S^3-Prune: Stability-Aware Token Budgeting for Long-Form Video-Language Models

Gilha Lee*[1], Seungil Lee*[2], and Hyun Kim†[1]

[1] Department of Electrical and Information Engineering and Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul, Republic of Korea
[2] Korea Electronics Technology Institute, Seongnam, Gyeonggi-do, Republic of Korea

Abstract. The deployment of long-form video-language models (Video-VLMs) is fundamentally constrained by the prohibitive scaling of attention complexity and KV-cache overhead as temporal and spatial resolutions increase. Existing pruning methods rely on local frame-level heuristics, which fail to capture shifting information density and often result in cumulative budget misallocation over long horizons. To address these challenges, we propose S^3-prune, a stability-aware token budgeting framework that jointly models spatio-temporal demand while ensuring consistent allocation over time. S^3-prune characterizes token demand by integrating spatial uncertainty (SU) from patch-level embedding shifts and segment transition (ST) from inter-segment semantic variations. To mitigate temporal instability, we introduce stability accumulation (SA) via a Kalman-filtered latent estimator, which smoothens stochastic spikes and provides uncertainty-aware margins for adaptive budgeting. Guided by these stabilized budgets, S^3-prune executes a hierarchical two-stage selection process. Extensive evaluations across seven benchmarks and various backbones demonstrate that S^3-prune significantly advances the efficiency-accuracy frontier, reducing inference latency while remaining robust even under aggressive token retention settings.

Keywords: Video Language Model · Token Pruning · Efficient Inference · Temporal Stabilization

1 Introduction

Understanding videos that encompass complex spatio-temporal information is a core challenge in modern artificial intelligence. Recently, video-language models (Video-VLMs) [6,7,16,18,19,32,35,37,39] have emerged as a dominant paradigm, integrating the robust reasoning capabilities of large language models (LLMs) [2, 10, 12, 15] with visual sequences to comprehend intricate event contexts and causal relationships within videos. However, maintaining high-resolution frame inputs to capture fine-grained visual features leads to an exponential expansion of the total input sequence, denoted by the product of the number of visual tokens

*Equal contribution. †Corresponding author: hyunkim@seoultech.ac.kr

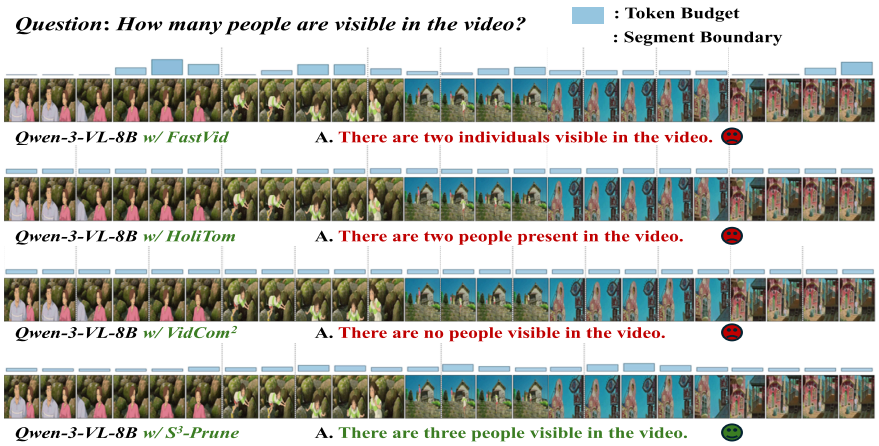

Fig. 1: Token allocation analysis. Baselines misallocate budgets (over/under), while S^3-prune adapts token allocation by modeling stochastic dynamics.

per frame N and the total number of frames F [4, 26]. In particular, since the self-attention mechanism of LLMs exhibits a quadratic complexity of $\mathcal{O}((F \cdot N)^2)$ with respect to the sequence length, the accumulation of tokens results in context window overflow and substantial computational and memory overhead [10]. To address this, compression techniques such as quantization [8,13,14,30] have been introduced; however, these methods merely reduce the precision of individual tokens and are inherently limited in resolving the fundamental computational bottleneck caused by the increasing number of tokens. Consequently, there is an imperative need for research that directly controls computational complexity by efficiently compressing or selecting the visual tokens themselves.

In response, numerous studies have sought to mitigate the computational burden of Video-VLMs by either thinning visual tokens at the encoder stage [22,24, 28] or adopting cascading compression strategies within the LLM [5,11,27,29,31]. These approaches effectively reduce redundant token processing by exploiting the high spatio-temporal correlations inherent in adjacent frames. However, most existing methods rely on heuristics derived from static observation data, failing to adequately address the *stochastic instability* inherent in long-form video environments. In this paper, we identify three critical control limitations faced by conventional deterministic pruning methods when processing long-term video streams: ***Lim.1: Observational stochasticity.*** As video duration increases, transient measurement noise—such as occlusions or sudden lighting fluctuations—is inevitably introduced. Relying naively on these unstable signals often leads to the misinterpretation of crucial visual information. ***Lim.2: Accumulated decision bias.*** A myopic allocation strategy reacting to localized fluctuations tends to amplify errors. Without a global perspective, a single misallocation at a noisy timestep constrains the available budget for subsequent critical frames, leading to a cascading failure in resource distribution (Fig. 1 first row).

Lim.3: Non-stationarity of information density. Despite the non-uniform and temporally evolving concentration of salient visual information during scene transitions or key events, conventional uniform pruning methods fail to effectively track and adapt to these dynamic state changes (Fig. 1 second-third rows).

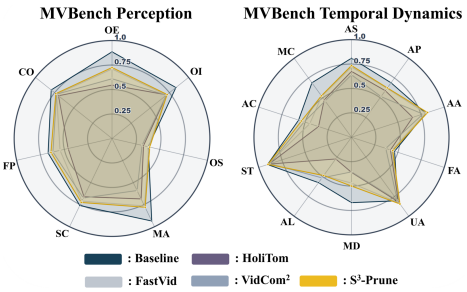

Fig. 2: MVBench radar plots grouped into perception and temporal dynamics.

To address these limitations, we propose S^3-*prune*, a token budgeting framework that integrates local observational uncertainty, information demand at segment transitions, and temporally stabilized budget control. As shown in Fig. 2, our approach achieves superior performance across various video benchmarks by effectively managing the visual sequence. Our contributions are:

- **Joint Spatio-Temporal Demand Modeling:** S^3-*prune* introduces a unified demand signal that jointly models intra-segment spatial uncertainty and inter-segment transition, enabling adaptive tracking of non-stationary information density (**solution for *Lim.1* & *Lim.3***).
- **Stability-Aware Budget Allocation:** To mitigate observational stochasticity, we develop an allocation mechanism that integrates stability accumulation with Kalman-based control. By effectively filtering transient noise from the underlying information state, our controller ensures a robust and smooth budget distribution over long horizons (**targets *Lim.2***).
- **Hierarchical Token Selection Strategy:** We design a hierarchical token selection strategy that first prunes candidates using vision-only cues and then refines them with task-aware relevance at the LLM stage, achieving a strong efficiency–accuracy trade-off.

2 Related Works

2.1 Token Pruning in Image Understanding

Initial token pruning research primarily focused on single-image inputs, discarding redundant or non-salient tokens to significantly enhance computational efficiency [1, 23, 25, 33, 38]. For instance, FastV [4] leverages LLM-based vision-language attention for informed token selection, while PyramidDrop [34] and VTW [20] employ hierarchical pruning strategies based on the increasing representational redundancy observed in deeper layers. Further extensions [17, 21] integrate multi-stage pruning mechanisms to more effectively balance the trade-off between efficiency and task performance. However, these static single-timestep designs face fundamental structural limitations in video streams where information accumulates temporally, necessitating dynamic and context-aware importance estimation.

2.2 Video Understanding: Token Pruning for Video-VLMs

Segment Partitioning. Beyond single image-centric methods, recent literature has shifted toward exploiting the intrinsic spatio-temporal correlations in video. Unlike independent frame-wise processing, most Video-VLM efficiency frameworks partition the video into semantically coherent segments based on temporal proximity. Formally, given a sequence of F frames $\mathcal{T} = \{1, \ldots, F\}$, the video is divided into S disjoint segments $\{\mathcal{T}_1, \ldots, \mathcal{T}_S\}$. Existing frameworks define these boundaries through various strategies: FastVID [28] employs a dynamic strategy (DySeg) that adapts boundaries based on inter-frame dissimilarity, while HoliTom [27] optimizes temporal partitioning via dynamic programming to maximize the ability to merge intra-segments. By performing token aggregation or selection within each $\mathcal{T}_s$, these methods condense the sequence length while preserving the overarching temporal context. This pruning process typically follows two primary architectural paradigms based on the execution stage: (i) one-stage pruning: Directly reducing visual tokens at the vision encoder output or the pre-LLM stage. (ii) two-stage pruning: Applying additional compression within the LLM layers following the initial pre-LLM reduction.

One-stage Pruning. One-stage pruning reduces the LLM sequence length by merging or pruning vision encoder outputs before they are fed into the LLM. FastVID [28] performs spatio-temporal pruning at the pre-LLM stage, while Entropy-Select [24] introduces training-free compression by filtering low-information tokens based on local entropy. Similarly, VidCom^2 [22] quantifies frame uniqueness to apply adaptive compression rates, preserving salient features while discarding redundancy. UniComp [36] further emphasizes informational uniqueness by grouping frames and allocating token budgets accordingly before LLM inference. These methods effectively reduce self-attention prefill costs, which is a significant bottleneck in long-video processing. However, since one-stage strategies finalize token selection without LLM feedback, they heavily rely on local observational statistics, often resulting in task-agnostic compression. This reliance reveals critical weaknesses in long-form video environments. Short-term observational signals are susceptible to noise, which destabilizes representation reliability and leads to either the loss of dynamic information or the excessive retention of non-essential tokens (***Lim.1***). Budgeting based on intra-segment redundancy fails to capture transition phases with surging information density(*e.g.*, scene changes or rapid motion), causing budget misallocation during these critical intervals (***Lim.2***). Tokens discarded prematurely at the pre-LLM stage are unrecoverable. Such cumulative budget mismatches lead to irreversible information loss across extended sequences (***Lim.3***).

Two-stage Pruning. In contrast, two-stage pruning enables task-aware compression by leveraging text-vision interaction signals within the LLM to identify query-relevant tokens. PruneVid [11] merges redundant tokens at the vision encoder and uses LLM attention for selective retention. DyCoke [29] applies intra-frame merging during prefilling and dynamic KV-cache reduction during decoding. To manage cumulative costs in streaming, StreamingTOM [5] combines pre-LLM reduction with bounded KV-cache management, while HoliTom [27]

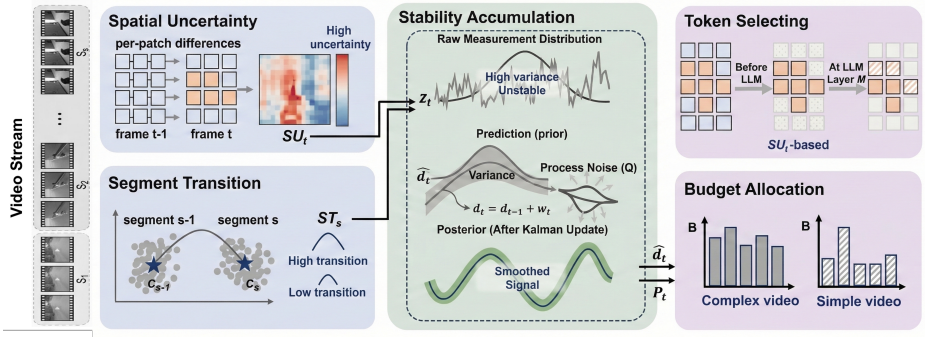

Fig. 3: Overall architecture of S^3-prune. The framework consists of three main stages: (i) spatio-temporal demand modeling in blue blocks, (ii) stability accumulation via Kalman filtering in green blocks, and (iii) adaptive budget allocation and hierarchical token selection in purple blocks.

integrates global temporal contraction with intra-LLM merging. Furthermore, METok [31] introduces a multi-stage framework that progressively compresses tokens across the entire pipeline—from encoding and prefilling to decoding. FlashVID [9] adopts a hybrid strategy that performs spatio-temporal compression before the LLM and further prunes tokens at an intermediate LLM layer to reduce both prefill and decoding costs. Despite these performance gains, these methods inherit fundamental vulnerabilities from their initial pre-LLM reduction steps. Since they still rely on observational signals for early-stage pruning, myopic decision-making remains inevitable. Consequently, budget misallocations during scene transitions accumulate over time, leading to significant performance degradation in long-form sequences.

3 Proposed Method: *S^3-prune*

3.1 Overall Architecture

As illustrated in Fig. 3, *S^3-prune* regulates the visual token stream through an integrated pipeline of demand modeling, budget stabilization, and hierarchical selection. The process begins with *joint spatio-temporal demand modeling* (Sec. 3.2), which characterizes video dynamics by fusing intra-segment spatial uncertainty (SU) and inter-segment segment transition (ST) into a unified demand signal. These signals are subsequently refined through *stability-aware budget allocation* (Sec. 3.3), where Kalman-based control and stability accumulation (SA) filter transient observational noise to ensure a robust and temporally consistent budget distribution. Finally, *hierarchical token selection* (Sec. 3.4) executes the pruning by progressively distilling visual information—first through vision-only coarse filtering and then via task-aware refinement at the LLM stage—to achieve a high-density representation of the most salient context.

3.2 Spatio-Temporal Demand Estimation

The distribution of information density within video streams is inherently non-uniform, governed by both localized spatial fluctuations and broader temporal context shifts. To precisely diagnose this varying information demand, we propose a multi-scale estimation framework that integrates SU for detecting fine-grained patch variations and ST for capturing macroscopic scene shifts.

Spatial Uncertainty. In long-form video streams, the increasing frame count inevitably introduces transient noise and subtle representational discrepancies into patch-level signals. This often leads to the erroneous pruning of critical dynamic features, which are misinterpreted as static redundancies (***Lim.1***). To precisely identify pruning candidates amidst these stochastic fluctuations, we utilize $\mathrm{SU}_{t,i}$ to quantify the temporal stability of patch i, serving as a core probabilistic signal for state estimation. Let $\mathbf{x}_{t,i} \in \mathbb{R}^d$ $(i = 1, \ldots, N)$ be the patch token embedding of frame t extracted by the vision encoder. We define the relative change in token representation between adjacent frames as follows:

$$\mathrm{SU}_{t,i} = \frac{\|\mathbf{x}_{t,i} - \mathbf{x}_{t-1,i}\|_2}{\|\mathbf{x}_{t,i}\|_2 + \epsilon}, \quad i = \{1, \ldots, N\}, \tag{1}$$

where ϵ is a small constant introduced for numerical stability. The normalization in the denominator of Eq. (1) mitigates the influence of the token embedding scale, ensuring consistent sensitivity across different feature spaces. While SU remains low in stable intervals, it reacts sharply to observational uncertainty or salient visual changes, effectively capturing signal volatility.

Token-level SU functions as a priority index for frame-wise pruning while simultaneously establishing the foundation for the frame-level observation m_t used to estimate latent demand. To capture the distributional characteristics of observations in long-form sequences, we define m_t by leveraging the mean μ_t and standard deviation σ_t of the ensemble $\{\mathrm{SU}_{t,i}\}_{i=1}^{N}$ as follows:

$$m_t = \mu_t + \sigma_t, \qquad \mu_t = \frac{1}{N}\sum_{i=1}^{N} \mathrm{SU}_{t,i}, \quad \sigma_t = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\mathrm{SU}_{t,i} - \mu_t)^2}. \tag{2}$$

In this capacity, m_t functions as a stochastic measurement encapsulating both the complexity and uncertainty of frame t, which is subsequently utilized to modulate the pruning intensity within our Kalman filter framework.

Segment Transition. While SU evaluates the pruning suitability of individual tokens, ST dictates the flexible allocation of the token budget according to the temporal context. When adjacent segments share a similar visual context, more aggressive pruning can be performed by recycling prior information. However, at scene transitions, the sudden surge in information influx requires a conservative preservation strategy. To overcome the constraints of conventional approaches that rely on fixed pruning ratios (***Lim.3***), we define the ST to measure the magnitude of change between consecutive segments.

To characterize segment-level semantic states, we first extract the CLS embedding $\mathbf{v}_t \in \mathbb{R}^d$ from the vision encoder for each frame t. We then compute the segment centroid $\mathbf{c}_s$—the representative state vector for segment s—by averaging all frame embeddings within the corresponding interval as follows:

$$\mathbf{c}_s = \frac{1}{T_s} \sum_{t \in \mathcal{F}_s} \mathbf{v}_t, \tag{3}$$

while $\mathbf{c}_s$ encapsulates the overarching semantic context of the segment, ST is formulated as the cosine distance between the centroid representations of consecutive segments $(s{-}1, s)$:

$$\mathrm{ST}_s = 1 - \frac{\mathbf{c}_s^\top \mathbf{c}_{s-1}}{\|\mathbf{c}_s\|_2 \, \|\mathbf{c}_{s-1}\|_2 + \epsilon}. \tag{4}$$

An increase in ST_s indicates a significant visual discrepancy between adjacent segments, implying reduced temporal redundancy and the influx of novel information. Accordingly, *S^3-prune* reserves additional information capacity for high-ST segments. This ST_s signal integrates with frame-level observations m_t to weight the Kalman-based budget allocation. While SU and ST provide effective multi-scale indicators, their raw measurements are susceptible to noise that can destabilize distribution. To address this, the next subsection details a stabilization process that refines these signals into a robust token budget via time-series filtering.

3.3 Stability Accumulation

While token-level *SU* (m_t) and segment transition ST individually reflect intra- and inter-segment information demands, directly mapping these raw signals to a token budget triggers misallocation from transient noise. In long-form sequences, such errors accumulate, leading to significant performance degradation (***Lim.2***). To mitigate this, we introduce **stability accumulation**, a Kalman filter-based state-space model designed to track and stabilize the latent demand of the video stream. SA offers several control-theoretic advantages: (i) recursive updates without offline training, (ii) explicit separation of observational noise from system dynamics to suppress temporal fluctuations, and (iii) quantification of estimation uncertainty to serve as a reliability metric for subsequent budget allocation.

Measurement fusion. The observation vector at frame t is defined as $[m_t, \mathrm{ST}_t]^\top \in \mathbb{R}^2$, combining m_t and ST_t. Through normalization utilizing online means $(\bar{m}_t, s_{m,t})$ and standard deviations $(\overline{\mathrm{ST}}_t, s_{\mathrm{ST},t})$, these two distinct signals are fused into a unified integrated measurement z_t as follows:

$$\tilde{m}_t = \frac{m_t - \bar{m}_t}{s_{m,t} + \epsilon}, \qquad \widetilde{\mathrm{ST}}_t = \frac{\mathrm{ST}_t - \overline{\mathrm{ST}}_t}{s_{\mathrm{ST},t} + \epsilon}, \qquad z_t = \tilde{m}_t + \widetilde{\mathrm{ST}}_t. \tag{5}$$

State-space modeling. SA regards the observation z_t as the result of latent demand $d_t \in \mathbb{R}$ corrupted by observation noise. Reflecting the temporal continuity of video, we assume that d_t evolves gradually rather than abruptly and

construct the following state-space model:

$$d_t = d_{t-1} + w_t, \qquad w_t \sim \mathcal{N}(0, Q), \tag{6}$$

$$z_t = d_t + n_t, \qquad n_t \sim \mathcal{N}(0, R). \tag{7}$$

Eq. (6) is the system transition model indicating that latent demand changes gradually based on the previous state, where w_t is the process noise capturing unpredictable fluctuations in demand. Eq. (7) is the observation model expressing that the actual observation z_t is obtained in the form of true demand d_t plus noise n_t. Here, Q is the process variance determining the tolerance for dynamic changes in the system, where higher values allow for rapid changes in demand, and R is the measurement variance, where higher values strongly suppress observation noise to keep the estimation conservative. We heuristically determine Q and R to reflect video dynamics (see Sec. 2.3 in Supplementary Materials for details).

Recursive Estimation. At each frame, SA yields the optimal demand estimate $\hat{d}_t$ and variance P_t through prediction and update stages. First, the current state is predicted based on prior information as follows:

$$\hat{d}_t^- = \hat{d}_{t-1}, \qquad P_t^- = P_{t-1} + Q, \tag{8}$$

where $\hat{d}_t^-$ is the predicted demand before observation, and P_t^- is the prediction uncertainty calculated by adding process noise Q to P_{t-1}. The subsequent update stage then refines the posterior estimate by incorporating observation z_t:

$$K_t = \frac{P_t^-}{P_t^- + R}, \tag{9}$$

$$\hat{d}_t = \hat{d}_t^- + K_t(z_t - \hat{d}_t^-), \tag{10}$$

$$P_t = (1 - K_t)P_t^-. \tag{11}$$

The Kalman gain $K_t \in [0, 1]$ is a key control variable modulating the relative reliability between the prediction $\hat{d}_t^-$ and observation z_t. In Eq. (10), $\hat{d}_t$ is updated by weighting the innovation $(z_t - \hat{d}_t^-)$ by K_t; a smaller K_t suppresses spikes under high noise, whereas a larger K_t enables rapid tracking of consistent signals. Eq. (11) shows that uncertainty decreases by $(1 - K_t)$ after the update. These refined estimates $(\hat{d}_t, P_t)$ provide the fundamental control basis for the adaptive resource allocation and token filtering detailed in the next section.

3.4 S^3-guided Adaptive Token Pruning

Translating the latent demand $\hat{d}_t$ into computational resources requires an adaptive mapping to a physical token budget. This section details our strategy for allocating resources based on SA-refined demand and uncertainty, while hierarchically pruning tokens by considering both information sparsity and query relevance.

Uncertainty-aware budget allocation. To mitigate irreversible information loss in long-form video streams, we define a unified demand signal ϕ_t integrating the Kalman-derived demand estimate $\hat{d}_t$ and its uncertainty P_t:

$$\phi_t = \hat{d}_t \cdot \left(1 + \sqrt{P_t}\right), \tag{12}$$

where $\sqrt{P_t}$ acts as an uncertainty compensation term, providing a safety margin that conservatively expands the budget during volatile intervals to mathematically suppress the risk of critical information loss from instantaneous misjudgments. Based on the derived ϕ_t, the token budget B_t for each frame is dynamically determined as follows:

$$B_t = \left\lceil \rho \cdot \frac{\phi_t}{\Phi_t} \cdot N \right\rceil, \quad \text{where } \Phi_t = \max(\Phi_{t-1}, \phi_t), \tag{13}$$

where Φ_t denotes the dynamic maximum of the demand signals observed up to the current time, defined as $\max_{1 \le \tau \le t} \phi_\tau$. This relative allocation mechanism maintains causality while elastically managing resources in response to runtime variations in the video stream. Consequently, S^3-*prune* achieves demand-responsive control by adaptively scaling token resources according to complexity.

Two-stage hierarchical selection. To extract the optimal information subset within the dynamically allocated budget B_t, we employ a two-stage hierarchical selection strategy considering both visual dynamics and query context. First, the top B_t tokens with the highest spatial uncertainty ($\mathrm{SU}_{t,i}$) are selected as the candidate set $\mathcal{C}_t$ before LLM processing:

$$\mathcal{C}_t = \mathrm{TopK}\big(\{\mathrm{SU}_{t,i}\}_{i=1}^{N},\ B_t\big)\,. \tag{14}$$

This stage leverages vision-level signals to primarily eliminate massive spatio-temporal redundancy, securing fundamental efficiency. Subsequently, self-attention maps from an intermediate LLM layer M are utilized to precisely analyze the relevance between the query and visual tokens. The relevance score $a_{t,j}$ is derived by aggregating the maximum attention weights from N_q query tokens toward each candidate $j \in \mathcal{C}_t$:

$$a_{t,j} = \max_{1 \le i \le N_q} \mathbf{A}_{q,t}^{(M)}(i,j), \quad j \in \{1, \ldots, |\mathcal{C}_t|\}. \tag{15}$$

Accordingly, the final set $\mathcal{R}_t$ is constructed by selecting the top $\lfloor \alpha \cdot B_t \rfloor$ tokens based on $a_{t,j}$:

$$\mathcal{R}_t = \big\{ j \in \mathcal{C}_t \;\big|\; \mathrm{Rank}(a_{t,j}) \le \lfloor \alpha \cdot B_t \rfloor \big\}, \tag{16}$$

where $\alpha \in [0,1]$ is the token retention ratio for task-aware refinement. Consequently, S^3-*prune* integrates stochastic control-based stabilization with query-driven semantic distillation. This hierarchical pruning transcends simple compression by densely preserving inference-critical information even under extreme sparsification, achieving an optimal balance between accuracy and efficiency.

Table 1: Performance comparison results for LLaVA-OV-7B and LLaVA-VID-7B under different token-retention ratios. Best-performing results are shown in **bold**, and the second-best results are indicated by underlined.

Method	MVBench	EgoSchema	LongVideoB	VideoMME		MLVU	NextQA	Avg.	
				w/o subtitle	w/ subtitle			Score	%
LLaVA-OV-7B Vision Tokens Retained Ratio (100%)									
Vanilla	58.28	60.34	42.18	58.48	61.81	64.81	79.31	60.7	100.0
LLaVA-OV-7B Vision Tokens Retained Ratio (↓ 25.0%)									
VisionZip (CVPR'25)	57.83	60.32	42.55	58.66	61.88	64.02	77.97	60.5	99.5
Dycoke (CVPR'25)	57.95	59.98	42.42	57.77	56.81	62.83	78.43	59.5	97.9
FastVid (NeurIPS'25)	58.23	59.23	42.92	57.85	61.11	62.87	78.35	60.1	98.9
VidCom^2 (EMNLP'25)	56.95	59.86	42.48	58.37	61.25	62.75	78.01	60.0	98.7
HoliTom (NeurIPS'25)	58.25	60.94	42.55	58.77	61.96	64.35	78.75	60.8	100.1
S^3-prune (Ours)	**58.50**	**61.11**	**43.12**	**59.01**	**62.34**	**64.71**	**78.91**	**61.1**	**100.6**
LLaVA-OV-7B Vision Tokens Retained Ratio (↓ 10.0%)									
VisionZip (CVPR'25)	53.43	58.14	42.11	53.44	57.71	60.17	75.14	57.2	94.1
Dycoke (CVPR'25)	56.93	58.24	42.15	56.51	56.03	61.22	**78.21**	58.5	96.3
FastVid (NeurIPS'25)	57.50	58.61	42.35	56.96	60.15	**61.96**	76.97	59.2	97.5
VidCom^2 (EMNLP'25)	57.30	57.56	42.03	56.51	57.74	58.03	74.34	57.6	94.9
HoliTom (NeurIPS'25)	57.40	59.96	**42.63**	57.18	60.33	61.64	77.76	59.6	98.0
S^3-prune (Ours)	**57.63**	**60.25**	42.44	**57.28**	**62.03**	60.45	77.91	**59.7**	**98.3**
LLaVA-VID-7B Vision Tokens Retained Ratio (100%)									
Vanilla	60.43	57.18	57.58	64.42	70.14	80.18	71.51	65.9	100.
LLaVA-VID-7B Vision Tokens Retained Ratio (↓ 25.0%)									
VisionZip (CVPR'25)	58.38	56.15	56.66	62.59	71.22	68.62	78.79	64.6	98.0
FastVid (NeurIPS'25)	60.83	55.45	57.74	63.62	71.22	68.53	79.37	65.3	99.0
VidCom^2 (EMNLP'25)	59.05	55.69	57.21	62.01	70.74	68.72	76.73	64.3	97.6
HoliTom (NeurIPS'25)	58.55	56.05	57.51	63.07	71.59	67.98	77.96	64.7	98.1
S^3-prune (Ours)	**61.25**	**56.41**	**58.37**	**63.84**	**71.96**	**68.89**	**79.96**	**65.8**	**99.8**
LLaVA-VID-7B Vision Tokens Retained Ratio (↓ 10.0%)									
VisionZip (CVPR'25)	56.05	53.98	54.21	59.07	68.92	65.22	76.47	62.0	94.0
FastVid (NeurIPS'25)	59.65	53.72	57.59	60.62	69.92	65.45	78.83	63.7	96.6
VidCom^2 (EMNLP'25)	57.05	51.08	56.91	57.41	66.88	64.63	71.85	60.8	92.3
HoliTom (NeurIPS'25)	56.63	53.78	57.51	**61.62**	69.96	65.68	76.66	63.1	95.8
S^3-prune (Ours)	**59.95**	**54.09**	**57.65**	60.63	**70.04**	**65.72**	**79.07**	**63.9**	**96.9**

4 Experiments

To ensure a fair comparison, all methods were evaluated under the same configuration. We evaluate performance on representative video understanding benchmarks widely used for Video-VLM evaluation, including temporal reasoning, multimodal alignment, and caption-based question answering. We conduct experiments across multiple Video-VLM architectures and benchmark suites to comprehensively assess both effectiveness and generalizability under diverse evaluation settings. All experiments were performed with a batch size of 1 in a training-free manner, where vision token pruning was applied exclusively during the inference stage without any additional training or fine-tuning. Furthermore, by maintaining consistent token retention ratios for each Video-VLM model, we evaluated their robustness from multiple perspectives, ranging from moderate to extreme compression conditions. In addition to accuracy, we analyze inference efficiency and memory consumption to provide a comprehensive evaluation of the trade-off between computational cost and task performance. Detailed implementation specifications are provided in Sec. 1 of Supplementary Materials.

Table 2: Performance comparison of our method against state-of-the-art methods across different backbones under a token retention ratio of 10%.

	Method	MVBench Efficiency ↓			Performance ↑				
		LLM time(s)	Total time(%)	Peak mem.(MB)	MVBench	EgoSchema	LongVideoB	VideoMME	Avg.(%)
Qwen-2.5 VL-7B	Vanilla	100%	100%	100%	69.63	59.41	55.63	67.41	100
	FastViD (NeurIPS'25)	27.9%	73.2%	96.0%	67.85	58.04	52.34	59.22	94.2
	VidCom^2 (EMNLP'25)	**21.1%**	**68.3%**	**93.7%**	66.55	53.31	50.41	59.01	91.0
	HoliTom (NeurIPS'25)	38.7%	76.6%	96.0%	58.35	57.14	51.23	60.22	90.0
	S^3-prune (Ours)	26.0%	71.7%	95.9%	**68.35**	**58.36**	**52.45**	**61.47**	**95.5**
Qwen-3 VL-8B	Vanilla	100%	100%	100%	68.63	65.13	61.23	64.47	100
	FastViD (NeurIPS'25)	41.1%	77.2%	96.9%	61.30	61.32	56.92	57.59	91.4
	VidCom^2 (EMNLP'25)	**21.1%**	**68.4%**	**94.9%**	56.85	60.58	53.85	57.96	88.4
	HoliTom (NeurIPS'25)	24.4%	70.8%	96.9%	54.83	63.56	54.31	56.00	88.1
	S^3-prune (Ours)	22.8%	70.4%	95.9%	**61.33**	**64.00**	**57.57**	**59.10**	**93.3**

4.1 Video Understanding Results

LLaVA models. Table 1 compares visual token pruning methods on seven video understanding benchmarks using LLaVA-OV-7B [16] and LLaVA-VID-7B [39] as backbones. At a 25.0% token retention ratio, *S^3-prune* achieves the best overall performance, preserving 100.6% and 99.8% of the average performance of the corresponding vanilla models, respectively. Methods relying on intra-segment redundancy or frame-level statistics (*e.g.*, HoliTom [27] and FastVid [28]) often suffer from unstable performance during scene transitions and dynamic events. In contrast, *S^3-prune* accurately tracks changes in information demand across scenes, achieving 58.50% and 43.12% accuracy on MVBench and LongVideoBench, respectively, demonstrating effective token allocation throughout the video stream. The advantage becomes even larger under extreme compression (↓ 10.0% tokens). Existing methods become increasingly sensitive to local noise and budget misallocation, whereas *S^3-prune* remains robust by jointly integrating SU, ST, and SA to prevent cumulative information loss. See Sec. 2.1 of Supplementary Materials for additional quantitative results and per-benchmark analyses.

Qwen models. Table 2 presents an analysis of efficiency and performance under a 10% vision token retention setting using Qwen-2.5-VL-7B and Qwen-3-VL-8B [3]. By substantially reducing visual tokens, *S^3-prune* suppresses the computational cost of self-attention and KV-cache accumulation, effectively alleviating the computational bottleneck in the LLM stage while reducing memory consumption. Unlike prior approaches such as FastVid [28] and HoliTom [27], which require parsing the entire video for segment allocation, *S^3-prune* adopts a cascading budget allocation strategy without explicit segmentation, similar to VidCom^2 [22]. This design removes indexing and segment management overhead while preserving information more effectively than one-stage pruning, resulting in superior inference acceleration without sacrificing task performance (see Sec. 2.2 of Supplementary Materials). Furthermore, *S^3-prune* maintains robust performance even under aggressive reduction, demonstrating its scalability across diverse Video-VLM architectures. See Sec. 2.1 of Supplementary Materials for additional results.

4.2 Stochastic Control and Demand Analysis

Kalman-based Stability Analysis. We analyze the proposed stability accumulation mechanism for budget allocation in Fig. 4. As shown in the top plot, raw demand measurements exhibit sharp spikes during scene transitions or abrupt motions. Directly reflecting this transient noise in allocation leads to unstable budget concentration or omissions (lower plot), increasing the risk of irreversible information loss. In contrast, Kalman-predicted demand smooths these fluctuations, tracking latent information demand as a stable trajectory. This results in a systematic and robust budget distribution. Notably, the uncertainty margin in high-variance regions provides a safety buffer that preserves critical events while maintaining stable dynamics, effectively mitigating observational variability (***Lim. 1***) and decision bias accumulation (***Lim. 2***).

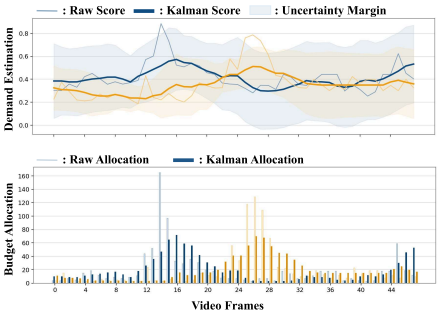

Fig. 4: Comparison of refined state estimation and budget allocation w/ and w/o Kalman filtering for Action Sequence and Character Order.

Ablation of Observational Measurements. Table 3 validates our joint spatio-temporal demand modeling, showing that each signal captures distinct yet complementary dynamical properties. The results indicate that each signal captures distinct dynamical properties of video content, playing a complementary role in overall performance. Leveraging patch-level SU achieves 59.9%—the strongest individual contribution. By distinguishing local motion from static backgrounds, SU prevents the erroneous removal of critical information caused by redundancy misidentification (***Lim. 1***). This patch-level correction ensures stable reasoning accuracy throughout long video streams. Meanwhile, segment-level ST yields 59.7% by identifying non-stationary regions with high information density, such as scene transitions, and adaptively allocating additional budget to those segments. This mechanism effectively mitigates the information loss inherent in previous approaches (***Lim. 3***). Finally, combining SU and ST achieves the highest performance of 60.2%, demonstrating that the synergistic modeling of fine-grained local variations and large-scale contextual transitions enables more accurate demand estimation. These multi-scale signals provide a robust control basis for the Kalman filter, ensuring dense information extraction even under extreme token reduction settings.

Table 3: Impact of individual observational signals in Qwen-3-VL-8B.

SU	ST	MVBench	VideoMME	Avg.
✓		61.00	58.75	59.9
	✓	60.83	58.53	59.7
✓	✓	**61.33**	**59.10**	**60.2**

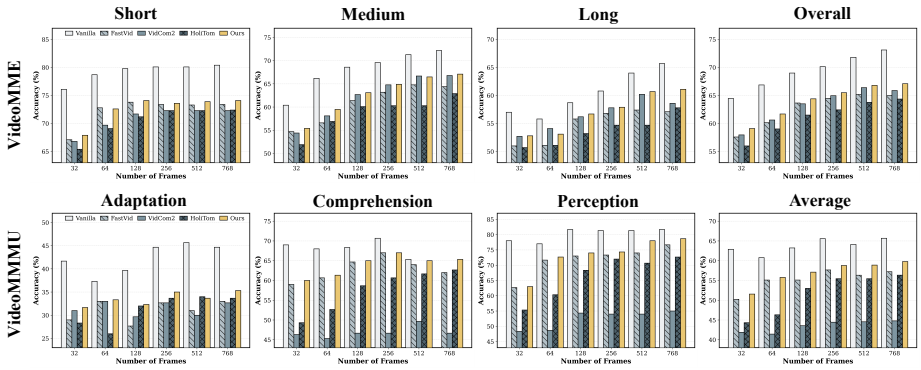

Fig. 5: Scalability analysis on long-form video benchmarks. Performance comparison across varying input frame counts (from 32 to 768) on Video-MME (top) and Video-MMMU (bottom) using Qwen-3-VL-8B. We evaluate our method against different token compression strategies under a fixed 15% token retention ratio.

4.3 Additional Analysis

Long-Frame Scalability Analysis As shown in Fig. 5, we comprehensively evaluate the scalability of our framework by scaling the number of input frames from 32 to 768, encompassing a wide range of temporal contexts. As video understanding models increasingly target complex real-world scenarios—such as feature-length movie analysis and long-term surveillance—the ability to process ultra-long frame sequences with consistent accuracy has become a definitive indicator of practical utility. All experiments were conducted under an extreme token retention ratio of 10% using Qwen-3-VL-8B on the Video-MME and Video-MMMU benchmarks. While increasing the frame count provides richer visual evidence, it simultaneously introduces higher inherent non-stationarity in information density due to frequent scene transitions and amplified observational noise. Existing static or greedy pruning methods fail to adapt to such stochastic variations; consequently, early-stage budget misallocation or decision bias tends to accumulate as the sequence length grows, leading to the irreversible loss of critical information. In contrast, S^3*-prune* consistently improves performance across all frame scales and significantly widens the performance gap with competing methods in the ultra-long regime (*i.e.*, 512–768 frames), demonstrating superior scalability. This sustained advantage stems from our integration of multi-scale dynamic signals and reliability-aware state estimation that stabilizes the entire inference process. Specifically, the Kalman filter-based mechanism effectively smooths local observation fluctuations and tracks the latent information demand, enabling stable budget control while proactively securing information capacity to preserve essential events. This mechanism is particularly effective for benchmarks like Video-MMMU, where capturing *sparse yet critical events* is paramount, thereby facilitating robust and reliable long-range reasoning over extended temporal horizons.

Table 4: Impact of Kalman filtering across different Video-VLM backbones under a token retention ratio of 10%. KF stabilizes the estimated temporal demand signal by mitigating noisy observations and accumulated bias, leading to consistent performance improvements across multiple benchmarks.

Model	MVBench		EgoSchema		LongVideoB		VideoMME		Avg.	
	w/o	w/ KF	w/o	w/ KF	w/o	w/ KF	w/o	w/ KF	w/o	w/ KF
LLaVA-OV	56.4	**57.6**	60.1	**60.3**	41.9	**42.4**	55.9	**57.3**	53.6	**54.4**
LLaVA-ViD	58.7	**60.0**	53.9	**54.1**	57.5	**57.7**	59.1	**60.6**	57.3	**58.1**
Qwen-2.5-VL-7B	67.1	**68.4**	58.2	**58.4**	52.3	**52.5**	60.8	**61.5**	59.6	**60.2**

Impact of Kalman Filtering The proposed stability accumulation mechanism is implemented using a Kalman filter, which smooths the estimated temporal demand signal and mitigates the effect of noisy observations and accumulated decision bias. To evaluate its effectiveness, we compare the performance of our method with and without Kalman filter across multiple Video-VLM backbones under a token retention ratio of 10%. Table 4 summarizes the results on four representative video understanding benchmarks. The results show that incorporating Kalman filter consistently improves performance across all models and datasets. For example, on LLaVA-OV-7B, the average score increases from 53.6 to 54.4, while LLaVA-ViD-7B improves from 57.3 to 58.1. Similarly, Qwen-2.5-VL-7B shows a consistent gain from 59.6 to 60.2. These results demonstrate that the proposed stability accumulation mechanism effectively stabilizes temporal demand estimation, leading to more reliable token allocation and improved overall performance across diverse Video-VLM architectures.

Effectiveness of Pre-LLM Selection. We further isolate the performance of the pre-LLM pruning stage in the hierarchical design of S^3-*Prune*, excluding the refinement process inside the LLM, and compare it with the selection modules of existing one-stage methods. Prior approaches such as HoliTom, FastVid, and VidCom^2 typically rely on fixed heuristics based on redundancy across adjacent frames or local statistics. In contrast, S^3-*Prune* dynamically allocates the token budget by refining SU and ST, extracted at the vision encoder level, through the Kalman filter-based SA module. Experimental results show that, even when only the pre-LLM stage is used, S^3-*Prune* achieves an average performance gain of +0.5% over the strongest baseline (FastVid), as shown in Table 5. These results demonstrate that stable budget allocation effectively preserves critical information during preprocessing. Unlike conventional methods with fixed compression ratios, S^3-*Prune* adaptively tracks underlying video dynamics and allocates resources according to information demand, providing stronger representations even without the task-aware refinement stage of the LLM.

Table 5: Comparison of pre-LLM selection strategies on **Qwen-3-VL-8B** across multiple benchmarks under a token retention ratio of 15%.

Method	MVBench	VideoMME	Avg. (%)
FastVid (NeurIPS'25)	<u>63.33</u>	59.48	<u>61.4</u>
VidCom^2 (EMNLP'25)	58.63	<u>59.59</u>	59.1
HoliTom (NeurIPS'25)	56.94	56.41	56.7
S^3-Prune (Ours)	**63.41**	**59.95**	**61.9**

Table 6: Performance comparison under temporal segmentation strategies from FastVID and HoliTom on the Qwen-3-VL-8B model at a token retention ratio of 10%.

Method	MVBench		EgoSchema		LongVideoB		VideoMME		Avg.	
	w/o S^3	w/ S^3	w/o S^3	w/ S^3	w/o S^3	w/ S^3	w/o S^3	w/ S^3	w/o S^3	w/ S^3
FastVID (NeurIPS'25)	61.3	**61.5**	61.3	61.6	56.9	<u>57.4</u>	57.6	<u>58.4</u>	59.3	<u>59.7</u>
HoliTom (NeurIPS'25)	54.8	55.13	63.6	<u>64.0</u>	54.3	55.0	56.0	56.7	57.2	57.7
S^3-prune (Ours)	-	<u>61.3</u>	-	**64.0**	-	**57.6**	-	**59.1**	-	**60.5**

Compatibility with Existing Segmentation Methods S^3-prune estimates temporal information demand using the spatial uncertainty (SU) signal, which naturally captures temporally coherent regions with similar information density. Although it does not rely on an explicit temporal segmentation algorithm, we evaluate its robustness under representative segmentation schemes from FastVID and HoliTom while keeping the proposed demand estimation and stability accumulation modules unchanged. As shown in Table 6, applying S^3-prune consistently improves performance across all benchmarks. For example, combining S^3-prune with FastVID segmentation improves the average score from 59.3 to 59.7, with notable gains on LongVideoBench and VideoMME. Similar improvements are observed with HoliTom segmentation. These results indicate that the performance gains primarily stem from the proposed stability-aware budgeting mechanism rather than a specific segmentation heuristic.

5 Conclusion

In this paper, we present *S^3-Prune*, a stability-aware token-budgeting framework for long-form video-language models. Unlike existing pruning methods that rely on noisy local heuristics, our approach models token demand as a latent state and stabilizes it through Kalman-based estimation. By integrating spatial uncertainty and segment transitions with stability accumulation, *S^3-Prune* effectively mitigates observational stochasticity and cumulative decision bias over long horizons. Extensive experiments across seven benchmarks demonstrate that *S^3-Prune* consistently improves the efficiency–accuracy trade-off, maintaining robust performance under aggressive 10% token retention while scaling to input sequences of up to 768 frames. Overall, *S^3-Prune* provides a robust and architecture-agnostic solution for efficient long-form video understanding.

Acknowledgements

This research was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481790) and (R&D on Hardware and Software Hybrid Memory Solution for processing engine and hybrid memory architecture, RS-2022-00155731) funded By the Ministry of Trade, Industry and Resources (MOTIR, Korea).

References

1. Arif, K.H.I., Yoon, J., Nikolopoulos, D.S., Vandierendonck, H., John, D., Ji, B.: Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1773–1781 (2025)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. 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)
4. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: 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. pp. 19–35. Springer (2024)
5. Chen, X., Tao, K., Shao, K., Wang, H.: Streamingtom: Streaming token compression for efficient video understanding. arXiv preprint arXiv:2510.18269 (2025)
6. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
7. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
8. Choi, D., Kim, H.: Gradq-vit: Robust and efficient gradient quantization for vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16019–16027 (2025)
9. Fan, Z., Chen, K., Xing, R., Li, Y., Jiang, L., Tian, Z.: Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. arXiv preprint arXiv:2602.08024 (2026)
10. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
11. Huang, X., Zhou, H., Han, K.: Prunevid: Visual token pruning for efficient video large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19959–19973 (2025)
12. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. ArXiv **abs/2310.06825** (2023), `https://api.semanticscholar.org/CorpusID:263830494`
13. Kang, B.J., Kim, N., Kim, H.: Lra-qvit: Integrating low-rank approximation and quantization for robust and efficient vision transformers. In: Forty-second International Conference on Machine Learning (2025)
14. Kim, N.J., Lee, J., Kim, H.: Hyq: Hardware-friendly post-training quantization for cnn-transformer hybrid networks. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization. vol. 8, pp. 4291–4299 (2024)
15. Kim, Y.C., Yoon, S.K., Han, S.S., Park, C.W., Park, J.O., Ko, J.H., Kim, H.: Accelerating language giants: A survey of optimization strategies for llm inference on hardware platforms. Journal of Systems Architecture p. 103690 (2026)

16. 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)
17. Li, K., Chen, X., Gao, C., Li, Y., Chen, X.: Balanced token pruning: Accelerating vision language models beyond local optimization. arXiv preprint arXiv:2505.22038 (2025)
18. 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)
19. 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)
20. Lin, Z., Lin, M., Lin, L., Ji, R.: Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5334–5342 (2025)
21. Liu, T., Shi, L., Hong, R., Hu, Y., Yin, Q., Zhang, L.: Multi-stage vision token dropping: Towards efficient multimodal large language model. arXiv preprint arXiv:2411.10803 (2024)
22. Liu, X., Wang, Y., Ma, J., Zhang, L.: Video compression commander: Plug-and-play inference acceleration for video large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 1910–1924 (2025)
23. Pei, R., Sun, W., Fu, Z., Wang, J.: Greedyprune: Retenting critical visual token set for large vision language models. arXiv preprint arXiv:2506.13166 (2025)
24. Qin, X., Niu, Z., Kang, D., Wang, H., Wang, J., Liu, Q., Dong, K., Jia, J.: Entropy-select: Training-free local entropy token compression for video llms
25. Ranjbar Alvar, S., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. arXiv e-prints pp. arXiv–2503 (2025)
26. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388 (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, L., Gong, G., He, T., Zhang, Y., Liu, P., Zhao, S., Ding, G.: Fastvid: Dynamic density pruning for fast video large language models. arXiv preprint arXiv:2503.11187 (2025)
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. Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-vlm: Post-training quantization for large vision-language models. Advances in Neural Information Processing Systems **37**, 114553–114573 (2024)
31. Wang, M., Chen, S., Kersting, K., Tresp, V., Ma, Y.: Metok: Multi-stage event-based token compression for efficient long video understanding. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 18881–18895 (2025)
32. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)

33. Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., Zhang, L.: Stop looking for important tokens in multimodal language models: Duplication matters more. arXiv preprint arXiv:2502.11494 (2025)
34. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024)
35. 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)
36. Yuan, C., Chen, S., Lin, M., Qiao, L., Wan, G., Ma, L.: Unicomp: Rethinking video compression through informational uniqueness. arXiv preprint arXiv:2512.03575 (2025)
37. 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)
38. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:2410.04417 (2024)
39. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data, 2024. URL https://arxiv. org/abs/2410.02713 **17**