Small Vision-Language Models are Smart Compressors for Long Video Understanding

Junjie Fei[1,2*,†], Jun Chen[1], Zechun Liu[1], Yunyang Xiong[1], Chong Zhou[1], Wei Wen[1], Junlin Han[1], Mingchen Zhuge[1,2], Saksham Suri[1], Qi Qian[1], Shuming Liu[1,2], Lemeng Wu[1], Raghuraman Krishnamoorthi[1], Vikas Chandra[1†], Mohamed Elhoseiny[2†], and Chenchen Zhu[1†]

[1] Meta AI, Menlo Park, USA
[2] KAUST, Thuwal, Saudi Arabia
https://feielysia.github.io/tempo-page/

Abstract. Adapting Multimodal Large Language Models (MLLMs) for hour-long videos is bottlenecked by context window limits. Existing efficiency heuristics blindly sacrifice fidelity. They frequently discard transient decisive moments, blur fine-grained evidence, and waste representational bandwidth on irrelevant backgrounds. We propose **Tempo**, an efficient, query-aware framework that compresses long videos for downstream understanding. Tempo unifies a Small Vision-Language Model (SVLM) to act as a local temporal compressor. It casts visual token reduction as an early cross-modal distillation process, generating compact, intent-aligned video representations in a single forward pass. To enforce strict inference budgets without breaking temporal causality, we introduce Adaptive Token Allocation (ATA). Exploiting the SVLM's inherent zero-shot relevance prior and empirical semantic front-loading, ATA acts as a training-free, $O(1)$ dynamic router. It allocates dense bandwidth to query-critical segments while compressing redundancies down to minimal *temporal anchors* to maintain the global storyline. Extensive experiments demonstrate that our compact 6B architecture achieves highly competitive performance with aggressive dynamic compression (0.5–16 tokens/frame). On the extreme-long LVBench (4101s), Tempo scores **52.3** under a strict 8K visual budget, outperforming proprietary models like GPT-4o and Gemini 1.5 Pro. Crucially, empirical profiling reveals that Tempo frequently compresses hour-long videos to token counts below theoretical limits, proving that true long-form understanding relies on intent-driven efficiency rather than greedily padded context windows.

Keywords: Long Video Understanding · Query-Aware Token Compression · Adaptive Token Allocation

1 Introduction

The advancement of Multimodal Large Language Models (MLLMs) has significantly transformed visual understanding, empowering systems to perform com-

* Work done during an internship at Meta, †Project lead.
Correspondence to: junjiefei@outlook.com, chenchez@alumni.cmu.edu

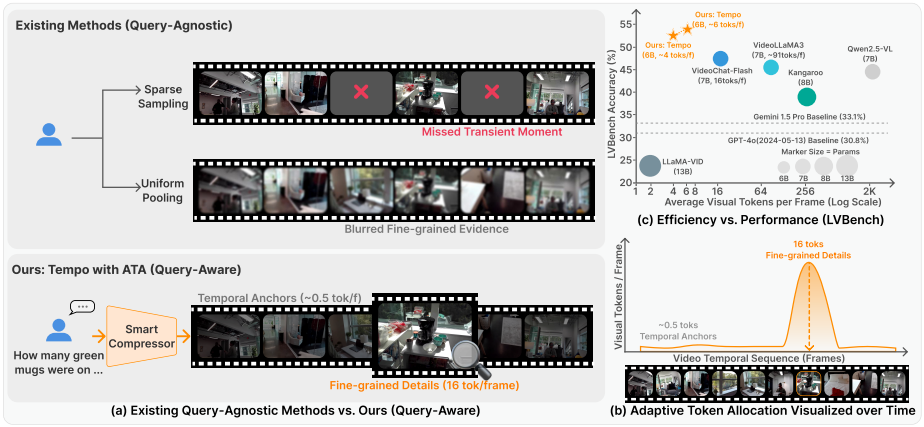

Fig. 1: Tempo achieves superior long video understanding via query-aware Adaptive Token Allocation (ATA). (a) Motivation: Query-agnostic methods either miss transient moments (sparse sampling) or blur details (uniform pooling). Tempo instead utilizes a small vision-language model as a *smart compressor* for query-aware cross-modal distillation. **(b) Mechanism:** ATA dynamically allocates high bandwidth (16 tokens/frame) to relevant segments for fine-grained details, while compressing redundant contexts down to minimal *temporal anchors* (∼0.5 tokens/frame) to maintain temporal causality. **(c) Result:** Leading performance on LVBench. Tempo-6B achieves superior accuracy at extreme compression rates (*i.e.*, 4 or 6 tokens/frame), outperforming open-source models and proprietary baselines with a fraction of the context budget.

plex semantic analysis over images and short video clips [2, 13, 14, 21, 22, 42, 44, 46]. However, adapting these capabilities to hour-long videos remains challenging [7, 32]. The core difficulty lies in the structural mismatch between the massive, continuous visual stream of long videos and the bounded context windows of downstream LLMs [26]. As temporal duration expands, raw visual tokens quickly overwhelm the input capacity, diluting attention mechanisms and causing models to fail at retrieving sparse evidence buried within extensive contexts [24].

To fit long videos into limited contexts, existing methods typically force one of two compromises. A common approach is sparse frame sampling [14,19,37], which reduces compute but inevitably risks skipping the transient yet decisive moments required to answer a specific query. Alternatively, some methods retain more frames but apply query-agnostic compression, such as uniform spatiotemporal pooling [11,25] or token merging [3,12,16]. By compressing without knowing *what the user will ask*, these heuristics often blur fine-grained evidence in query-critical segments while wasting representational bandwidth on irrelevant backgrounds. In essence, most existing pipelines reduce visual evidence before interacting with the language model, preventing the dynamic allocation of bandwidth to query-critical segments. Even pioneering query-aware methods (*e.g.*, LongVU [28]) rely on disjoint auxiliary modules that decouple visual encoding from compression, reducing the process to a passive dropping of pre-extracted representations.

We introduce **Tempo**, an efficient query-aware framework that natively learns to compress long videos for downstream understanding. As its name suggests, Tempo acts as an intelligent temporal compressor that dynamically modulates the *rhythm* of the video: it allocates high token bandwidth to semantic beats relevant to the query while swiftly fast-forwarding through temporally redundant background video segments. Rather than treating visual compression as a purely visual, query-agnostic operation [11, 16], Tempo casts this reduction as an early cross-modal semantic distillation process. Concretely, Tempo leverages a **Small Vision-Language Model (SVLM)** as a local compressor, seamlessly bridging it with an LLM for global understanding and response generation. By prepending the user query to the SVLM input, Tempo performs a preliminary cross-modal distillation pass that produces compact video memory tokens aligned with the user's intent, optimized end-to-end via standard auto-regressive objectives.

A practical challenge is enforcing a strict token budget at inference time (*e.g.*, representing a 1024-frame video under an 8K visual token budget) without sacrificing either fine-grained evidence or global causal structure. To this end, we propose **Adaptive Token Allocation (ATA)**, a training-free inference strategy guided by two key empirical properties of the Tempo architecture. (i) *Zero-shot relevance prior and temporal anchors.* Inheriting from the base model's extensive multimodal pre-training, the local compressor exhibits a zero-shot ability to estimate query-video relevance without auxiliary supervision. ATA exploits this prior to allocate budgets segment-wise, enabling an aggressive dynamic compression range (0.5–16 tokens per frame). Crucially, instead of hard pruning, which breaks temporal causality, ATA preserves dense representational bandwidth for relevant segments while compressing redundant contexts down to minimal *temporal anchors* (*i.e.*, a minimum of 4 tokens) to maintain the global storyline. (ii) *Semantic front-loading.* Our ablations reveal that under the SVLM's causal attention, salient visual semantics natively concentrate in the earliest video memory tokens. Consequently, a simple head truncation effectively isolates high-value evidence, avoiding lossy spatial blurring with zero overhead.

In summary, our contributions are:

- **Tempo:** an end-to-end, query-aware compression framework for long video understanding. It addresses the context window bottleneck by unifying an SVLM-based local compressor and an LLM-based global decoder, performing early query-conditioned cross-modal distillation in a single forward pass.
- **ATA:** a training-free, budget-aware inference strategy leveraging the local compressor's inherent zero-shot relevance prior and semantic front-loading. ATA dynamically dictates the optimal token allocation, preserving fine-grained evidence for query-critical moments while compressing redundancies down to minimal *temporal anchors* to maintain global causal structure.
- **Scaling Behaviors:** an empirical analysis revealing that optimal resource allocation varies with task complexity and video duration. While a 4K visual token budget acts as a *sweet spot* for standard long video tasks (*e.g.*, Video-MME Long, **30–60 mins**), restrictive budgets limit performance on extreme-long videos (*e.g.*, LVBench, $>$ **1 hour**). Scaling to larger capacities

unlocks new performance peaks. Notably, empirical profiling demonstrates that Tempo allocates tokens strictly based on semantic necessity, frequently compressing hour-long videos to token counts far below the available budget.
– **Leading Performances:** despite being a compact 6B model, Tempo establishes a new state-of-the-art among specialized long video MLLMs. On the challenging LVBench, it scores **52.3** under a strict 8K visual token budget, outperforming proprietary baselines (*e.g.*, GPT-4o, Gemini 1.5 Pro) and open-source counterparts (*e.g.*, VideoChat-Flash). Scaling to 2048 frames with a 12K budget further pushes performance to **53.7**, empirically validating Tempo's robust capability for hour-long video understanding.

2 Related Work

2.1 Multimodal Large Language Models for Videos

The rapid evolution of MLLMs has established a dominant paradigm: aligning pre-trained visual encoders with powerful LLMs. Recent state-of-the-art models, such as VideoChat2 [14], VILA [20], LLaVA-OneVision [13], VITA-1.5 [8], Kimi-VL [31], InternVL3.5 [33], and the Qwen-VL series [2], excel in short video understanding by directly mapping sampled frames into the LLM's context. However, extending this dense representation to hour-long videos results in a linear explosion of visual tokens, which overwhelms context limits, incurs prohibitive computational costs, and exacerbates the *lost-in-the-middle* phenomenon [24].

2.2 Context Extension and Token Reduction

To comprehend extended temporal horizons, existing methods typically follow two paradigms. The first extends context via algorithmic extrapolation [43], architectural innovations [35], or system-level parallelization [4] to support massive token sequences. While successfully pushing context boundaries, their strict reliance on processing dense visual streams incurs exorbitant memory footprints and computational overhead. The second employs query-agnostic token reduction via spatiotemporal pooling or sparse sampling [11, 12, 14, 16, 25]. However, being completely query-agnostic risks blurring semantic boundaries and discarding transient, fine-grained evidence crucial for downstream understanding.

2.3 Query-Aware Video Architectures

To overcome uniform processing limits, prior work either employs dual pathways to balance spatial and temporal resolutions [6,38,40,44] or adopts scene-richness-based token compression [34]. However, their resource allocation remains detached from user intent. Recent works explore query-aware processing via disjoint auxiliary modules [10, 18, 28], which decouple visual encoding from compression, reducing the process to a passive dropping or merging of pre-extracted features. Some works attempt dynamic query-aware compression *intra-LLM* for fast video

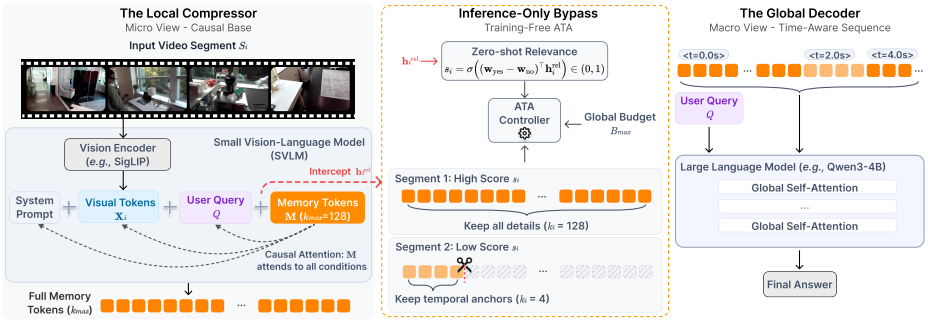

Fig. 2: Overview of the Tempo framework. Tempo casts long video understanding as an end-to-end, query-aware compression process. **The Local Compressor (Left).** For each segment, a Small Vision-Language Model (SVLM) acts as a semantic temporal compressor. Under causal attention, learnable memory tokens $\mathbf{M}$ distill the preceding visual tokens $\mathbf{X}_i$ and user query Q. **Inference-Only Bypass (Middle).** During a single forward pass, an Adaptive Token Allocation (ATA) controller intercepts the hidden state $\mathbf{h}_i^{\text{rel}}$ to compute a relevance score s_i. This enables an $\mathcal{O}(1)$ dynamic head truncation, allocating dense bandwidth to query-critical segments while compressing redundancies down to *temporal anchors* to satisfy a global budget $B_{\max}$. **The Global Decoder (Right).** The compressed memory tokens are assembled into a highly sparse, time-aware sequence using explicit temporal tags (*e.g.*, `<t=2.0s>`). A global LLM synthesizes this condensed multimodal context to generate the final response.

MLLMs [27, 29]. However, adapting these methods directly to long videos leads to severe attention dilution and memory bottlenecks due to the massive tokens in the early LLM layers. In contrast, Tempo shifts query-aware compression *pre-LLM*, casting it as an early cross-modal semantic distillation process. Rather than passively dropping or merging features, Tempo employs an SVLM to independently condense each segment into generative, query-aware representations. This approach dynamically modulates the video's *rhythm*, shielding the global LLM from massive token burdens without resorting to forced heuristic dropping.

3 TEMPO

3.1 Overall

We target the fundamental bottleneck in long video MLLMs: the downstream LLM can only attend to a limited number of visual tokens, while hour-long videos produce a massive, continuous stream. Tempo resolves this mismatch by turning visual token reduction into an early cross-modal distillation problem.

Problem Setup. Given a long video $\mathcal{V}$ and a user query Q, we uniformly partition $\mathcal{V}$ into N temporal segments $\mathcal{S} = \{S_1, \ldots, S_N\}$. Our objective is to convert each S_i into a compact set of query-conditioned video memory tokens, with the total sequence bounded by a global inference budget $B_{\max}$, enabling the downstream LLM to process the entire video and generate the final response efficiently.

Architecture. Tempo constitutes a two-level generative hierarchy (Fig. 2): (1) an SVLM-based local compressor $\mathcal{C}_\phi$, and (2) an LLM-based global decoder $\mathcal{D}_\theta$. Concretely, the SVLM's native vision encoder maps segment S_i to dense visual tokens $\mathbf{X}_i$. Its causal attention then performs query-conditioned distillation, integrating $\mathbf{X}_i$ and query Q into learnable memory tokens $\mathbf{M}$. This yields a fixed-capacity representation $\mathbf{H}_i$ of exactly $k_{\max}$ tokens. A linear projector maps $\mathbf{H}_i$ into the LLM's embedding space as $\tilde{\mathbf{H}}_i$. Finally, the global LLM $\mathcal{D}_\theta$ consumes all memory tokens $\{\tilde{\mathbf{H}}_i\}_{i=1}^N$ alongside Q to auto-regressively decode the response.

Training vs. Inference. Tempo is trained with a fixed per-segment capacity $k_{\max}$ to learn a strong query-aware local compressor $\mathcal{C}_\phi$. At inference, we additionally enforce a global budget $B_{\max}$. We therefore introduce ATA, which uses a zero-shot relevance prior extracted from the same SVLM forward pass to allocate per-segment budgets $k_i \in [k_{\min}, k_{\max}]$, followed by a constant-time head truncation.

3.2 Query-Aware Visual Compression

We cast segment compression as a query-driven sequence-to-sequence transformation. An explicit information bottleneck forces the compressor $\mathcal{C}_\phi$ to discard visual redundancies and distill semantic evidence relevant to user intent.

SVLM Input Construction. For each segment S_i, the SVLM constructs a single causal sequence comprising: (i) a system prompt, (ii) visual tokens $\mathbf{X}_i$ (extracted via its native vision encoder), (iii) user query Q, and (iv) learnable memory tokens $\mathbf{M}$. Placing $\mathbf{M}$ last is critical: under causal attention, each memory token inherently attends to all preceding visual and textual contexts. This conditions the SVLM to distill query-aligned evidence into $\mathbf{M}$. Extracting the final-layer hidden states of memory tokens yields the compressed representation $\mathbf{H}_i \in \mathbb{R}^{k_{\max} \times d_s}$.

Sequence Assembly & Temporal Grounding. To preserve temporal identity and causal order across the entire video, we prepend an explicit textual timestamp $\mathbf{t}_i$ (*e.g.*, `<t=2.0s>`) to each segment when assembling the global context. In practice, these temporal tags significantly stabilize long-range attribution (i.e., *which evidence comes from where*) within the downstream global LLM.

End-to-End Learning. Let the ground-truth answer be $A = \{a_t\}_{t=1}^T$. The global LLM $\mathcal{D}_\theta$ receives the assembled sequence of all timestamps and projected segment memories $\{\mathbf{t}_i, \tilde{\mathbf{H}}_i\}_{i=1}^N$ in temporal order, and is optimized via:

$$\mathcal{L}_{\mathrm{AR}}(\theta, \phi) = -\sum_{t=1}^{T} \log p_\theta\big(a_t \mid a_{<t}, Q, \{\mathbf{t}_i, \tilde{\mathbf{H}}_i\}_{i=1}^N\big) \tag{1}$$

Crucially, we do not impose auxiliary compression losses or heuristic token-dropping regularizations during training. The fixed capacity of $k_{\max}$ memory tokens acts as a hard structural bottleneck. The gradients back-propagated from $\mathcal{L}_{\mathrm{AR}}$ naturally compel the compressor $\mathcal{C}_\phi$ to discard query-irrelevant backgrounds and pack the most predictive visual evidence into this bounded space.

3.3 Zero-Shot Relevance Prior

A core insight driving Tempo is that modern multimodal foundation models (*e.g.*, Qwen3-VL [2]) inherently possess a robust zero-shot capability to evaluate semantic alignment between a visual sequence and a text query (Refer to Sec. 4.3 – D). We harness this foundational prior to extract a highly accurate relevance signal without introducing or training any auxiliary routing modules.

Logit-Based Relevance Score. To explicitly elicit this prior during inference, we slightly augment our system prompt (compression instruction) by appending a binary directive: *"Now, before compressing, answer exactly 'Yes' or 'No': is this segment relevant to the Query?"* Let $\mathbf{h}_i^{\text{rel}} \in \mathbb{R}^{d_s}$ be the final hidden state immediately preceding the model's binary response. Using the SVLM's frozen language modeling head weights for the vocabulary tokens `Yes` ($\mathbf{w}_{\text{yes}}$) and `No` ($\mathbf{w}_{\text{no}}$), we compute a continuous relevance probability s_i via logit difference [15]:

$$s_i = \sigma\Big((\mathbf{w}_{\text{yes}} - \mathbf{w}_{\text{no}})^\top \mathbf{h}_i^{\text{rel}}\Big) \in (0, 1), \tag{2}$$

where $\sigma(\cdot)$ is the `Sigmoid` function. This $O(1)$ projection avoids auto-regressive decoding overhead while yielding a highly stable ranking signal.

Single-Pass Design. The score s_i and the compressed memory tokens $\mathbf{H}_i$ are extracted within a single forward pass of $\mathcal{C}_\phi$. As illustrated in the *Inference-Only Bypass* (Fig. 2), we simply intercept the hidden state $\mathbf{h}_i^{\text{rel}}$ to compute the zero-shot score and then seamlessly continue the forward pass to extract $\mathbf{H}_i$. This architectural elegance guarantees that both the relevance routing signal and the compressed representations are rigorously conditioned on the exact same multimodal context, achieving adaptive evaluation with effectively zero latency.

3.4 Adaptive Token Allocation (ATA)

At inference, the total visual context provided to the global LLM must strictly satisfy a bounded capacity $B_{\max}$. As illustrated in Fig. 2, the ATA controller translates the zero-shot scores $\{s_i\}$ into dynamic per-segment token budgets k_i, executing the physical compression via zero-overhead head truncation.

Stage 1: Contrastive Linear Allocation. To guarantee causal continuity across the entire video sequence, we enforce a minimal *temporal anchor* for every segment, regardless of its relevance. We first normalize the raw scores via Min-Max scaling: $\hat{s}_i = (s_i - \min(\mathbf{s}))/(\max(\mathbf{s}) - \min(\mathbf{s}) + \epsilon)$, where ϵ is a constant to prevent zero division. To maximize the contrast between query-critical events and irrelevant backgrounds, we linearly map these normalized scores to a target capacity:

$$k_i^{\text{ideal}} = k_{\min} + \lfloor (k_{\max} - k_{\min}) \cdot \hat{s}_i \rfloor. \tag{3}$$

Algorithm 1: Adaptive Token Allocation (ATA) at inference

Input: Segment memories $\{\mathbf{H}_i\}_{i=1}^{N}$, relevance scores $\{s_i\}_{i=1}^{N}$, budget $B_{\max}$, bounds $k_{\min}, k_{\max}$
Output: Budgeted memories $\{\mathbf{H}_i^{\mathrm{ATA}}\}_{i=1}^{N}$
Normalize $\{s_i\} \to \{\hat{s}_i\}$ by min-max scaling;
Compute k_i^{ideal} via Eq. (3);
if $\sum_i k_i^{\mathrm{ideal}} \le B_{\max}$ **then**
  $k_i \leftarrow k_i^{\mathrm{ideal}}$;
else
  $k_i \leftarrow$ Eq. (4) with $B_{\mathrm{base}} = N k_{\min}$;
Discretize $\{k_i\}$ to integers s.t. $\sum_i k_i \le B_{\max}$;
$\mathbf{H}_i^{\mathrm{ATA}} \leftarrow \mathbf{H}_i[1{:}k_i], \quad \forall i \in \{1, \ldots, N\}$;
return $\{\mathbf{H}_i^{\mathrm{ATA}}\}_{i=1}^{N}$;

Stage 2: Capacity-Aware Protection. Let $B_{\mathrm{base}} = N \cdot k_{\min}$ represent the foundational cost required to maintain the global temporal anchors. If the sum of ideal allocations satisfies the global limit ($\sum_i k_i^{\mathrm{ideal}} \le B_{\max}$), we directly adopt $\{k_i^{\mathrm{ideal}}\}$ to maximize sparsity. Otherwise, we distribute the residual budget $B_{\mathrm{res}} = B_{\max} - B_{\mathrm{base}}$ proportionally based on the normalized scores:

$$k_i = k_{\min} + \left\lfloor B_{\mathrm{res}} \cdot \frac{\hat{s}_i}{\sum_{j=1}^{N} \hat{s}_j + \epsilon} \right\rfloor. \tag{4}$$

We then discretize $\{k_i\}$ and distribute any fractional remainders to strictly ensure $\sum_i k_i \le B_{\max}$ (see Alg. 1 for the complete allocation procedure).

Head Truncation: Zero-Overhead Token Selection. Once the dynamic budget k_i is allocated, we compress the segment by simply slicing the memory sequence, *i.e.*, $\mathbf{H}_i^{\mathrm{ATA}} = \mathbf{H}_i[1{:}k_i]$. Driven by the auto-regressive nature of the SVLM's causal attention, we empirically observe a semantic front-loading phenomenon: the local compressor packs the most salient global evidence into the earliest generated memory tokens (refer to Sec. 4.3 – C). Consequently, this $O(1)$ tensor slice naturally isolates high-value semantics without introducing lossy spatiotemporal pooling. The final global sequence $\{\tilde{\mathbf{H}}_i^{\mathrm{ATA}}\}_{i=1}^{N}$ strictly conforms to $B_{\max}$, rendering memory footprints entirely predictable even for hour-long videos.

4 Experiments

4.1 Experimental Setup

Architecture & Implementation. Tempo's local SVLM is initialized from Qwen3-VL-2B-Instruct [2], while the global LLM uses Qwen3-LM-4B [39]. A linear projector bridges the SVLM's memory space to the LLM, yielding a compact

6B-parameter architecture. We extract frames at 2 FPS via Decord, applying uniform subsampling if limits are exceeded. During training, continuous videos are partitioned into 4-frame segments, each compressed by the SVLM into $k_{\max} = 128$ memory tokens. During inference, we expand the segment window to 8 frames. ATA (Sec. 3.4) strictly enforces a global visual budget $B_{\max}$ (4K or 8K) via head truncation. Models are trained on a 64-GPU cluster with FSDP. Additional hyper-parameters are provided in the supplementary material.

Progressive Training Curriculum. We adopt a rigorous four-stage progressive training curriculum to ensure stable optimization and context extrapolation:

- **Stage 0 (Modality Alignment):** We freeze both the SVLM and the LLM, exclusively optimizing the linear projector on the standard LCS-558K dataset [22]. This establishes the fundamental vision-language alignment, bridging the SVLM's visual representation with the LLM's text embedding.
- **Stage 1 (Pre-training):** We unfreeze the entire architecture and optimize it on a large-scale, curated multimodal corpus comprising ∼2M images, ∼1.38M videos, and ∼143K pure texts. During this phase, videos are sparsely sampled at 8 frames, endowing the model with initial temporal perception.
- **Stage 2 (Broad Supervised Fine-Tuning):** To develop robust instruction-following and temporal reasoning capabilities, we perform comprehensive SFT using a diverse data mixture (∼0.93M images, ∼2.25M videos, and ∼71K text samples). In this stage, the temporal context is expanded, with the maximum number of sampled frames per video strictly capped at 128.
- **Stage 3 (Long-Context SFT):** To effectively extrapolate the context window, we freeze the SVLM and exclusively fine-tune the global LLM on a subset of ∼384K samples from Stage 2. Here, the maximum frame limit is extended to 384, enabling the LLM to handle long temporal sequences.

To curate our training data, we primarily follow the data mixtures established by LLaVA-OneVision-1.5 [1] and VideoChat-Flash [16]. All datasets utilized throughout our training are publicly accessible, ensuring full reproducibility.

Evaluation. To evaluate Tempo's long video understanding, we conduct comprehensive experiments across four prominent benchmarks, *i.e.*, MLVU [45], LongVideoBench [36], Video-MME [7], LVBench (extreme-long video) [32], spanning standard long-form tasks to hour-long stress tests. We benchmark Tempo against widely adopted proprietary models (*e.g.*, GPT-4o, Gemini 1.5 Pro), general open-weight MLLMs (*e.g.*, Qwen-VL), and specialized long-video MLLMs (*e.g.*, VideoChat-Flash, LongVA). All evaluations are conducted using `lmms-eval`.

4.2 Long Video Understanding

Tab. 1 summarizes the evaluation of Tempo (capped at 1024 frames) against state-of-the-art MLLMs across four major benchmarks. Despite a compact 6B-parameter architecture and aggressive token compression (0.5–16 tokens/frame),

Table 1: Comparison with SOTA MLLMs on long video benchmarks. Bold and underline denote the best and second-best results among specialized long video MLLMs. "-" indicates unavailable results. * indicates the average tokens per frame are dynamically adjusted. For our model, we report the theoretical range (0.5–16) alongside the actual empirical average tokens per frame (**gray rows**), demonstrating that Tempo inherently operates substantially below the maximum budget limits in practice.

Model	Size	Tokens per frame	LongVideoBench (473s)	MLVU (651s)	Video-MME (w/o sub.) Overall (1010s)	Video-MME (w/o sub.) Long (2386s)	LVBench (4101s)
Proprietary Models							
GPT-4o [9]	-	-	66.7	64.6	71.9	65.3	30.8
Gemini 1.5 Pro [30]	-	-	64.0	-	75.0	67.4	33.1
General Open-Source MLLMs							
VideoChat2-HD [14]	7B	72	-	47.9	45.3	39.8	-
LLaVA-OneVision [13]	7B	196	56.4	64.7	58.2	-	-
LLaVA-Video [44]	7B	676	58.2	70.8	63.3	-	-
VideoLLaMA3* [41]	7B	≤ 91	59.8	73.0	66.2	54.9	45.3
InternVL3.5 [33]	8B	256	62.1	70.2	66.0	-	-
Molmo2 [5]	8B	83	67.5	-	69.9	-	52.8
Qwen2.5-VL [2]	7B	1924	56.0	70.2	65.1	-	45.3
Qwen3-VL* [2]	2B	≤ 640	-	68.3	61.9	-	47.4
Qwen3-VL* [2]	8B	≤ 640	-	78.1	71.4	-	58.0
Specialized Long Video MLLMs							
LLaMA-VID [17]	7B	2	-	33.2	25.9	-	23.9 (13B)
LongVA [43]	7B	144	-	56.3	52.6	46.2	-
Kangaroo [23]	8B	256	54.8	61.0	56.0	46.7	39.4
LongLLaVA [35]	A13B	144	53.5	-	53.8	46.4	-
LongVILA [4]	7B	196	57.1	-	60.1	47.0	-
LongVU [28]	7B	64	-	65.4	60.6	59.5	-
Storm [11]	7B	64	60.5	72.9	63.4	53.4	-
BIMBA [10]	7B	36	59.5	71.4	64.7	-	-
VideoChat-Flash [16]	7B	16	<u>64.7</u>	74.7	65.3	55.4	48.2
Tempo* (4K Budget)	**6B**	**0.5–16**	64.5	**75.6**	**67.8**	**57.8**	**52.7**
↪ *actual avg. toks/frame*			*2.8*	*2.8*	*3.6*	*3.4*	*2.9*
Tempo* (8K Budget)	**6B**	**0.5–16**	**65.1**	<u>75.2</u>	<u>67.7</u>	<u>57.0</u>	<u>52.3</u>
↪ *actual avg. toks/frame*			*3.1*	*3.3*	*4.3*	*4.1*	*3.5*

Tempo achieves highly competitive performance. While larger open-weight models (*e.g.*, Qwen3-VL-8B, Molmo2) yield strong absolute scores via exorbitant visual token consumption, Tempo operates under extreme efficiency. By routing evidence through ATA, Tempo strictly bounds visual tokens to a 4K or 8K budgets. In practice, ATA dynamically distributes bandwidth so efficiently that the actual consumption falls well below these limits (*e.g.*, 2.9 tokens/frame on LVBench under the 4K budget). Remarkably, its comparative advantage over specialized long video MLLMs amplifies as the temporal span extends.

Dominance in Ultra-Long Video Understanding. The most notable results emerge on the extreme-long benchmark LVBench, a rigorous stress test for long-term memory and evidence retrieval. Operating strictly within a 4K visual budget, Tempo achieves **52.7**, outperforming the strongest 7B specialized MLLM, VideoChat-Flash (48.2), by 4.5 points. Impressively, despite its compact capacity, Tempo eclipses proprietary models in this ultra-long setting, surpassing GPT-4o (30.8) and Gemini 1.5 Pro (33.1) by massive margins. This proves that explicit

query-aware compression is vastly superior to blindly feeding raw frames into expansive LLM context windows, which often suffer from attention dilution.

Robustness Across Varied Temporal Contexts. This dominance consistently extends across other benchmarks. On Video-MME, Tempo secures **67.8** under the 4K budget, exceeding VideoChat-Flash (65.3) and showing massive improvement over its base model Qwen3-VL-2B (61.9). On the challenging Video-MME *Long* subset (2386s), Tempo achieves **57.8**. Similarly, Tempo delivers SOTA-level performance on MLVU (**75.6**) and LongVideoBench (**65.1** under 8K), asserting its robust generalization across diverse temporal scales and tasks.

The "Less is More" Phenomenon. Crucially, Tempo's performance under the 4K budget frequently matches or exceeds the 8K budget (*e.g.*, 52.7 vs. 52.3 on LVBench; 57.8 vs. 57.0 on Video-MME *Long* subset). This counter-intuitive phenomenon powerfully validates the proposed ATA strategy. Enforcing a stricter information bottleneck filters out background distractors, forcing the LLM to focus purely on high-value semantic beats. This actively mitigates the *lost-in-the-middle* phenomenon without requiring additional inference compute.

Qualitative and Error Analysis. We provide comprehensive qualitative results in the supplementary material to further analyze Tempo's adaptive behavior. We contrast localized queries (requiring pinpoint accuracy) with global queries (requiring holistic understanding). These comparisons explicitly demonstrate how Tempo dynamically shifts its compression rhythm—allocating high-fidelity bandwidth to query-critical moments while applying extreme sparsity to irrelevant backgrounds. Crucially, for overarching queries lacking singular salient events, ATA gracefully defaults to a smooth, low-variance token allocation, ensuring global temporal comprehension remains intact. We also include a discussion of representative failure cases to highlight current limitations and future directions.

4.3 Ablation Studies

We systematically decompose Tempo's core components in Tab. 2. Unless otherwise specified, all variants process videos uniformly sampled at 2 FPS for a maximum of 1024 frames, strictly bounded by an 8K visual token budget.

A. Progressive Training Curriculum. We first evaluate our training stages in Tab. 2A. Stopping after Stage 2 (w/o Long-Context SFT) yields sub-optimal performance on extreme-long benchmarks (*e.g.*, 47.3 on LVBench), as the LLM fails to extrapolate its temporal window. Introducing Stage 3 without adaptive allocation improves this to 51.1, despite a 16K budget. Crucially, combining the full curriculum with ATA achieves peak performance (52.3 on LVBench) while consuming only half the visual budget (8K). This confirms that extending context is not merely about scaling length, but maximizing information density.

Table 2: Ablation studies on Tempo's core components. We decompose our framework across five dimensions: (A) progressive training curriculum, (B) segment-level budget allocation, (C) intra-segment token reduction scheme, (D) relevance scoring source, and (E) temporal continuity. Unless otherwise specified, all variants process a maximum of 1024 frames bounded by an 8K token budget for fair comparison. The default Tempo configuration is highlighted in **gray**. LongVB denotes LongVideoBench.

Ablation Setting	LongVB (473s)	MLVU (651s)	Video-MME (w/o sub.)		LVBench (4101s)
			Overall (1010s)	Long (2386s)	
A. Progressive Training Curriculum					
w/o Stage 3 - Long-Context SFT (16K Budget)	61.4	67.2	66.1	56.3	47.3
w/o Adaptive Token Allocation (16K Budget)	62.8	73.5	67.0	56.2	51.1
Tempo Default (8K Budget)	**65.1**	**75.2**	**67.7**	**57.0**	**52.3**
B. Segment-Level Budget Allocation					
Uniform Subsampling (Equal tokens per segment)	61.9	74.0	66.3	55.2	49.9
Random Drop (Uniform random segment selection)	59.3	70.9	63.6	55.2	49.8
Adversarial Routing (Keep lowest-scoring segments)	50.7	59.3	52.4	47.8	36.9
Hard Top-K Routing (Keep highest-scoring segments)	63.5	73.9	66.7	56.2	52.7
ATA (Adaptive token allocation, Alg. 1)	**65.1**	**75.2**	**67.7**	**57.0**	**52.3**
C. Intra-Segment Token Reduction Scheme					
Uniform Tail Truncation (Fixed 64 tokens)	59.5	71.6	64.1	54.8	41.8
Uniform Head Truncation (Fixed 64 tokens)	63.2	73.4	66.9	56.2	51.5
Token Merging (Merge visual features to k_i tokens)	63.6	74.9	66.3	55.4	53.0
Dynamic Tail Truncation (Keep last k_i tokens)	61.9	73.4	64.8	54.2	50.5
Dynamic Head Truncation (Keep first k_i tokens)	**65.1**	**75.2**	**67.7**	**57.0**	**52.3**
D. Relevance Scoring Source & Zero-Shot Prior					
Base Model Prior (Standard prompt)	64.6	75.1	67.2	56.1	52.6
Base Model Prior (Explicit routing prompt)	65.7	76.3	67.6	57.3	52.7
External Dense Retriever (Qwen3-VL Reranker)	64.3	75.4	67.2	57.0	51.8
Tempo SVLM Prior (Standard prompt)	64.1	75.4	67.2	57.0	53.4
Tempo SVLM Prior (Explicit routing prompt)	**65.1**	**75.2**	**67.7**	**57.0**	**52.3**
E. Temporal Continuity (Minimum Token Guarantee)					
Hard Pruning (0 tokens for irrelevant segments)	63.9	74.8	67.4	56.3	52.3
Minimal Temporal Anchors ($k_{\min} = 4$)	**65.1**	**75.2**	**67.7**	**57.0**	**52.3**

B. Segment-Level Budget Allocation. To satisfy the 8K token budget, all variants fix the segment capacity at $k_{\max}$ (128) tokens. Given our sampling rate of 2 FPS (capped at a maximum of 1024 frames), Uniform Subsampling halves the initial input frames (processing up to 512 frames). In contrast, Random Drop and the routing variants process all sampled frames (up to 1024) but discard 50% of the segments to meet the budget (Tab. 2B). Notably, when we adversarially route by retaining only the lowest-scoring 50% of the segments (Adversarial Routing), performance catastrophically collapses (*e.g.*, 65.1 $\rightarrow$ 50.7 on LongVideoBench). This validates that our zero-shot relevance scores accurately isolate query-critical evidence. Furthermore, while Hard Top-K Routing (dropping the lowest-scoring half) performs competitively, ATA outperforms it across most benchmarks by dynamically scaling budgets and preserving overarching causality.

C. Intra-Segment Token Reduction. We next evaluate the mechanism for token reduction (Tab. 2C). Whether applying a fixed 64-token limit (Uniform) or utilizing our ATA-assigned budgets (Dynamic), Head Truncation consistently eclipses Tail Truncation (*e.g.*, 63.2 vs. 59.5 for Uniform, and 65.1 vs. 61.9 for Dynamic on LongVideoBench). This corroborates our *semantic front-loading* hypothesis (Sec. 3.4): the causal SVLM natively packs the most critical semantics into its earliest memory tokens. While Token Merging (ToMe) yields a marginal gain on

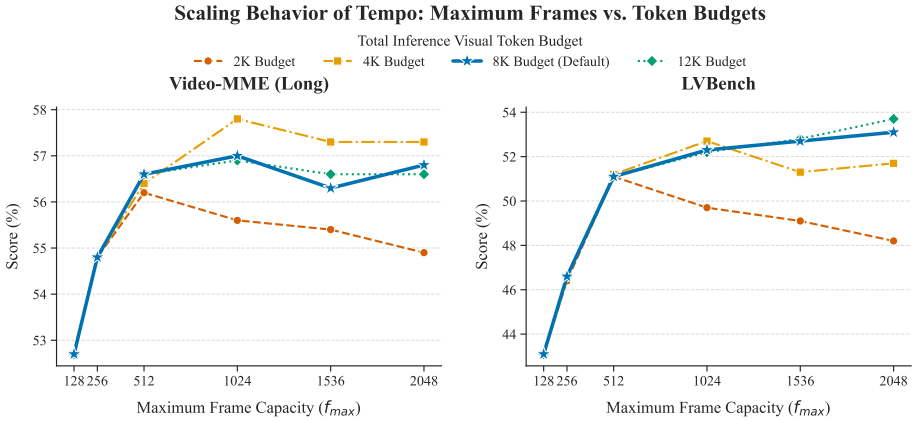

Fig. 3: Scaling behavior of Tempo. We investigate the interplay between maximum frame capacity ($f_{\max}$) and visual token budgets. **(Left)** On Video-MME (Long), a 4K budget acts as an optimal *sweet spot* by filtering redundancy, whereas larger budgets (8K/12K) introduce marginal noise. **(Right)** On the extreme-long LVBench, restrictive budgets constrain performance, whereas expanded capacities unlock new peaks at higher frame densities, proving the necessity of scaled context for hour-long videos.

LVBench (53.0 vs. 52.3), it introduces $O(N^2)$ spatial clustering overhead and degrades performance on shorter benchmarks (*e.g.*, 66.3 vs. 67.7 on Video-MME). The $O(1)$ dynamic head truncation strikes the optimal balance, offering superior generalizability with strictly zero computational overhead.

D. Relevance Scoring Source & Zero-Shot Prior. We systematically investigate the origin of the relevance scores (Tab. 2D). To isolate the routing effect, all variants here utilize Tempo's compressed memory tokens for final generation, differing only in how the ATA score s_i is computed. Surprisingly, we observe that the official Qwen3-VL-2B checkpoint (Base Model Prior) possesses a latent yet strong zero-shot capability for relevance alignment, achieving 76.3 on MLVU. However, extracting this prior from the base model, or utilizing an External Dense Retriever (which natively scores segments but yields sub-optimal performance), necessitates a redundant, isolated forward pass per segment. In contrast, our Tempo SVLM Prior extracts these accurate routing logits simultaneously during the visual compression pass. When guided by an explicit binary instruction (full prompt can be found in the supplementary material), it provides superior comprehensive performance (*e.g.*, 67.7 on Video-MME). This single-pass design capitalizes on the foundational prior with zero latency.

E. Temporal Continuity. We ablate the $k_{\min}$ constraint (Tab. 2E). Compared to a Hard Pruning strategy that aggressively drops irrelevant segments to 0 tokens, enforcing Minimal Temporal Anchors (*e.g.*, 4 tokens/segment) consistently improves performance (*e.g.*, 63.9 $\rightarrow$ 65.1 on LongVideoBench). This demonstrates

that maintaining a continuous, highly compressed timeline is essential for long-form video understanding, preventing the LLM from losing its temporal orientation and causal tracking in hour-long videos.

Scaling Behavior of Inference Context. Fig. 3 details Tempo's scaling dynamics. At low-frame regimes, all budget curves coincide. As the uncompressed token counts remain below the budget thresholds, ATA allocations are identical. As $f_{\max}$ expands, restrictive budgets (2K, 4K) naturally degrade: enforcing the $k_{\min} = 4$ temporal anchor across too many segments prematurely exhausts the capacity, starving query-critical segments of representational bandwidth. At high frame counts, behavior diverges by video length. Standard long-form tasks (*e.g.*, Video-MME Long) peak at $f_{\max} = 1024$ under a 4K budget. Conversely, the extreme-long LVBench necessitates scaled capacities, where expansive 8K and 12K budgets allow performance to scale monotonically, reaching **53.7** at $f_{\max} = 2048$ ($B_{\max} = 12$K). Notably, our empirical profiling (see the supplementary material for token distributions) reveals that even under 8K or 12K ceilings, ATA's dynamic sparsity often compresses LVBench videos ($f_{\max} = 1024$) to token counts substantially below the allocated budget in practice. This demonstrates a profound property: Tempo allocates bandwidth driven by semantic necessity, rather than greedily padding to fill the available context window.

5 Conclusion

We present Tempo, an efficient 6B-parameter framework that addresses the structural mismatch between massive video streams and bounded LLM context windows. Departing from query-agnostic heuristics like sparse sampling or spatiotemporal pooling, Tempo natively unifies a local SVLM and a global LLM. It casts visual token reduction as an early cross-modal distillation process, generating highly compressed, intent-aligned video representations in a single forward pass. To enforce strict visual budgets at inference without sacrificing fine-grained evidence or overarching causality, we propose Adaptive Token Allocation (ATA). Driven by the SVLM's *zero-shot relevance prior* and empirical *semantic front-loading*, ATA functions as a training-free, $O(1)$ dynamic router. By dynamically modulating the video's *rhythm*, it allocates dense bandwidth to query-critical semantic beats while compressing redundancies down to minimal *temporal anchors* to maintain the global storyline, followed by constant-time head truncation. Tempo achieves highly competitive performance across diverse benchmarks, notably outperforming specialized long video MLLMs and proprietary baselines on the extreme-long LVBench. Crucially, our scaling analysis reveals that optimal resource allocation depends on both the task and video duration. While a strict 4K budget acts as a highly efficient denoiser for standard long video tasks, mastering hour-long narratives necessitates scaled contextual capacities. By compressing videos to token counts substantially below theoretical limits in practice, Tempo demonstrates a profound property: true long-form multimodal understanding is best achieved not by greedily padding expansive contexts, but through intent-driven, dynamic allocation based on semantic necessity.

6 Discussion and Future Work

While Tempo successfully demonstrates that high-efficiency, end-to-end multimodal compression can resolve the context bottleneck of hour-long videos, this query-aware paradigm opens several critical frontiers for the community. We discuss the current limitations of our framework and possible future directions.

Eliciting Inherent Relevance Priors via Post-Training. Currently, the Adaptive Token Allocation (ATA) mechanism leverages the SVLM's zero-shot capability to evaluate whether a video segment is relevant for answering a given query. While our ablations demonstrate that this latent prior is already highly effective and accurate, it operates entirely in a zero-shot setting. A promising future direction is to explicitly elicit and further enhance this inherent capability through post-training. Standard supervised fine-tuning may introduce inductive biases or lead to overfitting on heuristic labels, whereas reinforcement learning can directly optimize the SVLM's routing policy with respect to downstream generation performance. By formally optimizing the local compressor to sharpen its internal relevance judgments, we can further elevate this routing precision, potentially driving substantial performance gains across the entire framework.

Autoregressive, Reasoning-Driven Compression. To ensure a highly efficient, single forward pass, Tempo currently compresses video segments using a fixed number of learnable memory tokens. Inspired by recent advancements in reasoning models that dynamically allocate test-time compute (*e.g.*, generating internal thought tokens before halting), a more intelligent paradigm would allow the SVLM to autoregressively generate compressed tokens for a video segment, autonomously deciding when sufficient visual evidence has been gathered to stop. However, adapting this autoregressive extraction without bottlenecking inference latency remains a profound optimization challenge for long video compressors.

On-Demand Distillation for Multi-Turn Dialogue. While Tempo efficiently compresses videos via query-aware distillation, adapting to shifting user intents in multi-turn dialogues still requires re-extracting visual features from the entire video. A highly promising direction is to transition toward an on-demand routing paradigm. By decoupling a query-agnostic global representation from the user-intent features for each segment, the global LLM can act as an active routing agent. Rather than passively consuming pre-extracted features, it can dynamically identify which temporal segments require deeper inspection, invoking the SVLM to distill high-fidelity anchors exclusively for those targeted moments.

Acknowledgements

This work was conducted during a research internship at Meta. Junjie Fei, Mingchen Zhuge, Shuming Liu, and Mohamed Elhoseiny were supported by funding from the King Abdullah University of Science and Technology (KAUST) Center of Excellence for Generative AI.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
4. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al.: Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188 (2024)
5. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
6. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
7. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
8. Fu, C., Lin, H., Wang, X., Zhang, Y.F., Shen, Y., Liu, X., Cao, H., Long, Z., Gao, H., Li, K., et al.: Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. arXiv preprint arXiv:2501.01957 (2025)
9. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
10. Islam, M.M., Nagarajan, T., Wang, H., Bertasius, G., Torresani, L.: Bimba: Selective-scan compression for long-range video question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29096–29107 (2025)
11. Jiang, J., Li, X., Liu, Z., Li, M., Chen, G., Li, Z., Huang, D.A., Liu, G., Yu, Z., Keutzer, K., et al.: Storm: Token-efficient long video understanding for multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5830–5841 (2025)
12. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
13. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
14. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
15. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified

framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
16. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., et al.: Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574 (2024)
17. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
18. Li, Y., Gui, H., Fan, Z., Wang, J., Kang, B., Chen, B., Tian, Z.: Less is more, but where? dynamic token compression via llm-guided keyframe prior. Advances in Neural Information Processing Systems **38**, 156861–156904 (2026)
19. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
20. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
23. Liu, J., Wang, Y., Ma, H., Wu, X., Ma, X., Wei, X., Jiao, J., Wu, E., Hu, J.: Kangaroo: A powerful video-language model supporting long-context video input. arXiv preprint arXiv:2408.15542 (2024)
24. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the association for computational linguistics **12**, 157–173 (2024)
25. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
26. Peng, B., Quesnelle, J., Fan, H., Shippole, E.: Yarn: Efficient context window extension of large language models. In: International Conference on Learning Representations. vol. 2024, pp. 31932–31951 (2024)
27. Shao, K., Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Holitom: Holistic token merging for fast video large language models. arXiv preprint arXiv:2505.21334 (2025)
28. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)
29. Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: Dycoke: Dynamic compression of tokens for fast video large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18992–19001 (2025)
30. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
31. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)

32. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
33. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
34. Wang, X., Zhang, J., Wang, T., Zhang, H., Zheng, F.: Seeing more, saying more: Lightweight language experts are dynamic video token compressors. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 541–558 (2025)
35. Wang, X., Song, D., Chen, S., Zhang, C., Wang, B.: Longllava: Scaling multimodal llms to 1000 images efficiently via hybrid architecture. arXiv preprint arXiv:2409.02889 **2**(5), 6 (2024)
36. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
37. Xu, L., Zhao, Y., Zhou, D., Lin, Z., Ng, S.K., Feng, J.: Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994 (2024)
38. Xu, M., Gao, M., Gan, Z., Chen, H.Y., Lai, Z., Gang, H., Kang, K., Dehghan, A.: Slowfast-llava: A strong training-free baseline for video large language models. arXiv preprint arXiv:2407.15841 (2024)
39. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
40. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl 1.5 technical report. arXiv preprint arXiv:2509.01563 (2025)
41. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
42. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. pp. 543–553 (2023)
43. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
44. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
45. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
46. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)