence modeling as the universal interface for multimodal efficiency.

3.3 Structural Sparsity

Structural sparsity enforces efficiency by utilizing conditional computation to activate only a subset of parameters during training or inference. Unlike pruning (which permanently removes weights), structural sparsity embeds dynamic routing directly into the model architecture, allowing systems to decouple total capacity from active computation. In multimodal contexts, this enables models to scale to billions of parameters while maintaining the inference footprint of much smaller networks, dynamically allocating resources based on input complexity.

Modality-Specialized Expert Routing. Early application of Mixture-of-Experts (MoE) in multimodal learning focus on mitigating cross-modal interference. Model architectures like VLMo (Bao et al., 2022) and LIMoE (Mustafa et al., 2022) introduce modality-aware routing, coupling shared attention mechanisms with modality-specific experts. By directing image and text tokens to distinct feed-forward networks (FFNs), these models achieve disentangled representation learning, proving that sparsity can enhance expert specialization while reducing the computational redundancy of monolithic transformers.

Unified Semantic Routing. Subsequent frameworks, such as Uni-MoE (Li et al., 2025f), have advanced from rigid modality partitioning to unified, content-driven routing. Here, experts are shared across modalities and activated dynamically based on token-level semantic complexity. This shift allows the architecture to adaptively allocate capacity—using more experts for complex reasoning tokens and fewer for simple patches—transforming sparsity from a static routing rule into a responsive, content-aware mechanism.

Budget-Aware Elasticity. Recent frameworks such as LEO-MINI (Wang et al., 2025f) and Flex-MoE (Yun et al., 2024) extend sparsity to resource-constrained deployment. Rather than maximizing capacity, they prioritize compute elasticity, employing hierarchical routing or mixed-rank experts to satisfy strict memory or latency budgets. Models like NVILA (Liu et al., 2025b) and SmolVLM (Marafioti et al., 2025) further optimize this by pruning token pathways, effectively creating any-budget architectures that dynamically adjust their active parameter set to fit the available hardware envelope.

3.4 Structural Decoding

Structural decoding enhances efficiency by architecturally constraining the interface between high-resolution perception and autoregressive generation. Instead of allowing the language model to attend directly to dense, variable-length feature maps—which incurs prohibitive quadratic costs and modality misalignment—recent architectures introduce a learnable bottleneck that decouples decoding complexity from input dimensionality. Perceiver-style decoders like Flamingo (Alayrac et al., 2022) employ a resampler mechanism where fixed latent queries cross-attend to visual inputs, compressing spatiotemporal features into a constant number of visual tokens. Refining this principle, query-based transformers like BLIP-2 (Li et al., 2023c) utilize a lightweight Q-Former to act as a semantic bridge, actively distilling dense encoder outputs into a compact, text-aligned token set. Ultimately, structural decoding reframes efficiency as a problem of interface design: replacing exhaustive cross-attention with a bounded, fixed-capacity channel that effectively isolates the generative engine from the raw scale of sensory data while aligning heterogeneous modalities.

3.5 Modular Adaptation

Modular Adaptation mainly consists of bottleneck adapters(Houlsby et al., 2019) and low-rank adaptation (LoRA)(Hu et al., 2022). Adapters enhance efficiency by inserting lightweight, trainable modules between or within frozen pretrained components, enabling rapid specialization and cross-modal alignment without the cost of full-model retraining(Sung et al., 2022b;a). Functionally, these architectures operate at two distinct structural levels: inter-module connection and intra-module tuning. For cross-modal alignment, projection adapters—such as the simple linear layers in LLaVA (Liu et al., 2023a) or the multi-layer perceptrons in MiniGPT-4 (Zhu et al., 2023)—act as minimal connectors that project sensory features directly into the language model's embedding space. Complementing these connectors, parameter-efficient tuning methods like LoRA (Hu et al., 2022; Guo et al., 2025b) introduce low-rank decompositions as bypass pathways inside transformer layers, allowing the model's internal reasoning to adapt to new modalities using a fraction of the original parameter count. Recent modular advances, such as MMA (Yang et al., 2024b) and PaLM2-VAdapter (Xiao et al., 2024), extend this paradigm by creating plug-and-play adapter banks that can be dynamically swapped for different tasks. Taken together, adapter-based strategies redefine efficiency as the structural decoupling of capability expansion from backbone maintenance—achieving scalable alignment and transfer under strict memory and compute budgets.

3.6 Discussion and Key Insights

Model-level efficiency establishes the architectural foundation of EML by redesigning how computation is organized within and across modalities. Beyond the individual techniques cataloged above, our analysis reveals three fundamental paradigm shifts that define the next generation of efficient multimodal architectures:

- **From Explicit Perception to Latent-Core Abstraction:** The evolution from modality-specific pipelines to shared Transformer trunks and latent-core bottlenecks reflects a strategic shift toward the information bottleneck principle. By forcing heterogeneous signals through a fixed-size latent array or a learnable resampler, architectures effectively decouple the quadratic cost of perception from the autoregressive complexity of reasoning. This structural constraint serves as a "semantic filter", ensuring that the generative engine processes only the most salient cross-modal alignments.
- **Addressing Representational Asymmetry via Sparsity:** Multimodal data exhibits inherent representational asymmetry in information density; for instance, visual tokens often contain significantly higher spatial and temporal redundancy compared to the dense semantics of text. Structural sparsity, particularly through MoE and conditional routing, enables architectures to address this asymmetry dynamically. Instead of a monolithic processing pass, modern models utilize modality-aware or content-driven routing to allocate high-capacity experts only to "hard" tokens while bypassing redundant patches.
- **Efficiency as a Structural Prerequisite Not a Trade-off:** A pivotal insight is that efficiency has evolved from a post-hoc optimization into an intrinsic design primitive. Through mechanisms like modular adaptation and parameter-efficient tuning, efficiency is no longer viewed merely as a compromise on capability. Rather, it is a structural property that enables unprecedented model scalability—allowing systems to sustain effectively infinite context windows or perform real-time, omni-sensory reasoning that would be physically prohibitive under traditional dense architectures.

Ultimately, these model-level innovations culminate in a paradigm shift: the transition from modular concatenation toward natively unified foundations. By moving beyond the assembly of modality-specific encoders to form a cohesive world model, the architectural objective evolves from post-hoc alignment toward structural parsimony. Within this unified regime, efficiency is redefined not as a secondary trade-off or adjustment, but as an intrinsic, emergent property of the architecture's fundamental design, where computational capacity is natively and dynamically modulated by the latent semantic complexity of multimodal signals.

4 Algorithm

Algorithm-level efficiency defines *how* computation executes, compressing information flow within fixed architectural topologies. Unlike structural redesigns, these strategies target operation-level reductions in computation and memory footprint. As illustrated in Fig. 4, key techniques—ranging from token compression

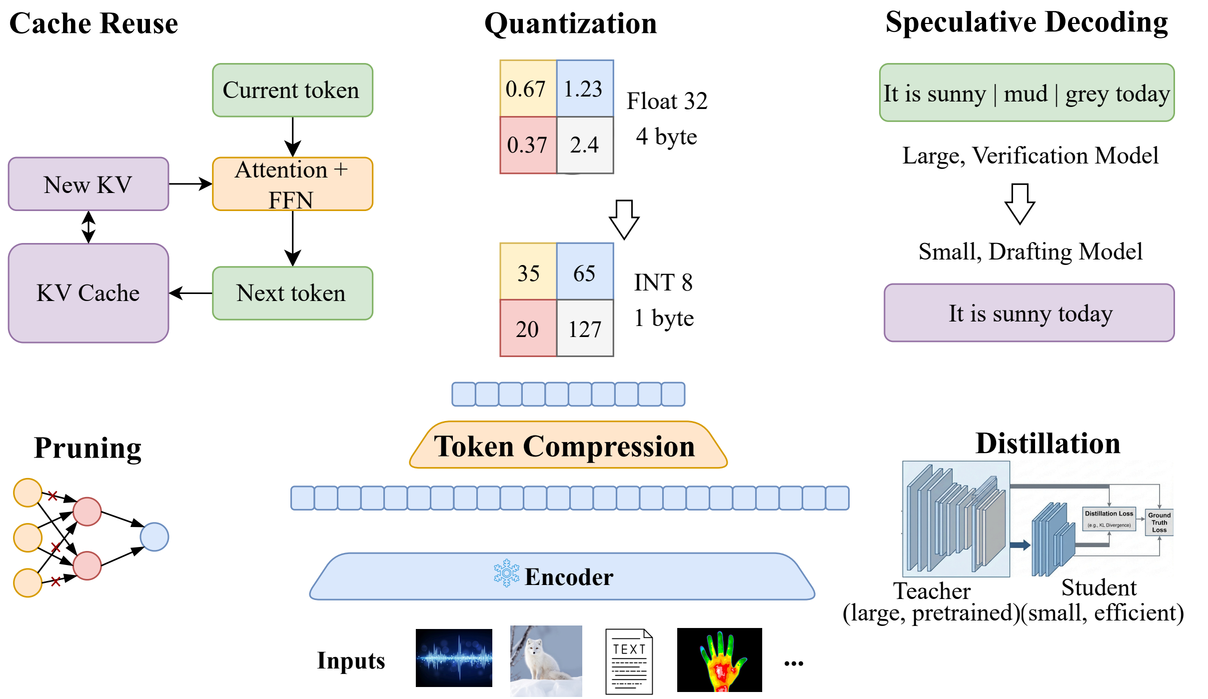

Figure 4: Algorithm-level efficiency for refining multimodal execution dynamics. This taxonomy illustrates the modulation of information flow across the EML pipeline through seven primary axes: (i) **Token compression and selective computing** to filter spatial redundancy and retain informative semantic regions; (ii) **Pruning** to eliminate structural redundancy within backbone architectures; (iii) **Quantization** to minimize memory bandwidth via precision discretization; (iv) **Knowledge distillation** to transfer reasoning behaviors and cognitive patterns to compact learners; (v) **Prompting and speculative decoding** to streamline input adaptation and parallelize generation; (vi) **Caching and reuse** to amortize prefill costs through temporal state persistence; and (vii) **Runtime sparsity** to enable adaptive computation based on input complexity. These strategies transform multimodal execution from static processing to a dynamic, information-flow-aware pipeline.

and quantization to state caching—systematically minimize data redundancy and arithmetic precision. By streamlining processing across both training-free and training-aware regimes, algorithm-level efficiency enhances runtime resource economy while preserving the model's structural integrity.

4.1 Token Compression & Selective Computing

Reducing redundant visual tokens has become a central mechanism for accelerating multimodal transformers, where the computation cost grows quadratically with token count. The objective is not mere token removal, but to identify and retain semantically informative regions that drive multimodal reasoning. This area has evolved along two complementary lines—training-free and training-based compression—each reflecting a balance between practicality and adaptivity.

Training-free Compression. Training-free approaches perform pruning or merging during inference without altering pretrained parameters. Early works exploit intrinsic attention maps or saliency patterns within vision encoders to prune tokens of low importance (Zhang et al., 2024b; Zhuang et al., 2025; Arif et al., 2025). Redundancy-aware methods extend this by measuring feature correlation or similarity to eliminate overlapping representations (Wen et al., 2025; Yang et al., 2025b;d; Tan et al., 2025). More recent advances adopt layer-adaptive or progressive pruning schedules that gradually reduce token counts across layers or decoding steps (Liu et al., 2024a; Zhuang et al., 2025; Wang et al., 2025e), mitigating abrupt information loss. Structural and optimization-aware frameworks further refine this process by modeling token relationships as graphs (Jiang et al., 2025; Xing et al., 2024; Arif et al., 2025) or formulating selection as constrained optimization (Ye et al., 2025b; Omri et al., 2025). Representative systems such as ST3 (Zhuang et al., 2025) and LLaVolta (Chen et al., 2024a) show that over 70% of visual tokens can be pruned while preserving multimodal accuracy, highlighting that meaningful compression is possible even without retraining.

Training-based Compression. Training-based methods embed learnable token selectors or merging modules into the model to co-optimize compression with downstream supervision. Learnable pruners such as P-Mod (Zhang et al., 2025a), FAST (Pertsch et al., 2025), and TokenPacker (Li et al., 2025d) introduce lightweight scoring networks that dynamically drop or aggregate tokens based on a learnable importance $\mathbf{s}_i$. The model is trained with a task loss $\mathcal{L}_{\text{task}}$ with sparsity weight controlling factor λ

$$\mathcal{L}_{\text{select}} = \mathcal{L}_{\text{task}} + \lambda \left\|\mathbf{s}\right\|_1, \tag{3}$$

where $\mathbf{s} = (s_1, s_2, ..., s_N), s_i \in [0, 1]$ denotes the vector of learnable importance scores, N is the token length, $\lambda \left\|\mathbf{s}\right\|_1$ encourages sparsity in these scores. Pooling- and clustering-based frameworks (Wang et al., 2025b; Yang et al., 2025d) further abstract semantically similar tokens into compact latent embeddings, effectively reducing sequence length while preserving semantics. Curriculum-based approaches (Chen et al., 2024a) progressively decrease token budgets throughout training, encouraging models to adapt to increasingly compressed representations. By integrating token selection into optimization, these approaches achieve higher accuracy under heavy compression and enable controllable trade-offs between efficiency and semantic fidelity.

4.2 Pruning

Model pruning enhances multimodal efficiency by structurally compressing large architectures through the removal of redundant layers, neurons, or attention heads, thereby reducing computation without extensive retraining. Early multi-stage frameworks (Wang et al., 2024c) jointly prune visual and textual branches via layer-wise(Sung et al., 2023b) and hidden-dimension reduction, while EfficientLLaVA (Liang et al., 2025a) formulates pruning as a generalization-aware search, automatically selecting attention and MLP weights using small proxy data. UPop (Shi et al., 2023a) and MoPE-CLIP (Lin et al., 2024b) advance this by performing unified, progressive pruning across modalities, dynamically allocating pruning ratios and refining masks during training to maintain convergence under high compression. More recent approaches such as MANU (Liu et al., 2025a) adopt modality-aware neuron pruning to remove cross-modal redundancies, improving both compactness and representational disentanglement. At a higher level, these methods shift pruning from static compression toward adaptive, semantically guided reduction—enabling deployable multimodal models that balance accuracy, scalability, and hardware efficiency.

4.3 Quantization

Quantization compresses model footprint and accelerates inference by mapping high-precision floating-point weights and activations into lower-bit discrete representations (e.g., INT8). Unlike pruning, which reduces the number of operations, quantization reduces the bit-width of operands, thereby lowering memory bandwidth and energy consumption. Formally, uniform affine quantization maps a real-valued tensor x to an integer q via a scaling factor Δ and a zero-point z:

$$q = \text{clip}\left(\left\lfloor \frac{x}{\Delta} \right\rceil + z,\ q_{\min},\ q_{\max}\right), \qquad \hat{x} = \Delta\,(q - z), \tag{4}$$

where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer, $[q_{\min}, q_{\max}]$ defines the discrete range, and $\hat{x}$ represents the dequantized approximation.

Post-Training Quantization (PTQ). PTQ applies low-bit mapping directly to pretrained models without requiring extensive retraining, making it highly deployable. However, multimodal models pose unique challenges due to the divergent statistical distributions of visual and textual features (Bhatnagar et al., 2025). Recent research addresses this via modality-aware calibration, where visual tokens—which often exhibit distinct outlier patterns compared to text—receive differentiated scaling or grouping. Approaches like Q-VLM (Wang et al., 2024b), MBQ (Li et al., 2025c), VLM-Q (Xue et al., 2025), and MQuant (Yu et al.,

2025) introduce sensitivity-based mixed precision, assigning higher bit-widths to varying channels or tokens that are critical for cross-modal alignment. Complementing weight-centric methods, STaMP (Federici et al., 2025) targets activation quantization, employing sequence transformations to suppress outlier activations that typically destabilize low-bit inference in large vision-language models.

Quantization-Aware Training (QAT). Aggressive compression regimes (e.g., sub-4-bit) often incur catastrophic discretization errors that PTQ fails to mitigate. QAT addresses this by simulating quantization-induced noise during the optimization trajectory, enabling the network to re-learn representational fidelity within a constrained bit-width. To overcome the prohibitive memory overhead of applying QAT to large-scale MLLMs, EfficientQAT (Chen et al., 2025b) introduces a tiered optimization paradigm: it utilizes block-wise training of all parameters followed by end-to-end refinement of quantization scales, effectively circumventing the "memory wall" while preserving the accuracy of 70B-scale models. In the context of multimodal reasoning, specialized challenges such as cross-modal outlier distributions are addressed by QSLAW (Xie et al., 2024), which integrates modality-aware warm-up and learnable step sizes into the instruction-tuning phase. By co-optimizing discretization parameters with neural weights, these methodologies ensure that the delicate semantic alignment between vision and language is preserved even in highly discretized spaces, bridging the gap between algorithmic compression and system-level inference throughput.

4.4 Caching and Reuse

The Key-Value (KV) cache is a critical bottleneck in autoregressive generation, where memory consumption grows linearly with sequence length and batch size. Optimization strategies aim to decouple memory growth from context length, ensuring long-horizon generation remains feasible within fixed hardware budgets.

Dynamic Eviction and Compression. Importance-based methods mitigate cache redundancy by selectively identifying and discarding non-essential tokens. Leveraging the high spatial redundancy of visual data relative to dense textual semantics, VL-Cache (Tu et al., 2024) introduces modality-aware pruning that aggressively evicts visual patches while shielding critical textual context. To enhance precision, ElasticCache (Liu et al., 2024c) and LOOK-M (Wan et al., 2024) employ attention entropy and anchor-based merging to differentiate between immutable semantic "anchors" and compressible repetitive patterns. Structural refinements, such as CSP (Pei et al., 2024), further stabilize this process by disentangling self-attention from cross-attention channels. On the deployment frontier, systems like FastCache (Zhu et al., 2025) and AirCache (Huang et al., 2025) implement retrieval-augmented hierarchies, offloading "cold" states to secondary storage while retaining "hot" states in GPU memory to balance extensive context horizons with limited hardware capacity.

Temporal and Cross-Session Reuse. Complementary to compression, reuse-oriented methods exploit the temporal coherence of multimodal signals to amortize computation. In dynamic scenarios like robotics or video streaming, the visual scene changes slowly. VLA-Cache (Xu et al., 2025) leverages this by reusing static background tokens across frames, recomputing only the dynamic patches relevant to the task. Similarly, Inf-MLLM (Ning et al., 2024) manages streaming inputs by identifying "attention saddle" points—tokens that sustain long-term dependencies—to maintain context over effectively infinite streams. Moving beyond single sessions, frameworks like MPIC (Zhao et al., 2025c) and Cache-of-Thought (Wu et al., 2025a) enable position-independent reuse. By projecting KV states into a transferable space or retrieving semantically related past states, these methods allow the model to reuse expensive prefill computations across different users or requests, significantly reducing latency for shared multimodal prompts.

4.5 Knowledge Distillation

Knowledge distillation (KD) has emerged as a pivotal technique for enhancing efficiency and reducing complexity in multimodal learning by transferring rich representations from sophisticated teacher models to lighter student models. We group KD for multimodal systems into (i) prediction-level (logits/soft labels), (ii) representation-level (feature/attention/relational alignment), and (iii) behavior-level (reasoning traces, preferences, or policy signals)

Prediction-level. Prediction-level distillation transfers multimodal knowledge by aligning the output probability distributions of teacher and student models. It serves as the most implementation-friendly strategy, enabling efficiency gains without altering the student's internal structure. Formally, the objective minimizes the divergence between the teacher's soft logits z_T and the student's logits z_S, often combined with ground-truth supervision:

$$\mathcal{L}_{\mathrm{KD}} = \alpha T^2 \, \mathrm{KL}\big(\sigma(z_T/T) \,\|\, \sigma(z_S/T)\big) + (1-\alpha)\, \mathrm{CE}\big(y,\, \sigma(z_S)\big), \tag{5}$$

where σ denotes the softmax function, T is the temperature parameter controlling distribution smoothness, and α balances the Kullback–Leibler (KL) divergence against the standard Cross-Entropy (CE) loss. While early approaches focus on simple logit matching for domain adaptation (Miech et al., 2021; Kang et al., 2025), recent works extend this to semantic grounding. PromptKD (Li et al., 2024d) employs prompt-based supervision to distill task priors without labeled data, while FDBPL (Zhang et al., 2025c) introduces region-aware binary prompts to transfer fine-grained spatial decision signals. These methods provide a lightweight baseline for replicating reasoning behavior through output imitation.

Representation-level. Going beyond final predictions, representation-level distillation aligns intermediate features—such as attention maps, token embeddings, and relational matrices—to transfer the teacher's internal knowledge. Early works like CLIP-KD (Yang et al., 2024a) and TinyCLIP (Wu et al., 2023) validate direct feature matching to compress vision-language backbones. Recent advances target deeper structural alignment: DSMD (Liang et al., 2025b) employs dynamic scheduling to synchronize feature evolution, while SF-CLIP (Sameni et al., 2024) uses masked distillation to focus learning on salient spatial regions. Optimization also extends to architectural adaptation; MobileCLIP (Vasu et al., 2024) introduces dataset-level caching for efficient training, and mm-Mamba (Liao et al., 2025) utilizes progressive alignment to transfer Transformer-based knowledge into linear-time SSMs. Collectively, these strategies elevate KD from output mimicry to representational geometry transfer, preserving alignment fidelity under strict compute constraints.

Behavior-level. The most advanced form of KD transfers the teacher's underlying reasoning behaviors—how it decomposes problems, ranks alternatives, or formulates chains of thought (CoT). This paradigm captures decision dynamics rather than static snapshots, making it essential for complex instruction following. Systems such as Skywork R1V (Peng et al., 2025) introduce adaptive rationale-length supervision to balance completeness and conciseness in CoT generation, while VPD (Hu et al., 2024) employs visual–programmatic distillation to transfer explicit reasoning traces for structured tasks. Furthermore, preference-based methods like LLaVA-MoD (Shu et al., 2024) apply ranking distillation, where the student learns from the teacher's comparative judgments rather than full text reconstruction. Together, these methods mark a shift from predictive mimicry to cognitive emulation, allowing compact learners to approximate the deliberative, preference-driven reasoning of large VLMs.

4.6 Prompting and Speculative Decoding

This category of methods accelerates multimodal generation by optimizing the input conditioning logic rather than pruning the model structure. It encompasses two complementary strategies: learning compact prompts to adapt frozen backbones (reducing training/storage cost) and employing speculative drafting to parallelize decoding (reducing inference latency).

Parameter-Efficient Prompting. Prompt learning turns adaptation into a continuous optimization problem over the input space. Early approaches treat prompts as static global adapters, injecting a small set of learnable tokens into frozen backbones to enable task transfer with negligible parameter cost (Jia et al., 2022; Li & Liang, 2021). Prefix Tuning (Li & Liang, 2021) formalizes this by attaching task-specific key–value prefixes at each Transformer layer, effectively steering the attention mechanism without modifying weights. Addressing the limitations of static prompts, CoCoOp (Zhou et al., 2022) introduces context- and instance-aware prompting, where the conditioning tokens evolve dynamically based on image features. This evolution from static to dynamic prompting allows models to achieve specialized performance with extreme parameter efficiency, often updating less than 1% of the total weights.

Multimodal Speculative Acceleration. Speculative decoding accelerates inference by breaking the memory-bound sequential dependency of autoregressive generation. It employs a lightweight draft model (or a prompt-conditioned head) to propose multiple tokens cheaply, which are then verified in parallel by the full target model (Yang et al., 2025a; Hu et al., 2025; Wang et al., 2025d). In multimodal contexts, efficiency is maximized by identifying shared resources: since the visual encoder is computationally heavy but static during decoding, systems share the visual KV cache between the drafter and the verifier. This allows the draft model to focus solely on linguistic prediction, creating a draft-and-verify loop that amortizes the high cost of loading the full model parameters over multiple accepted tokens per step.

4.7 Runtime Sparsity

Runtime sparsity optimizes efficiency by dynamically pruning the computation graph during inference based on input complexity. Unlike static model pruning, which permanently removes parameters, runtime sparsity exploits the variance in sample difficulty—allocating full compute only to hard samples while processing easy ones via lightweight pathways. Structurally, these methods operate along two primary axes: reducing network depth (layer skipping or early exit) and sparsifying token connectivity (attention masking).

Layer Skipping. Layer skipping modulates computational depth by conditionally bypassing redundant Transformer blocks for easy tokens or frames. Observations suggest that many visual tokens stabilize early in the network, rendering deep processing unnecessary. Heuristic approaches like Skip-Vision (Zeng et al., 2025) prune low-impact visual tokens and their KV entries based on accumulated attention scores. Policy-based methods, such as D-ViTDMETR (Uzkent et al., 2023), employ reinforcement learning to make discrete execute-or-skip decisions per layer, reducing FLOPs by over 50% with minimal degradation. Recent advances treat depth as a continuous routing dimension; MoLe-VL (Zhang et al., 2025b) and λ-MoD (Luo et al., 2024) learn sparse activation paths based on token entropy or spatiotemporal salience. Similarly, mPLUG (Li et al., 2022) utilizes structural shortcuts to enable speculative partial-depth execution, demonstrating that adaptive depth can effectively balance representation power with inference speed.

Early Exit and Adaptive Termination. Early exit mechanisms transform fixed-depth backbones into dynamic cascades, allowing inference to terminate at intermediate layers once a confidence threshold is met. In vision-language tasks, frameworks like DeeCap (Fei et al., 2022) demonstrate that shallow layers often contain sufficient semantic information for simple captioning instances. More rigorously, MuE (Tang et al., 2023a) formalizes this via convergence monitoring, halting computation when cross-modal representations saturate. In temporal or embodied contexts, efficiency is driven by stability; DeeR (Yue et al., 2024) halts processing for video frames or robotic actions once the policy distribution stabilizes over time. While some solutions like AdaInfer (Fan et al., 2024) rely on parameter-free statistical checks, others like CREMA (Yu et al., 2024) and CAPEEN (Bajpai & Hanawal, 2024) integrate exit decisions into the training loop, distilling knowledge from deep layers to shallow exits to ensure consistent performance regardless of termination depth.

Attention Sparsity and Merging. Attention sparsity mitigates the quadratic complexity of self-attention by restricting connection density. Static methods like ETESparse (Dai et al., 2021a) utilize fixed-pattern constraints, while content-adaptive approaches such as LoRA-Sparse (Song et al., 2024) dynamically compute attention only over top-ranked keys to minimize redundant computation. Complementing these structural reductions, merging paradigms offer a training-free pathway for efficiency through representation aggregation. At the model level, TIES-Merging (Yadav et al., 2023) and related benchmarks (Sung et al., 2023a) consolidate diverse parameter sets from multiple checkpoints to enhance performance without adding inference overhead. At the token level, LLaVA-PruMerge (Shang et al., 2025) introduces an adaptive framework that synergistically combines cross-modal attention-based pruning with similarity-driven merging. By dynamically identifying task-relevant visual tokens and aggregating redundant features, it significantly compresses the input space for the LLM while maintaining high-fidelity multimodal reasoning. Together, these methods optimize connectivity by either pruning non-essential interactions or aggregating similar representations into dense, informative states.

4.8 Discussion and Key Insights

Algorithm-level efficiency optimizes the execution dynamics of multimodal models by actively modulating the density and precision of information flow. Our synthesis of recent advancements reveals three critical insights into how algorithmic strategies move beyond post-hoc compression toward intelligent, adaptive computation:

- **Exploiting Asymmetric Redundancy for Token Economy:** A core pillar of algorithmic efficiency is the identification of asymmetric redundancy across modalities. While textual tokens are semantically dense and sequential, visual tokens often exhibit high spatial and temporal correlation. Effective token compression and selective computing (Wen et al., 2025) succeed by treating visual patches not as independent units, but as a hierarchical graph of information. This allows models to achieve high pruning rates by focusing computational budgets on "high-entropy" regions while aggressively aggregating static background tokens.
- **Sensitivity-Aware Discretization in Heterogeneous Spaces:** The primary challenge in multimodal quantization lies in the divergent statistical distributions of visual and textual features. Our analysis suggests that the most robust quantization strategies—such as sensitivity-based mixed precision and outlier suppression—succeed by acknowledging this heterogeneity (Xue et al., 2025). Rather than applying a uniform bit-width, these methods protect the delicate cross-modal alignment by maintaining higher precision in channels or tokens that act as "semantic anchors", while quantizing redundant activations to ultra-low-bit levels.
- **From Predictive Mimicry to Cognitive Emulation:** The paradigm shift in KD from simple logit-matching to behavior-level emulation (Peng et al., 2025) highlights a deeper goal: transferring the reasoning process itself. By distilling reasoning traces, preference signals, and chain-of-thought (CoT) behaviors, algorithmic efficiency enables compact student models to approximate the deliberative reasoning of large-scale teachers. This transforms efficiency from a reduction of FLOPs into a maximization of cognitive throughput.

Ultimately, the trajectory of algorithm-level innovation heralds a decisive paradigm shift toward cognitive-aware dynamic orchestration. By transcending the rigid constraints of static execution through the strategic synergy of runtime sparsity and speculative decoding, next-generation EML frameworks are poised to resolve the fundamental "Efficiency-Utility-Alignment" trilemma. Within this nascent regime, algorithms will function as elastic reasoning engines, autonomously and natively modulating computational expenditure in direct response to the latent semantic complexity of the multimodal stream.

5 System

Distinct from model and algorithm optimizations, system-level approaches focus on resource orchestration—determining *where* and *when* workloads should execute under physical constraints. This layer operationalizes efficiency by balancing infrastructure trade-offs between latency, memory, and energy, as shown in Fig. 5. We examine critical strategies for scalable deployment: context-aware memory management, edge–cloud collaboration, latency-sensitive scheduling, hardware–software co-design, and federated learning.

5.1 KV Cache Management and Serving

In production environments, KV cache management is the primary determinant of system throughput and maximum concurrent users. Profiling studies by Lee et al. (Lee et al., 2025b) reveal that for long-context workloads, end-to-end latency is governed by memory bandwidth and kernel launch overheads rather than arithmetic intensity. To address this, modern serving engines integrate `torch.compile`, CUDA graph execution, and FlashAttention to fuse kernels and minimize scheduling gaps. Building on this, research focuses on orchestrating the lifecycle of KV states to maximize hardware utilization.

High-Throughput Memory Pooling. Instead of viewing compression merely as a way to reduce FLOPs, system-level approaches treat it as a mechanism to increase batch size and GPU occupancy. FastCache (Zhu et al., 2025) introduces a stream-aware memory pool that dynamically manages KV blocks across concurrent decoding sessions. By optimizing the physical layout of cached states, it mitigates memory fragmentation—a common issue in dynamic length generation—thereby removing I/O bottlenecks to support high-throughput serving. Similarly, MEDA (Wan et al., 2025) utilizes per-layer entropy to guide adaptive memory allocation, ensuring that scarce VRAM is physically reserved for information-dense layers, increasing the maximum supported sequence length on a single device.

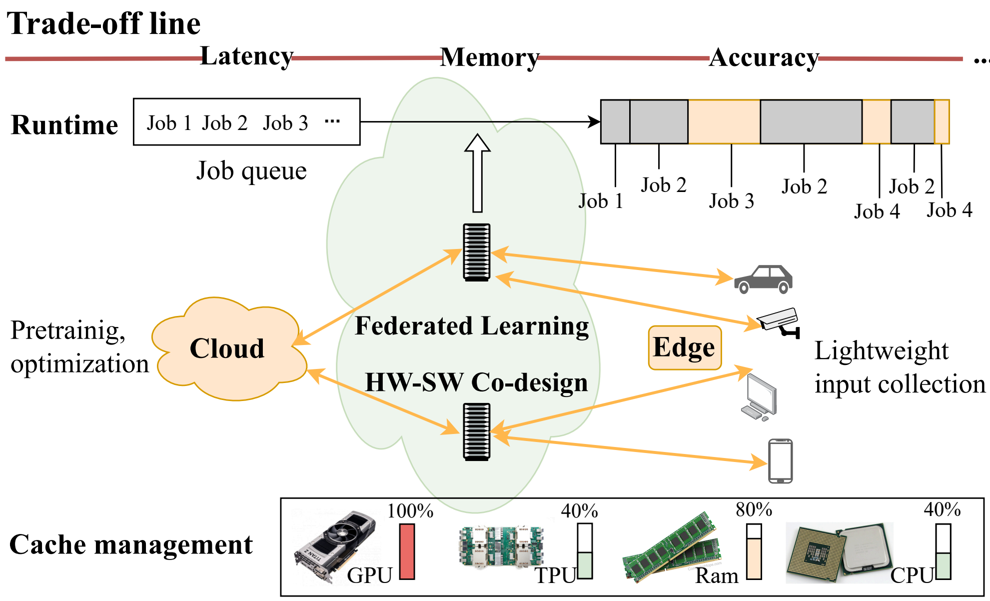

Figure 5: System-level efficiency for elastic resource orchestration. This framework illustrates the final operationalization of EML, where theoretical gains are translated into realized performance across five primary axes defined in our MAS taxonomy: (i) **KV Cache Management and Serving** to decouple memory growth from sequence length and optimize throughput; (ii) **Edge-cloud Collaboration** for establishing hierarchical cognitive pipelines and uncertainty-guided offloading; (iii) **Latency-Aware Scheduling and Pipelining** to maximize hardware utilization by reordering and overlapping cross-modal requests; (iv) **Hardware-software (HW-SW) Co-design** to natively align model architectural topology with the physical constraints of heterogeneous accelerators; and (v) **Federated Learning** to enable privacy-preserving, communication-efficient training across distributed, heterogeneous clients. Collectively, these strategies transform static multimodal execution into a dynamic, hardware-aware ecosystem.

Shared State Serving and Persistence. For multi-turn or multi-user scenarios, the system bottleneck shifts to the repeated loading of redundant contexts. Reuse-oriented architectures decouple the logical KV state from its physical storage, enabling zero-copy sharing across requests. MPIC (Zhao et al., 2025c) implements position-independent caching, allowing the serving backend to map multiple user prompts to the same shared physical memory block, regardless of their position indices. This significantly reduces the memory bandwidth requirement for prefix loading. At the industrial scale, Gemini 1.5 (Team et al., 2024) demonstrates the necessity of unified state management, where a centralized cache orchestration layer enables persistent context reuse across massive, heterogeneous multimodal streams. This transition from per-request computation to persistent shared memory transforms the KV cache from a temporary buffer into a globally managed system asset.

5.2 Edge–Cloud Collaboration

Real-world multimodal deployment operates under a fundamental tension: edge devices suffer from compute limitations, while cloud servers face latency and bandwidth bottlenecks. Edge–cloud collaboration resolves this by establishing a hierarchical computing architecture that dynamically partitions workloads based on resource availability and privacy concerns.

Adaptive Inference Offloading. This paradigm treats the edge as a low-latency semantic filter and the cloud as a high-capacity reasoning engine. System-level frameworks like EdgeCloudAI (Ghasemi et al., 2024) and MoA (Yang et al., 2025f) implement split-computing architectures. Instead of transmitting raw video streams, lightweight edge modules (e.g., CNNs or small VLMs) perform preliminary analytics to filter redundant frames or extract feature embeddings. Only high-value, hard-to-resolve samples are transmitted to the cloud-hosted large model. Evaluations on testbeds like NSF COSMOS (Raychaudhuri et al., 2020) demonstrate that this adaptive partitioning significantly reduces bandwidth consumption and end-to-end latency while maintaining near-cloud accuracy.

Collaborative Continuous Learning. Beyond static inference, system efficiency also entails maintaining model relevance over time without incurring massive data transfer costs. Frameworks for Cloud–Device Collaborative Adaptation (Gan et al., 2023) introduce uncertainty-guided interaction mechanisms. Here, edge devices utilize uncertainty quantification to identify out-of-distribution or ambiguous samples. Only these "hard" examples are uploaded for cloud-based labeling and training, after which a distilled, lightweight update is synchronized back to the edge. This bidirectional loop minimizes communication overhead and preserves privacy by keeping routine data local, ensuring that the system continuously adapts to evolving environments with minimal operational cost.

5.3 Latency-Aware Scheduling and Pipelining

Achieving low-latency multimodal inference requires moving beyond simple sequential execution to sophisticated scheduling that maximizes hardware utilization. Multimodal models introduce a unique system challenge: the "stalled pipeline" problem, where the large language model (LLM) often sits idle waiting for the visual encoder to process high-resolution images. System-level scheduling addresses this by optimizing the timeline of execution—reordering, overlapping, and routing requests to hide latency bubbles.

Asynchronous Pipelining and Overlap. To alleviate the computational stalls inherent in sequential cross-modal dependencies, recent system-level frameworks prioritize fine-grained temporal parallelism. RServe (Guo et al., 2025a) introduces a split-scheduling architecture that actively interleaves compute-intensive vision encoding with the language prefill phases of concurrent requests. By decoupling these execution stages and managing per-request embeddings asynchronously, it effectively masks the latency of visual encoding, yielding a 2× throughput improvement over traditional sequential serving. For high-throughput streaming environments, Inf-MLLM (Ning et al., 2024) employs a streaming-aware scheduler that dynamically governs the token memory lifecycle through adaptive KV retention. This methodology enables sustained long-context generation on a single GPU while significantly curtailing the latency spikes characteristic of heuristic eviction policies such as H2O (Zhang et al., 2023c).

Heterogeneous Routing and Tiering. In distributed environments, latency is optimized by routing requests to the most appropriate resource tier. Yuan et al. (Yuan et al., 2025) propose a decoupled architecture for edge-cloud systems, where lightweight encoders run on edge devices and LLMs on servers. They employ a Gaussian Process-Upper Confidence Bound (GP-UCB) scheduler to dynamically select the optimal offloading target and power configuration, minimizing energy consumption under strict latency constraints. At the production scale, Amazon Nova (Langford et al., 2025) adopts a tiered deployment strategy (Micro, Lite, Pro, Premier). By routing queries based on complexity—sending simple captioning tasks to cheaper models and complex reasoning to larger ones—the system satisfies strict Service Level Agreements (SLAs) while optimizing the global cost-latency trade-off.

5.4 Hardware-Software Co-design

As multimodal models scale, system efficiency is increasingly constrained by the mismatch between algorithmic requirements and hardware capabilities. Specifically, multimodal workloads exhibit diverse *arithmetic intensities*: dense layers in LLMs are typically compute-bound, while high-resolution visual encoders or attention mechanisms are often memory-bound. Hardware-software co-design addresses this by aligning the model's operational structure with the physical topology of the underlying accelerators.

Workload Characterization and Mapping. The first strategy involves systematically profiling operators to map them onto the most suitable heterogeneous hardware units. Golden et al. (Golden et al., 2024) demonstrate that treating all layers uniformly leads to resource underutilization. By characterizing the compute intensity and bandwidth demands of multimodal kernels, they propose a heterogeneity-aware mapping strategy. Compute-dense modules (e.g., large MLPs) are routed to high-FLOP units (like GPUs), while bandwidth-bound operations (e.g., normalization) are allocated to memory-rich processors or near-memory compute units. This approach minimizes data movement—the primary energy consumer in modern chips—by ensuring that each hardware component processes the workload type it was designed for.

Joint Architecture-Hardware Search. Instead of fitting a fixed model onto hardware, the second strategy uses Neural Architecture Search (NAS) to co-optimize the model structure and its deployment parameters. Harmonic-NAS (Ghebriout et al., 2024) exemplifies this joint optimization paradigm. It searches for modality-specific backbones and fusion networks under strict hardware constraints (e.g., latency, energy, or peak memory on an edge device). By incorporating hardware feedback directly into the search loop, it discovers a Pareto frontier of architectures that maximize accuracy within specific physical budgets. This effectively shifts the design process from a sequential "train-then-deploy" workflow to a unified co-design loop, producing hybrid architectures natively efficient for their target deployment platforms.

5.5 Federated Learning

In scenarios involving sensitive data—such as medical imaging or personal sensor streams—centralized training is often precluded by privacy regulations and bandwidth constraints. Federated Multimodal Learning (FML) addresses this by training models across distributed clients while keeping raw data local. From a system perspective, the primary bottleneck in FML is the communication overhead of synchronizing large multimodal backbones. To mitigate this, the standard aggregation protocol (FedAvg) is often adapted to exchange only lightweight updates. Formally, in communication round t, the server aggregates updates from K clients as:

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_t^{(k)}, \qquad n = \sum_{k=1}^{K} n_k, \tag{6}$$

where w represents the learnable parameters (often restricted to adapters or prompts rather than full weights), and n_k denotes the local sample count.

Communication-Efficient Tuning. To avoid transmitting massive gradients of foundational models, recent systems focus on parameter-efficient federated tuning. FedCLIP (Lu et al., 2023) and PFedPrompt (Guo et al., 2023) freeze heavy multimodal backbones and exchange only lightweight adapters or prompt vectors. This reduces communication payload by orders of magnitude (from GBs to MBs per round) while utilizing the generalization power of pre-trained models. Similarly, CreamFL (Yu et al., 2023) and FedMEKT (Le et al., 2025) perform aggregation at the representation level (knowledge distillation) rather than the parameter level, further decoupling communication cost from model size.

Heterogeneity and Missing Modalities. A unique system challenge in FML is device heterogeneity: different clients may possess different sensor configurations (e.g., some have cameras, others only IMUs). Frameworks like FedCMR (Zong et al., 2021) and PmcmFL (Bao et al., 2023) address this via missing-modality robustness mechanisms. PmcmFL aligns local embeddings to a shared prototype library, enabling the global model to aggregate knowledge from heterogeneous clients even when specific modalities are absent locally. Additionally, participant selection strategies like mmFedMC (Yuan et al., 2024) estimate the Shapley value of each client's modality contribution, ensuring that the system prioritizes high-quality, diverse updates to maximize convergence speed under bandwidth limits.

Modality-Specific Privacy Disentanglement. For sensor-rich environments, system efficiency also involves disentangling private user attributes from shared task features (Dai et al., 2024). Approaches like FDARN (Yang et al., 2022) and MCARN (Yang et al., 2024c) employ dual-encoder architectures to separate modality-agnostic (shared) features from modality-specific (private) noise. By aggregating only the shared components, these systems reduce the dimensionality of the uploaded update and enhance privacy guarantees, ensuring that the system learns generalizable patterns without overfitting to device-specific noise.

5.6 Discussion and Key Insights

System-level efficiency marks the final operationalization of EML, where theoretical algorithmic gains are translated into realized performance under physical constraints. Our synthesis reveals that at this layer, efficiency has evolved from a localized optimization of FLOPs into a global orchestration of data movement, temporal alignment, and trust:

- **Transitioning from Arithmetic to I/O-Bound Resource Management:** In production-scale MLLM serving, the primary bottleneck has shifted from arithmetic intensity to memory bandwidth and I/O overheads. Modern system efficiency is defined by the fluidity of the KV cache lifecycle. By treating the KV cache as a globally managed, persistent asset rather than a transient per-request buffer, systems can decouple memory growth from context length, transforming memory fragmentation into a manageable pool of high-throughput tokens.
- **Hierarchical Cognitive Pipelining via Edge-Cloud Symbiosis:** Edge-cloud collaboration transcends simple workload partitioning; it establishes a hierarchical cognitive pipeline mirroring biological sensory systems. By utilizing the edge as a low-latency semantic filter to prune modal redundancy, the system reserves cloud-scale reasoning for high-value outliers. This synergy ensures that efficiency is maintained through continuous, uncertainty-guided adaptation across the device-cloud continuum, effectively balancing the trilemma of latency, bandwidth, and accuracy.
- **Structural-Temporal Symbiosis as a Design Necessity:** The "stalled pipeline" problem in multimodal workloads exposes the mismatch between fixed-topology accelerators and dynamic cross-modal dependencies. True efficiency emerges when the model's structural topology is natively aligned with its temporal execution. Through joint architecture-hardware co-design and modality-aware scheduling, the design process evolves into a unified co-design loop that eliminates pipeline bubbles and minimizes data movement—the primary energy consumer in modern infrastructure.
- **Privacy-Efficiency-Utility Equilibrium in Decentralized EML:** As EML extends to sensitive domains, privacy is no longer a post-hoc constraint but a systemic dimension of efficiency. FML requires a delicate equilibrium: optimizing communication costs through parameter-efficient tuning while maintaining robustness against missing modalities. The future lies in privacy-aware resource economy, where modality-specific feature disentanglement enables secure, high-fidelity collaboration without the overhead of centralized data aggregation.

Ultimately, these insights posit that the future of EML lies in *elastic resource orchestration*—a self-regulating infrastructure that dynamically reconfigures its compute, memory, and communication pathways to satisfy the volatile, multi-objective demands of real-world multimodal intelligence.

6 Efficient MLLMs

Rather than an isolated research niche, Efficient MLLMs serve as the ultimate proving ground for our MAS framework. The evolution of this field, as visualized in Fig. 6, reveals a distinct maturity curve: a trajectory that begins with **architectural consolidation**, pivots toward **algorithmic refinement**, and is currently converging on **system-level orchestration**.

6.1 The Evolutionary Trajectory: From Models to Systems

The chronological progression of MLLM efficiency (see Fig. 6) reflects the shifting bottlenecks of large-scale efficient multimodal intelligence:

- **The Foundational Era (Model-Centric):** Initial efforts primarily targeted the training bottleneck. Early frameworks like BLIP-2 (Li et al., 2023c) and LLaVA (Liu et al., 2023a) focused on minimizing the cost of cross-modal alignment through modular adapters and frozen backbones. This era established the principle of "architectural frugality", where efficiency was synonymous with parameter-efficient fine-tuning.
- **The Inference Pivot (Algorithm-Centric):** As MLLMs moved toward deployment, the bottleneck shifted to memory and latency limits. Algorithm-level strategies matured from static pruning

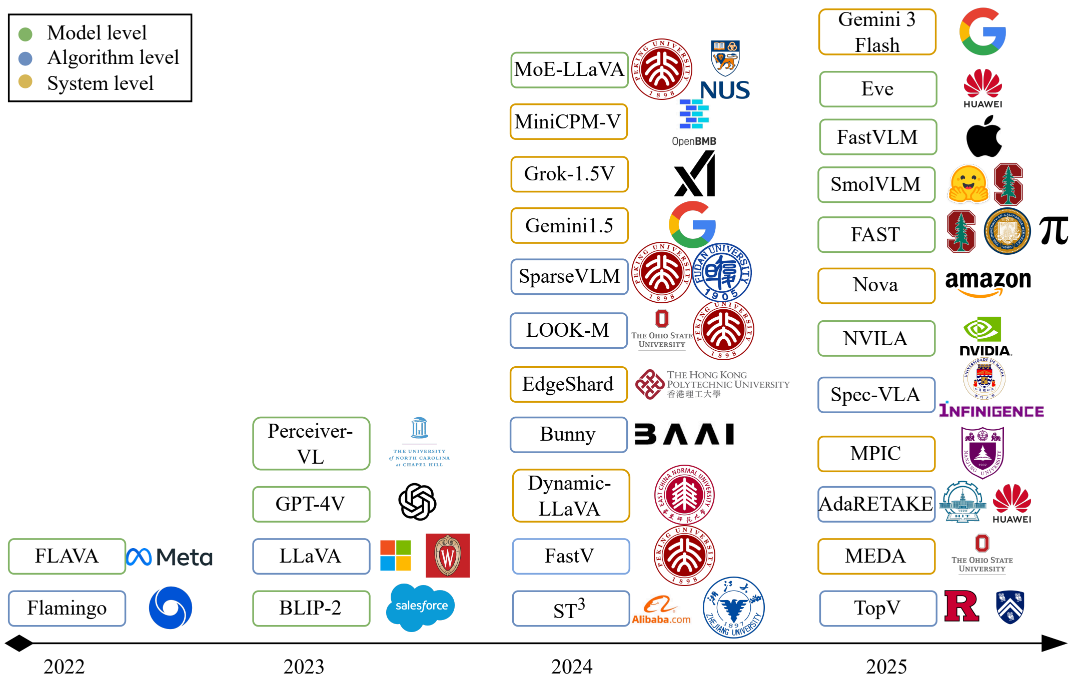

Figure 6: Chronological overview of representative efficient MLLMs. Models are categorized by primary optimization level: **model-level** (green), **algorithm-level** (blue), and **system-level** (orange). This distribution highlights a distinct paradigm shift, where model- and algorithm-level optimizations are dominant in the early stages, while system-level resource orchestration has gained significant prominence in recent years.

to dynamic runtime reduction. Innovations like LOOK-M (Wan et al., 2024) and Spec-VLA (Wang et al., 2025d) demonstrate a move toward "semantic-aware execution", where computation is selectively allocated to critical tokens (e.g., text-guided KV merging) to sustain long-context reasoning within fixed hardware envelopes.

- **The Deployment Frontier (System-Centric):** Most recently, we observe a significant surge in system-level optimizations—a trend driven by the transition from lab prototypes to high-throughput production. Works such as MiniCPM-V (Yao et al., 2024), EdgeShard (Zhang et al., 2024a), MPIC (Zhao et al., 2025c), and the speed-oriented co-design of Gemini 3-Flash (Google, 2025) highlight that the current frontier lies in resource orchestration. This shift acknowledges that even the most efficient model can fail in real-world scenarios without asynchronous pipelining and elastic memory management.

6.2 Discussion: Vertical Synergy as the Final Frontier

Our analysis of the MLLM landscape reveals three key insights regarding the synergy between MAS layers:

- **Decoupling Capacity from Cost:** Structural sparsity (e.g., MoE-LLaVA (Lin et al., 2024a)) represents a successful synthesis of model-level topology and algorithmic routing. By decoupling total parameters from active FLOPs, MLLMs are evolving into "sparse-active" systems that can scale in knowledge without scaling in inference latency.
- **The Convergence of Perception and Scheduling:** Recent systems like FastVLM (Vasu et al., 2025) and NVILA (Liu et al., 2025b) suggest that the boundary between "visual perception" and "system scheduling" is blurring. By designing architectures that are natively compatible with CUDA-graph execution and kernel fusion, these models ensure that architectural efficiency translates directly into wall-clock speed gains.

- **Towards Self-Regulating Computation:** The emerging trend of adaptive efficiency (e.g., MiniCPM-V (Yao et al., 2024)) suggests a future where MLLMs learn not only to reason, but to *self-regulate*. In this paradigm, efficiency becomes an internal optimization objective where the model autonomously decides the optimal path across the MAS stack—choosing *what* (Model), *how* (Algorithm), and *where* (System) to compute based on real-time constraints.

To conclude, efficient MLLMs serve as the genesis of multimodal systems where efficiency is no longer an afterthought, but an intrinsic, self-regulating property of intelligent computation.

7 Applications and Benchmarks

After detailing the methodological advances in the preceding sections, we now turn to the diverse application domains and benchmarking ecosystems that serve as the primary drivers for EML. As summarized in Table 1, these domains are not merely downstream tasks but represent distinct frontiers where the efficiency-performance trade-off is governed by specific physical and operational constraints. From the millisecond-latency requirements of autonomous driving to the memory-intensive horizons of long-video understanding, these applications necessitate a tight integration of MAS optimizations.

7.1 Affective Computing

Affective computing serves as a critical frontier for EML, where the objective of inferring nuanced human emotions from tri-modal cues—audio, visual, and linguistic—is fundamentally constrained by temporal transience and stringent privacy imperatives (Wang et al., 2025c). The core efficiency challenge in this domain lies in the high-frequency and ephemeral nature of affective signals, which necessitates a transition from computationally expensive monolithic encoders toward modular, on-device adaptation. Within our MAS framework, affective computing exemplifies a classic case of cross-layer synergy: model-level adapter tuning minimizes the memory footprint of backbone networks to enable rapid task specialization, while algorithm-level innovations—such as the dynamic token pruning and quantized inference utilized in UGotMe (Li et al., 2025b)—exploit the inherent spatiotemporal redundancy in facial and vocal features to achieve real-time responsiveness. This structural decoupling ensures that the generative reasoning engine is only invoked for high-entropy emotional transitions, while routine behavioral monitoring is offloaded to efficient primitives. Ultimately, these advancements move toward a paradigm of empathetic edge intelligence, where system-level constraints on latency and power consumption dictate the selection of lightweight fusion strategies, ensuring that affective intelligence remains continuous, responsive, and privacy-preserving without reliance on cloud-based orchestration.

7.2 Embodied AI and Robotics

Embodied AI and robotics represent the physical manifestation of EML, where the sensory-motor loop demands absolute synchronization between high-dimensional perception and real-time action generation under strict latency envelopes. The primary efficiency bottleneck in this domain arises from the explosive computational complexity of Vision-Language-Action (VLA) trajectories, necessitating a transition from static mapping to resource-aware execution (Chen et al., 2025c). Within our MAS framework, foundational systems like PaLM-E (Li et al., 2025e) and RT-2 (Zitkovich et al., 2023) exemplify how model-level structural priors and system-level asynchronous pipelining coalesce to hide the latency of compute-intensive visual encoding behind lightweight action decoding loops. This evolution highlights a broader movement toward adaptive sensor orchestration, where the agent dynamically modulates its perceptual resolution or sensor sampling rate based on task uncertainty. By utilizing the MAS system layer to prioritize computational budgets for high-stakes manipulation or navigation sub-goals, these systems effectively bridge the gap between abstract understanding and embodied reality, ensuring that multimodal intelligence remains responsive and scalable within the physical bounds of mobile hardware.

Table 1: A Taxonomy of application domains and benchmarks in EML. This table highlights the diverse modality combinations—ranging from canonical audio-visual pairs to complex sensor streams—that necessitate domain-specific MAS optimization strategies to balance computational fidelity with deployment constraints. V: Visual, T: Text, A: Audio, S: Spatial (LiDAR+GPS+Radar), D: Depth, Act: Action, Sen: Sensor, Tab: Tables, Cha: Charts.

Applications	**Task**	**Benchmark**	**Modalities**
Affective Computing	Sentiment Analysis, Emotion & Behavior Recognition	CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), CH-SIMS (Yu et al., 2020)	V+T+A
Embodied AI & Robotics	VLN, Manipulation, Embodied Agents	R2R (Anderson et al., 2018), REVERIE (Qi et al., 2020), ALFRED (Shridhar et al., 2020), Ego4D (Grauman et al., 2022), TEACh (Padmakumar et al., 2022), Habitat (Savva et al., 2019)	V+D, T+S+Act
Media Understanding & Generation	Video QA, Streaming Analysis, Retrieval & Generation	MSR-VTT (Xu et al., 2016), TVR (Lei et al., 2020b), Video-MME (Fu et al., 2024), MVBench (Li et al., 2024b), LAION-5B (Schuhmann et al., 2022), WebVid2M (Bain et al., 2021), TVQA+ (Lei et al., 2020a), AVA ActiveSpeaker (Roth et al., 2020)	V+T+A
Healthcare	Radiology Analysis, Clinical Speech, Digital Twin	MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2015), PPMI (Marek et al., 2011), MedVidQA (Gupta et al., 2023), mPower (Bot et al., 2016)	V+T+A+Sen
Spatial Understanding	Autonomous Driving, 3D Scene Reconstruction	nuScenes (Caesar et al., 2020), Waymo Open Dataset (Sun et al., 2020), KITTI (Geiger et al., 2012), Argoverse (Chang et al., 2019), Drive&Act (Martin et al., 2019)	V+S+IMU
Multimodal Reasoning	VQA, Math Reasoning, Visual Dialogue	VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), VCR (Zellers et al., 2019), ScienceQA (Lu et al., 2022), MathVista (Lu et al., 2024b), ChartQA (Masry et al., 2022)	Tab+Cha+V+T

7.3 Media Understanding and Generation

Media understanding and generation drive the frontiers of EML by interpreting and synthesizing content across vast spatiotemporal horizons, where efficiency is governed by the quadratic complexity of multimodal attention (Ye et al., 2025a; He et al., 2024a). The core challenge in this domain is to exploit the massive inherent redundancy in high-resolution video and audio streams to enable long-context reasoning without memory collapse. Recent innovations pivot from dense global modeling toward content-aware execution, utilizing algorithmic token pruning and temporal windowing—as demonstrated in Video-ChatGPT (Maaz et al., 2024) and CLIPER (Wu et al., 2024b)—to linearize the computational footprint of high-resolution media. This domain serves as a primary proving ground for MAS system-level orchestration, where the reuse of KV cache states and the dynamic offloading of cold tokens allow MLLMs to sustain effectively infinite context windows for streaming analysis. By aligning model-level structural sparsity with system-level memory persistence, media systems are evolving into responsive, long-form interpreters capable of real-time retrieval and generative QA across massive, heterogeneous archives.

7.4 Healthcare

In clinical intelligence, efficiency is a structural necessity due to massive imaging data, strict privacy regulations, and limited on-site compute. The goal is to maximize diagnostic fidelity while minimizing the computational footprint of multimodal stacks. Recent advances demonstrate a shift toward model-level specialization (Zhao et al., 2026); for instance, pretraining frameworks like BioViL-T (Bannur et al., 2023) extend beyond static image-report pairs to model longitudinal radiology studies, improving report grounding through the structural alignment of temporal data. Complementing this, algorithm-level innovations focus on augmenting compact backbones with structured expertise. Knowledge-enhanced systems like KARGEN (Li et al., 2024c) fuse disease graphs with frozen LLMs to mitigate hallucinations, proving that specialized, instruction-tuned architectures like CXR-LLaVA (Lee et al., 2025a) can outperform massive general-purpose VLMs in radiology reporting tasks while maintaining a fraction of the inference cost. Furthermore, EML operationalizes efficiency via cost-effective curriculum learning, as evidenced by LLaVA-Med (Li et al., 2023a), which facilitates the rapid training of biomedical assistants in under 15 hours by transitioning from basic vocabulary alignment to complex reasoning. Beyond imaging, multimodal pathology models enable slide-level diagnosis, while speech- and sensor-based systems (Ji & Zhou, 2024; Ji et al., 2025a) support clinical monitoring. Ultimately, these efforts converge on a privacy-centric system orchestration, where system-level deployment on secure, on-site hardware necessitates the MAS synergy of lightweight fusion and verifiable reasoning to ensure scalable, interpretable, and ethically-compliant clinical intelligence.

7.5 Spatial Understanding

Spatial understanding, even physical understanding, represents a critical efficiency frontier where high-fidelity geometric perception—integrating LiDAR, camera, and radar signals—must operate within the rigid latency and energy envelopes of edge devices. The bottleneck in this domain lies in the inherent dimensionality of 3D sensor streams, necessitating a paradigm shift from dense volumetric fusion to projection-based efficiency. Frameworks like PointPillars (Lang et al., 2019) and BEVFusion (Liu et al., 2023b) operationalize this by projecting sparse 3D point clouds into compact 2D Bird's-Eye-View (BEV) grids, significantly reducing memory bandwidth and kernel launch overheads while preserving the geometric fidelity essential for safe navigation. This evolution highlights a deeper synergy between model topology and hardware-software co-design, where the architectural layout is natively aligned with the memory-access patterns of automotive-grade accelerators to prevent the "stalled pipeline" problem in multi-sensor synchronization. Furthermore, algorithmic innovations like FusionPainting (Xu et al., 2021) refine this process through sparse hybrid fusion, utilizing cross-modal semantic "painting" to minimize redundant computation by focusing resources on informative regions rather than exhaustive global processing. Together, these advancements move toward a heterogeneity-aware paradigm, where the system orchestrates the mapping of compute-dense modules to high-FLOP units and memory-bound operations to near-memory compute units, ultimately enabling real-time, efficient spatial intelligence that maintains high-resolution grounding under tight compute budgets.

7.6 Multimodal Reasoning

Multimodal reasoning marks the critical transition from sensory perception to symbolic intelligence, necessitating the synthesis of heterogeneous evidence for multi-step decision-making. The primary bottleneck of monolithic transformers in this domain is the propensity for semantic hallucinations and logic collapses during multi-step inference, which traditionally necessitates massive parameter redundancy to maintain stability. To address this, the field is shifting from dense statistical mapping—exemplified by early cross-attention models like UNITER (Chen et al., 2020)—toward modular functional decoupling. Frameworks such as ViperGPT (Surís et al., 2023) and VisProg (Gupta & Kembhavi, 2023) operationalize this by redefining reasoning as a programmatic execution problem, where the LLM serves as a high-level planner that invokes specialized, low-latency sub-routines only for necessary perceptual sub-goals. This structural evolution ensures that expensive generative resources are reserved for deliberative logic while routine sensory grounding is offloaded to efficient primitives, effectively embodying a "Thinking Fast and Slow" paradigm within the MAS system layer. Furthermore, in high-stakes domains such as mathematics and physics, specialized architectures like MathGLM-vision (Yang et al., 2024d) and ScienceQA (Lu et al., 2022) integrate verifiable process rewards to prune erroneous reasoning paths early, demonstrating how structured decomposition allows compact models to achieve high symbolic fidelity while minimizing the cumulative computational footprint of long-horizon tasks. Ultimately, these advancements move toward a neuro-symbolic synergy where efficiency is realized through the strategic orchestration of a hierarchical model stack, balancing expressive power with verifiable resource economy.

8 Holistic Discussion: Methodological Synthesis and Application-Driven Insights

The preceding analysis of model, algorithm, and system levels establishes that EML is no longer a collection of isolated optimization tricks, but a sophisticated exercise in cross-layer co-design. This section synthesizes these dimensions into a unified methodological framework, exploring how vertical synergy and application-specific constraints redefine the boundaries of multimodal intelligence.

8.1 The Mechanics of Vertical Synergy

The fundamental insight of the MAS framework lies in the realization that efficiency gains are non-linear; the most profound accelerations occur at the intersection of layers rather than within them. At the **model-algorithm interface**, we observe that structural decisions—such as the transition from dual-stream encoders to unified, sequence-based foundations—do not merely simplify the architecture but fundamentally reshape the optimization landscape for algorithmic compression. Unified foundations provide a homogeneous semantic space that mitigates the modality-specific outlier distributions that historically plagued quantization and pruning. Furthermore, the **algorithm-system nexus** reveals that theoretical compression (e.g., sub-4-bit quantization or token eviction) only translates into realized speedups when natively supported by hardware-aware execution kernels. The evolution from generic matrix multiplication toward specialized tensor-core utilization and kernel fusion highlights that algorithmic sparsity must be structured and hardware-aligned to bypass the memory-wall constraints of modern accelerators. Ultimately, true efficiency emerges when the model's structural topology is designed to be natively compatible with the system's execution dynamics, ensuring that every theoretical efficiency manifests as a measurable gain in wall-clock throughput.

8.2 Application-Driven Tactical Blueprints

The deployment of EML systems is rarely a pursuit of universal optimality; instead, it is a pragmatic navigation of the "Efficiency-Utility-Privacy" trilemma, dictated by the unique physical and regulatory constraints of specific domains. In **Latency-Critical environments** such as embodied AI and autonomous driving, the blueprint prioritizes structural-temporal symbiosis. Here, the system must utilize the edge-cloud continuum not as a simple storage tier, but as a hierarchical cognitive pipeline where low-level sensory filtering at the edge prevents bandwidth-induced decision stalls. Conversely, in **Fidelity-Critical domains** like medical imaging and scientific discovery, the optimization focus shifts toward precision-preserving algorithms and federated system architectures. In these contexts, the methodological goal is to ensure that aggressive discretization does not erode the delicate semantic nuances required for diagnostic accuracy, while simultaneously utilizing decentralized learning to bypass the overhead of massive data aggregation. Finally, **Throughput-Oriented cloud services** demand a decoupling of capacity from cost, favoring structural sparsity (e.g., MoE) and aggressive cache management. By aligning the resource lifecycle with the momentary semantic complexity of user queries, these systems can sustain massive concurrent horizons within finite hardware budgets, demonstrating that the "optimal" MAS configuration is a dynamic, domain-specific equilibrium.

8.3 Reframing Efficiency: Toward Self-Regulating Intelligence

Looking beyond current methodologies, the synthesis of MAS layers points toward a fundamental reframing of efficiency: the transition from post-hoc adjustments to intrinsic, self-regulating properties. Historically, efficiency was treated as a secondary constraint—a "patch" applied to a pre-trained model for deployment. However, the emergence of natively efficient foundations and adaptive execution graphs suggests a future where intelligence and efficiency are inseparable. We envision a regime of cognitive-aware orchestration, where multimodal systems possess an internal representation of their own computational cost and hardware environment. In this paradigm, the model autonomously decides *what* information to process (Model), *how* to compress the computation (Algorithm), and *where* to execute the workload (System) based on real-time uncertainty and resource availability. This evolution toward self-regulating computation marks the final stage of the transition from models to systems, where efficiency is no longer an external metric to be minimized, but an emergent property of the model's fundamental structural design—an inherent parsimony that mirrors the biological efficiency of the human brain in processing the multimodal complexity of the physical world.

9 Open Challenges and Future Directions

Despite rapid progress in EML domain, several challenges remain unresolved. We summarize key questions and outline promising directions to drive the next generation of scalable, deployable, and intelligent multimodal systems—across both understanding and generation tasks.

9.1 Unified Tokenization Across Modalities

Unified multimodal scaling faces a critical bottleneck in *tokenization*—the translation of continuous signals into discrete transformer-compatible units. Current modality-specific tokenizers often suffer from semantic misalignment and disparate sequence lengths, destabilizing compute budgets and hindering efficient unified pretraining. A pivotal frontier is universal tokenization, which aligns diverse modalities within a shared semantic space while dynamically adapting granularity to satisfy strict latency and memory constraints. Promising directions include shared discrete codebooks for semantic alignment, variable-rate hierarchical strategies to compress redundancy, and tokenizer-scheduler co-design to optimize KV-cache reuse and streaming inference. Future benchmarks must standardize evaluation under fixed token budgets to rigorously assess these unified efficiency gains.

9.2 Multimodal Multi-Task Generalization & Robustness

Current efficient models are often optimized for specific modality pairs (e.g., vision–language), leaving their transferability to new tasks or missing-modality scenarios unproven. A critical challenge is ensuring that efficiency gains are intrinsic to the architecture rather than overfitted to a specific dataset configuration. Future research must move beyond narrow, single-domain evaluations to establish cross-modal efficiency benchmarks. These should rigorously stress-test transferability—such as training on one modality set and evaluating resource-accuracy trade-offs on another—to ensure robust deployment across heterogeneous real-world conditions.

9.3 Hardware-Software Co-Design and Deployment

Bridging the gap between algorithmic efficiency and physical realization requires a paradigm shift toward hardware-software co-design. While optimization kernels for unimodal tasks (e.g., CNNs, LLMs) are mature, hardware support for multimodal interaction mechanisms—such as high-bandwidth cross-attention and dynamic modality switching—remains sparse. Future research must move beyond generic optimizations to develop multimodal-native accelerators and compiler-aware strategies, including operator fusion for fusion layers and hardware-friendly sparsity patterns. Furthermore, deploying on constrained platforms (mobile, AR/VR, robotics) necessitates robust edge–cloud orchestration to dynamically balance on-device latency with cloud-scale reasoning.

9.4 Human-Centric and Perceptual Efficiency

True efficiency extends beyond minimizing computational overhead to maximizing Quality of Experience (QoE) for the end-user. Current metrics (e.g., FLOPs, latency) often fail to capture human-centric constraints such as cognitive load, perceptual latency, and fairness. For instance, in real-time multimodal interaction, users are often sensitive to response fluidity and initial time-to-token rather than total throughput. A critical frontier lies in aligning system-level optimization with human perception thresholds. This involves redefining loss functions to penalize latency spikes that disrupt cognitive flow, and designing adaptive systems that trade off imperceptible fidelity drops for gains in interactivity. Future research must bridge the gap between system efficiency (resource usage) and user efficacy (satisfaction and interpretability).

9.5 Privacy-Aware Efficiency and Security

The pursuit of efficiency often necessitates trade-offs that jeopardize security, such as offloading inference to the cloud or employing aggressive compression that may inadvertently widen the attack surface for model inversion. Current research on jointly optimizing efficiency and privacy—such as secure aggregation or compressed encrypted inference—remains sparse, particularly for high-dimensional, sensitive modalities like medical imaging and voice biometrics. A critical future direction is the co-design of efficiency and privacy, moving beyond post-hoc defenses. This includes developing architectural sparsification (e.g., pruning, routing) that is intrinsically resistant to membership inference attacks, and designing multi-objective systems that dynamically optimize latency, energy, and privacy budgets. Ultimately, the field must address the "Privacy-Efficiency-Utility" trilemma to enable trusted multimodal deployment.

10 Conclusion

This survey presents a comprehensive roadmap for Efficient Multimodal Learning, synthesizing over 300 studies into a unified Model–Algorithm–System taxonomy. We demonstrate that efficiency arises from the synergistic co-design of compact architectures, adaptive algorithms, and hardware-aware orchestration, rather than isolated optimizations. This reveals a critical paradigm shift: efficiency is evolving from a post-hoc constraint into an intrinsic design primitive. This work guides future work in navigating performance–cost trade-offs, paving the way for multimodal systems that are capable, robust, sustainable, and deployable across real-world applications.

Broader Impact Statement

The rapid proliferation of multimodal learning, especially MLLMs, has ushered in a critical "efficiency wall", where the exponential demand for computational power and memory threatens the scalability, accessibility, and sustainability of multimodal intelligence. This survey, through the systematic lens of the MAS taxonomy, delivers a multi-faceted impact on the research community and broader society.

1. Advancing the Academic Paradigm Toward Full-stack Co-design. The primary intellectual merit of this work lies in transforming Efficient Multimodal Learning (EML) from a collection of isolated, modality-specific heuristics into a structured, engineering discipline. By establishing the MAS framework, this paper provides a unified blueprint for "vertical synergies"—encouraging researchers to move beyond single-layer optimizations toward hardware-aware architectural search and algorithm-system co-design. This holistic perspective is essential for demystifying the complexity of multimodal interactions and establishing a theoretical foundation for where and how efficiency can be injected without collapsing semantic fidelity.

2. Democratization of High-Capability AI. Multimodal intelligence is currently concentrated within resource-rich industrial laboratories due to the prohibitive costs of training and serving massive models. This survey fosters the democratization of AI by systematizing strategies like parameter-efficient fine-tuning (PEFT), knowledge distillation, and modular adaptation. These methodologies lower the barrier to entry, empowering researchers and developers with limited compute budgets to build and specialize high-performing multimodal systems. Furthermore, the emphasis on on-device EML ensures that advanced AI capabilities are no longer tethered to massive cloud infrastructures, allowing for pervasive, decentralized intelligence in various local environments.

3. Environmental Sustainability and "Green AI". As the carbon footprint of training and deploying large-scale AI becomes a global concern, the transition toward EML is a structural necessity for environmental sustainability. By promoting techniques that linearize quadratic complexity ($O(N^2) \rightarrow O(N)$) and maximize hardware utilization per watt, this work directly supports the global effort toward "Green AI". The focus on KV-cache reuse, token compression, and persistent state management reduces the cumulative energy consumption of long-horizon multimodal generation, making massive-scale deployment socially and environmentally responsible.

4. Societal Safety, Privacy, and Human-Centric AI. The MAS framework's application-specific analysis ensures that efficiency serves human-centric goals. In the domain of healthcare and affective computing, the shift toward on-device inference facilitated by quantization and modular adaptation is a fundamental enabler for "privacy-by-design", keeping sensitive user data local and secure. Moreover, in high-stakes fields like embodied AI and autonomous systems, efficiency is synonymous with safety. By drastically reducing the "perception-action latency", the methodologies discussed in this paper ensure that intelligent agents can react to real-world stimuli in near-real-time, mitigating the risk of system failure in unpredictable environments.

In summary, this work offers a roadmap for the next generation of multimodal systems where efficiency is not a post-hoc optimization, but an intrinsic, self-regulating property that aligns artificial intelligence with the physical and ethical constraints of the human world.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738, 2021.

Divya Jyoti Bajpai and Manjesh Kumar Hanawal. Capeen: Image captioning with early exits and knowledge distillation. *arXiv preprint arXiv:2410.04433*, 2024.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Shubhang Bhatnagar, Andy Xu, Kar-Han Tan, and Narendra Ahuja. Luq: Layerwise ultra-low bit quantization for multimodal large language models. *arXiv preprint arXiv:2509.23729*, 2025.

Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Dorsey, et al. The mpower study, parkinson disease mobile data collected using researchkit. *Scientific Data*, 3(1):1–9, 2016.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 239–249, 2025.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19846–19856, 2025.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. Pumer: Pruning and merging tokens for efficient vision language models. *arXiv preprint arXiv:2305.17530*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757, 2019.

Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. Comodo: Cross-modal video-to-imu distillation for efficient egocentric human activity recognition. *arXiv preprint arXiv:2503.07259*, 2025a.

Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. Efficient large multimodal models via visual context compression. *Advances in Neural Information Processing Systems*, 37: 73986–74007, 2024a.

Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1110–1119, 2024b.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. Efficientqat: Efficient quantization-aware training for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10081–10100, 2025b.

Mingfei Chen, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Ziyi Kou, Yuan Tian, Eric Whitmire, Rajinder Sodhi, et al. Flowing from reasoning to motion: Learning 3d hand trajectory prediction from egocentric human interaction videos. *arXiv preprint arXiv:2512.16907*, 2025c.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024c.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Zizhao Chen, Yeqiang Qian, Xiaoxiao Yang, Chunxiang Wang, and Ming Yang. Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2025d.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Qian Dai, Dong Wei, Hong Liu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1445–1453, 2024.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5305–5316, 2021a.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021b.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sayna Ebrahimi, Sercan O Arik, Tejas Nama, and Tomas Pfister. Crome: cross-modal adapters for efficient multimodal llm. *arXiv preprint arXiv:2408.06610*, 2024.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*, 2024.

Marco Federici, Riccardo Del Chiaro, Boris van Breugel, Paul Whatmough, and Markus Nagel. Stamp: Sequence transformation and mixed precision for low-precision activation quantization. *arXiv preprint arXiv:2510.26771*, 2025.

Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. Deecap: Dynamic early exiting for efficient image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12216–12226, 2022.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2024.

Yulu Gan, Mingjie Pan, Rongyu Zhang, Zijian Ling, Lingran Zhao, Jiaming Liu, and Shanghang Zhang. Cloud-device collaborative adaptation to continual changing environments in the real-world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12157–12166, 2023.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

Mahshid Ghasemi, Zoran Kostic, Javad Ghaderi, and Gil Zussman. Edgecloudai: Edge-cloud distributed video analytics. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pp. 1778–1780, 2024.

Mohamed Imed Eddine Ghebriout, Halima Bouzidi, Smail Niar, and Hamza Ouarnoughi. Harmonic-nas: Hardware-aware multimodal neural architecture search on resource-constrained devices. In *Asian Conference on Machine Learning*, pp. 374–389. PMLR, 2024.

Alicia Golden, Samuel Hsia, Fei Sun, Bilge Acun, Basil Hosmer, Yejin Lee, Zachary DeVito, Jeff Johnson, Gu-Yeon Wei, David Brooks, et al. Generative ai beyond llms: System implications of multi-modal generation. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 257–267. IEEE, 2024.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, pp. 571–575, 2021.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022.

Google. Gemini 3 flash: The next generation of fast, efficient multimodal models. https://blog.google/products/gemini/gemini-3-flash/, December 2025. Accessed: 2025-12-18.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9233–9243, 2023.

Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pp. 1364–1374, 2023.

Tianyu Guo, Tianming Xu, Xianjie Chen, Junru Chen, Nong Xiao, and Xianwei Zhang. Rserve: Overlapping encoding and prefill for efficient lmm inference. *arXiv preprint arXiv:2509.24381*, 2025a.

Zirun Guo, Xize Cheng, Yangyang Wu, and Tao Jin. A wander through the multimodal landscape: Efficient transfer learning via low-rank sequence multimodal adapter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16996–17004, 2025b.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37:67850–67900, 2024.

Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. In *the Eleventh International Conference on Learning Representations*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Xuehai He, Jian Zheng, Jacob Zhiyuan Fang, Robinson Piramuthu, Mohit Bansal, Vicente Ordonez, Gunnar A Sigurdsson, Nanyun Peng, and Xin Eric Wang. FlexEControl: Flexible and efficient multimodal control for text-to-image generation. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856.

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipvl: Efficient large vision-language models with dynamic token sparsification. *arXiv preprint arXiv:2410.08584*, 2024b.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 131–135. IEEE, 2017.

Musashi Hinck, Matthew L Olson, David Cobbley, Shao-Yen Tseng, and Vasudev Lal. Llava-gemma: Accelerating multimodal foundation models with a compact language model. *arXiv preprint arXiv:2404.01331*, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *the Tenth International Conference on Learning Representations*, 2022.

Yunhai Hu, Zining Liu, Zhenyuan Dong, Tianfan Peng, Bradley McDanel, and Sai Qian Zhang. Speculative decoding and beyond: An in-depth survey of techniques. *arXiv preprint arXiv:2502.19732*, 2025.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

Kai Huang, Hao Zou, Bochen Wang, Ye Xi, Zhen Xie, and Hao Wang. Aircache: Activating inter-modal relevancy kv cache compression for efficient large vision-language model inference. *arXiv preprint arXiv:2503.23956*, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs. In *the Tenth International Conference on Learning Representations*, 2022.

Hui Ji and Pengfei Zhou. Advancing ppg-based continuous blood pressure monitoring from a generative perspective. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pp. 661–674, 2024.

Hui Ji, Wei Gao, and Pengfei Zhou. Translation from wearable ppg to 12-lead ecg. *arXiv preprint arXiv:2509.25480*, 2025a.

Yuheng Ji, Yue Liu, Zhicheng Zhang, Zhao Zhang, Yuting Zhao, Xiaoshuai Hao, Gang Zhou, Xingwei Zhang, and Xiaolong Zheng. Enhancing adversarial robustness of vision-language models through low-rank adaptation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 550–559, 2025b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Bootstrapping vision-language learning with decoupled language pre-training. *Advances in Neural Information Processing Systems*, 36:57–72, 2023.

Hongcheng Jiang and Zhiqiang Chen. Flexible window-based self-attention transformer in thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3076–3085, 2024.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4075–4083, 2025.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Seongjae Kang, Dong Bok Lee, Hyungjoon Jang, and Sung Ju Hwang. Simple semi-supervised knowledge distillation from vision-language models via dual-head optimization. *arXiv preprint arXiv:2505.07675*, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Anas Anwarul Haq Khan, Utkarsh Verma, Prateek Chanda, and Ganesh Ramakrishnan. Early exit and multi stage knowledge distillation in vlms for video summarization. *arXiv preprint arXiv:2504.21831*, 2025.

Junhan Kim, Chungman Lee, Eulrang Cho, Kyungphil Park, Ho-young Kim, Joonyoung Kim, and Yongkweon Jeon. Towards next-level post-training quantization of hyper-scale transformers. *Advances in Neural Information Processing Systems*, 37:94292–94326, 2024.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

Aaron Langford, Aayush Shah, Abhanshu Gupta, Abhimanyu Bhatter, Abhinav Goyal, Abhinav Mathur, Abhinav Mohanty, Abhishek Kumar, Abhishek Sethi, Abi Komma, et al. The amazon nova family of models: Technical report and model card. *arXiv preprint arXiv:2506.12103*, 2025.

Huy Q Le, Minh NH Nguyen, Chu Myaet Thwal, Yu Qiao, Chaoning Zhang, and Choong Seon Hong. Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning. *Neural Networks*, 183:107017, 2025.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pp. 1–13, 2025a.

Yejin Lee, Alicia Golden, Anna Sun, Basil Hosmer, Bilge Acun, Can Balioglu, Changhan Wang, Charles David Hernandez, Christian Puhrsch, Daniel Haziza, et al. Characterizing and efficiently accelerating multimodal generation model inference. *IEEE Micro*, 2025b.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, 2020a.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pp. 447–463. Springer, 2020b.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7241–7259, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023a.

DaiXun Li, Weiying Xie, Yunsong Li, and Leyuan Fang. Fedfusion: Manifold-driven federated learning for multi-satellite and multi-modality fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2023b.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:131224–131246, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023c.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Peizhen Li, Longbing Cao, Xiao-Ming Wu, Xiaohan Yu, and Runze Yang. Ugotme: An embodied system for affective human-robot interaction. In *2025 IEEE International Conference on Robotics and Automation*, pp. 5542–5548. IEEE, 2025b.

Shiyao Li, Yingchun Hu, Xuefei Ning, Xihui Liu, Ke Hong, Xiaotao Jia, Xiuhong Li, Yaqi Yan, Pei Ran, Guohao Dai, et al. Mbq: Modality-balanced quantization for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4167–4177, 2025c.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, 2018.

Tianxing Li, Jin Huang, Erik Risinger, and Deepak Ganesan. Low-latency speculative inference on distributed multi-modal data streams. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 67–80, 2021.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, pp. 1–19, 2025d.

Xiang Li, Tao Qin, Jian Yang, and Tie-Yan Liu. Lightrnn: Memory and computation-efficient recurrent neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Xiaoqi Li, Jingyun Xu, Mingxu Zhang, Jiaming Liu, Yan Shen, Iaroslav Ponomarenko, Jiahui Xu, Liang Heng, Siyuan Huang, Shanghang Zhang, et al. Object-centric prompt-driven vision-language-action model for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27638–27648, 2025e.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 382–392. Springer, 2024c.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025f.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024d.

Yinan Liang, Ziwei Wang, Xiuwei Xu, Jie Zhou, and Jiwen Lu. Efficientllava: Generalizable auto-pruning for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9445–9454, 2025a.

Zhengyang Liang, Meiyu Liang, Wei Huang, Yawen Li, Wu Liu, Yingxia Shao, and Kangkang Lu. Dynamic self-adaptive multiscale distillation from pre-trained multimodal large model for efficient cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2851–2859, 2025b.

Bencheng Liao, Hongyuan Tao, Qian Zhang, Tianheng Cheng, Yingyue Li, Haoran Yin, Wenyu Liu, and Xinggang Wang. Multimodal mamba: Decoder-only multimodal state space model via quadratic to linear distillation. *arXiv preprint arXiv:2502.13145*, 2025.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024a.

Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muyi Sun, Linqi Song, Ying Wei, and Zhenan Sun. Mope-clip: Structured pruning for efficient vision-language models with module-wise pruning error metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27370–27380, 2024b.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024c.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Federated learning for vision-and-language grounding problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11572–11579, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023a.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. *arXiv preprint arXiv:2502.15910*, 2025a.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation*, pp. 2774–2781. IEEE, 2023b.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4122–4134, 2025b.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024b.

Zuyan Liu, Benlin Liu, Jiahui Wang, Yuhao Dong, Guangyi Chen, Yongming Rao, Ranjay Krishna, and Jiwen Lu. Efficient inference of vision instruction-following models with elastic cache. In *European Conference on Computer Vision*, pp. 54–69. Springer, 2024c.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *the Twelfth International Conference on Learning Representations*, 2024b.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36:29615–29627, 2023.

Yaxin Luo, Gen Luo, Jiayi Ji, Yiyi Zhou, Xiaoshuai Sun, Zhiqiang Shen, and Rongrong Ji. $\gamma-$mod: Exploring mixture-of-depth adaptation for multimodal large language models. *arXiv preprint arXiv:2410.13859*, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in Neurobiology*, 95(4):629–635, 2011.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2801–2810, 2019.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2021.

Stavros Mitsis, Ermos Hadjikyriakos, Humaid Ibrahim, Savvas Neofytou, Shashwat Raman, James Myles, and Eiman Kanjo. Transformer redesign for late fusion of audio-text features on ultra-low-power edge hardware. *arXiv preprint arXiv:2510.18036*, 2025.

Shentong Mo, Paul Pu Liang, Russ Salakhutdinov, and Louis-Philippe Morency. Multiiot: Towards large-scale multisensory learning for the internet of things. *arXiv preprint arXiv:2311.06217*, 2023.

Shentong Mo, Louis-Philippe Morency, Ruslan Salakhutdinov, and Paul Pu Liang. Iot-lm: Large multisensory language models for the internet of things. *arXiv preprint arXiv:2407.09801*, 2024.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. Inf-mllm: Efficient streaming inference of multimodal large language models on a single gpu. *arXiv preprint arXiv:2409.09086*, 2024.

Yasmine Omri, Parth Shroff, and Thierry Tambe. Token sequence compression for efficient multimodal computing. *arXiv preprint arXiv:2504.17892*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2017–2025, 2022.

Xiaohuan Pei, Tao Huang, and Chang Xu. Cross-self kv cache pruning for efficient vision-language inference. *arXiv preprint arXiv:2412.04652*, 2024.

Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, 2018a.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, 2018b.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Miao Rang, Zhenni Bi, Chuanjian Liu, Yehui Tang, Kai Han, and Yunhe Wang. Eve: Efficient multimodal vision language models with elastic visual experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6694–6702, 2025.

Dipankar Raychaudhuri, Ivan Seskar, Gil Zussman, Thanasis Korakis, Dan Kilper, Tingjun Chen, Jakub Kolodziejski, Michael Sherman, Zoran Kostic, Xiaoxiong Gu, et al. Challenge: Cosmos: A city-scale programmable testbed for experimentation with advanced wireless. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pp. 1–13, 2020.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4492–4496. IEEE, 2020.

Sepehr Sameni, Kushal Kafle, Hao Tan, and Simon Jenni. Building vision-language models on solid foundations with masked distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14216–14226, 2024.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 42048–42070, 2024.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

Dachuan Shi, Chaofan Tao, Ying Jin, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Upop: Unified and progressive pruning for compressing vision-language transformers. In *International Conference on Machine Learning*, pp. 31292–31311. PMLR, 2023a.

Dachuan Shi, Chaofan Tao, Anyi Rao, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. *arXiv preprint arXiv:2305.17455*, 2023b.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(3):e70036, 2025.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. Taid: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. *arXiv preprint arXiv:2501.16937*, 2025.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10740–10749, 2020.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. Low-rank approximation for sparse attention in multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13763–13773, 2024.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *the Twelfth International Conference on Learning Representations*, 2024.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022a.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022b.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging. *arXiv preprint arXiv:2304.14933*, 2023a.

Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. *arXiv preprint arXiv:2310.02998*, 2023b.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025.

Hongxuan Tang, Hao Liu, and Xinyan Xiao. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv preprint arXiv:2503.21193*, 2025.

Shengkun Tang, Yaqing Wang, Zhenglun Kong, Tianchi Zhang, Yao Li, Caiwen Ding, Yanzhi Wang, Yi Liang, and Dongkuan Xu. You need multiple exiting: Dynamic early exiting for accelerating unified vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10791, 2023a.

Zineng Tang, Jaemin Cho, Jie Lei, and Mohit Bansal. Perceiver-vl: Efficient vision-and-language modeling with iterative latent attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4410–4420, 2023b.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. Flashsloth: Lightning multimodal large language models via embedded visual compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14570–14581, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. *arXiv preprint arXiv:2410.23317*, 2024.

Burak Uzkent, Amanmeet Garg, Wentao Zhu, Keval Doshi, Jingru Yi, Xiaolong Wang, and Mohamed Omar. Dynamic inference with grounding based vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2624–2633, 2023.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15963–15974, 2024.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19769–19780, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*, 2024.

Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang. Meda: Dynamic kv cache allocation for efficient multimodal long-context inference. *arXiv preprint arXiv:2502.17599*, 2025.

Ao Wang, Hui Chen, Jiaxin Li, Jianchao Tan, Kefeng Zhang, Xunliang Cai, Zijia Lin, Jungong Han, and Guiguang Ding. Prefixkv: Adaptive prefix kv cache is what vision instruction-following models need for efficient generation. *arXiv preprint arXiv:2412.03409*, 2024a.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Q-vlm: Post-training quantization for large vision-language models. *Advances in Neural Information Processing Systems*, 37: 114553–114573, 2024b.

Chuanming Wang, Yuxin Yang, Mengshi Qi, Huanhuan Zhang, and Huadong Ma. Towards efficient object re-identification with a novel cloud-edge collaborative framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7600–7608, 2025a.

Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quétu, Shuai Xiao, and Enzo Tartaglione. Folder: Accelerating multi-modal large language models with enhanced performance. *arXiv preprint arXiv:2501.02430*, 2025b.

Maolin Wang, Yao Zhao, Jiajia Liu, Jingdong Chen, Chenyi Zhuang, Jinjie Gu, Ruocheng Guo, and Xiangyu Zhao. Large multimodal model compression via iterative efficient pruning and distillation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 235–244, 2024c.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21180–21188, 2025c.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Songsheng Wang, Rucheng Yu, Zhihang Yuan, Chao Yu, Feng Gao, Yu Wang, and Derek F Wong. Spec-vla: speculative decoding for vision-language-action models with relaxed acceptance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26916–26928, 2025d.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5417–5432, 2025e.

Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Leo-mini: An efficient multimodal large language model using conditional token reduction and mixture of multi-modal experts. *arXiv preprint arXiv:2504.04653*, 2025f.

Yongxian Wei, Zixuan Hu, Li Shen, Zhenyi Wang, Chun Yuan, and Dacheng Tao. Open-vocabulary customization from clip via data-free knowledge distillation. In *the Thirteenth International Conference on Learning Representations*, 2025.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14205–14215, 2024a.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023.

Mingyuan Wu, Jize Jiang, Haozhen Zheng, Meitang Li, Zhaoheng Li, Beitong Tian, Bo Chen, Yongjoo Park, Minjia Zhang, Chengxiang Zhai, et al. Cache-of-thought: Master-apprentice framework for cost-effective vision language model inference. *arXiv preprint arXiv:2502.20587*, 2025a.

Penghao Wu, Lewei Lu, and Ziwei Liu. Streamline without sacrifice - squeeze out computation redundancy in lmm. In *Forty-second International Conference on Machine Learning*, 2025b.

Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *the Thirteenth International Conference on Learning Representations*, 2025c.

Xinglong Wu, Anfeng Huang, Hongwei Yang, Hui He, Yu Tai, and Weizhe Zhang. Towards bridging the cross-modal semantic gap for multi-modal recommendation. *arXiv preprint arXiv:2407.05420*, 2024b.

xAI / Grok team. Grok-1.5 vision: a preview of xai's multimodal model, 2024.

Junfei Xiao, Zheng Xu, Alan Yuille, Shen Yan, and Boyu Wang. Palm2-vadapter: progressively aligned language model makes a strong vision-language adapter. *arXiv preprint arXiv:2402.10896*, 2024.

Jingjing Xie, Yuxin Zhang, Mingbao Lin, Liujuan Cao, and Rongrong Ji. Advancing multimodal large language models with quantization-aware scale learning for efficient adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10582–10591, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022.

Jingyu Xu and Yang Wang. Fmt: A multimodal pneumonia detection model based on stacking moe framework. In *2025 8th International Conference on Information and Computer Technologies*, pp. 517–521. IEEE, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.

Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin, and Liangjun Zhang. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *2021 IEEE International Intelligent Transportation Systems Conference*, pp. 3047–3054. IEEE, 2021.

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation. *arXiv preprint arXiv:2502.02175*, 2025.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10637–10647, 2023.

Yufei Xue, Yushi Huang, Jiawei Shao, and Jun Zhang. Vlmq: Efficient post-training quantization for large vision-language models via hessian augmentation. *arXiv preprint arXiv:2508.03351*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115, 2023.

Sarthak Yadav and Zheng-Hua Tan. Audio mamba: Selective state spaces for self-supervised audio representations. In *Proceedings of Interspeech 2024*, pp. 552–556, 2024.

Chaoqun Yang, Ran Chen, Muyang Zhang, Weiguang Pang, Yuzhi Chen, Rongtao Xu, Kexue Fu, Changwei Wang, and Longxiang Gao. Aasd: Accelerate inference by aligning speculative decoding in multimodal large language models. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pp. 1–7. IEEE, 2025a.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19803–19813, 2025b.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15952–15962, 2024a.

Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. Clip-cid: Efficient clip distillation via cluster-instance discrimination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21974–21982, 2025c.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23826–23837, 2024b.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025d.

Xiaoda Yang, JunYu Lu, Hongshun Qiu, Sijing Li, Hao Li, Shengpeng Ji, Xudong Tang, Jiayang Xu, Jiaqi Duan, Ziyue Jiang, et al. Astrea: A moe-based visual understanding model with progressive alignment. *arXiv preprint arXiv:2503.09445*, 2025e.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3063–3071, 2022.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5345–5361, 2024c.

Zheming Yang, Qi Guo, Yunqing Hu, Chang Zhao, Chang Zhang, Jian Zhao, and Wen Ji. Moa-off: Adaptive heterogeneous modality-aware offloading with edge-cloud collaboration for efficient multimodal llm inference. *arXiv preprint arXiv:2509.16995*, 2025f.

Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, and Jie Tang. Mathglm-vision: solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*, 2024d.

Ziyi Yang, Mahmoud Khademi, Yichong Xu, Reid Pryzant, Yuwei Fang, Chenguang Zhu, Dongdong Chen, Yao Qian, Xuemei Gao, Yi-Ling Chen, et al. i-code v2: An autoregressive generation framework over vision, language, and speech data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1615–1627, 2024e.

Ziyi Yang, Mahmoud Khademi, Yichong Xu, Reid Pryzant, Yuwei Fang, Chenguang Zhu, Dongdong Chen, Yao Qian, Xuemei Gao, Yi-Ling Chen, et al. i-code v2: An autoregressive generation framework over vision, language, and speech data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1615–1627, 2024f.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Fanjiang Ye, Zepeng Zhao, Yi Mu, Jucheng Shen, Renjie Li, Kaijian Wang, Saurabh Agarwal, Myungjin Lee, Triston Cao, Aditya Akella, et al. Supergen: An efficient ultra-high-resolution video generation system with sketching and tiling. *arXiv preprint arXiv:2508.17756*, 2025a.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025b.

JiangYong Yu, Sifan Zhou, Dawei Yang, Shuoyu Li, Shuo Wang, Xing Hu, Chen Xu, Zukang Xu, Changyong Shu, and Zhihang Yuan. Mquant: Unleashing the inference potential of multimodal large language models via static quantization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1783–1792, 2025.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *the Eleventh International Conference on Learning Representations*, 2023.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *arXiv preprint arXiv:2402.05889*, 2024.

Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10819–10829, 2022.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Liangqi Yuan, Dong-Jun Han, Su Wang, Devesh Upadhyay, and Christopher G Brinton. Communication-efficient multimodal federated learning: Joint modality and client selection. *arXiv preprint arXiv:2401.16685*, 2024.

Xingyu Yuan, He Li, Mianxiong Dong, and Kaoru Ota. Adaptive scheduling of multimodal large language model in intelligent edge computing. *ACM Transactions on Autonomous and Adaptive Systems*, 2025.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6731, 2019.

Weili Zeng, Ziyuan Huang, Kaixiang Ji, and Yichao Yan. Skip-vision: Efficient and scalable acceleration of vision-language models via adaptive token skipping. *arXiv preprint arXiv:2503.21817*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *European Conference on Computer Vision*, pp. 698–714. Springer, 2020.

Jun Zhang, Desen Meng, Zhengming Zhang, Zhenpeng Huang, Tao Wu, and Limin Wang. p-mod: Building mixture-of-depths mllms via progressive ratio decay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3705–3715, 2025a.

Mingjin Zhang, Xiaoming Shen, Jiannong Cao, Zeyang Cui, and Shan Jiang. Edgeshard: Efficient llm inference via collaborative edge computing. *IEEE Internet of Things Journal*, 2024a.

Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18537–18546, 2023a.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510. Springer, 2022.

Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9155–9166, 2023b.

Rongyu Zhang, Menghang Dong, Yuan Zhang, Liang Heng, Xiaowei Chi, Gaole Dai, Li Du, Yuan Du, and Shanghang Zhang. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. *arXiv preprint arXiv:2503.20384*, 2025b.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023c.

Zherui Zhang, Jiaxin Wu, Changwei Wang, Rongtao Xu, Longzhao Huang, Wenhao Xu, Wenbo Xu, Li Guo, and Shibiao Xu. Fdbpl: Faster distillation-based prompt learning for region-aware vision-language models adaptation. *Expert Systems with Applications*, pp. 128577, 2025c.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10421–10429, 2025a.

Kun Zhao, Siyuan Dai, Pan Wang, Jifeng Song, Hui Ji, Chenghua Lin, Liang Zhan, and Haoteng Tang. Aligning findings with diagnosis: A self-consistent reinforcement learning framework for trustworthy radiology reporting. *arXiv preprint arXiv:2601.03321*, 2026.

Qiyan Zhao, Yue Yan, and Da-Han Wang. Textmamba: Scene text detector with mamba. *arXiv preprint arXiv:2512.06657*, 2025b.

Shiju Zhao, Junhao Hu, Rongxiao Huang, Jiaqi Zheng, and Guihai Chen. Mpic: Position-independent multimodal context caching system for efficient mllm serving. *arXiv preprint arXiv:2502.01960*, 2025c.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jianian Zhu, Hang Wu, Haojie Wang, Yinghui Li, Biao Hou, Ruixuan Li, and Jidong Zhai. Fastcache: Optimizing multimodal llm serving through lightweight kv-cache compression framework. *arXiv preprint arXiv:2503.08461*, 2025.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024a.

Yi Zhu, Zhou Yanpeng, Chunwei Wang, Yang Cao, Jianhua Han, Lu Hou, and Hang Xu. Unit: Unifying image and text recognition in one vision encoder. *Advances in Neural Information Processing Systems*, 37:122185–122205, 2024b.

Jiedong Zhuang, Lu Lu, Ming Dai, Rui Hu, Jian Chen, Qiang Liu, and Haoji Hu. St3: Accelerating multimodal large language model by spatial-temporal visual token trimming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11049–11057, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Linlin Zong, Qiujie Xie, Jiahui Zhou, Peiran Wu, Xianchao Zhang, and Bo Xu. Fedcmr: Federated cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1672–1676, 2021.

Appendix

A Case Study

To validate the descriptive and prescriptive power of the Model–Algorithm–System (MAS) framework, we conduct a retrospective analysis of a real-world edge multimodal system: the ultra-low-power audio–text emotion recognition pipeline by Mitsis et al. (Mitsis et al., 2025). While this system predates our framework, its design trajectory offers a compelling validation: under strict deployment constraints, the optimization choices naturally converge toward the principles codified in MAS. This case study demonstrates how the MAS framework serves as both an explanatory lens for existing successful designs and a principled blueprint for future resource-constrained multimodal systems.

A.1 Model Level: Topology and Primitive Selection

Hardware-Aligned Encoder Specialization. The MAS model level dictates that architectural topology must be intrinsically decoupled from redundancy. Consistent with this, the case system avoids generic large-scale backbones, opting for *modality-specific specialization*: a compact transformer with a CNN front-end for acoustics, and a lightweight keyword-spotting encoder for text. These choices reflect core MAS design primitives: early dimensionality reduction, the use of depthwise separable convolutions, and unified embedding dimensions to facilitate low-cost fusion.

Minimalist Late Fusion. To balance expressivity with latency, the system employs a fusion strategy that strictly separates representation learning from cross-modal reasoning—a key MAS recommendation for edge devices. The fusion head is implemented as a lightweight two-layer MLP:

$$h_{\text{cat}} = \text{Concat}(h^{(a)}, h^{(t)}), \quad \hat{y} = \text{Softmax}(W_2, \sigma(W_1 h_{\text{cat}})), \tag{7}$$

where σ denotes the activation function. This design minimizes parameters and avoids the quadratic complexity of cross-attention, ensuring the fusion mechanism remains agnostic to input sequence length and friendly to cache-limited hardware.

Constraint-Aware Primitive Selection. The MAS framework emphasizes that "efficiency starts at design time". This is evident in the system's architectural constraints: the exclusive use of ReLU6 activations, single-head attention mechanisms, and static tensor shapes. These are not merely algorithmic preferences but model-level structural decisions explicitly chosen to map efficiently to the specific instruction set architecture (ISA) of the target Edge TPU.

A.2 Algorithm Level: Flow Modulation and Robustness

Pre-Deployment Adaptation. The training pipeline exemplifies the MAS principle of "optimizing information flow before physical mapping". Rather than relying solely on post-training compilation, this work integrates algorithmic constraints—such as quantization-aware training (QAT)—directly into the learning loop. This ensures that the model's weights learn to accommodate the precision loss inherent to 8-bit integers.

Algorithmic Substitution for Model Capacity. Instead of scaling up model depth to handle noisy real-world data, the system leverages algorithm-level robustness strategies. Aggressive domain-targeted augmentations (e.g., frequency masking, shifting) and label smoothing are employed to mitigate the domain shift between PC-recorded training data and MCU-recorded inference data. From an MAS perspective, this demonstrates how *algorithmic complexity* (during training) can effectively substitute for *model capacity* (during inference), maintaining high accuracy without inflating the parameter budget.

A.3 System Level: Execution Orchestration

Train–Inference Consistency. A recurring failure mode in multimodal deployment is the misalignment between offline data loaders and online feature extractors. The case system adheres to the MAS requirement for *pipeline coherence* by utilizing the Silicon Labs MLTK MicroFrontend for both training and on-device inference. This ensures bit-exact consistency in spectrogram generation, eliminating a common source of system-level degradation.

Hardware-Software Co-Design. The system demonstrates the tight coupling between model architecture and hardware constraints characterized by MAS. The Edge TPU imposes strict limits: INT8-only execution, unbatched 3D tensors, and specific supported operations. The design process reflects a bidirectional optimization loop: system constraints dictated the removal of complex attention heads, while model requirements drove the selection of specific compiler directives.

Quantifiable Efficiency Gains. The final deployed system satisfies multidimensional resource budgets, validating the effectiveness of the MAS-aligned approach:

- **Latency:** ≈22 ms per inference (Real-time),
- **Storage:** 1.8 MB (INT8 quantized),
- **Memory footprint:** ≈1.4 MB peak RAM,
- **Energy:** ≈2.5 W active power.

A.4 Conclusion: MAS as a Blueprint

This analysis confirms that high-efficiency multimodal systems are not products of isolated optimizations but of vertical synergy. The case study illustrates the hierarchical reasoning inherent to MAS, which naturally yields a deployable multimodal system under real hardware constraints.

- **Model:** Define *what* to compute by selecting hardware-friendly, modality-specific topologies.
- **Algorithm:** Determine *how* to compute by embedding quantization and robustness constraints into the learning objective.
- **System:** Decide *where* to compute by enforcing pipeline consistency and respecting accelerator-specific operation sets.

B Subset of Recent Efficient Multimodal Models

Tables 2 and 3 present a curated subset of recent efficient multimodal models organized under the MAS framework. Our goal is not to enumerate all existing systems, but to highlight representative architectures that capture the dominant design patterns in practice.

Table 2: Representative efficient multimodal models summarized under the MAS framework in recent years. V: Video/Image, T: Text, A: Audio, MSM: multiple sensor modalities, Act: Action, Clin: Clinical, TS: Time series, Geo: location coordinates.

	Model	Backbone	Year	Parameters	Efficient strategy	Modalities
Model Level	Cobra (Zhao et al., 2025a)	Mamba, DINOv2-L, SigLIP	2025	2.8B/7B	Modality Specific	V+T
	LLaVA-Gemma (Hinck et al., 2024)	CLIP ViT-L+Gemma	2024	2B/7B	Modality Specific	V+T
	FW-SAT (Jiang & Chen, 2024)	Swin-Transformer–style	2024	-	Modality Specific	V+Thermal
	UGen (Tang et al., 2025)	TinyLlama	2025	1.1B	Unified	V+T
	i-Code V2 (Yang et al., 2024f)	OmniVL, WavLM Large, Z-Code	2024	1.4B	Unified	V+T+Speech
	Unified-IO 2 (Lu et al., 2024a)	Transformer+ViT-based	2024	1.1B/3.2B/6.8B	Unified	V+T+A+Act
	UNIT (Zhu et al., 2024b)	ViT-H	2024	632M	Unified	V+T
	Emu (Sun et al., 2024)	EVA-01-CLIP + LLaMA	2024	14B	Unified	V+T
	Grok-1.5V (xAI / Grok team, 2024)	Grok-1.5 LLM	2024	-	Unified	V+T
	Astrea (Yang et al., 2025e)	Vicuna-1.5, Hermes2-Yi	2025	13B/34B	Structural Sparsity	V+T
	FMT (Xu & Wang, 2025)	ResNet-50, BERT	2025	-	Structural Sparsity	V+T
	LEO-MINI (Wang et al., 2025f)	Llama3	2025	8B	Structural Sparsity	V+T
	NVILA (Liu et al., 2025b)	SigLIP, Qwen2	2025	8B/15B	Structural Sparsity	V+T
	SmolVLM (Marafioti et al., 2025)	SigLIP, SmolLM2	2025	up to 2.2B	Structural Sparsity	V+T
	Elastic EVE (Rang et al., 2025)	PanGu, ResNet-50, SigLIP, ViT-L	2025	1.8B	Structural Sparsity	V+T
	MoMa (Lin et al., 2024c)	Early-fusion multimodal Transformer	2024	1.4B/2.3B	Structural Sparsity	V+T
	MoME (Shen et al., 2024)	Vicuna-7B, CLIP, DINO, Pix2Struct	2024	7B	Structural Sparsity	V+T
	LLaVA-MoLE (Chen et al., 2024c)	LLaVA-1.5	2024	7B	Structural Sparsity	V+T
	MoE-LLaVA (Lin et al., 2024a)	MoE-LLaVA	2024	1.6B/1.8B/2.7B	Structural Sparsity	V+T
	EVE (Chen et al., 2024b)	BEiTv2-initialized	2024	-	Structural Sparsity	V+T
	RoE (Wu et al., 2025c)	LLaVA-1.5, LLaVA-HR, VILA	2024	7B	Structural Sparsity	V+T
	CuMo (Li et al., 2024a)	Mistral/Mixtral-8, CLIP-L	2024	7B–13B	Structural Sparsity	V+T
	Flex-MoE (Yun et al., 2024)	Transformer	2024	37M	Structural Sparsity	Clin
	Fuse-MoE (Han et al., 2024)	Longformer, Transformer, CNN, DenseNet	2024	-	Structural Sparsity	V+Clin+TS
	Omni-SMoLA (Wu et al., 2024a)	PaLI-X/PaLI-3	2024	5B/55B	Structural Sparsity	V+T
	Wander (Guo et al., 2025b)	BERT-base+ViT	2025	80-220M	Adapters	V+T+A
	Enhancing-LoRA (Ji et al., 2025b)	BLIP	2025	223M	Adapters	V+T
	CROME (Ebrahimi et al., 2024)	ViT-G, Vicuna, Flan-T5-XXL	2024	7B/13B/11B	Adapters	V+T
	MMA (Yang et al., 2024b)	CLIP	2024	-	Adapters	V+T
	PaLM2-VAdapter (Xiao et al., 2024)	CoCa ViT, PaLM 2	2024	1.8B/2.0B/2.8B/10.8B	Adapters	V+T
Algorithm Level	ST3 (Zhuang et al., 2025)	LLaVA-1.5	2025	7B/13B	Token Compression	V+T
	DART (Wen et al., 2025)	LLaVA-1.5/NEXT, Video-LLaVA, Qwen2-VL, MiniCPM	2025	7B/8B	Token Compression	V+T
	TopV (Yang et al., 2025b)	LLaVA-1.5, Inern-VL2, Video-LLaVA	2025	2B/7B/13B/26B	Token Compression	V+T
	VisionZip (Yang et al., 2025d)	LLaVA-1.5, LLaVA-NEXT	2025	7B/13B	Token Compression	V+T
	TokenCarve (Tan et al., 2025)	LLaVA-1.5	2025	7B/13B	Token Compression	V+T
	AdaRETAKE (Wang et al., 2025e)	LLaVA-Video, Qwen2/2.5-VL	2025	7B/72B	Token Compression	V+T
	HIReD (Arif et al., 2025)	ShareGPT4V, LLaVA-Next, LLaVA-1.5	2025	7B/13B	Token Compression	V+T
	Fit-and-Prune (Ye et al., 2025b)	LLaVA-1.5, LLaVA-NEXT, LLaVA-HR	2025	7B	Token Compression	V+T
	TOKEN (Omri et al., 2025)	LLaVA-1.5, VILA	2025	7B/13B/8B	Token Compression	V+T
	Folder (Wang et al., 2025b)	LLaVA-1.5, MiniGPT4v2, MMVP, Video-LLaVA, BLIP	2025	7B/13B	Token Compression	V+T
	FAST (Pertsch et al., 2025)	π0 VLA, OpenVLA	2025	3B/7B	Token Compression	V+T+Act
	ZipVL (He et al., 2024b)	LLaVA-Next/1.5, LongVA, Qwen-VL	2024	7B/13B	Token Compression	V+T
	MUST-Drop (Liu et al., 2024a)	LLaVA-1.5, LLaVA-Next, Video-LLaVA	2024	7B	Token Compression	V+T
	LLaVolta (Chen et al., 2024a)	CLIP ViT-L/14, Vicuna-v1.5	2024	7B	Token Compression	V+T
	PyramidDrop (Xing et al., 2024)	LLaVA-NeXT, LLaVA-1.5	2024	7B	Token Compression	V+T
	GPrune (Jiang et al., 2025)	LLaVA-NeXT	2024	8B	Token Compression	V+T
	P-Mod (Zhang et al., 2025a)	LLaVA-1.5, LLaVA-NeXT	2024	7B	Token Compression	V+T
	TokenPacker (Li et al., 2025d)	LLaVA-1.5	2025	7B/13B	Token Compression	V+T
	DeepStack (Meng et al., 2024)	Vicuna, CLIP	2024	7B/13B	Token Compression	V+T
	MoPE-CLIP (Lin et al., 2024b)	CLIP-ViT-B/32, SE-CLIP	2024	194M	Pruning	V+T
	MANU (Liu et al., 2025a)	LLaVA-1.5, Idefics2	2025	7B/8B	Pruning	V+T
	EfficientLLaVA (Liang et al., 2025a)	LLaVA-v1.5/LLaVA-SQA	2025	7B+	Pruning	V+T
	Q-VLM (Wang et al., 2024b)	MoE-LLaVA, LLaVA	2025	7B/13B/1.6B	Quantization	V+T
	MBQ (Li et al., 2025c)	LLaVA-onevision, InternVL2, Qwen2-VL	2025	7B/8B/26B/72B	Quantization	V+T
	MQuant (Yu et al., 2025)	InternVL2, Qwen/2-VL, MiniCPM, GLM,	2025	8B/9.6B/7B/9B/72B	Quantization	V+T
	VLMQ (Xue et al., 2025)	Qwen2-VL, Qwen2.5-VL, LLaVA-onevision	2025	7B/2B	Quantization	V+T
	STaMP (Federici et al., 2025)	Qwen 2.5, Llama3/3.2, PixArt-Σ, SANA	2025	0.6B/1.6B/1B/3B/8B	Quantization	V+T
	QSLAW (Xie et al., 2024)	CLIP-ViT-L, LLaMA, Vicuna	2024	7B–13B	Quantization	V+T

Table 3: Representative efficient multimodal models summarized under the MAS framework in recent years. V: Video/Image, T: Text, A: Audio, MSM: multiple sensor modalities, Act: Action, Clin: Clinical, TS: Time series, Geo: location coordinates.

	Model	Backbone	Year	Parameters	Efficient strategy	Modalities
Algorithm Level	DHO (Kang et al., 2025)	CLIP ViT-B/L, ResNet, MobileNetV2, DFN ViT-H	2025	3.5M-304M	Knowledge Distillation	V+T
	FDBPL (Zhang et al., 2025c)	CLIP-style, ViT-L/14, ViT-B/32	2025	-	Knowledge Distillation	V+T
	Comodo (Chen et al., 2025a)	TimeSformer, Mantis, MOMENT	2025	150M	Knowledge Distillation	V+Motion
	MoveKD (Cao et al., 2025)	LLaVA-1.5, LLaVA-NeXT	2025	1.7B/7B/13B	Knowledge Distillation	V+T
	mm-Mamba (Liao et al., 2025)	mmMamba-linear, mmMamba-hybrid	2025	2.7B	Knowledge Distillation	V+T
	OpenVoca (Wei et al., 2025)	ResNet-50, ResNet-152, EfficientNet	2025	-	Knowledge Distillation	V+T
	CLIP-CID (Yang et al., 2025c)	ViT-B/32, ViT-B/16, OPENCLIP ViT-bigG/1	2025	-	Knowledge Distillation	V+T
	Skywork R1V (Peng et al., 2025)	Skywork R1	2025	38B	Knowledge Distillation	V+T
	TAID (Shing et al., 2025)	Qwen2, InternVL2	2025	8B/ 72B	Knowledge Distillation	V+T
	PromptKD (Li et al., 2024d)	ViT-B/16, ViT-L/14, CLIP	2024	-	Knowledge Distillation	V+T
	DSMD (Liang et al., 2025b)	VIT, BERT	2024	197M	Knowledge Distillation	V+T
	AMFD (Chen et al., 2025d)	ResNet-18	2024	-	Knowledge Distillation	V+MSM
	CLIP-KD (Yang et al., 2024a)	ViT-B, CLIP-ViT-L	2024	350M	Knowledge Distillation	V+T
	LLavaKD (Cai et al., 2025)	SigLIP-B, Qwen1.5	2024	7B	Knowledge Distillation	V+T
	LLaVA-MoD (Shu et al., 2024)	CLIP ViT-L/14, Qwen-1.5/2	2024	8B	Knowledge Distillation	V+T
	VPD (Hu et al., 2024)	PaLI-3/PaLI-X	2024	5B/55B	Knowledge Distillation	V+T
	MEDA (Wan et al., 2025)	LLaVA-family, InternVL, LongVA	2025	7B–32B	Caching & Reuse	V+T
	FastCache (Zhu et al., 2025)	LLaVA-1.5	2025	7B	Caching & Reuse	V+T
	AirCache (Huang et al., 2025)	LLaVA-OneVision, InternVL2, Qwen2-VL	2025	1B/4B/7B/8B/26B	Caching & Reuse	V+T
	VLA-Cache (Xu et al., 2025)	OpenVLA, OpenVLA-OFT, CogAct	2025	-	Caching & Reuse	V+T+Act
	VL-Cache (Tu et al., 2024)	llava-v1.6-mistral, llava-v1.6	2024	7B/34B	Caching & Reuse	V+T
	PrefixKV (Wang et al., 2024a)	LLaVA-1.5	2024	7B/13B	Caching & Reuse	V+T
	ElasticCache (Liu et al., 2024c)	LLaVA-1.5, Qwen2-VL	2024	7B/13B	Caching & Reuse	V+T
	CSP (Pei et al., 2024)	InceptionV3, ResNet-50, Space2Vec	2024	40M	Caching & Reuse	V+Geo
	LOOK-M (Wan et al., 2024)	LLaVA-v1.5, InternVL-v1.5, MobileVLM-v2	2024	3B/7B/13B	Caching & Reuse	V+Chat
	AASD (Yang et al., 2025a)	LLaVA-1.5, LLaVA-NeXT, InternVL-1.5	2024	7B/13B	Speculative Decoding	V+T
	SpecVLA (Wang et al., 2025d)	OpenVLA	2025	7B	Speculative Decoding	V+T+ACT
	λ-MoD (Luo et al., 2024)	LLaVA-1.5, LLaVA-HR, Mini-Gemini-HD	2024	7B/13B	Runtime Sparsity	V+T
	LoRA-Sparse (Song et al., 2024)	LLaVA-1.5	2024	7B/13B	Runtime Sparsity	V+T
	LLaVA-PruneMerge (Shang et al., 2025)	LLaVA-1.5, Video-LLaVA	2025	7B/13B	Runtime Sparsity	V+T
	SkipVision (Zeng et al., 2025)	LLaVA, LLaVA-HD, CoS	2025	8B	Runtime Sparsity	V+T
	MoLe-VL (Zhang et al., 2025b)	OpenVLA, CogAct VLA	2025	7B	Runtime Sparsity	V+T+Act
	DeeVISum (Khan et al., 2025)	PaLI-Gemma2 VLMs	2025	3B/10B/28B	Runtime Sparsity	V+T+A
System Level	MEDA (Wan et al., 2025)	LLaVA-family, InternVL, LongVA	2025	7B–32B	Cache Management	V+T
	FastCache (Zhu et al., 2025)	LLaVA-1.5	2025	7B	Cache Management	V+T
	MPIC (Zhao et al., 2025c)	LLaVA-1.6	2025	7B	Cache Management	V+T
	Cache-of-Thought (Wu et al., 2025a)	GPT-4o, Qwen-VL-2, OpenFlamingo	2025	3B/7B/9B	Cache Management	V+T
	CEAM-GM (Lee et al., 2025b)	Code Llam, SeamlessM4T, Chameleon	2025	34B	Cache Management	V+T+Speech
	Gemini 1.5 (Team et al., 2024)	Gemini 1.5 Pro/Flash	2024	-	Cache Management	V+T+A
	CloudEdgeCo (Wang et al., 2025a)	any ReID, CNN	2025	-	Edge–cloud Collaboration	V+TS
	MoA (Yang et al., 2025f)	Qwen2-VL-2B, Qwen2.5-VL	2025	2B/7B	Edge–cloud Collaboration	V+T
	EdgeCloudAI (Ghasemi et al., 2024)	cloud VLM, CNN	2024	-	Edge–cloud Collaboration	V+T
	RServer (Guo et al., 2025a)	Qwen2.5-VL	2025	72B	Job Scheduling	V+T
	Amazon Nova (Langford et al., 2025)	Nova	2025	-	Job Scheduling	V+T
	ProxyV (Wu et al., 2025b)	Vicuna-1.5 InternLM2.5	2025	7B	Job Scheduling	V+T
	DivPrune (Alvar et al., 2025)	LLaVA-1.5, LLaVA-1.6	2025	7B	Job Scheduling	V+T
	Inf-MLLM (Ning et al., 2024)	Vicuna, LLaMA-2, Pythia, Chat-UniVi, Flash-VStream	2024	7B	Job Scheduling	V+T
	MCARN (Yang et al., 2024c)	CNN/RNN/GNN-style	2024	-	Federated Learning	V+Sensor
	mmFedMC (Yuan et al., 2024)	LSTM, CNN	2024	-	Federated Learning	MSM
	FedMEKT (Le et al., 2025)	MLP, LSTM	2025	-	Federated Learning	V+TS