Efficient Inference for Large Vision-Language Models: Bottlenecks, Techniques, and Prospects

Jun Zhang[1,2]✌, Yicheng Ji[1,2]✌, Feiyang Ren[1,2]✌, Yihang Li[1,2]✌, Bowen Zeng[1,2]✌, Zonghao Chen[1,2]✌, Ke Chen[1,2], Lidan Shou[1,2], Gang Chen[1], Huan Li[1,2]✉

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University

[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

{zj.cs, jiyicheng.cs, feiyangren, zbw.cs, 22521269, chenk, should, cg, lihuan.cs}@zju.edu.cn

Abstract

Large Vision-Language Models (LVLMs) enable sophisticated reasoning over images and videos, yet their inference is hindered by a systemic efficiency barrier known as *visual token dominance*. This overhead is driven by a multi-regime interplay between high-resolution feature extraction, quadratic attention scaling, and memory bandwidth constraints. We present a systematic taxonomy of efficiency techniques structured around the inference lifecycle, consisting of *encoding*, *prefilling*, and *decoding*. Unlike prior reviews focused on isolated optimizations, we analyze the end-to-end pipeline to reveal how upstream decisions dictate downstream bottlenecks, covering compute-bound visual encoding, the intensive prefilling of massive contexts, and the "visual memory wall" in bandwidth-bound decoding. By decoupling the efficiency landscape into the axes of shaping information density, managing long-context attention, and overcoming memory limits, this work provides a structured analysis of how isolated optimizations compose to navigate the trade-off between visual fidelity and system efficiency. The survey concludes by outlining four future frontiers supported by pilot empirical insights, including hybrid compression based on functional unit sensitivity, modality-aware decoding with relaxed verification, progressive state management for streaming continuity, and stage-disaggregated serving through hardware-algorithm co-design. Our literature repository is at https://github.com/SuDIS-ZJU/Efficient-LVLMs-Inference.

1 Introduction

Large Vision-Language Models (LVLMs) (Wang et al., 2024d; An et al., 2025; Wang et al., 2025d) have evolved from research artifacts into the infrastructure for complex multimodal reasoning. However, as these models scale to process fine-grained visual inputs and long-form video streams, they encounter a systemic efficiency barrier: *visual token dominance* (Yang et al., 2024b; Tao et al., 2025a; Liu et al., 2025b). Unlike text-only inputs, visual data yields orders of magnitude more tokens, pushing inference into a regime constrained not merely by compute cycles, but by the quadratic scaling of attention and the "visual memory wall" [1] (Wan et al., 2024b; Li et al., 2025d; Wang et al., 2025b).

The central thesis of this survey is that LVLM inference is not a monolithic workload, but a dynamic pipeline traversing three distinct hardware regimes: i) ***Encoding*** (specifically *visual* encoding) is compute-bound by high-resolution feature extraction; ii) ***Prefilling*** suffers from the quadratic complexity of massive visual contexts; and iii) ***Decoding*** hits the memory wall due to static, bandwidth-consuming Key-Value (KV) caches. Optimizing one stage in isolation often shifts the bottleneck elsewhere without improving end-to-end latency.

Despite the surge in interest, the current literature remains fragmented. Prior reviews have predominantly focused on isolated verticals, such as token compression techniques (Shao et al., 2025b) or efficient architectures for specific modalities (Zhou et al., 2024; Zhang et al., 2024a) [2]. These works, however, overlook the systemic interconnectivity of the inference pipeline. They lack a holistic view of how upstream decisions (e.g., encoder resolution) dictate downstream bottlenecks (e.g., decoding bandwidth), leaving a gap in understanding end-to-end efficiency.

This survey bridges this gap by advancing a unified, ***stage-wise taxonomy*** of efficient LVLM inference. We decouple the efficiency landscape into three critical axes: *shaping information density* (encoding), *managing long-context attention* (prefilling), and *overcoming memory bandwidth limits* (decoding). This framework provides a structured lens to evaluate how isolated optimizations compose, helping researchers navigate the trade-off between visual fidelity and system efficiency.

✌Equal contribution. ✉Corresponding author.

[1]For instance, a `Qwen2.5-VL-72B` processing 20 images already exceeds 40K tokens and 13 GB of cache, while a 5-second 720p video surpasses 50K tokens and 16 GB.

[2]Section 8 provides a detailed related survey discussion.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21036–21066

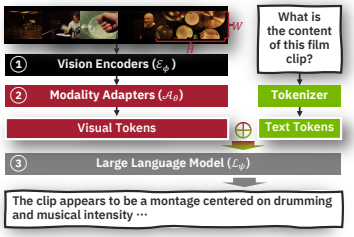

Figure 1: Three-stage pipeline for LVLM inference.

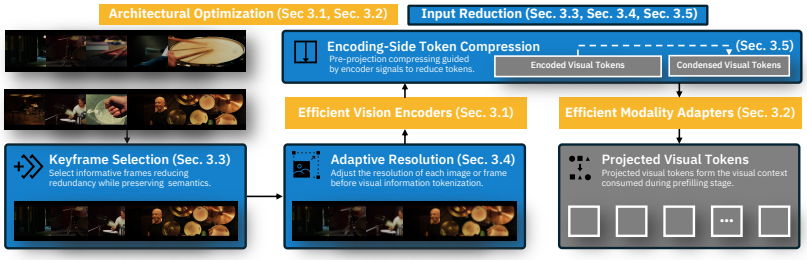

Figure 2: Efficient encoding workflow: architectural optimization (Section 3.1, Section 3.2) and input reduction (Section 3.3, Section 3.4, Section 3.5).

2 Preliminaries and Inference Dynamics

LVLMs encounter unique efficiency bottlenecks compared to Large Language Models (LLMs), primarily due to the massive visual inputs. We formalize the canonical LVLM architecture (Section 2.1) and analyze its inference dynamics through a "physics of computing" lens (Section 2.2), mapping hardware bottlenecks to user-centric metrics to structure the survey (Section 2.3).

2.1 The Canonical LVLM Architecture

LVLMs typically adopt a *three-stage pipeline* (Figure 1) that connects a vision encoder to an LLM.[3] Given a multimodal tuple $(\mathbf{V}, \mathbf{T})$ comprising raw visual input[4] $\mathbf{V} \in \mathbb{R}^{F\times H\times W\times 3}$ and a text prompt $\mathbf{T}$ of N_t tokens, the pipeline is formalized as:

① *Visual Encoding.* The encoder $\mathcal{E}_\phi$ (with parameters ϕ) processes $\mathbf{V}$ into patch embeddings $\mathbf{X}_v \in \mathbb{R}^{N_p\times D_v}$ with N_p the output patch number[5] and D_v the vision channel dimension.

② *Modality Alignment.* A modality adapter $\mathcal{A}_\theta$ (with parameters θ) maps $\mathbf{X}_v$ into the LLM's latent space, yielding visual context $\mathbf{H}_v = \mathcal{A}_\theta(\mathbf{X}_v) \in \mathbb{R}^{N_v\times D_\mathcal{L}}$ with $D_\mathcal{L}$ the LLM hidden dimension. The effective token count N_v varies by projection strategy (e.g., pooling), defined by the compression ratio $r = N_v/N_p$.

③ *Autoregressive Generation.* The LLM backbone $\mathcal{L}_\psi$ (with parameters ψ) concatenates visual and text embeddings into a joint context $\mathbf{C} = [\mathbf{H}_v; \mathbf{H}_t]$ (where $\mathbf{H}_t \in \mathbb{R}^{N_t\times D_\mathcal{L}}$ represents the prompt of N_t textual tokens) to generate the output response $\mathbf{Y} = (y_1, \ldots, y_{N_o})$ of length N_o autoregressively:

$$p(\mathbf{Y} \mid \mathbf{C}) = \prod_{k=1}^{N_o} p(y_k \mid \mathbf{C}, y_{<k}; \psi). \quad (1)$$

[3]Detailed component implementations and model taxonomy are provided in Appendix B.

[4]F frames with $(H \cdot W)$ resolution and RGB channels.

[5]For single-frame inputs ($F = 1$), $N_p = (H \cdot W)/P^2$ with patch size $(P \times P)$; for videos with $F \geq 2$, N_p depends on keyframe selection (Section 3.3), adaptive resolution (Section 3.4), and other compression strategies (e.g., frame pooling and Q-Former).

Here, a defining characteristic is the ***visual token dominance***: the visual content ($N_v \approx 576$ – 4,000+) significantly exceeds standard text prompts ($N_v \gg N_t$). This structural imbalance dictates the inference bottlenecks analyzed below.

2.2 The Physics of Inference Bottlenecks

We model the end-to-end inference latency as:

$$\tau_{\text{total}} = \tau_{\text{ENC}} + \tau_{\text{PFL}} + N_o \cdot \tau_{\text{DEC}}, \quad (2)$$

where the first two terms contribute to Time-to-First-Token (TTFT) at encoding and prefilling, respectively, and τ_{DEC} determines Time-Per-Output-Token (TPOT) at decoding. To identify bottlenecks, we apply the *Roofline model* [6], which bounds performance based on the workload's *arithmetic intensity* $\mathcal{I}$ (FLOPs/Byte). A stage is ***compute-bound*** if $\mathcal{I} \geq \pi_{\text{peak}}/\beta_{\text{mem}}$, saturating the peak compute throughput π_{peak}; otherwise, it is ***memory-bound***, throttled by the memory bandwidth β_{mem}. As summarized in Table 1, encoding is compute-bound with high arithmetic intensity from dense matrix operations; prefilling exhibits mixed behavior where both computation (quadratic attention) and memory I/O (KV cache materialization) can dominate; and decoding is strictly memory-bound due to the low arithmetic intensity of autoregressive token generation. Understanding these stage-specific bottlenecks is crucial for targeting optimization efforts.

Encoding Stage: *Compute-Bound*. This stage executes dense matrix multiplications over N_p patches, a high-intensity workload (see Table 1) strictly bounded by compute throughput:

$$\tau_{\text{ENC}} \approx \text{FLOPs}_{\text{ENC}}/\pi_{\text{peak}}. \quad (3)$$

The encoder produces N_p patch embeddings, which are then projected by $\mathcal{A}_\theta$ to yield N_v visual tokens entering the LLM. While encoding cost is constant per request (independent of N_t or N_o), reducing N_v yields *cascading benefits*: it lowers prefilling complexity from $\mathcal{O}((N_v + N_t)^2)$ to $\mathcal{O}((N_v' + N_t)^2)$ where $N_v' < N_v$ (see Table 1), and shrinks KV cache size linearly (Equation (5)).

[6]A detailed Roofline analysis is provided in Appendix E.

Table 1: Hardware bottlenecks, arithmetic intensity, and complexity dynamics across the three inference stages.

Characteristic	Encoding Stage	Prefilling Stage	Decoding Stage
Primary Metric	TTFT	TTFT	TPOT
Bottleneck	**Compute-Bound**	**Compute & Memory Bound**	**Memory-Bound**
Arithmetic Intensity	High ($\gg 1$)	Medium	Low ($\ll 1$)
Complexity (FLOPs)	$\mathcal{O}(N_p \cdot D_v^2)$	$\mathcal{O}((N_v + N_t)^2 \cdot D_\mathcal{L})$	$\mathcal{O}((N_v + N_t) \cdot D_\mathcal{L})$
LVLM Challenge	High-res inputs ($N_p \uparrow$) surge FLOPs	($N_v \gg N_t$) causes quadratic spikes	Static visual KV cache saturates VRAM

Prefilling Stage: *Compute & Memory Bound.* This stage processes the context $\mathbf{C}$ to populate the initial Key-Value (KV) cache. While attention computation is quadratic, the *materialization* of the KV cache for massive visual tokens creates a heavy memory write burden. The latency is determined by the bottleneck resource:

$$\tau_{\text{PFL}} \approx \max\left(\frac{\text{FLOPs}_{\text{attn}}}{\pi_{\text{peak}}}, \frac{|\mathcal{KV}|_{\text{PFL}}}{\beta_{\text{mem}}}\right), \quad (4)$$

where $|\mathcal{KV}|_{\text{PFL}} \approx 2 \cdot L \cdot (N_v + N_t) \cdot D_\mathcal{L} \cdot r_{\text{kv}} \cdot \mathcal{P}$ represents the bytes written to HBM. Here, L is the number of layers, $\mathcal{P}$ is the element size in bytes, and r_{kv} denotes the ratio of KV heads to Query heads (i.e., $r_{\text{kv}} = 1$ for MHA, $r_{\text{kv}} < 1$ for GQA/MQA). Unlike text-only prefilling, a large N_v can push this stage towards the memory wall.

Decoding Stage: *Memory-Bound.* Generating each output token necessitates streaming the model weights ψ and the accumulated KV cache from HBM to on-chip SRAM. The KV cache size at generation step i ($1 \le i \le N_o$) is dynamic:

$$|\mathcal{KV}|_i \approx 2 \cdot L \cdot (N_v + N_t + i) \cdot D_\mathcal{L} \cdot r_{\text{kv}} \cdot \mathcal{P}. \quad (5)$$

This stage is strictly memory-bound due to low arithmetic intensity (batch size $\approx$ 1), with the single-step latency $\tau_{\text{DEC}}^{(i)}$ and total decoding latency τ_{DEC} defined as [7]:

$$\tau_{\text{DEC}} = \sum_{i=1}^{N_o} \tau_{\text{DEC}}^{(i)}, \quad \text{where} \quad \tau_{\text{DEC}}^{(i)} \approx \left(|\psi| + |\mathcal{KV}|_i\right)/\beta_{\text{mem}}. \quad (6)$$

Here, $|\psi|$ is the model weights size. The *visual memory wall* arises because the visual component $|\mathcal{KV}|_v$ (where $|\mathcal{KV}|_v \propto N_v \cdot L \cdot D_\mathcal{L}$) necessitates the repeated loading of massive static states, dominating memory bandwidth consumption throughout the entire generation process (N_o generation steps).

2.3 Survey Organization

Given this bottleneck analysis, we organize the remainder of the survey around the stage-aware taxonomy illustrated in Figure 3: Section 3 examines upstream techniques on architectural optimization and input reduction to minimize τ_{ENC} and reduce N_v; Section 4 focuses on mitigating quadratic computation via token compression and sparse attention; and Section 5 addresses the memory-bound decoding stage via KV cache optimization, speculative execution, and efficient reasoning. Crucially, we distill empirical insights from each section into a set of **Key Takeaways** in Appendix A, which serve as the foundation for future directions discussed in Section 6. As essential supplementary references, Appendix B details architectural taxonomies while Appendix C presents system-level serving and evaluation frameworks.

3 Efficiency Techniques at Encoding

Guided by the workflow in Figure 2, this section surveys efficiency techniques for the LVLM encoding stage, structured into two strategic axes: i) *architectural optimization* focuses on designing vision encoders $\mathcal{E}_\phi$ (Section 3.1) and adapters $\mathcal{A}_\theta$ (Section 3.2) to minimize the on-model tokenization latency τ_{ENC}; and ii) *input reduction* explores optimized visual token representations to reduce the number of visual tokens N_v entering the downstream pipeline, including keyframe selection (Section 3.3), adaptive resolution (Section 3.4), and encoding-side token compression (Section 3.5).

3.1 Efficient Vision Encoders

The vision encoder $\mathcal{E}_\phi$ acts as the upstream efficiency regulator, governing the initial visual token density N_v that propagates through the pipeline.

Image-Related. Recent architectures optimize backbone efficiency through structural reparameterization (FastViT (Vasu et al., 2023)) and distillation (EfficientViT-SAM (Zhang et al., 2024f)). To mitigate token bloat from high-resolution inputs, ConvLLaVA (Ge et al., 2024) and FastVLM (Vasu et al., 2025) employ hierarchical compression and hybrid encoding to generate compact feature sets.

Video-Related. Approaches here focus on temporal adaptation and scalability. Video-LLaMA (Zhang et al., 2023) propose a video Q-Former to assemble a pre-trained image encoder into video encoder, while Qwen2-VL (Wang et al., 2024d) implements Native Dynamic Resolution for adaptive token generation. VideoChatGPT (Maaz

[7] We assume sufficient single-GPU memory capacity. Thus, Tensor Parallelism across N_{gpu} devices linearly scales the effective bandwidth to $N_{\text{gpu}} \cdot \beta_{\text{mem}}$. However, since the arithmetic intensity remains unchanged, the decoding process persists as strictly memory-bound on each individual device.

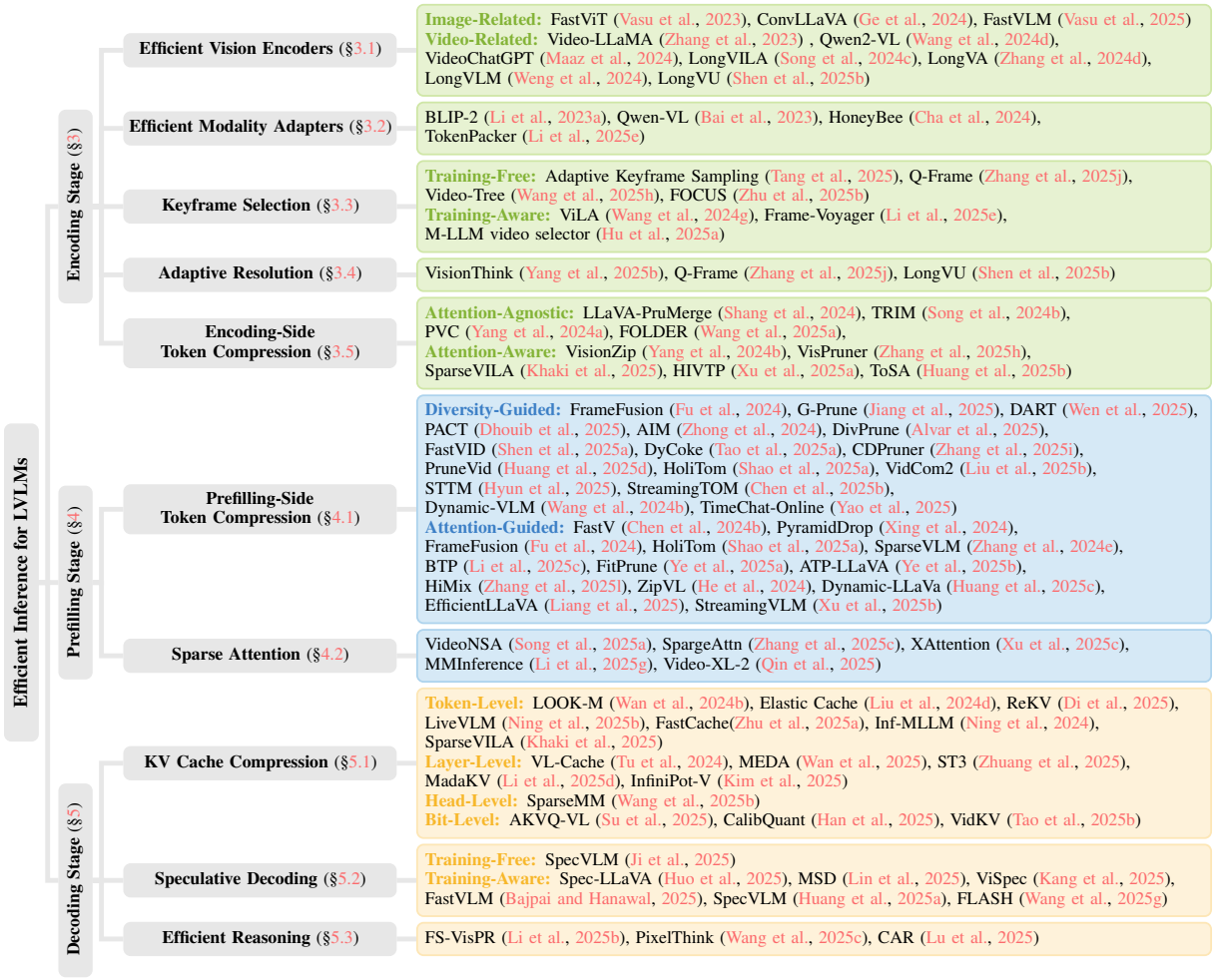

Figure 3: A stage-aware taxonomy of efficient LVLM inference. We categorize techniques by their intervention stage and optimization mechanism. This framework maps research to their target hardware bottlenecks, elucidating WHERE in the lifecycle and HOW via specific algorithms computational redundancy is reduced.

et al., 2024) enhances image encoders to capture spatiotemporal representations in videos. For long-context scenarios, MovieChat (Song et al., 2024c), LongVA (Zhang et al., 2024d), LongVLM (Weng et al., 2024), LongVILA (Chen et al., 2025c), and LongVU (Shen et al., 2025b) leverage context extension and supervised fine-tuning to support extended temporal encoding.

The field is pivoting from static, fixed-resolution backbones toward dynamic, density-aware architectures. Future architectures will likely favor end-to-end learnable compressors to maximize this upstream efficiency.

3.2 Efficient Modality Adapters

The modality adapter $\mathcal{A}_\theta$ semantically aligns the vision encoder's outputs with the LLM backbone. Baseline architectures like LLaVA (Liu et al., 2023b,a) employ simple MLPs. While computationally inexpensive, this one-to-one mapping prevents token reduction, causing visual token count N_v to scale linearly with input resolution. To tackle the token explosion problem, BLIP-2 (Li et al., 2023a) bridges the modality gap with a lightweight Q-Former. Recent works introduce resampler (Bai et al., 2023) or abstractor (Cha et al., 2024) to enforce compactness. TokenPacker (Li et al., 2025e) further refines this via a coarse-to-fine injection scheme, "packing" enriched visual semantics into fewer tokens.

The adapter is evolving from a passive bridge to an active information bottleneck, prioritizing high-density latent representations over raw feature preservation to mitigate downstream computational load.

3.3 Keyframe Selection

Keyframe selection acts as a pre-encoding filter, discarding redundant frames from $\mathbf{V}$ to minimize the computational load on the vision encoder $\mathcal{E}_\phi$. We categorize these strategies by their optimization substrate: training-free with heuristic metrics versus training-aware with learnable policies.

Training-Free Selection. This paradigm decouples selection from model training, deploying frozen encoders as plug-and-play scorers to rank

frames by semantic relevance (Yu et al., 2023; Ranasinghe et al., 2024; Liang et al., 2024). Beyond thresholding, recent works introduce structural priors: Adaptive keyframe sampling (Tang et al., 2025) jointly optimizes prompt relevance and temporal coverage via a split-and-judge policy. Q-Frame (Zhang et al., 2025j) employs the Gumbel-Max trick based on a text-image matching network for efficient probabilistic sampling; VideoTree (Wang et al., 2025h) constructs a hierarchical tree to extract query-relevant details from long videos in a coarse-to-fine manner; and FOCUS (Zhu et al., 2025b) formulates selection as a combinatorial pure-exploration problem in multi-armed bandits.

Training-Aware Selection. This paradigm, conversely, treats selection as a learnable policy optimized end-to-end for downstream performance. ViLA (Wang et al., 2024g) learns a text-guided "Frame-Prompter" to identify question-related frames that maximize video QA accuracy, while Frame-Voyager (Yu et al., 2024) minimizes the combination loss against ground-truth answers. Others, like the M-LLM video selector (Hu et al., 2025a), employ explicit cross-entropy-based supervision from spatial and temporal signals.

The trade-off is pragmatic: heuristics offer portability for open-domain deployment, whereas learnable policies can combat reasoning biases and select frames conditioned on the query.

3.4 Adaptive Resolution

Adaptive resolution optimizes the upstream information budget by modulating input fidelity prior to tokenization. For static visual inputs, methods like VisionThink (Yang et al., 2025b) and ViCO (Cui et al., 2025) implement complexity-aware scaling, dynamically adjusting resolution or selecting image compression ratios via multi-branch MLP connectors based on semantic difficulty of samples. This logic extends to query-conditional resolution for videos: Q-Frame (Zhang et al., 2025j) and LongVU (Shen et al., 2025b) maintain high-fidelity features strictly for query-relevant frames, while aggressively reducing background context via spatial pooling or downsampling.

This marks a shift toward content-adaptive encoding, where the computational budget is dynamically allocated to high-value signals rather than wasted on uniform processing.

3.5 Encoding-Side Token Compression

This category reduces the visual token count N_v immediately after encoding, operating independently of the LLM backbone. Techniques are categorized by their reliance on the encoder's internal signals.

Attention-Agnostic Compression. These methods exploit the inherent spatial redundancy of visual patches using lightweight similarity metrics. LLaVA-PruMerge (Shang et al., 2024) and TRIM (Song et al., 2024b) prune encoder-output tokens based on their similarities to the global [CLS] token or CLIP-based metrics. PVC (Yang et al., 2024a) adopts a progressive strategy by treating static images as pseudo-temporal sequences to filter redundant features. FOLDER (Wang et al., 2025a) integrates a plug-and-play merging module directly into the final blocks of the encoder.

Attention-Aware Compression. These methods (Han et al., 2026) utilize the encoder's self-attention maps as proxies for feature saliency. VisionZip (Yang et al., 2024b), VisPruner (Zhang et al., 2025h), and SparseVILA (Khaki et al., 2025) directly derive importance scores from attention matrices to retain high-value tokens. Extensions like HIVTP (Xu et al., 2025a) extract attention maps from intermediate layers of the vision encoder for early filtering, while ToSA (Huang et al., 2025b) combines semantic attention with spatial proximity to perform spatially-aware token merging.

Unlike prefilling-stage compression (Section 4), these techniques are upstream and prompt-agnostic, compressing visual data solely based on intrinsic visual properties without interaction from the textual query or LLM weights.

4 Efficiency Techniques at Prefilling

This section surveys efficiency techniques for LVLM prefilling, where causal self-attention processes massive visual contexts to materialize the KV cache. As the primary determinant of TTFT (see Equation (2)), prefilling latency τ_{PFL} acts as a hard gate on responsiveness. To mitigate this bottleneck, we structure the landscape into two strategic axes: i) *prefilling-side token compression* (Section 4.1) that aims to reduce the quantity of visual tokens (N_v) during prefilling, and ii) *sparse attention* (Section 4.2) that reduces computational complexity of the attention mechanism itself.

4.1 Prefilling-Side Token Compression

Unlike encoding-side compression, prefilling-side strategies operate within the LLM backbone's latent space ($\mathbf{H}_v$). By leveraging cross-modal semantic signals available only after projection, these methods achieve potentially higher compression ratios, directly mitigating the quadratic attention bottleneck during prefilling. As summarized in Table 2, we categorize techniques by their optimization signal: diversity-guided (minimizing redundancy) and attention-guided (maximizing saliency).

Category	Method	Input Modality	Training-Free	Key Strategy & Insight
Diversity-Guided	G-Prune (Jiang et al., 2025)	General	Yes	Similarity graph & information flow to retain representative tokens
	PACT (Dhouib et al., 2025)	General	Yes	Distance-bounded clustering & merging redundant tokens
	DivPrune (Alvar et al., 2025)	General	Yes	Max-Min diversity optimization for token subset selection
	CDPruner (Zhang et al., 2025i)	General	Yes	Determinantal Point Processes (DPP) & conditional diversity
	DART (Wen et al., 2025)	General	Yes	Pivot-based duplication pruning
	DyCoke (Tao et al., 2025a)	Video	Yes	Plug-and-play module for temporal token merging
	PruneVid (Huang et al., 2025d)	Video	Yes	Spatiotemporal token merging
	AIM (Zhong et al., 2024)	Video	Yes	Spatiotemporal token merging
	FrameFusion (Fu et al., 2024)	Video	Yes	Merges shallow-layer tokens based on adjacent frame similarity
	FastVID (Shen et al., 2025a)	Video	Yes	Temporal segmentation & density spatiotemporal pruning
	HoliTom (Shao et al., 2025a)	Video	Yes	Global redundancy-aware segmentation & spatiotemporal merging
	VidCom2 (Liu et al., 2025b)	Video	Yes	Dynamic compression based on frame uniqueness
	STTM (Hyun et al., 2025)	Video	Yes	Quadtree spatial transformation & directed pairwise merging
	Dynamic-VLM (Wang et al., 2024b)	Video	No	Dynamic compression architecture adapting to video length
	StreamingTOM (Chen et al., 2025b)	Streaming Video	Yes	Causal temporal reduction with fixed per-frame budget
	TimeChat-Online (Yao et al., 2025)	Streaming Video	No	Differential token drop for redundant content filtering
Attention-Guided	FastV (Chen et al., 2024b)	General	Yes	Learns attention patterns in early layers to prune in deep layers
	PyramidDrop (Xing et al., 2024)	General	No	Multi-stage pruning using attention score ranking
	HiMix (Zhang et al., 2025l)	General	No	Hierarchical vision injection via mixture attention
	ZipVL (He et al., 2024)	General	Yes	Dynamic token sparsification based on attention scores
	Dynamic-LLaVA (Huang et al., 2025c)	General	No	Dynamic vision-language context sparsification
	EfficientLLaVA (Liang et al., 2025)	General	No	Few-shot pruning policy search via structural risk minimization
	BTP (Li et al., 2025c)	General	Yes	Multi-stage pruning with diversity and attention ranking
	SparseVLM (Zhang et al., 2024e)	General	Yes	Pruning based on text-visual attention scores
	FitPrune (Ye et al., 2025a)	General	Yes	Minimizes divergence of attention distributions
	ATP-LLaVA (Ye et al., 2025b)	General	No	Learnable module for input-adaptive pruning
	FrameFusion (Fu et al., 2024)	Video	Yes	Pruning in deep layers based on cumulative attention scores
	HoliTom (Shao et al., 2025a)	Video	Yes	Uses cumulative attention scores for pruning inside LLM
	StreamingVLM (Xu et al., 2025b)	Streaming Video	No	Keeps attention sinks and aligns training with streaming inference

Table 2: Representative prefilling-side token compression methods, categorized by the optimization signal (Diversity or Attention) and input modality (general vision-language, video, and streaming video).

Within each category, we further classify these methods with the corresponding *input modality*.

Diversity-Guided Compression. These methods operate on the premise that visual tokens exhibit high spatial and temporal correlation (Chen et al., 2025a). The objective is to retain a subset of tokens that maximizes semantic coverage while minimizing embedding similarity. Techniques like G-Prune (Jiang et al., 2025), DivPrune (Alvar et al., 2025), and CDPruner (Zhang et al., 2025i) utilize clustering algorithms or Determinantal Point Processes to identify and merge redundant tokens based on geometric distance in the feature space. For videos with temporal dimensions, methods such as DyCoke (Tao et al., 2025a), FastVID (Shen et al., 2025a),HoliTom (Shao et al., 2025a) and VidCom^2 (Liu et al., 2025c) extend this logic to spatiotemporal merging. They fuse temporally adjacent or spatially similar patches across frames (Lin et al., 2026), preserving the "motion flow" while discarding static redundancies.

Attention-Guided Compression. This paradigm leverages LLMs' intrinsic self-attention weights as a proxy for token utility (Liu et al., 2026). FastV (Chen et al., 2024b) and PyramidDrop (Xing et al., 2024) observe that early-layer attention patterns are strong predictors of deep-layer relevance. They employ "early-exit" strategies, pruning tokens with low cumulative attention scores in initial layers to save compute in deeper layers. Advanced variants like FitPrune (Ye et al., 2025a) minimize the divergence between full and pruned attention distributions, while ATP-LLaVA (Ye et al., 2025b) introduces learnable gating modules. For video, StreamingVLM (Xu et al., 2025b) utilize "attention sinks" of both text and visual tokens to maintain reasoning stability over long contexts without quadratic computation overhead and linear memory growth.

Moving beyond binary choices, effective systems are hybridizing these paradigms, fusing outer-LLM diversity filtering with inner-LLM attention pruning to decouple geometric redundancy from semantic reasoning. This integration is crucial for addressing the inherent modal asymmetry of LVLMs, necessitating cross-modal signals that aggressively compress redundant visual states while preserving textual fidelity. Consequently, the video domain is rapidly pivoting toward real-time streaming, shifting focus from holistic offline processing to progressive, locality-aware mechanisms capable of handling infinite visual contexts.

4.2 Sparse Attention

To combat the quadratic complexity of prefilling, sparse attention mechanisms restrict computation to high-salience regions. Early generic approaches, such as XAttention (Xu et al., 2025c) (antidiagonal block scoring) and SpargeAttn (Zhang et al., 2025c) (two-stage online filtering), impose sparsity patterns derived from standard LLM heuristics. However, these methods often overlook the unique structural properties of visual tokens. Addressing this, MMInference (Li et al., 2025g) introduces *modality-aware permutation*, optimizing

sparse kernels by explicitly modeling the distinct attention signatures of visual versus textual data. For video, Video-XL-2 (Qin et al., 2025) introduces chunk-based prefilling that divides visual sequence into chunks where tokens attend only to their local chunk and coarse-grained historical timestamp tokens. Pushing this further, VideoNSA (Song et al., 2025a) shifts from post-hoc masking to *native sparse training*. VideoNSA employs Native Sparse Attention (NSA) (Yuan et al., 2025) for video tokens while retaining dense attention for text to preserve reasoning capability.

The evolution of sparse attention in LVLMs suggests the necessity to consider modality-aware architectures that integrate sparsity objectives directly into the training loop.

5 Efficiency Techniques at Decoding

This section surveys efficiency techniques for LVLM decoding, where textual output is generated token-by-token. Governed by TPOT (τ_{DEC} in Equation (2)), this stage is strictly memory-bound: latency is dominated by the limited bandwidth β_{mem} required to load model weights $|\psi|$ and the dynamic KV cache $|\mathcal{KV}|_i$ at each required step i. To address these constraints, we structure the landscape into three strategic axes: i) *KV cache compression* (Section 5.1) that reduces the memory footprint $|\mathcal{KV}|_i$, directly alleviating the bandwidth bottleneck; ii) *speculative decoding* (Section 5.2) that breaks the sequential dependency, amortizing the cost of large-model verification over rapid, lightweight draft steps; and iii) *efficient reasoning* (Section 5.3) that targets reducing the generation length N_o via optimizing the conciseness of reasoning chains.

5.1 KV Cache Compression

KV cache compression optimizes τ_{DEC} by minimizing the effective number of processed KV pairs (Feng et al., 2025a,c,b). Unlike generic compression, LVLM-specific methods exploit *modal asymmetry*, the observation that visual tokens exhibit far higher redundancy than textual tokens. As categorized in Table 3, techniques operate across four granularities: ***Token-Level***: Methods like LOOK-M (Wan et al., 2024b) and ReKV (Di et al., 2025) employ post-hoc pruning or retrieval strategies, decoupling the massive prefill context from the active working set by offloading or evicting non-salient visual states. ***Layer/Head-Level***: Methods like VL-Cache (Tu et al., 2024),SparseMM (Wang et al., 2025b), and MixKV (Liu et al., 2025a) optimize structural allocation, assigning larger cache budgets to "dense" layers or "heads" that handle cross-modal reasoning. ***Bit-Level***: Methods like VidKV (Tao et al., 2025b) push the limits of precision, utilizing sub-2-bit quantization for robust visual tokens while preserving precision for sensitive text tokens.

While current methods typically function in isolation, the distinct redundancy profiles of LVLMs demand hybrid frameworks that synergize retrieval (context), pruning (sparsity), and quantization (density). We conduct a pilot exploration of this unified paradigm in Appendix D.1.

5.2 Speculative Decoding

Speculative decoding (SD) accelerates inference by decoupling generation into rapid drafting (via a lightweight draft model and parallel verification (via the target model). While effective in LLMs (Xia et al., 2025; Zhang et al., 2024c; Song et al., 2025b), LVLMs introduce a unique bottleneck: the *visual memory wall*, where the computational cost of processing massive visual contexts ($|\mathcal{KV}|_i$) erodes the efficiency gains of the draft model.

Most existing SD adaptations are training-aware, focusing on visually specialized draft models. MSD (Lin et al., 2025) and Spec-LLaVA (Huo et al., 2025) utilize multi-stage training or distillation to align draft capabilities. To optimize visual processing, FLASH (Wang et al., 2025g), ViSpec (Kang et al., 2025) and SpecVLM (Huang et al., 2025a) introduce mechanisms like semi-autoregressive heads or adaptive visual compression. Alternatively, HiViS (Xie et al., 2025) and FastVLM (Bajpai and Hanawal, 2025) reduce computational costs by reusing the target model's hidden states or early layers, bypassing raw visual inputs. To bypass training overhead, training-free SD prioritizes direct deployment. In video scenarios, SpecVLM (Ji et al., 2025) exploits the draft model's insensitivity to visual density and performs visual token pruning for the draft model.

Despite validating SD's potential in LVLMs, existing frameworks suffer from a rigid verification bottleneck. They ignore the inherent semantic flexibility of visual descriptions. We argue that visual tasks admit substantial room for relaxed verification, a hypothesis we empirically validate in Appendix D.2.

5.3 Efficient Reasoning

Efficient reasoning targets the output horizon N_o, aiming to mitigate the latency cost of Chain-of-Thought (CoT) by dynamically aligning inference depth with problem complexity. Current strategies rely on adaptive computation length regulation in various multimodal scenarios. Pixel-Think (Wang et al., 2025c) leverages reinforcement learning to modulate reasoning length, while FS-VisPR (Li et al., 2025b) implements a "fast-

Granularity	Method	Scenario	Key Strategy & Insight
Token-Level	LOOK-M (Wan et al., 2024b)	Static	**Text-Prior Pruning**: Prioritizes textual KVs; evicts visual tokens based on attention scores.
	Elastic Cache (Liu et al., 2024d)	Static	**Merging**: Fuses less important KVs guided by distinct encoding/decoding metrics.
	FastCache (Zhu et al., 2025a)	Serving	**Self-supervised**: Uses a lightweight modality-specific compressor to reduce overhead.
	Inf-MLLM (Ning et al., 2024)	Streaming	**Bias Adjustment**: Maintains compact cache with adjustable attention bias for long-term dependency.
	SparseVILA (Khaki et al., 2025)	Streaming	**Decoupled Sparsity**: Decouples query-agnostic pruning (prefill) and query-aware retrieval (decoding).
	ReKV (Di et al., 2025)	Streaming	**Retrieval**: Offloads video chunks to external memory and selectively retrieves query-relevant KVs.
	LiveVLM (Ning et al., 2025b)	Streaming	**Dual-Memory**: Combines a short-term sliding window with retrieval from compressed long-term memory.
Layer-Level	VL-Cache (Tu et al., 2024)	Static	**Sparsity-based**: Allocates larger cache budgets to layers with denser attention patterns.
	MEDA (Wan et al., 2025)	Static	**Entropy-based**: Guided by cross-modal attention entropy to preserve complex interactions.
	ST3 (Zhuang et al., 2025)	Static	**Progressive Pruning**: Prunes more visual tokens in deeper layers based on decreasing visual importance.
	MadaKV (Li et al., 2025d)	Static	**Inter-layer Compensation**: Adjusts subsequent layer budgets based on current compression.
	InfiniPot-V (Kim et al., 2025)	Streaming	**Adaptive Pooling**: Uses varying pooling kernel sizes across layers to balance abstraction and detail.
Head-Level	SparseMM (Wang et al., 2025b)	Static	**Asymmetric Budget**: Identifies vital visual heads and allocates higher budgets to them.
Bit-Level	AKVQ-VL (Su et al., 2025)	Static	**Adaptive Mixed-Precision**: High bit-width for critical tokens, 2-bit for others.
	VidKV (Tao et al., 2025b)	Static	**Sub-2-bit**: Differential treatment for K (channel-wise) and V (1.58-bit + salient token preservation).
	CalibQuant (Han et al., 2025)	Static	**Calibrated 1-bit**: Channel-wise 1-bit quantization with post-calibration for extreme values.

Table 3: Representative KV cache compression methods, categorized by operational granularity (token, layer, head, and bit) and inference scenario (static, streaming, and serving).

slow" routing mechanism, dispatching queries between lightweight direct solvers and heavy programmatic workflows. Similarly, CAR (Lu et al., 2025) adopts an uncertainty-driven expansion, triggering extended reasoning chains only when initial confidence is low.

While effective, these methods rely on coarse instance-level mechanisms. New opportunities remain for step-level optimization: strategically pruning steps within a generated chain rather than simply processing the entire one.

6 Challenges and Future Directions

We identify three algorithmic frontiers targeting the distinct bottlenecks of representation, generation, and continuity. Crucially, we argue that their ultimate realization hinges on a fourth, integrative trajectory: end-to-end system co-design, which unifies these optimization primitives into a cohesive, hardware-aware deployment paradigm.

Representation: Hybrid Compression. Employ a uniform strategy (Wan et al., 2024b) or adjusting budget allocation alone (Li et al., 2025d; Wang et al., 2025b) are insufficient for the heterogeneous entropy of LVLMs. As preliminarily explored in Appendix D.1, the frontier may lie in *strategic orchestration*: assigning distinct operators (retrieval, pruning, and quantization) tailored to the specific sensitivity of each component.

Generation: Modality-Aware Decoding. To overcome the visual memory wall, current efficient decoding strategies (Xie et al., 2025; Ji et al., 2025; Gao et al., 2025) must abandon generic NLP heuristics. The path forward requires resolving two deficits: (i) Visual Draft Alignment, ensuring lightweight drafters can handle dense visual contexts, and (ii) Relaxed Verification, moving from rigid exact-match criteria to semantic-aware validation (as supported by Appendix D.2).

Continuity: The Streaming Pivot. The transition from offline processing to infinite-context streaming demands a shift from holistic analysis to progressive state management (Xu et al., 2025b). Future work should prioritize stage-specific optimizations, such as streaming visual memory management at encoding (Zhang et al., 2025b), progressive token compression at prefilling (Chen et al., 2025b; Xu et al., 2025b; Wang et al., 2025e), and locality-aware KV cache compression at decoding (Ning et al., 2025b). Sustaining unbounded throughput will require synergizing training-free heuristics (Chen et al., 2025b) with training-aware paradigms (Xu et al., 2025b; Zhang et al., 2025b) to prevent resource saturation.

The Unifying Imperative: End-to-End System Co-Design. Algorithm-level optimizations often falter against system-level bottlenecks (Zhang et al., 2025d) like bandwidth saturation and pipeline bubbles. Emerging disaggregated architectures (e.g., EPDServe (Singh et al., 2024), ModServe (Qiu et al., 2025)) demonstrate the necessity of mapping distinct inference stages to specialized hardware. The critical path forward lies in hardware-algorithm co-design, unifying architectural tailoring with semantic-aware predictive scheduling. We provide a detailed analysis of these serving architectures and their evaluation standards in Appendix C.

7 Literature Selection Protocol

We followed a systematic three-phase protocol to curate the literature included in this survey:

Broad Exploration. We began with a broad search on Google Scholar to identify the major research themes, representative architectures, and key terminology related to large vision-language models and efficient inference.

Targeted Filtering. Based on the initial candidate pool, we performed targeted screening over papers from major venues in NLP, machine learning, artificial intelligence, and computer vision, including ACL, EMNLP, NAACL, ICML, NeurIPS, ICLR, CVPR, and ICCV, as well as relevant arXiv preprints. We primarily focused on work published from 2020 to early 2026.

Bidirectional Citation Tracking. To further improve coverage, we applied bidirectional citation tracking. We traced backward from seminal papers such as LLaVA and BLIP-2 to identify foundational work, and traced forward to capture recent extensions and state-of-the-art systems, including representative models from the Qwen series.

8 Positioning in the Evolving Landscape

The surge in LVLMs has been accompanied by a proliferation of survey literature focusing on computational efficiency. To clarify the unique contributions of our work, we position this survey within the broader landscape of Large Language Model (LLM) and Multimodal Large Language Model (MLLM) research.

Comparison with LLM-Centric Surveys. Existing efficiency research has predominantly focused on the text modality, spanning the spectrum from algorithmic optimizations to system-level serving. Broad-spectrum surveys have systematized these efforts through data-, model-, and system-level perspectives (Zhou et al., 2024; Wan et al., 2024a), with recent comprehensive tutorials further establishing full-stack taxonomies that link algorithmic design directly to hardware bottleneck diagnosis (Ning et al., 2025a). Complementing these holistic views, specialized reviews delve into specific techniques like quantization and alternative architectures (Cheng et al., 2025a; Sun et al., 2025), while deployment-centric works emphasize MLSys challenges such as request scheduling and cluster-level load balancing (Zhen et al., 2025; Miao et al., 2025). While these works establish fundamental principles for text generation, they do not address the unique "visual memory wall" and the specific pipeline bottlenecks inherent in processing fine-grained visual inputs.

Comparison with MLLM-Centric Surveys. Surveys in the multimodal domain typically prioritize different thematic axes. *data-centric perspectives* focus exclusively on data preparation and post-training (Zhang et al., 2025e) techniques like synthesis and distillation (Bai et al., 2024; Luo et al., 2025). *architectural overviews* provide taxonomies of model structures and training recipes (Zhang et al., 2024a), often targeting edge computing scenarios (Jin et al., 2024; Zhang et al., 2025f) or resource-constrained devices (Shinde et al., 2025; Zhou et al., 2025c). Finally, *modality-specific* reviews focus narrowly on Vision-Language-Action (VLA) Models (Yu et al., 2025) or token compression across images and videos to mitigate quadratic attention (Shao et al., 2025b). Unlike these isolated optimizations, we focus on *stage-aware algorithmic optimizations* across the end-to-end inference pipeline.

Unique Contribution: End-to-End LVLMs Inference. In contrast to prior reviews that often focus on isolated optimizations, this survey provides a systematic analysis of the end-to-end LVLMs inference pipeline. We distinguish our contribution through three primary dimensions. First, we provide a *stage-specific taxonomy* along three execution stages: *encoding*, *prefilling*, and *decoding*. Second, we conduct a *bottleneck-aware analysis* to examine how overhead is shaped not only by compute but by memory traffic, cache locality, and sequence length, specifically addressing the transition from compute-bound encoders to bandwidth-bound decoding. Third, we offer a synthesis of *design principles and prospects*, identifying the pivotal shift toward dynamic, density-aware mechanisms and advocating for stage-disaggregated serving architectures to guide future research.

9 Conclusion

This survey systematizes efficient LVLM inference through a stage-aware taxonomy covering *encoding*, *prefilling*, and *decoding*. We identify the critical bottleneck shift from compute-bound visual encoding to memory-bound autoregression, showing that efficiency hinges on mitigating *visual token dominance* across the pipeline. Crucially, our analysis locates the algorithmic frontier in three modality-centric shifts: from uniform compression to hybrid orchestration, from rigid verification to semantic-aware relaxation, and from holistic processing to progressive state management. Ultimately, we argue that the advancement of this field necessitates a shift from isolated algorithmic enhancements to holistic, full-stack optimizations.

10 Acknowledgements

The work was supported by the Major Research Program of the Zhejiang Provincial Natural Science Foundation (Grant No. LD24F020015), CCF-Baidu Open Fund (No. 202509), and Zhejiang Province "Leading Talent of Technological Innovation Program" (No. 2023R5214).

11 Limitations

While this survey synthesizes efficient inference methodologies across encoding, prefilling, and decoding stages, the rapid release of proprietary models (e.g., GPT-4o) means some undocumented, closed-source optimizations may be omitted. Crucially, our analysis prioritizes the massive computational redundancy in image and video scenarios, where the visual memory wall is most acute. Consequently, domain-specific optimizations for document understanding (e.g., layout-driven cropping, OCR-aware patching) and heterogeneous multi-image scheduling receive less depth, as their discrete token structures diverge from the continuous temporal focus of this work. Finally, we concentrate on latency and memory throughput, leaving energy efficiency and theoretical compression bounds for future investigation. We advocate for standardized, hardware-agnostic benchmarks to further guide the deployment of next-generation LVLMs.

12 Ethical Considerations

This work synthesizes existing literature and involves no human subjects. The discussed methods aim to advance Green AI by reducing energy consumption and democratizing access to multimodal systems. However, we caution that efficiency-oriented optimizations, particularly lossy compression, pose risks, including the potential degradation of safety guardrails and increased hallucination rates. We urge the community to adopt robust evaluation protocols that monitor these ethical dimensions alongside latency metrics.

Regarding our pilot experiments, all evaluations are conducted using open-source models and datasets in strict compliance with their respective licenses (Apache 2.0 and CC-BY-4.0). We utilize MileBench, VideoChatGPT, and VideoDetailCaption benchmarks to ensure reproducibility.[8]

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Preprint*, arXiv:2404.14219.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, and 14 others. 2025. Yi: Open Foundation Models by 01.AI. *Preprint*, arXiv:2403.04652.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *Preprint*, arXiv:2204.14198.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-Based Visual Token Pruning for Large Multimodal Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, and 3 others. 2025. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. *Preprint*, arXiv:2509.23661.

Gregor Bachmann, Sotiris Anagnostidis, Albert Pumarola, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Edgar Schönfeld, Ali Thabet, and Jonas K Kohler. 2025. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment. In *The Thirteenth International Conference on Learning Representations*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Ping Huang, Jiulong Shan, Conghui He, Binhang Yuan, and Wentao Zhang. 2024. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *Preprint*, arXiv:2405.16640.

[8] Dataset sources: MileBench and VideoChatGPT are under Apache 2.0; VideoDetailCaption is under CC-BY-4.0. Access URLs are available at: https://github.com/MileBench/MileBench, https://huggingface.co/datasets/lmms-lab/VideoChatGPT, and https://huggingface.co/datasets/lmms-lab/VideoDetailCaption.

Divya Jyoti Bajpai and Manjesh Kumar Hanawal. 2025. FastVLM: Self-Speculative Decoding for Fast Vision-Language Model Inference. *arXiv preprint arXiv:2510.22641*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. InternLM2 Technical Report. *Preprint*, arXiv:2403.17297.

Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Chen Wang, Jiahua Dong, Wangbo Yu, Ge Zhang, and 1 others. 2025. Video Simpleqa: Towards Factuality Evaluation in Large Video Language Models. *arXiv preprint arXiv:2503.18923*.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-Enhanced Projector for Multimodal LLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. 2025. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. WebQA: Multihop and Multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2024a. CG-Bench: Clue-Grounded Question Answering Benchmark for Long Video Understanding. *arXiv preprint arXiv:2412.12075*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *Preprint*, arXiv:2310.09478.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. 2025a. Variation-aware vision token dropping for faster large vision-language models. *arXiv preprint arXiv:2509.01552*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024b. An Image Is Worth 1/2 Tokens After Layer 2: Plug-and-play Inference Acceleration for Large Vision-Language Models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024c. Are We On the Right Way for Evaluating Large Vision-Language Models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. 2025b. StreamingTOM: Streaming Token Compression for Efficient Video Understanding. *arXiv preprint arXiv:2510.18269*.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2025c. LongVILA: Scaling Long-Context Visual Language Models for Long Videos. In *The Thirteenth International Conference on Learning Representations*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *CoRR*, abs/2312.14238.

Jian Cheng, Haidong Kang, Yuxin Shao, Nan Li, Pengjun Chen, Rui Wang, Saiqin Long, Xiaochun Yang, and Lianbo Ma. 2025a. Survey on Efficient Large Language Models: Principles, Algorithms, Applications, and Open Issues. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025b. Video-Holmes: Can MLLM Think Like Holmes for Complex Video Reasoning? *arXiv preprint arXiv:2505.21374*.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025c. SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *Preprint*, arXiv:2402.03766.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2024. TVBench: Redesigning Video-Language Evaluation. *CoRR*, abs/2410.07752.

Long Cui, Weiyun Wang, Jie Shao, Zichen Wen, Gen Luo, Linfeng Zhang, Yanting Zhang, Yu Qiao, and Wenhai Wang. 2025. ViCO: A Training Strategy towards Semantic Aware Dynamic High-Resolution. *Preprint*, arXiv:2510.12793.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,

Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *Preprint*, arXiv:2305.06500.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. 2025. Pact: Pruning and Clustering-Based Token Reduction for Faster Visual Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14582–14592.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. 2025. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540*.

Xianzhe Dong, Tongxuan Liu, Yuting Zeng, Liangyu Liu, Yang Liu, Siyu Wu, Yu Wu, Hailong Yang, Ke Zhang, and Jing Li. 2025. HydraInfer: Hybrid Disaggregated Scheduling for Multimodal Large Language Model Serving. *arXiv preprint arXiv:2505.12658*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, and 4 others. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *Preprint*, arXiv:2401.16420.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2020. Counting Out Time: Class Agnostic Video Repetition Counting in the Wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10387–10396.

Yuan Feng, Haoyu Guo, JunLin Lv, S. Kevin Zhou, and Xike Xie. 2025a. Taming the fragility of kv cache eviction in llm inference. *Preprint*, arXiv:2510.13334.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. 2025b. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *Preprint*, arXiv:2407.11550.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. 2025c. Identify critical kv cache in llm inference from an output perturbation perspective. *Preprint*, arXiv:2502.03805.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. 2024. FrameFusion: Combining Similarity and Importance for Video Token Reduction on Large Visual Language Models. *arXiv preprint arXiv:2501.01986*.

Jun Gao, Qian Qiao, Tianxiang Wu, Zili Wang, Ziqiang Cao, and Wenjie Li. 2025. Aim: Let any multimodal large language models embrace efficient in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3077–3085.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. ConvLLaVA: Hierarchical Backbones as Visual Encoder for Large Multimodal Models. *arXiv preprint arXiv:2405.15738*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Tianyu Guo, Tianming Xu, Xianjie Chen, Junru Chen, Nong Xiao, and Xianwei Zhang. 2025. RServe: Overlapping Encoding and Prefill for Efficient LMM Inference. *arXiv preprint arXiv:2509.24381*.

Insu Han, Zeliang Zhang, Zhiyuan Wang, Yifan Zhu, Susan Liang, Jiani Liu, Haiting Lin, Mingjie Zhao, Chenliang Xu, Kun Wan, and 1 others. 2025. CalibQuant: 1-Bit KV Cache Quantization for Multimodal LLMs. *arXiv preprint arXiv:2502.14882*.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. 2026. Filter, correlate, compress: Training-free token reduction for

mllm acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4601–4609.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can MLLMs Reason in Multimodality? EMMA: An Enhanced Multimodal Reasoning Benchmark. *arXiv preprint arXiv:2501.05444*.

Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. 2024. ZipVL: Efficient Large Vision-Language Models with Dynamic Token Sparsification and KV Cache Compression. *CoRR*, abs/2410.08584.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8450–8460.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and 1 others. 2025a. M-LLM Based Video Frame Selection for Efficient Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025b. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv preprint arXiv:2501.13826*.

Haiduo Huang, Fuwei Yang, Zhenhua Liu, Xuanwu Yin, Dong Li, Pengju Ren, and Emad Barsoum. 2025a. SpecVLM: Fast Speculative Decoding in Vision-Language Models. *arXiv preprint arXiv:2509.11815*.

Hsiang-Wei Huang, Wenhao Chai, Kuang-Ming Chen, Cheng-Yen Yang, and Jenq-Neng Hwang. 2025b. ToSA: Token Merging with Spatial Awareness. *Preprint*, arXiv:2506.20066.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. 2025c. Dynamic-LLaVA: Efficient Multimodal Large Language Models via Dynamic Vision-language Context Sparsification. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Xiaohu Huang, Hao Zhou, and Kai Han. 2025d. PruneVID: Visual Token Pruning for Efficient Video Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19959–19973.

Mingxiao Huo, Jiayi Zhang, Hewei Wang, Jinfeng Xu, Zheyu Chen, Huilin Tai, and Yijun Chen. 2025. Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding. *arXiv preprint arXiv:2509.11961*.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. 2025. Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23990–24000.

Yicheng Ji, Jun Zhang, Jinpeng Chen, Cong Wang, Lidan Shou, Gang Chen, and Huan Li. 2026. See the forest for the trees: Loosely speculative decoding via visual-semantic guidance for efficient inference of video llms. *Preprint*, arXiv:2604.05650.

Yicheng Ji, Jun Zhang, Heming Xia, Jinpeng Chen, Lidan Shou, Gang Chen, and Huan Li. 2025. SpecVLM: Enhancing Speculative Decoding of Video LLMs via Verifier-Guided Token Pruning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7216–7230.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025. 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*, pages 4075–4083.

Liuyi Jin, Tian Liu, Amran Haroon, Radu Stoleru, Michael Middleton, Ziwei Zhu, and Theodora Chaspari. 2023. Emsassist: An end-to-end mobile voice assistant at the edge for emergency medical services. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services, MobiSys 2023, Helsinki, Finland, June 18-22, 2023*, pages 275–288.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient Multimodal Large Language Models: A Survey. *Preprint*, arXiv:2405.10739.

Jialiang Kang, Han Shu, Wenshuo Li, Yingjie Zhai, and Xinghao Chen. 2025. ViSpec: Accelerating Vision-Language Models with Vision-Aware Speculative Decoding. *arXiv preprint arXiv:2509.15235*.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N Plataniotis, Yao Lu, Song Han, and Zhijian Liu. 2025. SparseVILA: Decoupling Visual Sparsity for Efficient VLM Inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23784–23794.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. 2025. InfiniPot-V: Memory-Constrained KV Cache Compression for Streaming Video Understanding. *arXiv preprint arXiv:2506.15745*.

Quan Kong, Yuhao Shen, Yicheng Ji, Huan Li, and Cong Wang. 2026. ParallelVLM: Lossless Video-LLM Acceleration with Visual Alignment Aware Parallel Speculative Decoding. *arXiv preprint arXiv:2603.19610*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. LLaVA-OneVision: Easy Visual Task Transfer. *Trans. Mach. Learn. Res.*, 2025.

Chenglin Li, Feng Han, Feng Tao, Ruilin Li, Qianglong Chen, Jingqi Tong, Yin Zhang, and Jiaqi Wang. 2025b. Adaptive Fast-and-Slow Visual Program Reasoning for Long-Form VideoQA. *arXiv preprint arXiv:2509.17743*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. 2025c. Balanced Token Pruning: Accelerating Vision Language Models Beyond Local Optimization. *arXiv preprint arXiv:2505.22038*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. MVBench: A Comprehensive Multi-Modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Kunxi Li, Zhonghua Jiang, Zhouzhou Shen, ZhaodeWang ZhaodeWang, Chengfei Lv, Shengyu Zhang, Fan Wu, and Fei Wu. 2025d. MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13306–13318.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2025e. TokenPacker: Efficient Visual Projector for Multimodal LLM. *Int. J. Comput. Vis.*, 133(10):6794–6812.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025f. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling. *Preprint*, arXiv:2501.00574.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. MMCoQA: Conversational Question Answering over Text, Tables, and Images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4220–4231.

Yucheng Li, Huiqiang Jiang, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Amir H Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and 1 others. 2025g. MMInference: Accelerating Pre-filling for Long-Context VLMs via Modality-Aware Permutation Sparse Attention. *arXiv preprint arXiv:2504.16083*.

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. 2024. KeyVideoLLM: Towards Large-scale Video Keyframe Selection. *arXiv preprint arXiv:2407.03104*.

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

Luxi Lin, Zhihang Lin, Zhanpeng Zeng, and Rongrong Ji. 2025. Speculative Decoding Reimagined for Multimodal Large Language Models. *arXiv preprint arXiv:2505.14260*.

Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. 2026. V-cast: Video curvature-aware spatio-temporal pruning for efficient video large language models. *arXiv preprint arXiv:2603.27650*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*.

Xuyang Liu, Xiyan Gui, Yuchao Zhang, and Linfeng Zhang. 2025a. Mixing Importance with Diversity: Joint Optimization for KV Cache Compression in Large Vision-Language Models. *arXiv preprint arXiv:2510.20707*.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. 2025b. Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models. *arXiv preprint arXiv:2505.14454*.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. 2025c. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1910–1924.

Xuyang Liu, Ziming Wang, Junjie Chen, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Siteng Huang, and Honggang Chen. 2026. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7350–7358.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. MMBench: Is Your Multi-Modal Model An All-Around Player? In *European conference on computer vision*, pages 216–233. Springer.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024c. TempCompass: Do Video LLMs Really Understand Videos? *arXiv preprint arXiv:2403.00476*.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y Charles, Xinyu Zhou, and Xu Sun. 2025d. VideoReasonBench: Can MLLMs Perform Vision-Centric Complex Video Reasoning? *arXiv preprint arXiv:2505.23359*.

Zedong Liu, Shenggan Cheng, Guangming Tan, Yang You, and Dingwen Tao. 2025e. ElasticMM: Efficient Multimodal LLMs Serving with Elastic Multimodal Parallelism. *arXiv preprint arXiv:2507.10069*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, and 8 others. 2025f. NVILA: Efficient Frontier Visual Language Models. *Preprint*, arXiv:2412.04468.

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

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, and 1 others. 2025. Prolonged Reasoning Is Not All You Need: Certainty-Based Adaptive Routing for Efficient LLM/MLLM Reasoning. *arXiv preprint arXiv:2505.15154*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *arXiv preprint arXiv:2310.02255*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *Preprint*, arXiv:2405.20797.

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and Ming Zhang. 2025. "A Survey on Efficient Large Language Model Training: From Data-centric Perspectives". In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30904–30920, Vienna, Austria.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2025. Towards Efficient Generative Large Language Model Serving: A Survey from Algorithms to Systems. *ACM Computing Surveys*, 58(1):1–37.

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, and 1 others. 2025. Minerva: Evaluating Complex Video Reasoning. *arXiv preprint arXiv:2505.00681*.

Xuefei Ning, Guohao Dai, Haoli Bai, Lu Hou, Yu Wang, and Qun Liu. 2025a. Efficient Inference for Large Language Models –Algorithm, Model, and System. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 1–3, Suzhou, China. Association for Computational Linguistics.

Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. 2025b. LiveVLM: Efficient Online Video Understanding via Streaming-Oriented KV Cache and Retrieval. *arXiv preprint arXiv:2505.15269*.

Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. 2024. Inf-MLLM: Efficient Streaming Inference of Multimodal Large Language Models on a Single GPU. *arXiv preprint arXiv:2409.09086*.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2025. OmniDocBench: Benchmarking Diverse PDF Document

Parsing with Comprehensive Annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24838–24848.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. 2025. Omnia de EgoTempo: Benchmarking Temporal Understanding of Multi-modal LLMs in Egocentric Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24129–24138.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. VCR-Bench: A Comprehensive Evaluation Framework for Video Chain-of-Thought Reasoning. *arXiv preprint arXiv:2504.07956*.

Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. 2025. Video-XL-2: Towards Very Long-Video Understanding Through Task-Aware KV Sparsification. *arXiv preprint arXiv:2506.19225*.

Haoran Qiu, Anish Biswas, Zihan Zhao, Jayashree Mohan, Alind Khare, Esha Choukse, Íñigo Goiri, Zeyu Zhang, Haiying Shen, Chetan Bansal, and 1 others. 2025. ModServe: Modality-and Stage-Aware Resource Disaggregation for Scalable Multimodal Model Serving. *arXiv preprint arXiv:2502.00937*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision Language Models are Blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. 2024. Understanding Long Videos in One Multimodal Language Model Pass. *arXiv preprint arXiv:2403.16998*, 3(4):12.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, and 1 others. 2025. ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models. *arXiv preprint arXiv:2502.09696*.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. *Preprint*, arXiv:2403.15388.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. *arXiv preprint arXiv:2410.23266*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing Multi-Modal Language Models with A Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025a. HoliTom: Holistic Token Merging for Fast Video Large Language Models. *arXiv preprint arXiv:2505.21334*.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. 2025b. When Tokens Talk Too Much: A Survey of Multimodal Long-Context Token Compression across Images, Videos, and Audios. *Preprint*, arXiv:2507.20198.

Hui Shen, Xin Wang, Ping Zhang, Yunta Hsieh, Qi Han, Zhongwei Wan, Ziheng Zhang, Jingxuan Zhang, Jing Xiong, Ziyuan Liu, and 1 others. 2026a. MMSpec: Benchmarking Speculative Decoding for Vision-Language Models. *arXiv preprint arXiv:2603.14989*.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. 2025a. FastVID: Dynamic Density Pruning for Fast Video Large Language Models. *arXiv preprint arXiv:2503.11187*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2025b. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *Forty-second International Conference on Machine Learning*.

Yuhao Shen, Tianyu Liu, Junyi Shen, Jinyang Wu, Quan Kong, Li Huan, and Cong Wang. 2026b. Double: Breaking the Acceleration Limit via Double Retrieval Speculative Parallelism. *Preprint*, arXiv:2601.05524.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. 2025. A Survey on Efficient Vision-Language Models. *Preprint*, arXiv:2504.09724.

Gursimran Singh, Xinglu Wang, Yifan Hu, Timothy Yu, Linzi Xing, Wei Jiang, Zhefeng Wang, Xiaolong Bai, Yi Li, Ying Xiong, and 1 others. 2024. Efficiently Serving Large Multimodal Models Using EPD Disaggregation. *arXiv preprint arXiv:2501.05460*.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024a. MileBench: Benchmarking MLLMs in Long Context. *Preprint*, arXiv:2404.18532.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. 2024b. Less is More: A Simple yet Effective Token Reduction Method for Efficient Multi-modal LLMs. *Preprint*, arXiv:2409.10994.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024c. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Enxin Song, Wenhao Chai, Shusheng Yang, Ethan Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. 2025a. VideoNSA: Native Sparse Attention Scales Video Understanding. *arXiv preprint arXiv:2510.02295*.

Mingbo Song, Heming Xia, Jun Zhang, Chak Tou Leong, Qiancheng Xu, Wenjie Li, and Sujian Li. 2025b. KNN-SSD: Enabling Dynamic Self-Speculative Decoding via Nearest Neighbor Layer Set Optimization. *CoRR*, abs/2505.16162.

Zunhai Su, Wang Shen, Linge Li, Zhe Chen, Hanyu Wei, Huangqi Yu, and Kehong Yuan. 2025. AKVQ-VL: Attention-Aware KV Cache Adaptive 2-Bit Quantization for Vision-Language Models. *arXiv preprint arXiv:2501.15021*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*.

Weigao Sun, Jiaxi Hu, Yucheng Zhou, Jusen Du, Disen Lan, Kexin Wang, Tong Zhu, Xiaoye Qu, Yu Zhang, Xiaoyu Mo, Daizong Liu, Yuxuan Liang, Wenliang Chen, Guoqi Li, and Yu Cheng. 2025. Speed Always Wins: A Survey on Efficient Architectures for Large Language Models. *Preprint*, arXiv:2508.09834.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: Complex Question Answering over Text, Tables and Images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. 2025. ALLVB: All-in-One Long Video Understanding Benchmark. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pages 7211–7219.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13636–13645.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. 2025. Adaptive Keyframe Sampling for Long Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025a. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001.

Keda Tao, Haoxuan You, Yang Sui, Can Qin, and Huan Wang. 2025b. Plug-and-Play 1. x-Bit KV Cache Quantization for Video Large Language Models. *arXiv preprint arXiv:2503.16257*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint*, arXiv:2403.08295.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *Preprint*, arXiv:2406.16860.

Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. 2024. VL-Cache: Sparsity and Modality-Aware KV Cache Compression for Vision-Language Model Inference Acceleration. *arXiv preprint arXiv:2410.23317*.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document Understanding Dataset and Vvaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and 1 others. 2025. FastVLM: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19769–19780.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. 2023. FastViT: A Fast Hybrid Vision Transformer Using Structural Reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5785–5795.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision Language Models are Biased. *arXiv preprint arXiv:2505.23941*.

Zhongwei Wan, Hui Shen, Xin Wang, Che Liu, Zheda Mai, and Mi Zhang. 2025. MEDA: Dynamic KV Cache Allocation for Efficient Multimodal Long-context Inference. *arXiv preprint arXiv:2502.17599*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024a. Efficient Large Language Models: A Survey. *Preprint*, arXiv:2312.03863.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. 2024b. LOOK-M: Look-Once Optimization in KV Cache for Efficient Multimodal Long-Context Inference. *arXiv preprint arXiv:2406.18139*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. MuirBench: A Comprehensive Benchmark for Robust Multi-Image Understanding. *arXiv preprint arXiv:2406.09411*.

Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quétu, Shuai Xiao, and Enzo Tartaglione. 2025a. FOLDER: Accelerating Multi-Modal Large Language Models with Enhanced Performance. *Preprint*, arXiv:2501.02430.

Han Wang, Yuxiang Nie, Yongjie Ye, Guanyu Deng, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. 2024b. Dynamic-VLM: Simple Dynamic Visual Token Compression for VideoLLM. *CoRR*, abs/2412.09530.

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. 2025b. SparseMM: Head Sparsity Emerges from Visual Concept Responses in MLLMs. *arXiv preprint arXiv:2506.05344*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024c. Measuring Multimodal Mathematical Reasoning with Math-Vision Dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024d. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World At Any Resolution. *arXiv preprint arXiv:2409.12191*.

Song Wang, Gongfan Fang, Lingdong Kong, Xiangtai Li, Jianyun Xu, Sheng Yang, Qiang Li, Jianke Zhu, and Xinchao Wang. 2025c. PixelThink: Towards Efficient Chain-of-Pixel Reasoning. *arXiv preprint arXiv:2505.23727*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024e. LVBench: An Extreme Long Video Understanding Benchmark. *CoRR*, abs/2406.08035.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024f. CogVLM: Visual Expert for Pretrained Language Models. *Preprint*, arXiv:2311.03079.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025d. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *Preprint*, arXiv:2508.18265.

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C Lin, and Shan Yang. 2024g. ViLA: Efficient Video-Language Alignment for Video Question Answering. In *European Conference on Computer Vision*, pages 186–204. Springer.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. 2025e. Accelerating Streaming Video Large Language Models via Hierarchical Token Compression. *arXiv preprint arXiv:2512.00891*.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, and 1 others. 2025f. MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly. *arXiv preprint arXiv:2505.10610*.

Zihua Wang, Ruibo Li, Haozhe Du, Joey Tianyi Zhou, Yu Zhang, and Xu Yang. 2025g. FLASH: Latent-Aware Semi-Autoregressive Speculative Decoding for Multimodal Tasks. *arXiv preprint arXiv:2505.12728*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025h. VideoTree: Adaptive Tree-Based Video Representation for LLM Reasoning on Long Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Stop Looking for Important Tokens in Multimodal Language Models: Duplication Matters More. *arXiv preprint arXiv:2502.11494*.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. LongVLM: Efficient Long Video Understanding via Large Language Models. In *European Conference on Computer Vision*, pages 453–470. Springer.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. LongVideoBench: A Benchmark for Long-Context Interleaved Video-Language Understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024b. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *Preprint*, arXiv:2412.10302.

Heming Xia, Yongqi Li, Jun Zhang, Cunxiao Du, and Wenjie Li. 2025. SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration. In *The Thirteenth International Conference on Learning Representations*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. *arXiv preprint arXiv:2407.04973*.

Zhinan Xie, Peisong Wang, and Jian Cheng. 2025. HiViS: Hiding Visual Tokens from the Drafter for Speculative Decoding in Vision-Language Models. *arXiv preprint arXiv:2509.23928*.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and 1 others. 2024. Pyramiddrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. *arXiv preprint arXiv:2410.17247*.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying CLIP Data. *arXiv preprint arXiv:2309.16671*.

Jingqi Xu, Jingxi Lu, Chenghao Li, Sreetama Sarkar, and Peter A. Beerel. 2025a. HIVTP: A Training-Free Method to Improve VLMs Efficiency via Hierarchical Visual Token Pruning Using Middle-Layer-Based Importance Score. *Preprint*, arXiv:2509.23663.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *Preprint*, arXiv:2407.15841.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. 2025b. StreamingVLM: Real-Time Understanding for Infinite Video Streams. *arXiv preprint arXiv:2510.09608*.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. 2025c. Xattention: Block Sparse Attention with Antidiagonal Scoring. *arXiv preprint arXiv:2503.16428*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. 2024a. PVC: Progressive Visual Token Compression for Unified Image and Video Processing in Large Vision-Language Models. *Preprint*, arXiv:2412.09613.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2024b. VisionZip: Longer is Better but Not Necessary in Vision Language Models. *Preprint*, arXiv:2412.04467.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025b. Visionthink: Smart and Efficient Vision Language Model via Reinforcement Learning. *arXiv preprint arXiv:2507.13348*.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. 2025. TimeChat-Online: 80% Visual Tokens are Naturally Redundant in Streaming Videos. *CoRR*, abs/2504.17343.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025a. 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*, pages 22128–22136.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025b. ATP-LLaVA: Adaptive Token Pruning for Large Vision Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24972–24982.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-Chained Image-Language Model for Video Localization and Question Answering. *Advances in Neural Information Processing Systems*, 36:76749–76771.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, and 1 others. 2024. Frame-Voyager: Learning to Query Frames for Video Large Language Models. *arXiv preprint arXiv:2410.03226*.

Zhaoshu Yu, Bo Wang, Pengpeng Zeng, Haonan Zhang, Ji Zhang, Lianli Gao, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. 2025. A Survey on Efficient Vision-Language-Action Models. *Preprint*, arXiv:2510.24795.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, and 1 others. 2025. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23078–23097.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. MMMU-Pro: A More Robust Multi-Discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

Bowen Zeng, Feiyang Ren, Jun Zhang, Xiaoling Gu, Ke Chen, Lidan Shou, and Huan Li. 2026. Hybridkv: Hybrid kv cache compression for efficient multimodal large language model inference. *Preprint*, arXiv:2604.05887.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision*, pages 11941–11952.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *Preprint*, arXiv:2501.13106.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. 2025b. Flash-VStream: Efficient Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2506.23825*.

He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, and Xun Cao. 2024b. MMVP: A Multimodal Mocap Dataset with Vision and Pressure Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21842–21852.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. 2025c. Spargeattn: Accurate Sparse Attention Accelerating Any Model Inference. *arXiv preprint arXiv:2502.18137*.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra. 2024c. Draft & Verify: Lossless Large Language Model Acceleration via Self-Speculative Decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11263–11282.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Gang Chen, Qin Xie, Guiming Xie, and Xuejian Gong. 2025d. HMI: hierarchical knowledge management for efficient multi-tenant inference in pretrained language models. *The VLDB Journal*, 34(4):43.

Jun Zhang, Jue Wang, Huan Li, Lidan Shou, Ke Chen, Yang You, Guiming Xie, Xuejian Gong, and Kunlong Zhou. 2025e. Train Small, Infer Large: Memory-Efficient LoRA Training for Large Language Models. *The Thirteenth International Conference on Learning Representations*.

Jun Zhang, Jue Wang, Huan Li, Zhongle Xie, Ke Chen, and Lidan Shou. 2025f. CHASe: Client Heterogeneity-Aware Data Selection for Effective Federated Active Learning . *IEEE Transactions on Knowledge & Data Engineering*, 37(06):3088–3102.

Kaiyuan Zhang, Chenghao Yang, Zhoufutu Wen, Sihang Yuan, Qiuyue Wang, Chaoyi Huang, Guosheng Zhu, He Wang, Huawenyu Lu, Jianing Wen, and 1 others. 2025g. MME-CC: A Challenging Multi-Modal Evaluation Benchmark of Cognitive Capacity. *arXiv preprint arXiv:2511.03146*.

Libo Zhang, Zhaoning Zhang, Wangyang Hong, Peng Qiao, and Dongsheng Li. 2026. Sparrow: Text-Anchored Window Attention with Visual-Semantic Glimpsing for Speculative Decoding in Video LLMs. *arXiv preprint arXiv:2602.15318*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024d. Long Context Transfer from Language to Vision. *arXiv preprint arXiv:2406.16852*.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2025h. Beyond Text-Visual Attention: Exploiting Visual Cues for Effective Token Pruning in VLMs. *Preprint*, arXiv:2412.01818.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. 2025i. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. *arXiv preprint arXiv:2506.10967*.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. 2025j. Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs. *arXiv preprint arXiv:2506.22139*.

Shuoming Zhang, Jiacheng Zhao, Siqi Li, Xiyu Shi, Yangyu Zhang, Shuaijiang Li, Donglin Yu, Zheming Yang, Yuan Wen, Huimin Cui, and 1 others. 2025k. SpaceServe: Spatial Multiplexing of Complementary Encoders and Decoders for Multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xuange Zhang, Dengjie Li, Bo Liu, Zenghao Bao, Yao Zhou, Baisong Yang, Zhongying Liu, Yujie Zhong, Zheng Zhao, and Tongtong Yuan. 2025l. HiMix: Reducing Computational Complexity in Large Vision-Language Models. *CoRR*, abs/2501.10318.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024e. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. *arXiv preprint arXiv:2410.04417*.

Z Zhang, H Cai, and S EfficientViT-SAM Han. 2024f. Accelerated Segment Anything Model Without Performance Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA*, pages 16–22.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

Ranran Zhen, Juntao Li, Yixin Ji, Zhenlin Yang, Tong Liu, Qingrong Xia, Xinyu Duan, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2025. "Taming the Titans: A Survey of Efficient LLM Inference Serving". In *Proceedings of the 18th International Natural Language Generation Conference*, pages 522–541, Hanoi, Vietnam.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. 2024. AIM: Adaptive Inference of Multi-Modal LLMs via Token Merging and Pruning. *CoRR*, abs/2412.03248.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025a. MLVU: Benchmarking Multi-task Long Video Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.

Yuhao Zhou, Yiheng Wang, Xuming He, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, and 1 others. 2025b. Scientists' First Exam: Probing Cognitive Abilities of MLLM via Perception, Understanding, and Reasoning. *arXiv preprint arXiv:2506.10521*.

Yuxin Zhou, Zheng Li, Jun Zhang, Jue Wang, Yiping Wang, Zhongle Xie, Ke Chen, and Lidan Shou. 2025c. FloE: On-the-Fly MoE Inference on Memory-constrained GPU. *Forty-second International Conference on Machine Learning*.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A Survey on Efficient Inference for Large Language Models. *CoRR*, abs/2404.14294.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *Preprint*, arXiv:2304.10592.

Jianian Zhu, Hang Wu, Haojie Wang, Yinghui Li, Biao Hou, Ruixuan Li, and Jidong Zhai. 2025a. FastCache: Optimizing Multimodal LLM Serving Through Lightweight KV-Cache Compression Framework. *arXiv preprint arXiv:2503.08461*.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. 2025b. FOCUS: Efficient Keyframe Selection for Long Video Understanding. *arXiv preprint arXiv:2510.27280*.

Jiedong Zhuang, Lu Lu, Ming Dai, Rui Hu, Jian Chen, Qiang Liu, and Haoji Hu. 2025. St3: Accelerating Multimodal Large Language Model by Spatial-Temporal Visual Token Trimming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11049–11057.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. *arXiv preprint arXiv:2411.00836*.

A Takeaways

Through a comprehensive review of efficiency techniques for LVLM inference, we distill the following critical insights across the three execution stages. These takeaways highlight the shifting bottlenecks and the emerging design principles for the next-generation LVLM systems.

A.1 Efficiency Techniques at Encoding Stage

The encoding phase dictates the initial computational footprint of the entire inference lifecycle. Decisions made here regarding resolution and feature granularity set the baseline cost for downstream processing. Our analysis identifies a decisive shift from static, one-size-fits-all preprocessing to dynamic, density-aware mechanisms that align computational expenditure with information content.

Takeaway

❶ Encoding as the Primary Lever for Token Efficiency. Improving token efficiency at the encoding stage (e.g., via pooling or C-Abstractors) offers the highest leverage within the inference pipeline. Removing a single token at this point eliminates its computation across all subsequent LLM layers and prevents it from occupying space in the KV cache. Therefore, the encoding stage is decisive for downstream token efficiency. Future architectures will likely favor end-to-end learnable compressors over heuristic selection to maximize this upstream efficiency.

❷ Paradigm Shift from Fixed to Dynamic. Visual information naturally exhibits non-uniform density and substantial redundancy. Traditional preprocessing methods that treat all visual information uniformly are becoming increasingly obsolete. The SOTA models (e.g., Qwen3-VL (Yang et al., 2025a) and LLaVA-Next (Liu et al., 2024a)) suggest that Native Dynamic Resolution is a promising direction, as it adaptively allocates tokens based on the information density of the input, simultaneously preserving fine-grained visual details and reducing the total token count. Moreover, dynamic token budgeting across regions of images and frames of videos is becoming increasingly mainstream, further underscoring this trend.

❸ Encoding is Compute-Bound, Not Memory-Bound. Unlike the decoding stage, the visual encoding phase is characteristically compute-bound. As input resolutions scale (e.g., to 4K), the latency of the Vision Transformer (ViT) becomes a significant component of the (TTFT). Consequently, optimizations that target architectural efficiency, such as structural reparameterization (e.g., FastViT (Vasu et al., 2023)) and operator fusion, yield higher end-to-end acceleration gains than simple weight quantization in this stage.

A.2 Efficiency Techniques at Prefilling Stage

Optimizing the prefilling stage is fundamentally about mitigating the quadratic scaling of attention mechanisms ($\mathcal{O}(N^2)$) in the face of increasingly long visual contexts. As models scale to handle high-resolution imagery and long-form video, the latency of the first token (TTFT) becomes a primary bottleneck. The literature converges on the insight that visual data exhibits significantly higher redundancy than text, permitting aggressive, non-uniform compression strategies.

Takeaway

❶ The Trend toward Hybrid Token Compression Paradigms. Emerging trends in prefilling-side token compression indicate a shift from isolated processing toward hybrid, multi-stage architectures (e.g., FrameFusion (Fu et al., 2024), Holitom (Shao et al., 2025a)) that synergize diversity-guided and attention-guided mechanisms. This paradigm leverages a "coarse-to-fine" strategy: maximizing compression rates by first eliminating intrinsic visual redundancy via token diversity outside the LLM backbone, and subsequently refining the selection based on attention distribution inside the LLM backbone.

❷ Asymmetric Modality Redundancy. Visual tokens exhibit significantly higher spatiotemporal redundancy compared to the semantic redundancy of text tokens. Uniform pruning or sparse attention strategies often fail because they treat both modalities equally. Optimal prefilling requires modality-aware policies that apply aggressive compression (Zhang et al., 2024e) and transformation (Li et al., 2025g) to visual tokens (based on similarity or attention) while maintaining a conservative approach for text tokens to prevent semantic collapse.

❸ The Trajectory toward Real-time Streaming Videos. Research is pivoting from offline video understanding (e.g., DyCoke (Tao et al., 2025a), VidCom2 (Liu et al., 2025b))—characterized by static, fully known contexts—to online streaming paradigms (e.g., StreamingTOM (Chen et al., 2025b), StreamingVLM (Xu et al., 2025b)). Streaming scenarios require processing infinite visual sequences with minimal delay, redefining prefilling from a singular initialization event to a recurring operational cost. Technical approaches transition from compression based on holistic video information toward progressive and locality-aware compression mechanisms to effectively minimize the recurring prefill costs.

A.3 Efficiency Techniques at Decoding Stage

The decoding stage is characteristically memory-bound, defined by the operational intensity of loading massive Key-Value (KV) caches for autoregressive generation. In LVLMs, this bottleneck manifests as a “Visual Memory Wall”: for long-context multimodal inference, the KV cache footprint of-

ten exceeds model weights themselves, with visual tokens accounting for 80%-90% of the total memory usage. However, the generation phase relies predominantly on textual history and only sparse visual cues. The takeaways below distill the emerging design principles that exploit this asymmetry to maximize throughput and minimize latency.

Takeaway

❶ From Coarse to Multi-Granular Modality-Aware Compression. Traditional uniform compression is suboptimal for LVLMs as it ignores the extreme redundancy of visual data and the importance of textual history. The current trend is modality-adaptive compression that operates at finer granularities from layers and heads to bits. Techniques now allocate distinct budgets: applying aggressive compression to visual tokens (e.g., sub-2-bit quantization (Tao et al., 2025b), sparse head pruning (Wang et al., 2025b)) while retaining high precision for most text tokens. This asymmetric treatment reduces the "visual memory wall" payload without compromising much performance.

❷ Draft Model Adaptation for Multimodal Speculation. Applying generic speculative decoding to LVLMs is inefficient if the draft model is burdened by heavy visual processing. The trend is shifting towards visual-lightweight drafting, achieved via two main paths: training specialized lightweight draft models (e.g., MSD (Lin et al., 2025), Spec-LLaVA (Huo et al., 2025)) or pruning redundant visual tokens from the draft input in a training-free manner (e.g., SpecVLM (Ji et al., 2025)). Both approaches leverage language priors to accelerate candidate generation, effectively amortizing the high bandwidth cost of the target model's verification step.

❸ From Monolithic to Disaggregated Serving Architectures. The resource demands of LVLM inference are highly heterogeneous: Encoding is compute-intensive, while Decoding is bandwidth-intensive. Monolithic serving architectures struggle to balance these conflicting needs. The field is moving toward Stage-Disaggregated Serving (e.g. ModServe (Qiu et al., 2025)), where encoding and decoding are decoupled and scheduled onto specialized hardware configurations to maximize cluster-wide utilization and throughput.

A.4 Efficiency Techniques at the System Level

Complementing our granular analysis of the encoding, prefilling, and decoding stages, we extend our scope to the holistic serving ecosystem. In Appendix C.1, we survey the architectural landscape of efficient LVLM serving systems adopted by the community. By synthesizing the trade-offs identified across these isolated stages with broader system-level constraints, we distill the following key takeaways for optimization.

Takeaway

❶ Resource Decoupling via Spatial Multiplexing. Fundamentally, efficient LVLM serving relies on *spatial multiplexing* (Zhang et al., 2025k), which decouples inference components onto specialized resource groups. This separation allows each group to be independently configured and scaled, effectively resolving the conflicting resource affinities inherent to distinct inference stages and modalities.

❷ Optimal Architecture Selection. As synthesized in Table 4, the optimal topology is dictated by the workload shape: (1) **Stage-based** disaggregation isolates computational bursts, making it ideal for *Long-Video* tasks to stabilize P99 latency. (2) **Modality-based** partitioning simplifies operations while maximizing throughput for *Balanced* workloads. (3) **Resource-based** multiplexing eliminates communication overhead, serving as the preferred solution for *Latency-Critical* or *Edge* applications (Jin et al., 2023).

B Model Architecture

B.1 Overview

As formalized in Section 2.1, modern LVLMs converge on a unified three-component architecture: vision encoder $\mathcal{E}_\phi$, modality adapter $\mathcal{A}_\theta$, and LLM backbone $\mathcal{L}_\psi$. This appendix provides implementation details and a taxonomy of representative models organized by their efficiency-oriented design choices.

We detail the implementation variants of each component and categorize representative models by their efficiency strategies.

B.2 Core Components

Vision Encoder. The vision encoder $\mathcal{E}_\phi$ produces patch embeddings $\mathbf{X}_v \in \mathbb{R}^{N_p \times D_v}$ from raw visual input. Modern LVLMs typically reuse pretrained visual encoders such as CLIP (Radford et al., 2021), MetaCLIP (Xu et al., 2023), EVA-CLIP (Sun et al., 2023), SigLIP (Zhai et al., 2023), or ViT (Dosovitskiy et al., 2021) as general-purpose front ends. These encoders underpin many widely used systems. For instance, LLaVA (Liu et al., 2023b) and LLaVA-OneVision (An et al., 2025) leverage CLIP and SigLIP variants, respectively, while InternVL (Chen et al., 2023b) and Qwen-VL (Bai et al., 2023) utilize scale-up strategies based on powerful backbones such as InternViT and OpenCLIP. They convert images into patch-based token sequences whose granularity and resolution define the initial size of the multimodal context.

Video-centric LVLMs extend this paradigm to the temporal dimension. Early approaches like Video-ChatGPT (Maaz et al., 2024) apply average pooling over frame-level features to obtain compact representations. In contrast, models like Video-

LLaMA (Zhang et al., 2023) and VideoChat (Li et al., 2023b) utilize a Video Q-Former to aggregate temporal information. More recent efficient models, such as VideoLLaMA 3 (Zhang et al., 2025a) and SlowFast-LLaVA (Xu et al., 2024), employ hierarchical or dual-stream encoders to capture spatiotemporal dependencies without explicitly expanding the token count linearly with frame numbers.

Across these models, the vision encoder controls the number and density of visual tokens generated during the encoding stage and is therefore one of the primary factors shaping computational and memory cost throughout the inference pipeline.

Modality Adapter. The modality adapter $\mathcal{A}_\theta$ maps $\mathbf{X}_v$ to visual context $\mathbf{H}_v \in \mathbb{R}^{N_v \times D_\mathcal{L}}$. Modern implementations fall into two categories:

1. **Linear Projection.** Exemplified by LLaVA-1.5 (Liu et al., 2023a) and InternVL-3.5 (Wang et al., 2025d), this approach uses a simple MLP with compression ratio $r = N_v/N_p = 1$, preserving full visual granularity but incurring high prefilling cost.

2. **Learnable Query-Based Mechanisms.** Pioneered by models like BLIP-2 (Li et al., 2023a) and Video-LLaMA (Zhang et al., 2023), these methods utilize a fixed set of latent queries (e.g., via Q-Former or Video Q-Former) to extract semantic information from variable-length visual features. This process compresses dense visual inputs into a compact, fixed-length sequence of tokens regardless of the input resolution. Recent works such as Dynamic-VLM (Wang et al., 2024b) and TokenPacker (Li et al., 2025e) further refine this paradigm by introducing dynamic compression rates or coarse-to-fine injection schemes, aiming to balance high compression ratios with the preservation of fine-grained spatial details.

LLM Backbone. The LLM backbone $\mathcal{L}_\psi$ processes joint context $\mathbf{C}$ (concatenation of visual and text embeddings) to generate responses. Modern implementations build on pretrained LLMs: server-scale backbones like LLaMA (Grattafiori et al., 2024), Qwen (Yang et al., 2025a), Mistral (Jiang et al., 2023), and InternLM (Cai et al., 2024), or lightweight variants like Phi (Abdin et al., 2024) and Gemma (Team et al., 2024) for edge deployment.

LVLMs differ in how visual tokens are introduced into the backbone:

1. **Input Concatenation**: The dominant strategy, pioneered by LLaVA (Liu et al., 2023b), projects visual tokens into the textual embedding space and concatenates them directly with text tokens at the input layer. This allows visual information to flow through all self-attention layers, enabling deep multimodal interaction. Due to its architectural simplicity and training efficiency, this approach has become the mainstream strategy for recent open-source models.

2. **Cross-Attention Injection**: In contrast, architectures like LLaMA 3.2-Vision (Grattafiori et al., 2024) and Flamingo (Alayrac et al., 2022) inject visual information into intermediate layers via interleaved cross-attention modules. This approach typically keeps the pretrained LLM parameters frozen (or partially frozen) and uses these adapter layers to fuse visual features conditionally. While this avoids extending the input context length with dense visual tokens, it necessitates architectural modifications to the attention blocks and introduces additional parameters.

All multimodal reasoning ultimately occurs inside the backbone, its internal pathways determine how tokens are preserved, abstracted, or attenuated as computation proceeds. As a result, many inference-time efficiency techniques operate directly on this module, making it the central substrate governing both capability and efficiency in modern LVLMs.

B.3 Model Taxonomy

Although modern LVLMs broadly follow the unified architecture outlined above, existing research exhibits clear differentiation in how visual information is represented, injected, and managed. From an efficiency-oriented perspective, we categorize current LVLMs into three groups based on their approach to managing visual token count N_v and inference complexity:

Performance-Prioritized Models. These models aim to maximize multimodal capability. They primarily concentrate on designing refined training pipelines and curating high-quality training data to ensure robust multimodal alignment. Representative models include InternVL-3.5 (Wang et al., 2025d), Qwen2.5-VL (Wang et al., 2024d), DeepSeek-VL2 (Wu et al., 2024b), LLaVA-OneVision (Li et al., 2025a), Llama-3.2-Vision (Grattafiori et al., 2024), CogVLM2 (Wang et al., 2024f), Cambrian-1 (Tong et al., 2024), Yi-VL (AI et al., 2025), InternLM-XComposer2 (Dong et al., 2024) and Ovis (Lu et al., 2024). These models typically generate dense visual token sequences to preserve fine-

grained details, serving as high-computation baselines for efficiency studies.

Partially-Optimized Models. These models preserve the high-performance backbone of standard LVLMs but introduce specific optimization strategies targeting isolated bottlenecks, particularly in the adapter or token selection modules. They strive to balance performance and efficiency by reducing N_v through adapter-level compression or token selection strategies. Representative models include BLIP-2 (Li et al., 2023a), InstructBLIP (Dai et al., 2023), Qwen-VL (Bai et al., 2023), MiniGPT-4 (Zhu et al., 2023), MiniGPT-v2 (Chen et al., 2023a), mPLUG-Owl2 (Ye et al., 2024), and Honeybee (Cha et al., 2024).

Holistically-Optimized Models. Holistic models aim for end-to-end efficiency through system-level co-design or efficiency-native architectural innovations. A prime example of full-pipeline optimization is NVILA (Liu et al., 2025f), which introduces a "scale-then-compress" paradigm. It jointly optimizes the architecture by scaling up resolutions for precision while compressing visual tokens for efficiency, and further enhances the entire lifecycle from training to deployment with system-level accelerations. Other works achieve holistic efficiency by redesigning the architecture for specific constraints. MobileVLM V2 (Chu et al., 2024) co-designs a mobile-friendly vision encoder with a tailored small-scale LLM to achieve efficient inference on edge devices. In the video domain, VideoChat-Flash (Li et al., 2025f) implements a full-pipeline hierarchical compression strategy, progressively reducing redundancy from the visual encoder to the LLM to handle long contexts efficiently.

Discussion of Representative Architectures. Representative multimodal architectures also reflect several distinct pathways toward efficient inference. The Qwen series (Bai et al., 2025a; Qwen Team, 2026) exemplifies the shift toward Native Dynamic Resolution, while Qwen3.5 further integrates Hybrid Attention with Sparse MoE to substantially improve long-context efficiency, reportedly achieving up to 19× higher throughput. DeepSeek-VL2 (Wu et al., 2024b), by contrast, serves as a representative sparse-computation architecture whose Mixture-of-Experts (MoE) design effectively decouples overall model capacity from per-token inference cost. InternVL-3.5 (Wang et al., 2025d) highlights a system-oriented optimization perspective, where the combination of a Visual Resolution Router and Decoupled Deployment yields a 4.05× system-level inference speedup. Meanwhile, Llama-3.2-Vision (Grattafiori et al., 2024) adopts Cross-Attention Injection as a memory-efficient alternative to direct visual-text token concatenation, specifically aiming to alleviate the visual memory wall.

C System & Evaluation

This section surveys the architectural landscape of efficient LVLM serving and outlines the evaluation standards adopted by the community. We first categorize SOTA serving systems based on their decoupling paradigms. Subsequently, we systematize the evaluation landscape by compiling the industry-standard metrics and capability benchmarks commonly used to assess these systems.

C.1 System Architecture

The core challenge in serving LVLMs stems from the *conflicting resource affinities* inherent to distinct inference phases. The pipeline comprises three stages: Encoding (`E`), Prefilling (`P`), and Decoding (`D`), each exhibiting fundamentally conflicting resource requirements and performance characteristics. Specifically, the `E` and `P` stages are compute-intensive with low batch saturation, while the `D` stage is memory-intensive but supports high batch saturation. This fundamental conflict renders a monolithic, integrated service architecture inefficient. Depending on the specific decoupling granularity and resource allocation strategy, these serving systems can be categorized into three types: *Stage-based*, *Modality-based*, and *Resource-based*.

Stage-based. Stage-based strategies decompose model inference into distinct temporal stages. EPDServe (Singh et al., 2024) exemplifies this approach by using a black-box optimizer to identify the optimal configuration based on historical workload analysis and introduces dynamic role switching to enhance system adaptivity. RServe (Guo et al., 2025) adopts a fine-grained scheduling method to overlap `E` and `P` stages, thereby mitigating inter- and intra-request pipeline bubbles. HydraInfer (Dong et al., 2025) treats each stage as a composable object (e.g., `EP+D`, `E+P+D`) and utilizes a hybrid EPD disaggregation profiler to dynamically deploy the optimal decoupling topology according to Service Level Objective (SLO) requirements.

Modality-based. These strategies partition resource groups according to the type of request or functional module. ModServe (Qiu et al., 2025) segregates resources into a dedicated image pool for visual encoding and a text pool for the LLM backbone. ElasticMM (Liu et al., 2025e) similarly groups instances into text-only and multi-modal pools. However, unlike functional decoupling, ElasticMM executes the entire inference pipeline within each respective group and introduces elastic

Paradigm	Optimal Workload	Key Mechanism	Performance Impact		Communication Cost
			P99 TTFT	TPOT	
Stage-based *(e.g., EPDServe)*	**Long Video / Heavy Prefill** (Prefill-dominant)	Inter-device Disaggregation (Avoids Compute Blocking)	↓↓ (Best Stability)	↑ (Dedicated)	**High** (Context/KV Transfer)
Modality-based *(e.g., ModServe)*	**Balanced Multimodal** (General VQA)	Inter-device Partitioning (Resource Specialization)	– (Stable)	↑↑ (Pipeline)	**Medium** (Embedding Transfer)
Resource-based *(e.g., SpaceServe)*	**Latency-Sensitive / Edge** (Real-time Interaction)	SM-level Multiplexing (Zero Network Hops)	↓ (Fastest Avg.)	↑ (Utilization)	**None** (Intra-GPU Fusion)

Table 4: Qualitative comparison of LVLM serving paradigms, mapping architectural choices to workload characteristics. Legend: ↓/↑ denotes latency/throughput improvement; P99 TTFT indicates worst-case stability.

Metric	Category	Formulation	Definition
TTFT	Latency	$t_{\text{first}} - t_{\text{arr}}$	*Time to First Token*. The duration of the prefilling phase, measured from the request arrival time t_{arr} to the first token generation t_{first}.
TPOT	Latency	$\frac{t_{\text{end}} - t_{\text{first}}}{N}$	*Time Per Output Token*. The generation speed during decoding, measured from the first token t_{first} to the last token generation t_{end}, averaged over N tokens.
SLO Attainment	Reliability	$\frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(L_r \leq \tau)$	*Service Level Objective Attainment*. Proportion of total requests M where the end-to-end latency L_r of request r satisfies the threshold τ. $\mathbb{I}(\cdot)$ is the indicator function.
Goodput	Throughput	$\frac{1}{T}\sum_{i=1}^{M}\mathbb{I}(L_r \leq \tau)$	*Effective Throughput Rate*. The number of requests per second that strictly satisfy the SLO constraint τ, calculated over the total serving duration T.

Table 5: Taxonomy of efficiency metrics, categorized by performance dimension (Latency, Reliability, Throughput).

Domain	Task Competency	Representative Benchmarks
Image	Multimodal Reasoning	MathVista (Lu et al., 2023), MMMU (Yue et al., 2024), MathVision (Wang et al., 2024c), DynaMath (Zou et al., 2024), LogicVista (Xiao et al., 2024), VPCT (Shao et al., 2024), MMMU-Pro (Yue et al., 2025), EMMA (Hao et al., 2025), SFE (Zhou et al., 2025b), ZeroBench (Roberts et al., 2025), WebQA(Chang et al., 2022), MultiModalQA(Talmor et al., 2021)
	General Visual QA	HallusionBench (Guan et al., 2024), MMStar (Chen et al., 2024c), MMBench (Liu et al., 2024b), MUIRBench (Wang et al., 2024a), MMVP (Zhang et al., 2024b), VLMsAreBlind (Rahmanzadehgervi et al., 2024), VLMsAreBiased (Vo et al., 2025), SimpleVQA (Cheng et al., 2025c), MME-CC (Zhang et al., 2025g), MMCoQA(Li et al., 2022), SlideVQA(Tanaka et al., 2023)
	Long-Context Understanding	DUDE (Van Landeghem et al., 2023), MMLongBench (Wang et al., 2025f), OminiDocBench (Ouyang et al., 2025), MileBench(Song et al., 2024a)
Video	Long Video Understanding	CGBench (Chen et al., 2024a), LongVideoBench (Wu et al., 2024a), MLVU (Zhou et al., 2025a), LVBench (Wang et al., 2024e), ALLVB (Tan et al., 2025), VideoMME (Fu et al., 2025), VDC (Chai et al., 2025)
	Knowledge & Reasoning	VideoMMMU (Hu et al., 2025b), MMVU (Zhao et al., 2025), VCRBench (Qi et al., 2025), VideoReasonBench (Liu et al., 2025d), VideoHolmes (Cheng et al., 2025b), Minerva (Nagrani et al., 2025), VideoSimpleQA (Cao et al., 2025), NExT-QA(Xiao et al., 2021), Video-ChatGPT(Maaz et al., 2024)
	Motion & Perception	Countix (Dwibedi et al., 2020), TVBench (Cores et al., 2024), TempCompass (Liu et al., 2024c), TOMATO (Shangguan et al., 2024), MVBench (Li et al., 2024), EgoTempo (Plizzari et al., 2025), MotionBench (Hong et al., 2025), Vatex(Wang et al., 2019)

Table 6: Taxonomy of LVLM capability benchmarks, categorized by input domain (Image or Video) and chronological evolution from single-image perception to complex reasoning.

partition scheduling to dynamically reallocate instances or preempt decoding tasks for prefill bursts based on a gain-cost model. Both approaches rely on inter-device communication to coordinate the multi-modal data flow.

Resource-based. Unlike inter-device partitioning, this paradigm works at the hardware-resource granularity. SpaceServe (Zhang et al., 2025k) is a representative example via Streaming Multiprocessor (SM)-level partitioning. It logically separates modality encoders and the text decoder, yet runs them on the same GPU. By assigning SM resources to different tasks, it improves utilization and removes inter-device communication overhead that can bottleneck cross-node modality pools.

Our analysis of the trade-offs across different serving architectures is summarized in Table 4.

C.2 Evaluation Standards

To provide a structured understanding of the efficiency landscape, we systematize the evaluation standards into two complementary dimensions: *efficiency metrics* and *capability benchmarks*. This taxonomy serves as a guideline for analyzing the trade-offs between serving latency and multimodal generation quality. Validating optimization techniques requires a dual approach: quantifying speed-up gains using industry-standard metrics while ensuring, through rigorous benchmarking, that model utility is preserved across diverse contexts.

Efficiency Metrics. We categorize the metrics used to quantify inference efficiency into latency-oriented and throughput-oriented indicators, as defined in Table 5. These metrics constitute the standard framework for evaluating serving system performance in production environments.

Real-world Alignment: Cost, Scalability, and Energy Efficiency. To bridge the gap between technical metrics and production deployment, we formalize the transition of the metrics defined in Table 5 into economic and operational indicators:

- **Cost Savings (CPSR):** For service providers, raw throughput must be translated into the *Cost Per Successful Request* (CPSR, [USD / req]), representing unit economic efficiency. We derive this as:

$$CPSR = \frac{C_{\text{node}}}{G \cdot T_{\text{ref}}} \tag{7}$$

  where C_{node} denotes the operational expenditure (OpEx) of the hardware instance [USD / node] over a reference time window T_{ref} [s], and G represents the **Goodput** [req / s]. By maximizing Goodput, developers can increase request density per hardware unit, directly lowering the CPSR by reducing amortized infrastructure overhead.

- **Scalability Efficiency (η_{scale}):** We define scalability as the system's capacity to maintain **SLO Attainment** under an *iso-resource scaling* scenario, where both hardware capacity H (e.g., number of GPU nodes) and request volume M are increased by a factor of k:

$$\eta_{\text{scale}} = \frac{\text{SLO Attainment}(k \cdot M \mid k \cdot H)}{\text{SLO Attainment}(M \mid H)} \tag{8}$$

  The notation $(M \mid H)$ denotes the performance measured under workload M conditioned on hardware resources H. In large-scale clusters, an ideal system maintains $\eta_{\text{scale}} \approx 1$, signifying linear scalability. A significant degradation indicates systemic bottlenecks, such as VRAM saturation or interconnect contention, triggering the need for elastic resource orchestration.

- **Energy Efficiency (EPT):** Sustainability is quantified via *Energy Per Token* (EPT, [J / token]). While **TTFT** captures the compute-bound energy burst during prefilling, cumulative energy is primarily governed by **TPOT**. Given the low arithmetic intensity of decoding ($AI \ll 1$), EPT is formulated as:

$$EPT \approx P_{\text{TDP}} \times TPOT_{\text{effective}} \tag{9}$$

  where P_{TDP} is the Thermal Design Power [W, or J/s] and $TPOT_{\text{effective}}$ is the effective decoding latency [s / token]. Reducing TPOT minimizes the duration GPUs spend in high-power active states, serving as the primary driver for energy sustainability.

Capability Benchmarks. Optimization strategies must be validated against established performance standards to ensure that efficiency gains do not compromise model fidelity. In Table 6, we curate a taxonomy of prevailing benchmarks spanning static image reasoning and dynamic video understanding. We order these datasets chronologically to illustrate the shift of the community towards increasingly complex tasks, ranging from fine-grained visual perception to long-context temporal reasoning.

D Future-Forward Pilot Exploration

D.1 Hybrid KV Cache Compression

To validate the potential of *hybrid compression mechanisms*, we explore a differentiated strategy that moves beyond uniform compression with adaptation to budget allocation (Zeng et al., 2026). Specifically, we utilize text-visual information to categorize attention heads, thereby allocating varying budgets and orchestrating a hybrid compression mechanism combining pruning and retrieval. We conduct preliminary experiments to evaluate this approach on Qwen2.5-VL-7B [9] (Bai et al., 2025b) using NVIDIA L40S GPUs.

Performance Evaluation. We compare our hybrid scheme against SOTA KV compression methods in LVLMs (LOOK-M (Wan et al., 2024b), MadaKV (Li et al., 2025d), SparseMM (Wang et al., 2025b)) across four image benchmarks (SlideVQA (Tanaka et al., 2023), MMCoQA (Li et al., 2022), WebQA (Chang et al., 2022), MultiModalQA (MM-QA) (Talmor et al., 2021)) from Milebench and one video benchmark (VideoChatGPT (Maaz et al., 2024)). Table 7 shows that the hybrid compression consistently outperforms uniform compression baselines, achieving performance competitive with the full cache upper bound.

[9] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

Method	Image				Video-ChatGPT				
	SlideVQA	MMCoQA	WebQA	MM-QA	CI	DO	CU	TU	CO
Full Cache	83.50	66.50	76.50	75.00	3.06	3.15	3.52	2.24	2.69
LOOK-M	82.50	52.50	71.00	73.50	2.92	2.97	3.38	2.05	2.49
MadaKV	82.00	55.00	70.50	74.50	2.94	3.03	3.41	2.02	2.56
SparseMM	83.50	62.00	70.50	75.50	2.90	2.95	3.37	1.99	2.52
Hybrid	**83.50**	**63.00**	**76.00**	**76.00**	**2.99**	**3.05**	**3.47**	**2.15**	**2.57**

Table 7: Performance of four KV cache compression on Qwen2.5-VL-7B across image and video tasks. Image tasks use *exact match accuracy*. For Video-ChatGPT, scores (ranging from 0 to 5) are generated by `gpt-4o-mini` across five dimensions: Correctness of Information (CI), Detail Orientation (DO), Contextual Understanding (CU), Temporal Understanding (TU), and Consistency (CO).

Method	Budget	Accuracy (Avg.)	GPU Memory (GB)	Latency (ms/token)
Full Cache	100%	2.93	1.73	58.94
Hybrid	20%	**2.88**	0.40	42.08
Hybrid	10%	2.85	**0.22**	**38.65**

Table 8: KV cache GPU memory usage and decoding latency on Qwen2.5-VL-7B with Video-ChatGPT.

Efficiency Evaluation. Further, we assess the efficiency of the hybrid KV compression with Video-ChatGPT for real-world long video understanding scenarios. We randomly sample 20 data entries and set the maximum generation length to 128 tokens for evaluation. All experiments use FlashAttention (Dao, 2023). As shown in Table 8, our method markedly reduces both GPU memory and decoding latency relative to the Full Cache baseline.

D.2 Relaxed Speculative Decoding

As discussed in Section 5, speculative decoding for LVLMs still leaves substantial room for exploration, as many techniques in speculative decoding for LLMs (Zhang et al., 2024c; Xia et al., 2025; Shen et al., 2026b; Song et al., 2025b) remain largely unexplored. Beyond modality-aware draft models (Ji et al., 2025; Kong et al., 2026; Zhang et al., 2026; Shen et al., 2026a), *modality-aware verification strategies* constitute another promising direction. In prevailing benchmarks for visual captioning and open-ended VQA, the importance of the output tokens may be non-uniform. Intuitively, descriptively visual tokens are more critical and thus require strict verification, whereas some prepositions, conjunctions, and other function words can be verified with relaxation (Ji et al., 2026). This aligns with the idea that *exact match* is not always needed in the recent study of relaxed speculative decoding for LLMs (Bachmann et al., 2025). To validate this phenomenon, we conduct a simple experiment on Qwen2.5-VL and Video Detail Caption (Chai et al., 2025) benchmark, using two NVIDIA H200 GPUs. The dataset is divided into four subsets, from which we sample 4×30 instances for evaluation. We select the training-free speculative decoding method SpecVLM (Ji et al., 2025) as baseline, which applies visual token pruning (90%) to the draft model and adopts strict match verification. As a comparison, we construct a "random relaxation" method, where mismatched tokens during decoding are accepted with a random probability. The results are reported in Table 10 and Table 9.

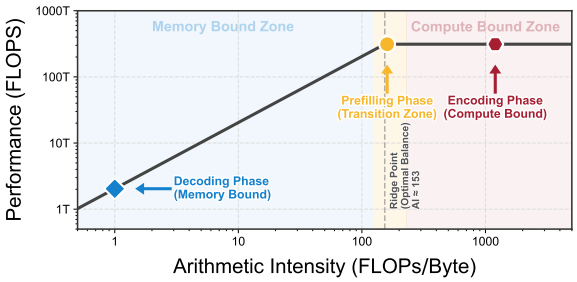

Figure 4: Stage-wise bottleneck analysis of generic LVLM inference on NVIDIA A100. We illustrate the operational intensity of distinct inference phases against the hardware Roofline limits. The *Decoding* phase is strictly memory-bound ($\mathcal{I}_a \approx 1$), constrained by bandwidth. In contrast, the *Visual Encoding* phase represents a compute-bound workload ($\mathcal{I}_a \approx 1200$). The *Prefilling* phase ($\mathcal{I}_a \approx 160$) occupies the transitional region near the hardware's ridge point ($\mathcal{I}_{ridge} \approx 153$ FLOPs/Byte), utilizing both compute and memory resources efficiently.

Efficiency Evaluation. By adopting random relaxation, the verification stage of speculative decoding is directly relaxed, resulting in a largely boosted mean accepted length. Compared with SpecVLM, the random relaxation variants increase the decoding speedup from 1.40×, 0.97×, and 1.26× to 2.61×, 2.11×, and 1.80×, respectively, effectively pushing the efficiency boundary of speculative methods.

Performance Evaluation. Remarkably, despite achieving the significant speedups mentioned above, random relaxation retains 81.4% to 97.6% of the original output quality across different model settings. This preservation of quality provides preliminary evidence that the output patterns of visual tasks exhibit certain sparsity, implying that many mismatches can be relaxed without severe detriment. Exploiting such output patterns, where visual entities co-occur with prepositions, conjunctions, and other function words, to perform adaptive relaxed speculative decoding remains an interesting direction for future work.

E Roofline Analysis Details

This section details the hardware profiling methodology and the theoretical framework underpinning the *arithmetic intensity* ($\mathcal{I}$) estimations used in our Roofline analysis (see Figure 4).

Target / Draft Model	Method	Video Detail Caption										
		Main Object		Detail		Camera		Background		Average		
		Acc(%)	Score	Acc(%)	Score	Acc(%)	Score	Acc(%)	Score	Acc(%)	Score	Ret.(%)
Qwen2.5-VL-32B / 7B	SpecVLM	34.52	2.11	31.12	1.93	28.59	1.78	30.85	1.91	31.27	1.93	100.0
	+Random Relaxation (50%)	31.85	1.94	32.63	2.00	28.37	1.82	25.13	1.70	29.50	1.87	**95.6**
	+Random Relaxation (75%)	31.23	1.97	31.87	1.90	26.22	1.73	25.63	1.75	28.74	1.84	**93.6**
Qwen2.5-VL-32B / 3B	SpecVLM	34.52	2.11	31.12	1.93	28.59	1.78	30.85	1.91	31.27	1.93	100.0
	+Random Relaxation (50%)	31.43	1.94	30.90	1.95	15.69	1.25	24.43	1.58	25.61	1.68	**84.5**
	+Random Relaxation (75%)	28.78	1.80	29.46	1.81	20.03	1.37	21.07	1.45	24.84	1.61	**81.4**
Qwen2.5-VL-7B / 7B	SpecVLM	30.79	1.95	33.44	2.02	25.80	1.69	24.77	1.56	28.70	1.81	100.0
	+Random Relaxation (50%)	29.22	1.85	28.37	1.85	26.39	1.78	26.15	1.66	27.53	1.79	**97.4**
	+Random Relaxation (75%)	27.93	1.81	31.30	1.94	25.63	1.63	26.96	1.70	27.96	1.77	**97.6**

Table 9: Performance metric using Qwen2.5-VL on three speculative decoding settings and Video Detail Caption benchmark. Ret.(%) refers to the performance retention ratio on average of accuracy and score, compared with autoregressive decoding.

Target / Draft Model	Method	Video Detail Caption	
		Mean Accepted Length	Speedup
Qwen2.5-VL-32B / 7B	SpecVLM	3.40	1.40×
	+Random Relaxation (50%)	5.55	2.04×
	+Random Relaxation (75%)	7.42	2.61×
Qwen2.5-VL-32B / 3B	SpecVLM	2.99	0.97×
	+Random Relaxation (50%)	5.30	1.64×
	+Random Relaxation (75%)	7.33	2.11×
Qwen2.5-VL-7B / 7B	SpecVLM	5.31	1.26×
	+Random Relaxation (50%)	7.20	1.61×
	+Random Relaxation (75%)	8.45	1.80×

Table 10: Efficiency metrics using Qwen2.5-VL on three speculative decoding settings and Video Detail Caption benchmark. Draft tokens per decoding step is set to 10. Decoding speedup is measured relative to autoregressive decoding.

E.1 Hardware Specifications

We employ the NVIDIA A100-SXM4-80GB GPU as the reference hardware platform. Performance limits are derived based on Half-Precision (FP16) tensor core operations, the standard precision for Large Vision-Language Model (LVLM) inference. The specifications are summarized in Table 11.

Parameter	Value
Peak Performance (π_{peak})	312 TFLOPS (FP16 Tensor Core)
Memory Bandwidth (β_{mem})	2,039 GB/s (HBM2e)
Ridge Point ($\mathcal{I}_{ridge}$)	$\approx$ 153 FLOPs/Byte

Table 11: Hardware specifications for the NVIDIA A100 (80GB) used in the Roofline model.

The *Ridge Point* ($\mathcal{I}_{ridge}$), delineating the boundary between memory-bound and compute-bound regimes, is calculated as:

$$I_{ridge} = \frac{\pi_{peak}}{\beta_{mem}} = \frac{312 \times 10^{12}}{2039 \times 10^{9}} \approx 153.0 \text{ FLOPs/Byte} \quad (10)$$

E.2 Workload Characterization by Stage

We characterize the three distinct stages of modern LVLM inference by analyzing their theoretical arithmetic intensity ($\mathcal{I}_a$).

Decoding ($\mathcal{I}_a \approx 1.0$). The decoding stage follows an autoregressive generation pattern, producing one token per step. This operation is dominated by Matrix-Vector multiplication (GEMV). For a model with parameters θ, generating a single token necessitates loading the entire weight matrix to perform the computation. Under FP16 precision (2 bytes per parameter), the intensity is derived as:

$$I_{dec} \approx \frac{2 \cdot |\theta| \cdot 1 \text{ (token)}}{2 \cdot |\theta| \text{(bytes)}} = 1.0 \text{ FLOPs/Byte} \quad (11)$$

Consequently, the decoding stage is strictly **memory-bound**, situated significantly to the left of the ridge point. Performance in this regime is solely determined by memory bandwidth utilization.

Prefilling ($\mathcal{I}_a \approx 160.0$). The prefilling stage processes the input prompt in parallel, relying on Matrix-Matrix multiplication (GEMM). Unlike decoding, the arithmetic intensity here scales with the input sequence length $(N_v + N_t)$ due to weight reuse. We visualize a representative operational point of $\mathcal{I}_a \approx 160.0$, which corresponds to moderate-to-long context lengths (e.g., $(N_v + N_t) \approx 512$). This value lies in the immediate vicinity of the hardware ridge point (153.0), indicating a mixed bottleneck regime. In this region, the workload simultaneously saturates memory bandwidth and approaches peak compute utilization, making performance highly sensitive to both data movement and arithmetic throughput optimization.

Encoding ($\mathcal{I}_a \approx 1200.0$). The encoding stage processes high-resolution image inputs via a Vision Transformer (ViT) backbone. Unlike the sparse memory access patterns in decoding, the vision encoder performs dense, highly parallel computations on image patches. We model this workload with an approximate intensity of $\mathcal{I}_a \approx 1200.0$, nearly an order of magnitude higher than the ridge point. This classifies the visual encoder as strictly compute-bound, implying that optimizations to memory bandwidth yield negligible performance gains in this stage.

F LLM Usage

Large Language Models (LLMs) were used to aid in code writing and manuscript polishing. Specif-

ically, the usage includes refining the language, improving readability, and ensuring clarity in the paper. It is important to note that LLMs were not involved in the ideation, research methodology, or experimental design.