POINTS-Long: Adaptive Dual-Mode Visual Reasoning in MLLMs

Haicheng Wang[1,2*], Yuan Liu[2*✉], Yikun Liu[1,2*], Zhemeng Yu[1], Zhongyin Zhao[2],
Yangxiu You[2], Zilin Yu[2], Le Tian[2], Xiao Zhou[2], Jie Zhou[2], Weidi Xie[1], Yanfeng Wang[1✉]
[1] SAI, Shanghai Jiao Tong University, China [2] WeChat AI, Tencent, China

Abstract

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable capabilities in cross-modal understanding and generation. However, the rapid growth of visual token sequences—especially in long-video and streaming scenarios—poses a major challenge to their scalability and real-world deployment. Thus, we introduce POINTS-Long, a native dual-mode MLLM featuring dynamic visual token scaling inspired by the human visual system. The model supports two complementary perception modes: focus mode and standby mode, enabling users to dynamically trade off efficiency and accuracy during inference. On fine-grained visual tasks, the focus mode retains the optimal performance, while on long-form general visual understanding, the standby mode retains 97.7-99.7% of the original accuracy using only 1/40-1/10th of the visual tokens. Moreover, POINTS-Long natively supports streaming visual understanding via a dynamically detachable KV-cache design, allowing efficient maintenance of ultra-long visual memory. Our work provides new insights into the design of future MLLMs and lays the foundation for adaptive and efficient long-form visual understanding. Model and code are available at Link.

1. Introduction

Multimodal Large Language Models (MLLMs) [3, 21, 49, 52, 63–65, 70, 79, 86] have recently achieved remarkable progress in cross-modal comprehension and reasoning. However, these remarkable abilities come at a steep computational cost when processing long visual content like videos. The root cause lies in the visual tokenization, which expands the total sequence length with video duration, resulting in quadratic growth of computation and memory costs. This inherent scalability bottleneck remains a critical challenge for real-world long-duration applications.

Extensive research has recently yielded sophisticated strategies for visual sequence compression [41, 72, 99].

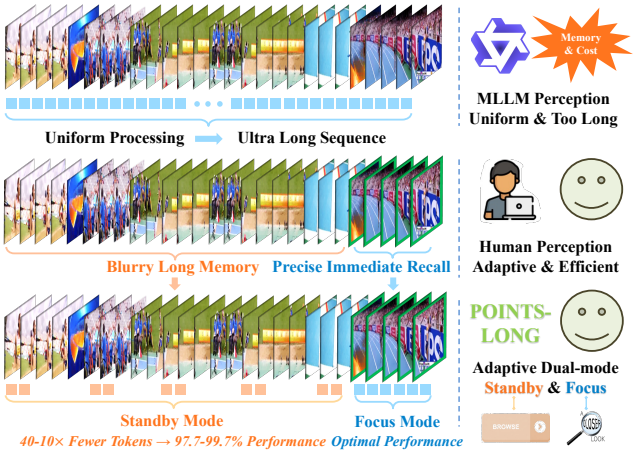

Figure 1. **POINTS-Long: Bridging the Gap between Human Visual Perception and MLLM Scalability.** Inspired by human's adaptive visual processing, POINTS-Long introduces a dual-mode system which switches between high-fidelity Focus Mode and efficient Standby Mode, enabling both detailed analysis and long-term streaming understanding with significantly reduced cost.

Nevertheless, most widely-used MLLMs still rely on simple methods like pixel-shuffle [69, 70] and pooling [21]. This gap between research and practice stems from three key challenges hindering the adoption of advanced techniques in production systems: (1) Insufficient Compression Ratio: The reduction ratio is inadequate for long-video applications (thousands of frames) without a significant drop in performance [62, 68, 83]. (2) Lack of Generality: Models are often forced into a trade-off, becoming either efficient long-video specialists [37, 42, 58] that sacrifice fine-grained reasoning, or capable reasoners that cannot scale, limiting their utility as all-in-one assistants. (3) Deployment Difficulty: Many methods [75, 80, 94] are incompatible with modern inference optimizations or frameworks (e.g., Flash-Attn [14], vLLM [31], SGLang [97]), preventing their theoretical efficiency from being realized in practice.

This suggests that a paradigm shift, rather than incremental improvements, may be necessary. We are thus motivated to ask: *Is the current monotonous approach to visual processing in MLLMs inherently flawed?* We draw inspira-

*: Core contributor. ✉: Corresponding author.

tion from the human visual system, which effortlessly processes a continuous stream of visual information without being overwhelmed. Human perception appears to operate in at least two distinct modes: a focused mode for high-fidelity details and a standby mode for low-effort, general awareness [18, 66]. This duality is also reflected in our hierarchical memory [2, 50]: precise immediate recall, blurry short-term memories, and semantic long-term recollections, like a textual summary. This reveals an efficient architecture: a precise buffer for the present, a compressed cache for short-term, and a conceptual archive for long-term.

Inspired by this human cognitive model, we introduce POINTS-Long, a MLLM built upon POINTS1.5 [47]. Its core innovation is a native dual-mode visual processing system: **Focus Mode**: Uses the complete visual sequence for tasks requiring fine-grained analysis, ensuring maximum performance. **Standby Mode**: Operates on a drastically reduced number of visual tokens for the holistic perception of long videos, with only a negligible drop in performance.

To implement this functionality without compromising the model's original strengths, we employ a two-stage post-training adaptation process. First, in a visual distillation stage, we freeze the original MLLM and train a small set of new parameters to distill the rich information from the full visual sequence into a compact set of "Standby tokens" (Sec. 3.2). This ensures that the Standby tokens are semantically aligned with the original "Focus tokens" while leaving the Focus mode's pathway entirely unaffected. In the second stage, we adapt the LLM by fine-tuning it with a small learning rate on high-quality data, enabling it to effectively interpret inputs from both modes (Sec. 3.3.3).

This strategic approach yields remarkable efficiency: on the OpenCompass video benchmark, our Standby mode retains 97.7%–99.7% of the original model's performance while using just 1/40th to 1/10th of the visual tokens. Crucially, this efficiency is achieved without compromise, as the Focus mode fully preserves the model's original fine-grained capacity. Furthermore, this dual-mode architecture enables a more effective approach to streaming vision. By dynamically combining modes, POINTS-Long emulates a human-like memory system—a high-fidelity "present" (Focus) and a compressed "short-term" (Standby)—through a novel detachable KV cache mechanism. This allows for native, long-term understanding without costly context re-prefills. Notably, POINTS-Long is designed for practical deployment; all evaluations were conducted using SGLang [97] inference framework. Overall, our contributions can be summarized as follows:

- We introduce POINTS-Long, a novel MLLM inspired by human cognition. It features a dual-mode visual system (Focus and Standby) that resolves the critical trade-off between fine-grained reasoning and long-vision scalability.
- We propose a generalizable two-stage post-training strategy that can efficiently equip a well-trained MLLM with the high-compression Standby mode while fully preserving its original performance in the Focus mode.
- We demonstrate the practical viability and state-of-the-art efficiency of our approach. POINTS-Long natively supports long-term streaming video understanding through a novel detachable KV cache mechanism and is fully compatible with modern inference frameworks, achieving up to 6.2× generation throughput with negligible loss.

2. Related Work

Video Large Language Models. MLLMs have demonstrated impressive capabilities in understanding multimodal information like video [3, 21, 32, 34, 63–65, 70, 79, 89]. However, the rapid growth in computational cost from the large number of visual tokens severely limits their scalability for practical, long-form video tasks. To address this bottleneck, some MLLMs [37, 42, 58, 62] employ visual token compression for efficient long-form understanding. However, they often result in highly specialized models: some sacrifice fine-grained image reasoning to become video experts, while others [53, 91] built for streaming video are even more task-specific. This specialization highlights a critical need for a native MLLM that can perform both long-video processing and precise image analysis.

Efficient MLLMs Inference. The practical deployment of MLLMs is dominated by inference frameworks like vLLM [31] and SGLang [97]. These systems achieve state-of-the-art throughput by leveraging kernel-level optimizations like FlashAttention [14] and PagedAttention [31]. However, many visual token reduction methods are incompatible with these frameworks (or hard to implement), *e.g.,* requiring explicit attention matrices or disrupting the uniform block structure of the KV cache. As a result, their theoretical efficiency doesn't translate to real-world performance, severely limiting their practical use.

Visual Token Reduction in MLLMs. Some preliminary studies mainly focus on Vision Transformers [4, 30, 55] and KV cache compression [38, 61, 96] for LLMs. In the context of MLLMs, common methods like Q-Former [33], resampler [12] and pooling [8] are widely used during the training phase to reduce visual tokens. Recently, some studies tried to handle the token reduction problem in more delicate ways [1, 23–25, 56, 67, 73, 84]. In particular, training-free methods mainly leverage task-orientated attention importance [10, 43, 75, 94], or inherent visual redundancy [28, 68, 81], compromising efficiency with performance. Methods that require additional training [36, 39, 40, 57, 93] can compress visual tokens more effectively, but they often enforce a fixed trade-off, leading to performance degradation and poor extensibility. We aim to build a natively adaptive MLLM that provides the flexibility to dynamically balance between computational efficiency and reasoning accuracy.

3. Method

3.1. Overview

Our dynamic visual understanding framework is inspired by the human visual system, incorporating both a Focus Mode and a Standby Mode. This design aims to selectively and drastically reduce computational load, and potentially maintain long visual memory, which is guided by four key principles: (P1) Performance Preservation: The Focus Mode remains equivalent to the original, well-trained MLLM. (P2) Optimized Standby Performance: The Standby Mode strives to approximate Focus Mode quality with drastically lower cost. (P3) Deployment Simplicity: The architecture should be easy to deploy and compatible with modern inference frameworks (e.g., vLLM [31], SGLang [97]) for real-world speed-ups. (P4) Extensibility: The training solution should be adaptable to a wide variety of existing MLLMs.

To adhere to these principles, we begin with an instruct model and introduce the Standby Mode capacity via one post-training phase. We add several learnable modules between the vision backbone and the projector (Sec. 3.3), a solution designed to satisfy (P1) and (P3) while maximizing the performance of (P2). We then propose a two-stage training strategy, including (1) Visual Distillation and Alignment (2) LLM Mode Adaptation, to efficiently integrate this new mode (Sec. 3.3.3). The resulting model natively supports both focus and standby modes. This dual-mode capability, with a novel detachable KV cache mechanism, allows the model to naturally support efficient, streaming visual understanding (Sec. 3.4) while keeping its full capacity.

3.2. Architecture of Base MLLM

We use POINTS1.5-8B-Instruct (improved version of POINTS1.5 [47]) for experiments, a highly competitive MLLM comparable to mainstream MLLMs like Qwen2.5-VL [3]. It is composed of a LLM initialized from Qwen3-8B-base [78] (1D RoPE [60] for visual inputs) and a native-resolution image encoder initialized from Qwen2-VL-ViT [69] (employing 2D RoPE). This base model has already undergone a comprehensive, multi-stage training pipeline, including multimodal alignment, continued pre-training, multimodal SFT, and post-training phase. (details are in supplementary material). Our proposed dynamic dual-mode scheme is applied as a post-training phase on top of this instruct model. Note that our approach can be applied to any MLLM following a similar architecture.

3.3. Native Visual Compression Structure

Starting from POINTS1.5-8B-Instruct architecture, we introduce a novel modification to the vision backbone (ViT) and projector, keeping the original inference path unchanged. The core objective is to distill the vast information from the original visual sequence into a small set of tokens, enabling a highly efficient "standby" mode without compromising the performance of the original "focus" mode.

3.3.1. Dual-Path ViT Architecture

Inspired by CLIP [54], we append n learnable tokens onto the patchified sequence, where n is significantly smaller than the average visual sequence length. These tokens are intended to act as a compressed representation of the full sequence. However, integrating these learnable tokens introduces a training dilemma: (1) If we freeze the ViT and train only the learnable tokens, the model lacks the fitting capability to distill complex visual information, leading to poor performance (Tab. 6). (2) Unfreeze ViT can improve the fitting capability, but the training dynamics are altered, impairing the model's original "focus mode" performance.

To resolve this, we re-architect the ViT by introducing a parallel processing path for new learnable tokens, shown in Fig 2. Similar to MoT [15], for each MLP layer in the ViT, we duplicate it to create a new one, which is initialized with the weights of the original MLP. The original visual sequence is processed by the original MLPs, while the new learnable tokens are processed exclusively by these new MLPs. This parallel structure is also mirrored in the final projector with the same operation. The key interaction between these two paths is the shared attention block. This simple design significantly boosts the performance (Tab. 6).

In addition, to preserve the invariability of the original "focus" path, we employ an asymmetric attention mask: the original patch tokens compute attention only among themselves (masking out new learnable tokens), ensuring the invariance of their representations. In contrast, the learnable tokens are allowed to attend to the entire sequence, enabling them to aggregate global visual information. This simple masking strategy is fully compatible with Flash Attention [14]. Finally, we assign positional embeddings to the learnable tokens by uniformly sampling the original 2D RoPE [60], an initialization technique that, as visualized in our supplementary material, encourages different tokens to specialize in different spatial regions of the image.

Discussion This methodology was designed to achieve the 4 principles outlined in Sec. 3.1: The parallel ViT architecture and asymmetric attention mask ensure the original path is undisturbed, maintaining focus mode performance and easing deployment. The added parameters enhance the representation ability, boosting the capacity of standby mode.

3.3.2. Temporal Modeling

The architecture so far originates from POINTS1.5 image encoder and, consequently, solely addresses intra-frame spatial redundancy, overlooking the significant temporal redundancy in video inputs, which can be more critical.

A naive application of our method would compress each frame into n tokens independently and then concate-

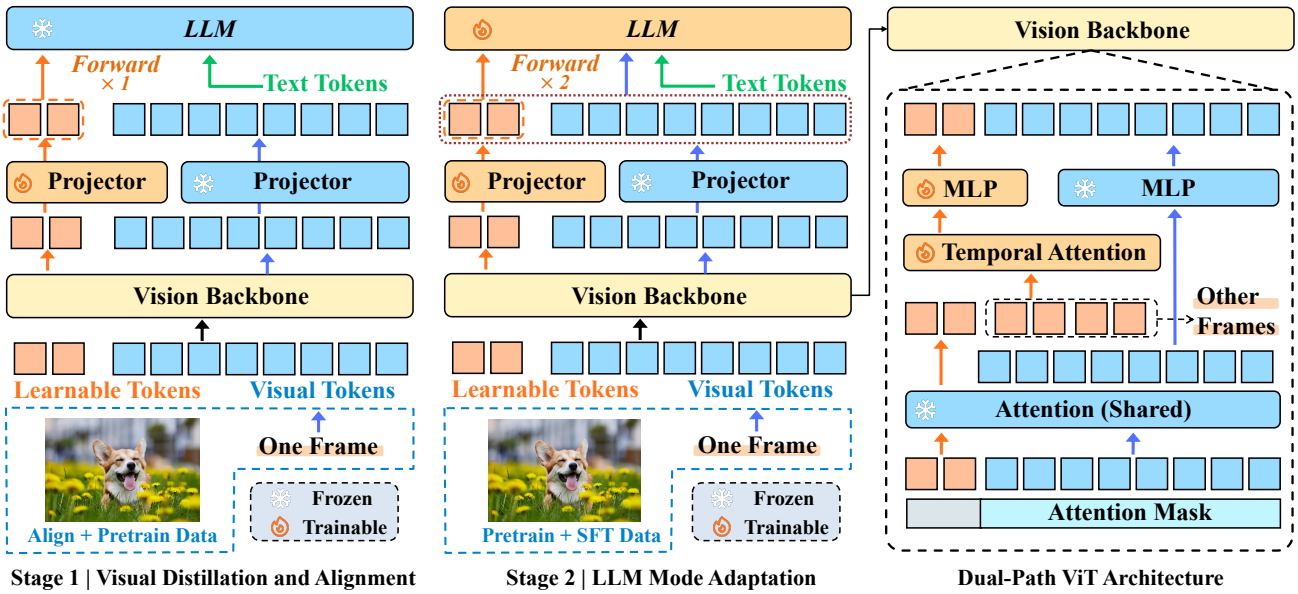

Figure 2. **POINTS-Long Architecture.** The original visual patch sequence (blue) is processed by the original ViT modules. We introduce n learnable tokens (orange) processed through duplicated learnable MLPs and projector, to act as the compressed representation of the full sequence. An additional temporal modeling allows better compression for video inputs. With symmetric attention mask, the original path is totally unaffected, thus preserving its performance. This dual-mode system is enabled by a two-stage post-training: Stage 1 (left) trains only the new parameters for visual distillation, while Stage 2 (middle) fine-tunes the LLM with a small learning rate for mode adaptation.

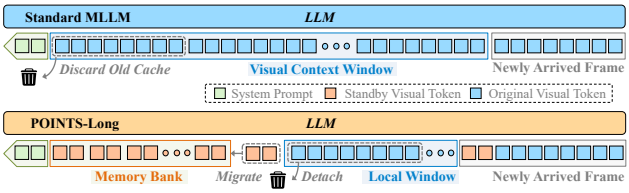

Figure 3. **Streaming Inference in LLM.** (↑) When handling streaming inputs, general MLLMs discard previous cached context when reaching maximum budget. (↓) POINTS-Long encodes new frames in Focus Mode. When local window is full, the original sequence's cache is detached, and the compact standby-sequence cache is migrated to a long-term "Memory Bank".

nate. However, a joint compression that models spatio-temporal relationships could achieve higher information fidelity. Therefore, to further enhance the standby mode's efficacy for video, we introduce an explicit temporal modeling component. As shown in Fig 2, we insert a temporal attention module into the final layers (last 5) of the ViT, positioning it between attention and MLP blocks. This module operates only on the compressed learnable token sequences. It concatenates the learnable tokens from k adjacent frames and applies causal attention across this new temporal sequence. We use standard 1D RoPE as position encoding.

Through this temporal attention layer, the compressed representations of neighboring frames can exchange and refine information, significantly raising the upper bound of information retention in the final standby sequence for video understanding. The use of causal attention is a deliberate design choice to ensure compatibility with streaming video encoding scenarios (as detailed in the supplementary material). We note that this explicit temporal module is only designed for MLLMs with image encoder as ViT, while for those using native video encoder, it's no longer necessary.

3.3.3. Two-Stage Dual-Mode Training

To adapt the MLLM to these two distinct modes, we propose a two-stage training pipeline (Fig. 2).

Stage 1: Visual Distillation and Alignment We freeze all parameters of the original POINTS1.5 model (ViT, projector, and LLM) and train only the newly introduced components: the learnable tokens, the duplicated MLP, projector and the temporal attention layers. During this stage, the LLM is fed only the compressed learnable token sequence. This stage functions similarly to the alignment phase in MLLM training, forcing the new modules to distill the essential visual information into the compact token sequence. For this stage, we use the POINTS1.5 alignment data and a subset of the multimodal continue-pretrain data.

Stage 2: LLM Mode Adaptation After Stage 1, the learnable tokens effectively carry the distilled visual information. However, the LLM has not been trained to understand this new, compressed sequence format. Therefore, in Stage 2, we unfreeze the LLM and fine-tune it with a small learning rate, training it jointly with the Stage 1 parameters. Meanwhile, a critical challenge arises: even low-LR fine-tuning can degrade the LLM's performance on the original focus mode. To mitigate this, we employ a 2-forward training

strategy: In each training step, we perform two forward passes. *Pass 1 (Standby)*: We feed LLM the short learnable token sequence. *Pass 2 (Focus)*: We feed the LLM the full sequence (learnable tokens + original tokens).

We average the losses from both passes and backpropagate the combined loss. This joint objective forces the LLM to adapt to the new standby mode while maintaining its original focus mode capabilities. As shown in Tab. 4 and Tab. 6, this method significantly improves standby mode performance while fully preserving focus mode accuracy.

Discussion While other token compression modules exist, e.g. resampler [26] or Q-Former [33], they often suffer from training instability due to the random initialization of new parameters. Our approach, by contrast, initializes all new modules from the pre-trained weights, ensuring a more stable training. Furthermore, our parallel MLP design maintains better computational parallelism than sequential cross-attention modules. Still, our primary contribution is not the specific compression module itself, but the introduction of a human-like, dual-mode paradigm that allows a model to switch between high-efficiency (standby) and high-fidelity (focus) visual processing at will.

3.4. Model Inference

Offline Inference Following the two-stage training, POINTS-Long can perform two distinct modes, Focus and Standby, which can be selected based on task requirements.
(1) Fine-grained Understanding (Focus Mode): For tasks demanding high-fidelity detail, Focus Mode is employed to achieve optimal performance (see Tab. 4).
(2) Holistic Long-sequence Understanding (Standby Mode): For tasks involving holistic comprehension or long visual sequences (e.g., video-QA), we switch to Standby Mode. This mode achieves nearly identical performance using drastically fewer tokens. For example, processing 64 frames of a 480p video, which originally required ≈20k visual tokens, now requires only 0.5k-2k tokens while retaining 97.7-99.7% of the full-sequence performance. Notably, Standby Mode effectively overcomes the context length limitations of most MLLMs (e.g., 32k). By representing each frame compactly, POINTS-Long can gain steadily with respect to sampled frame number (Tab. 3).

Streaming Inference POINTS-Long is inherently well-suited for the streaming scenario. Previous models face a critical limitation on streaming understanding: as new frames are encoded (prefilled), the context limit/KV cache budget will eventually be reached. At that time the oldest cached frame will be discarded, resulting in a short memory window, e.g., prefilling a 480p video at 2fps would retain only about 50 seconds of visual memory for 32K context.

Meanwhile, POINTS-Long enables a far more effective hybrid memory strategy. As shown in Fig 3, we can maintain a "local window" by Focus Mode (prefilling new frames using short+full sequence) and a "memory bank" in Standby Mode (retaining only the short-seq KV cache from older frames). When the local window limit is reached, we only discard the large full-sequence cache, migrating its compact standby-sequence cache into the long-term memory bank. This allows us to manage a 32k context budget dynamically, e.g., a 4k local window and a 28k memory bank would allow the model to maintain 6 seconds of complete current visual information (Focus) while preserving up to 30 minutes of compressed visual memory (Standby). This represents up to 40x increase in memory duration.

Discussion In real-world scenarios, many tasks prioritize efficiency over fine-grained detail, e.g., video tagging and security auditing. Concurrently, emerging applications like interactive livestreaming and multimodal assistants demand both long-term comprehension and high-fidelity analysis. Meanwhile in MLLM design, efficiency and granularity are always treated as a fixed trade-off, *i.e.*, either an efficient model or a performant one. Inspired by the observations of human visual processing, we argue that the two modes should be decoupled and fit in one single model, *i.e.*, rather than a **fixed trade-off**, they should represent a **choice**.

In this work, the choice between modes is predefined at inference time. We believe that an advanced model could learn to make this choice dynamically——learning which parts of a video to "glance" at (Standby) and which to "scrutinize" (Focus). This concept, which we term "Thinking with Videos" [90], will be explored in future work.

4. Experiments

4.1. Implementation Details

To balance between efficiency and fidelity, we compress single image into $n \in \{8, 16, 32\}$ tokens and set temporal $k = 8$. For stage 1, we use the alignment data of POINTS1.5 and a subset of pretrain data. The newly introduced parameters were trained with learning rate 5e-5. For stage 2, we use high-quality data from pretrain and SFT stage, where the LLM parameters are unfrozen and jointly trained with learning rate 1e-5 (details are in supplementary material). The two-stage training process required approximately 25,000 H20 GPU hours. Note on Reproducibility: This work is primarily aimed at MLLM pre-training teams. The computational cost is highly dependent on the scale of proprietary training data and the size of the model.

4.2. Evaluation & Benchmarks

Fine-grained Image Benchmarks We follow Opencompass [13] image leaderboard, evaluating on MMBench [44], MathVista [48], HallusionBench [20], OCRBench [46], AI2D [29], MMVet [87], MMStar [9], MMMU [88].

Video Benchmarks We evaluate on a wide range of video benchmarks, including Opencompass video

Table 1. **Opencompass Video Benchmark.** Under the same setting (64 frames), POINTS-Long achieves competitive performance (97.7%-99.7%) against original POINTS1.5-8B with drastically less tokens (2.5%-10%), and retains even better focus mode performance.

Model	Num Frame	Token/ Frame	Total Num of Token	MVBench	Video-MME	MMBench-Video	Tempcompass	MLVU	LongVideoBench	Avg
Qwen2-VL-7B [69]	64	-	>12k	66.0	59.7	48.3	69.6	66.4	55.6	60.9
MiMo-VL-7B-RL [63]	64	-	>12k	63.2	65.0	54.0	-	66.2	-	-
Qwen2.5-VL-7B [3]	64	-	>12k	67.5	62.8	53.0	72.0	70.2	56.0	63.6
VideoLLaMA3-7B [89]	-	-	>12k	69.7	66.2	-	68.1	73.0	59.8	-
InternVL2.5-8B [11]	64	-	>12k	70.5	63.7	56.0	68.7	68.5	-	-
Qwen2.5-Omni-7B [76]	64	-	>12k	69.0	64.1	55.0	70.7	67.5	-	-
InternVL3-8B [101]	64	-	>12k	73.2	66.0	56.3	70.4	71.4	58.8	66.0
GLM4.1V-9B [65]	-	-	>12k	68.2	68.4	54.3	-	71.5	65.7	-
Kimi-VL-A3B-2506 [64]	-	-	>12k	59.7	67.8	-	-	74.2	64.5	-
InternVL3.5-8B [70]	-	-	> 12k	72.1	66.0	55.7	-	70.2	62.1	-
POINTS1.5-8B (baseline)	64	324	≈ 20K	60.3	66.1	61.0	71.1	72.0	59.8	65.0
POINTS1.5-8B (low-res)	64	32	2048	54.9	61.2	51.0	67.1	67.3	53.9	59.2
POINTS1.5-8B (pooling)	64	32	2048	54.9	55.4	43.0	66.6	67.1	54.5	56.9
POINTS-Long (standby)	64	8	512 (2.5%)	59.4	63.5	58.0	69.9	71.9	58.2	63.5 (97.7%)
POINTS-Long (standby)	64	16	1024 (5%)	59.7	65.0	59.3	69.1	71.7	58.9	63.9 (98.3%)
POINTS-Long (standby)	64	32	2048 (10%)	60.8	65.7	60.9	70.3	71.6	59.5	64.8 (99.7%)
POINTS-Long (focus)	64	324+32	≈ 22K	61.0	66.1	60.3	71.3	73.2	59.4	65.2 (100.3%)

Table 2. **More Video Benchmarks.** We evaluate on more video benchmarks to prove universality.

Model	Num Frame	Token/ Frame	Total Num of Token	CG-Bench (long-acc)	MovieChat1k	Egoschema	LVBench	TemporalBench	Activitynet-qa	WorldSense	Avg
Moviechat [59]	-	-	>12k	-	62.3	53.5	22.5	-	45.7	-	-
LLaVA-OV-7B [32]	64	-	>12k	30.9	-	59.8	26.9	59.4	56.0	37.7	-
MiMo-VL-7B-RL [63]	64	-	>12k	-	-	59.4	37.1	-	-	-	
LLaVA-Video-7B [95]	-	-	>12k	-	-	57.3	-	63.6	56.5	40.2	-
InternVL3-8B [101]	-	-	>12k	38.6	-	-	44.1	-	-	-	-
POINTS1.5-8B (baseline)	64	324	≈ 20K	36.7	77.0	60.6	44.3	64.3	54.3	40.4	53.9
POINTS-Long (standby)	64	8	512 (2.5%)	33.6	76.0	59.7	40.4	64.4	55.4	39.4	52.7 (97.8%)
POINTS-Long (standby)	64	16	1024 (5%)	34.6	73.0	60.5	42.5	65.1	55.1	39.7	52.9 (98.1%)
POINTS-Long (standby)	64	32	2048 (10%)	35.7	77.0	62.6	42.6	64.1	54.7	40.0	53.8 (99.8%)
POINTS-Long (focus)	64	324+32	≈ 22K	35.4	79.0	62.5	44.5	64.5	54.7	40.4	54.4 (100.9%)

leaderboard: VideoMME [19], Tempcompass [45], MVBench [35], MMBench-Video [17], MLVU [100], LongVideoBench [74], and other commonly used video benchmarks: MovieChat1K [59], CG-Bench [7], EgoSchema [51], TemporalBench [6], Activitynet-qa [5], LVBench [71] and WorldSense [22]. For Streaming understanding, we choose LongVideoBench, VideoMME, MLVU, LVBench, EgoSchema and CG-Bench. We use VLMEvalKit [16] and lmms-eval [92] for evaluation.

4.3. Main Results

4.3.1. General Video Understanding

In Tab. 1 and 2, we compare POINTS-Long with base model POINTS1.5-8B-Instruct on a wide range of video benchmarks, under the same setting. As shown in Tab. 1, with only 2.5% to 10% of the original tokens, our Standby Mode retains 97.7% to 99.7% of the full performance. Similar results are observed in Tab. 2 on more benchmarks.

This high level of performance retention, achieved through our native dual-mode training, significantly outperforms all prior visual token compression schemes, e.g. PruneVid [24] retains only 96.9% at a 10% token ratio. In Tab. 1, we report results using avg-pooling and low-resolution. Even with 4 times fewer tokens, POINTS-Long outperforms the baselines by a large margin (+4.3%). Notably, when operating in Focus Mode, the model's performance is fully maintained (65.2 vs 65.0). This allows users to fully leverage the flexibility of our dual-mode system.

Furthermore, our model is designed for practical deployment: it requires no hyperparameter tuning, works out-of-the-box, and can be easily deployed in modern inference frameworks, making it ideal for industrial applications.

4.3.2. Scalability of Frames in Inference

We observe that for general MLLMs, video understanding performance stops increasing when exceeding certain number of frames (e.g. 64) [27, 102]. We attribute this phenomenon to the LLM's long-range decay, stemming from inherent limitations in context length and their training data.

To validate the assumption, we evaluated our Standby Mode on long-video understanding benchmarks. As shown in Tab. 3, the base model's performance doesn't improve much when scaling from 64 to 128 frames. In contrast, the Standby Mode of POINTS-Long shows continuously improving performance as the number of input frames increases. This allows it to achieve superior results on long-video tasks, all while using a much smaller token budget.

Note that POINTS1.5 was never trained on data over 128 frames. This zero-shot scalability is a remarkable property that only manifests in Standby Mode. We defer a detailed theoretical explanation for this phenomenon to future work.

4.3.3. Fine-grained Image Understanding

A core design principle of POINTS-Long is the preservation of its fine-grained understanding capabilities. Un-

Table 3. **Scalability of Inference Frame.** By drastically compressing visual tokens, POINTS-Long can process more frames without context length overflow. we witness a steady gain with respect to frame number, which is not always the case for general MLLMs.

Model	Num Frame	Token/ Frame	Total Num of Token	LVBench	VideoMME (Long/Overall)	MMBench-Video	CG-Bench (60+/Overall)	MLVU	LongVideoBench (3600+/Overall)	Avg
POINTS1.5-8B	64	324	≈ 20K	44.3	56.0/66.1	61.0	31.1/36.7	72.0	50.7/59.8	52.5
POINTS1.5-8B	128	144	≈ 18K	45.4	54.4/65.0	61.3	32.4/37.0	72.0	51.2/60.2	52.8
POINTS-Long	64	8	512 (2.5%)	40.4	54.0/63.5	58.0	29.5/33.6	71.9	49.3/58.2	50.5
POINTS-Long	128	8	1024 (5%)	42.9	56.4/64.4	59.0	31.1/35.3	71.9	49.5/59.6	51.8
POINTS-Long	256	8	2048 (10%)	43.6	57.1/66.1	60.0	30.4/36.0	72.4	50.7/59.7	52.4
POINTS-Long	64	16	1024 (5%)	42.5	55.3/65.0	59.3	30.7/34.6	71.7	48.4/58.9	51.3
POINTS-Long	128	16	2048 (10%)	43.1	56.4/66.4	61.0	33.2/36.2	72.7	51.6/60.3	53.0
POINTS-Long	256	16	4096 (20%)	44.1	58.0/66.9	61.3	33.6/37.4	72.2	50.7/59.5	53.3
POINTS-Long	64	32	2048 (10%)	42.6	55.9/65.7	60.9	32.0/35.7	71.6	48.6/59.5	51.9
POINTS-Long	128	32	4096 (20%)	45.3	56.9/66.9	62.0	32.0/37.3	72.5	51.1/60.4	53.3
POINTS-Long	256	32	8192 (40%)	46.9	58.0/66.5	61.3	34.4/37.4	72.5	49.8/59.8	53.8

Table 4. **Opencompass Image Benchmark.** We show that our two-stage training will not harm the fine-grained capacity of focus mode. Bonus: With simple training-free attention-based pruning, the focus mode can be more efficient, beating other training-free baselines.

Model	MMBench	MMStar	MMMU_val	MathVista	OCRBench	AI2D	HallusionBench	MMVet	Avg
Qwen2-VL-7B [69]	81.0	60.7	53.7	61.6	84.3	83.0	50.4	61.8	67.1
InternVL2.5-8B [11]	82.5	63.2	56.2	64.5	82.1	84.6	49	62.8	68.1
MiniCPM-o-2.6 [82]	80.6	63.3	50.9	73.3	88.9	86.1	51.1	67.2	70.2
Qwen2.5-VL-7B [3]	82.2	64.1	58	68.1	88.8	84.3	51.9	69.7	70.9
Ovis2-8B [49]	83.6	64.6	57.4	71.8	89.1	86.6	56.3	65.1	71.8
SAIL-VL1.6-8B [85]	84.0	69.5	55.4	74.2	90.5	87.5	54.4	73.3	73.6
InternVL3-8B [101]	82.1	68.7	62.2	70.5	88.4	85.1	49.0	82.8	73.6
POINTS1.5-8B (baseline)	81.9	65.7	53.2	70.9	85.8	83.9	50.1	64.7	69.5
POINTS-Long (focus)	82.1	66.1	53.7	70.6	85.5	84.2	48.3	66.7	69.7
+ Attn-prune 50%	82.0	64.5	52.4	69.0	83.5	84.2	47.2	66.7	68.7
+ Avg-pooling 50%	80.7	62.2	53.0	64.6	75.1	83.8	48.5	65.5	66.7
+ Folder [68] 50%	81.5	64.0	52.7	64.0	75.5	83.5	49.1	64.9	66.9

like specialized video models [37, 42, 58], POINTS-Long completely retains the original model's fine-grained image understanding abilities through its Focus Mode. As shown in Tab. 4, POINTS-Long (Focus Mode) matches the base model's performance (69.7 vs 69.5), proving that our dual-mode training process is strictly beneficial and non-destructive to the model's core capabilities.

Furthermore, as an extra bonus, the learnable tokens can be leveraged to perform training-free visual token pruning. Specifically, by using the average attention weights from the learnable tokens in the final ViT layer to all visual tokens, we retain only the top $m\%$ with the highest scores for the LLM (details in the supplementary material). This simple, training-free method yields impressive results. Compared to avg-pooling and other plug-and-play techniques [68], our attention-based pruning method achieves significantly better performance retention at the same compression ratios.

4.3.4. Streaming Video Inference

Streaming video understanding demands both fine-grained understanding of recent events and robust long-term memory. Standard MLLMs [3, 47, 70] fail the latter: as new frames are prefilled, the context window is exhausted, forcing the earliest KV cache to be discarded. Such "sliding window" methods [77] yield only short-term memory. Conversely, specialized streaming models [53, 91] sacrifice the former, lacking critical fine-grained understanding.

POINTS-Long, however, is inherently well-suited for such a scenario via its dual-mode system. To validate the claim, we compare two setups. The baseline model is limited to a 64-frame sliding window, discarding older frames' KV cache. POINTS-Long, by contrast, activates its dual-mode strategy (Sec. 3.4): the most recent 8 frames in Focus Mode (local window) and all preceding frames in Standby Mode (memory bank). We evaluate at the end of the video to test long-term recall. As shown in Tab. 5, the baseline fails due to information loss, whereas POINTS-Long achieves superior performance by its high-quality memory.

4.3.5. Ablation Study

Tab. 6 validates our key design choices. Duplicating the MLP layers enhances fitting capability, significantly boosting visual distillation. The temporal attention layer models temporal redundancy for more compact compression, further enhancing video understanding. Finally, our two-stage training is crucial: the second stage substantially improves Standby Mode while successfully preserving the Focus Mode's fine-grained understanding. Our final design, combining all components, yields the best performance.

4.3.6. Inference Efficiency and Performance

A primary motivator for POINTS-Long is computational efficiency. Prior research on visual token compression [10, 84, 94] focus solely on algorithms, overlooking the practi-

Table 5. **Streaming Understanding.** General MLLMs struggle at long-range streaming VQA, while POINTS-Long preserves ultra-long memory by detachable KV cache mechanism shown in Fig 3, resulting in much better performance.

Model	Num Frame	Token/ Frame	Total Num of Token	LVBench	Video-MME	CG-Bench	Egoschema	MLVU	LongVideoBench	Avg
VideoStreaming [53]	-	-	256	-	-	-	44.1	-	-	-
Qwen2-VL-online [69]	-	-	11520	39.8	59.4	-	64.0	62.9	-	-
Flash-Vstream [91]	-	-	11520	42.0	61.2	-	68.2	66.3	-	-
POINTS1.5-8B-online	64	324	20736	41.7	59.3	33.1	59.3	64.7	53.5	51.9
POINTS-Long	248+8	8	3200	44.2	65.4	36.4	61.6	71.8	59.3	56.5
POINTS-Long	504+8	8	5248	46.0	65.0	35.5	59.7	70.3	59.2	56.0
POINTS-Long	248+8	16	5248	46.3	65.8	37.0	60.7	72.1	59.1	56.8
POINTS-Long	504+8	16	9344	48.6	64.9	35.6	58.4	71.2	58.8	56.3

Table 6. **Ablation Study.** We ablate the training design in Sec. 3.1. All components are essential for obtaining the optimal result.

Num Frame	MLP	Temporal	Stage 2	MVBench	Video-MME	MMBench-Video	Tempcompass	MLVU	LongVideoBench	Avg
64	×	×	×	55.6	58.6	47.3	67.1	67.4	52.7	58.1
64	✓	✓	×	57.5	61.5	53.7	67.8	69.2	56.2	61.0
64	✓	×	✓	58.1	63.2	56.3	69.8	70.6	57.5	62.6
64	✓	✓	✓	**59.4**	**63.5**	**58.0**	**69.9**	**71.9**	**58.2**	**63.5**

Table 7. **Real-world Inference Speed-up.** We evaluate the speed-up of the LLM side prefilling and decoding using SGLang and Pytorch Profiler on H20. While ViT's cost grows linearly with the number of frames, LLM shows a quadratic increase in complexity.

Model	Num Frame	Token/ Frame	FLOPs (ViT+LLM)	LLM Prefill Latency (s)	Generation Throughput (token/s)
POINTS1.5-8B	128	144	99.7+456.7	3.23	311
POINTS1.5-8B	256	144	199.4+1314.5	8.95	240
POINTS-Long	128	8	106.2+14.8	0.21 (6.5%)	1887 (6.1×)
POINTS-Long	256	8	208.5+30.9	0.41 (4.6%)	1494 (6.2×)
POINTS-Long	128	16	110.8+30.9	0.41 (12.7%)	1447 (4.7×)
POINTS-Long	256	16	221.6+66.8	0.70 (7.8%)	1124 (4.7×)

cal deployment. Here, we analyze the acceleration benefits of Standby Mode from an infrastructure perspective. We divide MLLM into two components: ViT and LLM, which have very different computational and memory profiles.

Disparate Workloads The ViT-LLM computational balance is task-dependent. (1) For high-resolution images: For ∼10B models, compute is surprisingly comparable, as techniques like pixel-shuffle shortens the LLM's sequence. (2) For long videos: The bottleneck shifts to LLM. ViT compute scales linearly with frames, whereas the LLM's prefill scales quadratically, creating a dominant cost (Tab. 7).

Distinct Compute Phases The ViT encoding and LLM prefill phases are compute-intensive, while the LLM decode phase is I/O bound, as its speed is primarily limited by reading the KV cache. Infrastructure optimizations like continuous batching [31] maximize throughput by batching parallel decode requests. The primary factor limiting this batch size is the available VRAM for the KV cache.

Based on these characteristics, our Standby Mode provides two crucial acceleration benefits:

(1) Drastic Reduction in LLM Compute and Latency As shown in Tab. 7, Standby Mode slashes the LLM's compute by 30-40×. While ViT compute is not reduced, it can be overlapped with LLM, similar to "decoupled deployment" (PD) schemes [70, 98]. This optimization is paramount, as the LLM's total time (prefill + decode) typically exceeds the ViT's encode time, making it the primary bottleneck.

(2) Increased Generation Throughput via Batching By reducing the visual sequence length, the KV cache footprint per sample becomes drastically smaller. This allows the inference system to batch significantly more concurrent decode requests within the same VRAM budget. This directly translates to a massive improvement (6.2×) in overall generation throughput, a critical metric for production services.

We validated these claims using SGLang [97] and PyTorch Profiler. Despite our implementation being preliminary and not fully optimized, the practical benefits are already substantial. As shown in Tab. 7, Standby Mode significantly reduces LLM prefill latency and boosts generation throughput. These advantages are especially pronounced in multi-frame video scenarios, confirming our approach's effectiveness for processing long visual sequences.

5. Conclusion

We introduce POINTS-Long, a novel dual-mode MLLM addressing the trade-off between fine-grained performance and computational efficiency. Inspired by human cognition, POINTS-Long operates in a high-fidelity "Focus Mode" and a highly compressed "Standby Mode". Our two-stage post-training strategy effectively integrates Standby Mode while fully preserving fine-grained abilities. POINTS-Long achieves state-of-the-art efficiency, retaining 97.7%-99.7% performance with only 1/40-1/10th visual tokens. Its dual-mode architecture also enables an efficient detachable KV cache for long-term streaming video understanding. Compatible with modern inference frameworks like SGLang, POINTS-Long offers a practical and powerful solution to the challenging trade-off in MLLM visual understanding.

References

[1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025. 2

[2] Alan Baddeley. Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839, 2003. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3, 6, 7

[4] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 2

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 6

[6] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 6

[7] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024. 6

[8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 5

[10] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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, 2024. 2, 7

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 7

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[13] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 5

[14] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 1, 2, 3

[15] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3

[16] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 6

[17] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 6

[18] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1):10–10, 2007. 2

[19] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 6

[20] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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, 2024. 5

[21] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1, 2

[22] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025. 6

[23] Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, Yao Hu, and Shaohui Lin. Dynamic-llava: Efficient multimodal large

language models via dynamic vision-language context sparsification. *arXiv preprint arXiv:2412.00876*, 2024. 2

[24] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024. 6

[25] Ziyuan Huang, Kaixiang Ji, Biao Gong, Zhiwu Qing, Qinglong Zhang, Kecheng Zheng, Jian Wang, Jingdong Chen, and Ming Yang. Accelerating pre-training of multimodal llms via chain-of-sight. *Advances in Neural Information Processing Systems*, 37:75668–75691, 2024. 2

[26] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5

[27] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Token-efficient long video understanding for multimodal llms. *arXiv preprint arXiv:2503.04130*, 2025. 6

[28] Chen Ju, Haicheng Wang, Haozhe Cheng, Xu Chen, Zhonghua Zhai, Weilin Huang, Jinsong Lan, Shuai Xiao, and Bo Zheng. Turbo: Informativity-driven acceleration plug-in for vision-language large models. In *European Conference on Computer Vision*, pages 436–455. Springer, 2024. 2

[29] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 5

[30] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning, 2022. 2

[31] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 1, 2, 3, 8

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 5

[34] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 6

[36] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, pages 1–19, 2025. 2

[37] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 1, 2, 7

[38] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024. 2

[39] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2

[40] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2

[41] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024. 1

[42] Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025. 1, 2, 7

[43] Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, Siteng Huang, and Honggang Chen. Compression with global guidance: Towards training-free high-resolution mllms acceleration. *arXiv e-prints*, pages arXiv–2501, 2025. 2

[44] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5

[45] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 6

[46] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 5

[47] Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1. 5: Building a vision-language model towards real world applications. *arXiv preprint arXiv:2412.08443*, 2024. 2, 3, 7

[48] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang,

Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 5

[49] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 1, 7

[50] Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997. 2

[51] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 6

[52] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, et al. Gpt-4 technical report, 2024. 1

[53] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024. 2, 7, 8

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[55] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 2021. 2

[56] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 2

[57] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pages 444–462. Springer, 2024. 2

[58] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 1, 2, 7

[59] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 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, 2024. 6

[60] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 3

[61] Hanshi Sun, Li-Wen Chang, Wenlei Bao, Size Zheng, Ningxin Zheng, Xin Liu, Harry Dong, Yuejie Chi, and Beidi Chen. Shadowkv: Kv cache in shadows for high-throughput long-context llm inference. *arXiv preprint arXiv:2410.21465*, 2024. 2

[62] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001, 2025. 1, 2

[63] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report, 2025. 1, 2, 6

[64] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 6

[65] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multi-modal reasoning with scalable reinforcement learning, 2025. *URL https://arxiv.org/abs/2507.01006*. 1, 2, 6

[66] Rufin VanRullen. Perceptual cycles. *Trends in cognitive sciences*, 20(10):723–735, 2016. 2

[67] Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*, 2024. 2

[68] Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quétu, Shuai Xiao, and Enzo Tartaglione. Folder: Accelerating multi-modal large language models with enhanced performance. *arXiv preprint arXiv:2501.02430*, 2025. 1, 2, 7

[69] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3, 6, 7, 8

[70] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 2, 6, 7, 8

[71] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025. 6

[72] Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*, 2025. 1

[73] Zichen Wen, Shaobo Wang, Yufa Zhou, Junyuan Zhang, Qintong Zhang, Yifeng Gao, Zhaorun Chen, Bin Wang, Weijia Li, Conghui He, et al. Efficient multi-modal large language models via progressive consistency distillation. *arXiv preprint arXiv:2510.00515*, 2025. 2

[74] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 6

[75] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024. 1, 2

[76] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 6

[77] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025. 7

[78] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[79] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 1, 2

[80] Longrong Yang, Dong Shen, Chaoxiang Cai, Kaibing Chen, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Libra-merging: Importance-redundancy and pruning-merging trade-off for acceleration plug-in in large vision-language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9402–9412, 2025. 1

[81] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802, 2025. 2

[82] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[83] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24972–24982, 2025. 1

[84] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29836–29846, 2025. 2, 7

[85] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025. 7

[86] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 1

[87] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5

[88] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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, 2024. 5

[89] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2, 6

[90] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025. 5

[91] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. *arXiv preprint arXiv:2506.23825*, 2025. 2, 7, 8

[92] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024. 6

[93] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025. 2

[94] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 1, 2, 7

[95] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 6

[96] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023. 2

[97] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi

Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024. 1, 2, 3, 8

[98] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. {DistServe}: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 193–210, 2024. 8

[99] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024. 1

[100] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406, 2024. 6

[101] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6, 7

[102] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025. 6