DO WE STILL NEED TOKEN COMPRESSION IN VIDEO LARGE LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Token compression has been an effective strategy for accelerating video large language models (VideoLLMs) by exploiting the substantial spatial and temporal redundancy in video streams. Existing methods typically either leverage attention-based importance to retain informative tokens or exploit spatiotemporal similarity to merge redundant ones after ViT encoding, demonstrating favorable efficiency–accuracy trade-offs on widely adopted LLaVA and Qwen-VL models. However, modern architectures such as Qwen3-VL incorporate native-resolution visual encoding and built-in spatiotemporal compression, improving token efficiency and calling into question the additional benefits of specialized compression methods. In this work, we systematically revisit post-encoding token compression through extensive comparative experiments on four VideoLLMs with native-resolution design. Our results show that carefully designed compression methods do not consistently outperform simple random token dropping or uniform downsampling in these models. We further investigate the factors underlying this limited advantage through targeted diagnostic analyses. Motivated by these observations and the limited content adaptivity of these simple baselines, we advocate content-adaptive dynamic resolution as a promising alternative for future efficient VideoLLMs. By adjusting input resolution before visual encoding, this paradigm can reduce both vision encoder computation and the number of visual tokens processed by the language model, while exploiting native-resolution support for adaptive token allocation. Our study calls for rethinking the role of token compression and highlights adaptive resolution control as a promising future direction for efficient video understanding.

1 INTRODUCTION

The large number of visual tokens produced by video inputs makes inference efficiency a central challenge for video large language models (VideoLLMs). The substantial spatial and temporal redundancy in videos has motivated visual token compression as a widely adopted plug-in solution (Lin et al., 2026; Huo & LI, 2026; Li et al., 2026; Fan et al., 2026; Huang & Li, 2026; Dong et al., 2026; Shen et al., 2025; Tao et al., 2025; Shao et al., 2026). Existing approaches broadly follow two strategies: importance-driven pruning, which identifies and retains informative tokens, and redundancy-driven merging, which aggregates similar tokens across space or time, as illustrated in Figure 1. Their success, demonstrated primarily on LLaVA-family models (Li et al., 2024a;b), has established token compression as an effective approach to improving the accuracy–efficiency trade-off of video understanding.

However, the visual processing architectures underlying VideoLLMs have evolved. In fixed-resolution LLaVA-based video pipelines, each frame is mapped to a prescribed visual-token grid regardless of its original resolution. Native-resolution architectures instead support variable-sized visual inputs, allowing token counts to vary with input resolution. They also incorporate different mechanisms for reducing visual sequences: Qwen3-VL (Bai et al., 2025) and

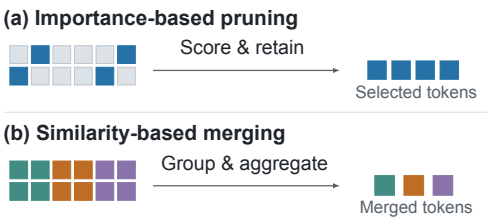

Figure 1: Two common compression strategies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

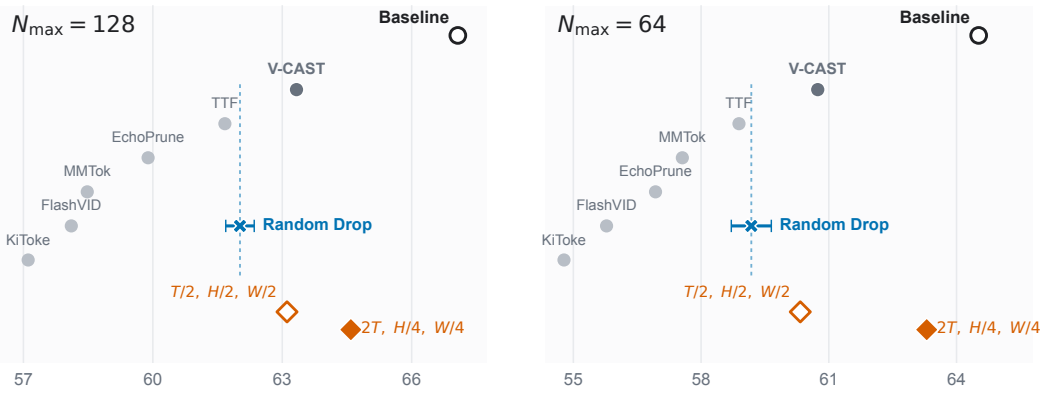

Figure 2: **VideoMME accuracy (%) on Qwen3-VL-4B-Instruct at a 12.5% visual-token budget.** The best compression method is only comparable to uniform spatiotemporal downsampling $(T/2, H/2, W/2)$ and is outperformed by $(2T, H/4, W/4)$. Random Drop markers and error bars show the mean and standard deviation over five independent runs.

GLM-4.6V-Flash (Hong et al., 2025) combine temporal patching with spatial compression; VideoLLaMA3 (Zhang et al., 2025a) integrates spatial downsampling and differential frame pruning; and Oryx (Liu et al., 2025) employs a dynamic spatial compressor. These designs change both the representations on which additional compression operates and the available ways to control visual-token budgets, motivating a reassessment of existing compression methods.

An initial evaluation on Qwen3-VL-4B-Instruct (Bai et al., 2025) reveals a surprising result. Under matched visual-token budgets, most evaluated compression methods fail to outperform simple random token dropping on VideoMME (Fu et al., 2025), as shown in Figure 2. Even V-CAST (Lin et al., 2026), the strongest compression method in this comparison, showcases only a small advantage over straightforward spatiotemporal downsampling ($(T/2, H/2, W/2)$). Moreover, a downsampling variant that increases temporal sampling density while reducing spatial resolution achieves higher performance within the same token budget. These observations suggest that the benefits of carefully designed token compression methods may be less consistent than previously assumed, while the allocation of the visual budget itself deserves closer attention. This raises a fundamental question: *Do we still need token compression in modern VideoLLMs?*

We investigate this question through a comprehensive evaluation across four native-resolution architectures Bai et al. (2025); Hong et al. (2025); Zhang et al. (2025a); Liu et al. (2025). Our study compares one of the strongest compression methods, V-CAST, with random dropping, input downsampling, and a simple dynamic-resolution baseline that adjusts frame resolutions based on CLIP-based frame–query similarity. We evaluate these strategies on four video benchmarks (Fu et al., 2025; Wu et al., 2024; Cheng et al., 2025; Deng et al., 2025) under multiple evaluation settings and visual-token budgets. The results reveal that the well-designed compression strategy does not consistently deliver better accuracy–efficiency trade-offs than the simple alternatives.

To understand these results, we further develop a deeper analysis of the architectural and data-related factors that govern the value of visual token compression. Our analysis examines how visual aggregation, spatiotemporal coverage, and the distribution of relevant information in videos affect the benefits of token selection. This perspective helps explain why residual video redundancy can coexist with limited gains from sophisticated compression strategies: the opportunity to reduce visual input does not, by itself, establish the advantage of a particular selection or merging rule.

These findings motivate dynamic input-resolution allocation as a promising direction for efficient VideoLLMs. Although random dropping and uniform downsampling provide strong baselines, their fixed configurations offer limited adaptation to the information requirements of different inputs. Native-resolution support provides a natural basis for allocating visual budgets according to content. By adjusting resolution before visual encoding, this approach can reduce both vision-encoder computation and the number of visual tokens processed by the LLM. The competitive performance of simple input-resolution strategies in our experiments supports further investigation of this direc-

tion. Our results therefore challenge the necessity of specialized, additional token compression and encourage greater attention to adaptive control of visual inputs.

Our contributions are threefold:

- We benchmark efficient VideoLLM inference across four native-resolution architectures, showing that a leading compression method does not consistently outperform random dropping, downsampling, or dynamic resolution.
- Through controlled ablations, we show that importance-guided selection can underperform random sampling under fixed temporal budgets. We further characterize backbone-dependent token geometry and selection preferences.
- We identify dynamic resolution as a promising alternative: a simple baseline achieves competitive accuracy while reducing computation in both the vision encoder and the LLM.

2 Related Work

Token Compression in VideoLLMs DyCoke (Tao et al., 2025) combining cross-frame merging and dynamic KV-cache pruning. Subsequent approaches develop two complementary operations: *pruning* less informative tokens and *merging* correlated representations. PruneVid (Huang et al., 2025) and FastVID (Shen et al., 2025) combine spatiotemporal aggregation with relevance- or attention-based selection. HoliTom (Shao et al., 2026), FlashVID (Fan et al., 2026), KiToke (Huang & Li, 2026), and TTF (Huo & LI, 2026) further exploit temporal or spatiotemporal redundancy through structured token merging. Recent pruning methods refine token utility and budget allocation: EchoPrune (Li et al., 2026) combines query relevance with temporal novelty, while V-CAST (Lin et al., 2026) uses semantic trajectory curvature to allocate frame-wise budgets and preserves token coordinates during spatial selection.

Existing work has also revisited the effectiveness of token compression relative to simple baselines. Hao et al. (2025) examine the competitiveness of random dropping in video action recognition; Wen et al. (2025a); Endo et al. (2025) identify positional biases in attention-based pruning and limitations of existing evaluations; Wen et al. (2025b) emphasize token duplication over importance; and Wang et al. (2026) investigate diminishing visual-token information in deeper VLLM layers. However, these analyses primarily concern image tasks, action recognition, or pruning within the language model, leaving unclear whether additional post-encoding compression consistently outperforms random dropping and input downsampling across native-resolution VideoLLMs. Our work reassesses whether their advantages remain consistent relative to random dropping and input downsampling under matched token budgets, and investigates dynamic resolution allocation as an alternative that can reduce computation in both the vision encoder and the LLM.

Native-Resolution Architectures and Dynamic Resolution Native-resolution designs enable variable-length visual token sequences according to input resolution and aspect ratio. NaViT (Dehghani et al., 2023) enables efficient variable-resolution encoding through sequence packing. To accommodate variable patch grids, Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025), and VideoLLaMA3 (Zhang et al., 2025a) adopt 2D-RoPE, while Oryx (Liu et al., 2025) uses positional embedding interpolation. Qwen3-VL (Bai et al., 2025) and GLM-4.6V-Flash (Hong et al., 2025) combine both techniques. These designs make input resolution a practical control over both visual-encoding cost and the LLM's visual-token budget, providing the architectural basis for adaptive resolution allocation. Building on this flexibility, dynamic-resolution methods adapt input fidelity to allocate limited visual computation. Q-Frame (Zhang et al., 2025b) combines query-aware frame selection with multi-resolution assignment, while F2C (Sun et al., 2025) extends selected frames into temporally coherent clips by trading spatial resolution for clip length. ResAdapt (Liao et al., 2026) learns per-frame resolution allocation through cost-aware reinforcement learning, and LDDR (Chen et al., 2026) combines diversity-aware frame selection with importance-guided resolution allocation. Our work evaluates the competitiveness of input-resolution allocation against sophisticated post-encoding token compression across modern VideoLLMs.

3 EXPERIMENTS

Native-resolution VideoLLMs already control visual-token counts through input resolution and built-in aggregation. This raises two questions: (1) whether additional token compression offers a consistent advantage over simple token dropping and downsampling? and (2) whether adaptive resolution allocation provides a better accuracy–efficiency trade-off? We investigate these questions across multiple architectures, token budgets, and benchmarks, assessing whether the conclusions generalize beyond a single model or task.

3.1 EXPERIMENT SETUP

Frame and token budgets. We evaluate 64- and 128-frame configurations with target visual-token retention ratios of 12.5% and 5%. For each backbone and frame configuration, the retention ratio is defined relative to the visual-token count of the corresponding uncompressed baseline after its native visual processing. We compare strategies under matched target visual-token budgets.

Benchmarks. We evaluate on four complementary benchmarks: Video-MME (Fu et al., 2025) for video understanding across diverse domains and durations; LongVideoBench (Wu et al., 2024) for long-context video understanding; Video-Holmes (Cheng et al., 2025) for complex reasoning over temporally distributed visual clues; and SciVideoBench (Deng et al., 2025) for scientific reasoning over real-world experimental videos.

3.2 COMPARED STRATEGIES

V-CAST. V-CAST (Lin et al., 2026) combines curvature-guided temporal budget allocation with spatial token selection based on feature diversity and magnitude. It reports state-of-the-art compression performance on Qwen3-VL-8B, LLaVA-OneVision-7B, and LLaVA-Video-7B. It also achieves the highest accuracy among the token compression methods in Figure 2. We therefore adopt it as a strong representative of post-encoding token compression.

Random Token Dropping. We uniformly sample visual tokens without replacement across all frames to meet the target budget, without importance scoring.

Spatiotemporal Downsampling. We reduce the input dimensions to $(T/2, H/2, W/2)$ and $(T/2, H/\sqrt{10}, W/\sqrt{10})$ for the 12.5% and 5% settings, respectively. The realized visual-token budgets may deviate from these targets due to rounding the resized dimensions to spatial grids.

Dynamic Resolution. We obtain frame relevance weights $\{p_t\}_{t=1}^{T}$ by applying softmax to CLIP ViT-B/16 frame–query cosine similarities with CLIP's learned scale, and use the exponential of Shannon entropy as a heuristic estimate of the effective support of the frame relevance distribution, which determines the initial number of retained frames as follows:

$$N_{\text{eff}} = \exp\left(-\sum_{t=1}^{T} p_t \log p_t\right).$$

We select the highest-scoring frames, with their count initialized to $\lceil N_{\text{eff}} \rceil$ and adjusted to satisfy the total budget B and per-frame limits. We first assign 64 tokens to each selected frame and distribute the remaining budget proportionally to frame relevance. Allocations are capped at 1024 tokens per frame, with any excess iteratively redistributed among uncapped frames. For each quota b_t, we search for a compatible spatial grid with $H_t = d\, g_{h,t}$, $W_t = d\, g_{w,t}$, and $\hat{b}_t = g_{h,t} g_{w,t}$, where d is the model-specific spatial token stride (e.g., $d = 32$ pixels for Qwen3-VL-4B-Instruct). The search approximates the target quota and original aspect ratio, with budget deviations arising from grid alignment. Selected frames are resized without cropping and supplied as timestamped image blocks in chronological order. Note that we do not propose a new approach here; a similar approach has been explored in prior work (Zhang et al., 2025b). Our goal is to provide a simple dynamic resolution baseline for comparing inference paradigms.

3.3 MODELS

We consider four native-resolution VideoLLMs with different visual encoding and token reduction mechanisms, summarized in Figure 3.

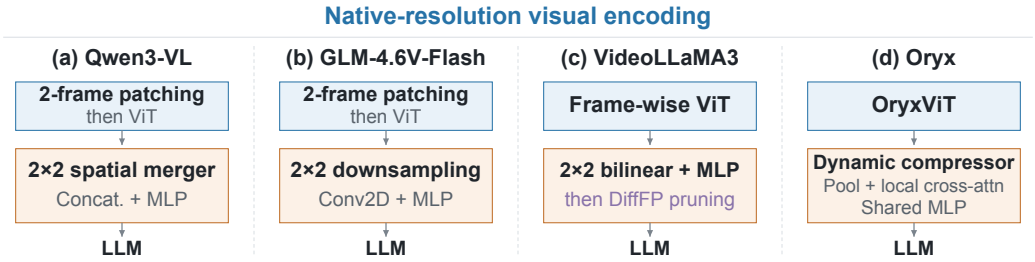

Figure 3: Illustration of the four evaluated VideoLLMs.

Qwen3-VL-4B-Instruct. Qwen3-VL (Bai et al., 2025) combines a dynamic-resolution SigLIP2-based vision encoder with temporal patch embedding over pairs of frames and a learned 2×2 spatial merger. Its DeepStack mechanism additionally injects intermediate visual features into early LLM layers.

GLM-4.6V-Flash. GLM-4.6V-Flash (Hong et al., 2025) supports variable-resolution visual encoding with two-frame temporal patch embedding. Encoded features undergo learned 2×2 convolutional downsampling followed by MLP projection into the LLM feature space.

VideoLLaMA3-7B. VideoLLaMA3 (Zhang et al., 2025a) uses frame-wise any-resolution visual encoding and 2×2 bilinear spatial downsampling. Its architecture also includes a differential frame pruner (DiffFP), which removes temporally redundant tokens based on pixel-level differences between adjacent frames.

Oryx-1.5-7B. Oryx-1.5 (Liu et al., 2025) combines OryxViT with a Qwen2.5 language backbone. Its dynamic compressor integrates spatial pooling and local cross-attention to support different compression ratios, followed by a shared MLP projector.

3.4 Main Results

Our experiments test whether sophisticated token compression provides a reliable advantage over simple alternatives in native-resolution VideoLLMs. Tables 1 and 2 show that at least one simple alternative—random dropping, downsampling, or CLIP-based dynamic resolution—achieves a higher four-benchmark average than V-CAST in 13 of the 16 model–frame-count–budget configurations. The comparisons below reveal where its advantage disappears, how this changes with compression strength, and which backbone and task differences require further explanation.

Token compression provides no consistent performance advantage over simple reduction. Random dropping and downsampling each exceed V-CAST in four-benchmark average accuracy in 10 of the 16 configurations. On GLM-4.6V-Flash and Oryx-1.5, both baselines outperform V-CAST across all evaluated frame counts and token budgets. For example, with 128 frames and a 5% budget on GLM-4.6V-Flash, V-CAST achieves an average score of 43.69, compared with 48.81 for random dropping and 49.07 for downsampling. Both baselines also achieve higher scores on every individual benchmark in this setting, so the deficit is not merely an artifact of averaging across tasks. These results establish content-agnostic reduction as a competitive alternative to V-CAST and challenge the necessity of sophisticated token selection for achieving favorable accuracy–budget trade-offs in native-resolution VideoLLMs.

Tighter budgets further weaken the competitiveness of token compression. A more restrictive token budget provides a stronger test of whether sophisticated selection preserves useful visual evidence. As we can observe, reducing retention from 12.5% to 5% produces a larger decline in four-benchmark average accuracy for V-CAST than for both random dropping and downsampling in 7 of the 8 settings. This pattern also holds on VideoMME individually, where V-CAST degrades more than both baselines in all 8 cases. For example, on GLM-4.6V-Flash with 128 frames, V-CAST loses 5.44 average-score points, compared with 1.96 for random dropping and only 0.53 for downsampling. Its deficit to downsampling consequently expands from 0.47 to 5.38 points. The deterioration also reverses previously favorable rankings: at 128 frames, both Qwen3-VL and VideoLLaMA3 shift from favoring V-CAST at 12.5% to favoring random dropping at 5%. Thus, tighter

Table 1: Comparison of efficient inference paradigms on Qwen3-VL-4B and GLM-4.6V-Flash. **Bold** and underlining indicate the best and second-best results within each setting, respectively.

Model	Size	$N_{\max}$	Setting	Token budget	VideoMME				LongVideo Bench	Video-Holmes	SciVideo Bench	Avg.	Ratio
					Short	Medium	Long	Overall					
Qwen3-VL	4B	64	Full Tokens	100%	76.00	64.56	53.00	64.52	61.56	41.37	30.20	49.41	100.00%
			V-CAST	12.5%	72.89	59.11	50.22	60.74	58.94	36.25	**26.80**	45.68	92.45%
			Random	12.5%	70.49	58.27	48.78	59.18	57.67	38.11	24.80	44.94	90.95%
			Downsample	12.5%	71.11	58.67	51.00	60.26	55.81	**38.98**	26.70	45.44	91.97%
			DynRes	12.5%	**74.00**	**59.89**	**51.11**	**61.67**	**63.87**	37.62	25.70	**47.22**	**95.56%**
			V-CAST	5%	67.11	55.67	47.33	56.70	56.02	37.13	24.40	43.56	88.16%
			Random	5%	66.44	54.56	49.56	56.85	55.37	36.31	23.20	42.93	86.89%
			Downsample	5%	69.00	**57.33**	**50.33**	**58.89**	55.88	**37.56**	**25.30**	**44.41**	**89.88%**
			DynRes	5%	**70.78**	55.67	49.89	58.78	**59.31**	36.31	22.80	44.30	89.66%
		128	Full Tokens	100%	77.00	66.56	57.67	67.07	62.53	44.20	30.90	51.18	100.00%
			V-CAST	12.5%	75.00	61.89	53.11	63.33	60.06	39.79	**28.30**	47.87	93.53%
			Random	12.5%	72.40	61.02	52.64	62.02	59.76	39.96	26.50	47.06	91.95%
			Downsample	12.5%	73.44	62.22	51.78	62.48	59.25	**40.07**	28.00	47.45	92.71%
			DynRes	12.5%	**75.22**	**65.67**	**57.00**	**65.96**	**65.52**	39.47	27.30	**49.56**	**96.83%**
			V-CAST	5%	70.11	58.11	50.22	59.48	56.84	37.78	24.30	44.60	87.14%
			Random	5%	68.22	57.33	51.00	58.85	56.92	**38.92**	24.60	44.82	87.57%
			Downsample	5%	72.11	**60.56**	50.56	61.07	57.23	38.76	**26.60**	45.92	89.72%
			DynRes	5%	**72.89**	**60.56**	**52.89**	**62.11**	**62.23**	37.83	25.40	**46.89**	**91.62%**
GLM-4.6V-Flash	9B	64	Full Tokens	100%	75.67	64.67	55.89	65.41	61.26	43.17	33.70	50.89	100.00%
			V-CAST	12.5%	69.67	58.33	51.44	59.81	60.28	40.77	30.30	47.79	93.91%
			Random	12.5%	73.00	61.44	54.22	62.89	59.84	43.33	29.80	48.97	96.23%
			Downsample	12.5%	71.33	58.33	53.11	60.93	57.82	**43.93**	30.50	48.30	94.91%
			DynRes	12.5%	**75.22**	**63.44**	**54.89**	**64.52**	**65.15**	40.77	**32.70**	**50.79**	**99.80%**
			V-CAST	5%	60.44	50.67	48.22	53.11	55.12	39.96	26.20	43.60	85.67%
			Random	5%	68.11	57.78	53.22	59.70	56.54	42.73	29.20	47.04	92.43%
			Downsample	5%	69.89	57.56	49.78	59.07	56.10	41.05	**30.40**	46.66	91.69%
			DynRes	5%	**72.78**	**60.11**	**54.78**	**62.56**	**59.99**	**42.90**	28.20	**48.41**	**95.13%**
		128	Full Tokens	100%	78.00	70.33	58.78	69.04	63.72	46.05	34.10	53.23	100.00%
			V-CAST	12.5%	70.67	62.78	**56.67**	63.37	59.31	42.84	31.00	49.13	92.30%
			Random	12.5%	74.44	62.44	56.11	64.33	62.60	**43.93**	32.20	50.77	95.38%
			Downsample	12.5%	73.33	65.11	54.78	64.41	59.54	**43.93**	30.50	49.60	93.18%
			DynRes	12.5%	**77.56**	**67.67**	56.22	**67.15**	**68.29**	43.71	**33.10**	**53.06**	**99.68%**
			V-CAST	5%	62.11	53.00	50.00	55.04	55.72	38.70	25.30	43.69	82.08%
			Random	5%	70.89	58.33	**56.22**	61.81	59.76	**43.77**	29.90	48.81	91.70%
			Downsample	5%	**73.00**	60.78	54.00	62.59	59.09	43.00	**31.60**	49.07	92.18%
			DynRes	5%	72.56	**63.44**	55.44	**63.81**	**63.28**	43.06	31.20	**50.34**	**94.57%**

budgets do not restore a reliable advantage for sophisticated selection; they generally amplify the tested policy's limitations relative to simpler reduction strategies.

Dynamic resolution provides a competitive alternative to pruning, particularly under tight budgets. DynRes exceeds V-CAST in four-benchmark average accuracy in 12 of the 16 configurations, including all eight Qwen3-VL and GLM-4.6V-Flash settings. Its advantage becomes more prevalent as the budget tightens: the number of configurations favoring DynRes increases from five of eight at 12.5% retention to seven of eight at 5%. For example, on GLM-4.6V-Flash with 128 frames, DynRes at 5% achieves an average score of 50.34, exceeding V-CAST at 12.5% (49.13) on all four benchmarks while using only 40% of its nominal visual-token budget. The long-video evaluations provide further evidence. On VideoMME-Long, DynRes outperforms V-CAST in three of eight configurations at 12.5%, increasing to seven of eight at 5%. On LongVideoBench, it achieves higher accuracy in 13 of 16 configurations; even its 5% results exceed those of V-CAST at 12.5% in five of eight model–frame-count pairs. Although V-CAST retains higher averages in three of four VideoLLaMA3 settings, these results demonstrate that input-level allocation can outperform sophisticated post-encoding selection at the same or substantially smaller visual-token budgets.

Table 2: Comparison of efficient inference paradigms on VideoLLaMA3 and Oryx-1.5. **Bold** and underlining indicate the best and second-best results within each setting, respectively.

Model	Size	$N_{\max}$	Setting	Token budget	VideoMME				LongVideo Bench	Video-Holmes	SciVideo Bench	Avg.	Ratio
					Short	Medium	Long	Overall					
VideoLLaMA3	7B	64	Full Tokens	100%	76.89	63.22	55.00	65.04	60.08	43.77	27.70	49.15	100.00%
			V-CAST	12.5%	**75.67**	59.22	**52.89**	**62.59**	**59.31**	**42.62**	**27.00**	**47.88**	**97.42%**
			Random	12.5%	72.78	**59.44**	51.44	61.22	56.69	42.03	26.90	46.71	95.04%
			Downsample	12.5%	73.56	57.11	51.56	60.74	56.25	40.94	26.60	46.13	93.86%
			DynRes	12.5%	72.33	58.78	52.11	61.07	56.10	42.30	25.50	46.24	94.08%
			V-CAST	5%	69.56	55.89	49.00	58.15	**58.41**	39.90	25.60	**45.52**	**92.61%**
			Random	5%	67.89	55.89	**51.44**	58.41	55.57	**41.43**	25.10	45.13	91.82%
			Downsample	5%	**70.44**	**57.44**	50.22	**59.37**	54.82	40.56	**25.80**	45.14	91.84%
			DynRes	5%	68.56	56.22	51.33	58.70	55.12	39.58	24.00	44.35	90.23%
		128	Full Tokens	100%	76.22	64.00	55.67	65.30	61.33	44.37	28.40	49.85	100.00%
			V-CAST	12.5%	**74.44**	**60.11**	54.00	**62.85**	**59.09**	42.13	**27.80**	**47.97**	**96.23%**
			Random	12.5%	74.22	60.00	**54.11**	62.78	56.62	**42.57**	26.40	47.09	94.46%
			Downsample	12.5%	72.89	57.78	51.33	60.67	55.87	42.35	26.10	46.25	92.78%
			DynRes	12.5%	72.33	58.78	53.11	61.41	56.69	41.48	24.50	46.02	92.32%
			V-CAST	5%	69.11	**57.56**	51.33	59.33	56.39	39.68	25.10	45.13	90.53%
			Random	5%	**70.00**	57.00	**53.00**	**60.00**	54.90	**41.43**	**26.10**	**45.61**	**91.49%**
			Downsample	5%	67.78	54.11	49.67	57.19	54.60	41.32	24.60	44.43	89.13%
			DynRes	5%	69.87	57.00	51.78	59.55	**56.77**	40.12	24.30	45.19	90.64%
Oryx-1.5	7B	64	Full Tokens	100%	70.89	56.00	49.44	58.78	57.52	36.20	17.10	42.40	100.00%
			V-CAST	12.5%	68.00	54.00	47.67	56.56	52.80	32.88	15.70	39.49	93.14%
			Random	12.5%	**70.22**	**54.22**	**49.11**	**57.85**	53.18	33.59	**16.90**	**40.38**	**95.24%**
			Downsample	12.5%	68.78	52.56	47.89	56.41	53.03	**33.70**	16.70	39.96	94.25%
			DynRes	12.5%	66.22	51.78	45.78	54.59	**54.23**	32.50	16.20	39.38	92.88%
			V-CAST	5%	62.44	50.44	45.33	52.74	50.93	32.12	**16.10**	37.97	89.55%
			Random	5%	**66.56**	**52.44**	**47.33**	**55.44**	51.53	**32.39**	15.80	**38.79**	**91.49%**
			Downsample	5%	64.00	49.78	45.56	53.11	52.21	31.85	15.30	38.12	89.91%
			DynRes	5%	64.78	51.44	44.89	53.70	**54.00**	32.01	15.30	38.75	91.39%
		128	Full Tokens	100%	69.78	57.11	50.56	59.15	56.10	35.66	17.60	42.13	100.00%
			V-CAST	12.5%	69.22	**55.22**	**49.22**	**57.89**	53.63	34.02	16.50	40.51	96.15%
			Random	12.5%	**69.78**	54.67	49.11	57.85	54.97	33.59	16.30	40.68	96.56%
			Downsample	12.5%	**69.78**	53.11	48.11	57.00	54.82	**34.68**	**17.30**	**40.95**	**97.20%**
			DynRes	12.5%	67.67	54.67	47.78	56.70	**55.12**	33.81	16.70	40.58	96.32%
			V-CAST	5%	64.44	50.33	47.11	53.96	52.58	33.04	14.70	38.57	91.55%
			Random	5%	**67.22**	**53.89**	46.78	55.96	53.18	**33.70**	15.20	39.51	93.78%
			Downsample	5%	65.33	52.44	46.56	54.78	53.63	32.28	15.60	39.07	92.74%
			DynRes	5%	**67.22**	53.78	**48.33**	**56.44**	**55.05**	31.68	**16.50**	**39.92**	**94.75%**

The benefits of dynamic resolution also depend on the task. Relative to uniform downsampling, DynRes improves LongVideoBench in 15 of 16 configurations but scores lower on Video-Holmes in 12. This contrast motivates further investigation of how input budgets should be allocated to task-relevant evidence. Together with the observed backbone dependence, these results suggest that realizing the potential of dynamic resolution requires allocation policies suited to both the model and the task. The following section examines how token selection, visual representations, and evidence requirements shape these outcomes.

4 Discussion

Our results motivate reconsidering sophisticated token compression as a default requirement for efficient VideoLLM inference. We first isolate the contribution of token selection under fixed temporal allocations, then compare visual-token geometry and selection preferences across architectures. These analyses inform the case for dynamic resolution, whose computational benefits we subsequently evaluate across the full inference pipeline.

Table 3: Different token selection strategies at 5% retention on GLM-4.6V-Flash using V-CAST. V and R denote V-CAST and random selection, respectively. Percentages in the selection row indicate the random replacement ratio, with V and R corresponding to 0% and 100%.

Temporal budget allocation	V	V	V	V	V	V	R
Within-group selection	V	10%	25%	50%	80%	R	R
VideoMME-sub Acc. (%)	59.17	61.73	62.95	63.38	63.96	64.58	**65.69**

4.1 Where Does the Performance Gap to Random Drop Arise?

Importance and diversity scores aim to identify visual tokens worth retaining (Lin et al., 2026; Li et al., 2026). Using V-CAST as a representative method, we freeze its token quota for each native temporal group and progressively replace its diversity–saliency selections with uniform random sampling. Replacement ratios range from 0% to 100%, isolating within-group selection from temporal allocation. We evaluate 120 VideoMME videos (360 questions) using GLM-4.6V-Flash with 128 frames and 5% retention. All conditions share sampled frames and actual retained-token counts, preserve original token coordinates, and average randomized results over two mask seeds. As shown in Table 3, accuracy increases monotonically with the random replacement ratio. Replacing just 10% of selections improves accuracy from 59.17% to 61.73%; replacing 50% reaches 63.38%, and fully random within-group selection achieves 64.58%. Global random dropping provides a further improvement to 65.69%, but most of the observed gain is already obtained with temporal quotas unchanged. This progressive improvement identifies within-group selection as a source of performance loss in this setting: greater reliance on the designed diversity–saliency ranking does not yield better accuracy. These results strengthen the evidence that sophisticated token selection is not a prerequisite for competitive token reduction.

4.2 Why Can Token Compression Underperform in Modern Architecture?

To interpret this result, we examine token geometry and selection preferences across LLaVA-Video-7B (Li et al., 2024b), Qwen3-VL-4B (Bai et al., 2025), and GLM-4.6V-Flash (Hong et al., 2025) on the same VideoMME subset with identical 128-frame inputs and 5% token budget. We extract visual embeddings before the first LLM layer, excluding special tokens. Selection statistics use the full candidate sequence at 5% retention, comparing V-CAST (V) with random drop in the same temporal budget allocation (H).

Feature Space. We measure the largest anchor-based cluster share C at cosine threshold 0.90 and the entropy-based effective rank of the cosine Gram matrix. Table 4 shows that LLaVA-Video has larger similarity clusters and a lower effective rank (34.36) than Qwen3-VL (52.29) and GLM (71.15), indicating greater directional concentration in its original feature space. These observations suggest that token compression methods may benefit from LLaVA's feature space, while native visual aggregation in modern architectures may alter the geometric cues these methods rely on, limiting the gains from further compression.

Selection Preferences. We characterize retained tokens by their within-group norm percentiles and deviation from the group mean, $1 - \cos(\mathbf{x}, \bar{\mathbf{x}}_g)$. On LLaVA-Video, V increases mean deviation from 0.422 to 0.899, while its average norm percentile is below H (0.437 vs. 0.500). On Qwen3-VL and GLM, V instead exhibits strong high-norm enrichment, with an average percentile of 0.859. On GLM, mean deviation also increases from 0.528 to 0.662. These measurements show that a common scoring principle induces different retention preferences across feature spaces, which explains why existing token compression paradigms behave differently across architectures.

Selection preferences need not yield predictive gains. The GLM results connect these preferences directly to Section 4.1: despite favoring high-norm, mean-deviating tokens, V achieves 59.17% accuracy, compared with 64.58% for H. Accuracy also increases monotonically across the tested random replacement ratios while temporal quotas remain fixed. Together, these observations reveal a mismatch between the evaluated geometric criterion and downstream predictive utility: stronger reliance on its ranking does not produce a more useful retained representation in this setting. This helps explain the competitiveness of random sampling and motivates re-evaluating geometric importance criteria across backbones.

Table 4: Visual-token geometry and selection preferences. V/H denote V-CAST and quota-matched random selection at 5% retention.

Metric	LLaVA-Video-7B	Qwen3-VL-4B	GLM-4.6V-Flash
Largest cluster (%)	1.890	0.633	0.498
Effective rank	34.36	52.29	71.15
Norm percentile (V / H)	0.437 / 0.500	0.859 / 0.500	0.859 / 0.500
Mean deviation (V / H)	0.899 / 0.422	0.596 / 0.449	0.662 / 0.528

Table 5: **Stage-wise inference latency (ms).** Means over 100 LongVideoBench cases using vLLM and Qwen3-VL-4B-Instruct with 64 candidate frames and 12.5% retention ratio.

Stage	Baseline	V-CAST	Random	Downsample	DynRes
Video Decoding	1304.2	1335.2	1343.9	752.6	1355.1
Method Overhead	—	14.3	—	—	463.5
Preprocessing	316.2	213.2	231.8	108.4	611.6
ViT Encoding	231.0	233.4	234.5	21.7	59.3
vLLM Preparation	2049.6	2019.0	2032.7	806.9	399.5
LLM Prefill-related Work	685.7	59.5	67.7	54.3	74.7
LLM Decoding	15.8	7.0	7.5	6.8	7.6
End-to-End	4602.4	3881.6	3918.2	**1750.7**	2973.3

4.3 Why Is Dynamic Resolution a Promising Alternative?

Random dropping and uniform downsampling are competitive with sophisticated token compression, but neither adapts its allocation to video content or the question. Since relevant events and required spatial detail vary across frames, nonuniform allocation offers opportunities to use a limited visual budget more effectively. Native-resolution support makes dynamic resolution a practical way to exploit these opportunities: frame-level token budgets are mapped to input resolutions before encoding, enabling adaptive allocation and computational savings in both the vision encoder and the LLM. The competitive performance of our simple CLIP-based baseline (Tables 1 and 2) supports this direction. Although gains depend on the backbone and allocation policy, these results motivate dynamic resolution as a promising alternative to sophisticated post-encoding selection.

Besides, as shown in Table 5, DynRes reduces end-to-end latency by 23.4% and 24.1% relative to V-CAST and random dropping, respectively, including adaptation overhead. By reducing visual input before encoding, DynRes lowers ViT latency to 59.3 ms, compared with 233.4 ms for V-CAST and 234.5 ms for random dropping. Its computational benefits therefore extend beyond the LLM to the vision encoder. Uniform downsampling achieves the lowest latency (1750.7 vs. 2973.3 ms for DynRes). However, DynRes achieves higher average accuracy in 11 of 16 configurations across Tables 1 and 2. Dynamic resolution thus offers a competitive accuracy–efficiency trade-off, with the flexibility to adapt frame retention and spatial resolution to query relevance.

5 Conclusion

We systematically evaluate efficient VideoLLM inference strategies across four native-resolution architectures and four video benchmarks. Our results show that a strong post-encoding token compression method does not consistently outperform simple random dropping, while a basic query-aware dynamic resolution baseline achieves the highest average accuracy in half of the evaluated configurations. Although its benefits vary across tasks and architectures, dynamic resolution offers a flexible way to allocate visual budgets according to frame–query relevance, with the potential to reduce both visual-encoding cost and the number of tokens processed by the LLM. These findings challenge the assumption that sophisticated token compression is necessary for competitive performance under limited visual-token budgets. Native-resolution support makes adaptive input allocation a practical alternative, while the persistent spatiotemporal redundancy of video leaves substantial room for further efficiency improvements. We therefore advocate greater emphasis on dynamic resolution and query-aware input allocation in future research on efficient VideoLLMs.

AI USE STATEMENT

Generative AI tools, including Codex and Claude Code, were used to assist with experimental design, manuscript organization, code debugging, and language editing. These tools were also used for multiple rounds of simulated manuscript review to identify potential weaknesses and improve the paper. All AI-generated suggestions and outputs were reviewed by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Jingfeng Chen, Jiawen Qian, Wendi Deng, Yinuo Guo, Jiaqi Yu, Sicong Leng, Raghuveer Thirukovalluru, and Bhuwan Dhingra. Lddr: Linear-dpp-based dynamic-resolution frame sampling for video mllms. *arXiv preprint arXiv:2605.11477*, 2026.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*, 2025.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, volume 2026, pp. 48075–48099, 2026.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835, 2025. doi: 10.1109/ICCV51701.2025.02119.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. FlashVID: Efficient video large language models via training-free tree-based spatiotemporal token merging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=H6rDX4w6Al.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Xinyue Hao et al. Principles of visual tokens for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Haifeng Huang and Yang Li. Kitoke: Kernel-based interval-aware token compression for video large language models. *arXiv preprint arXiv:2604.03414*, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Simin Huo and Ning LI. Ttf: Temporal token fusion for efficient video-language model. *arXiv preprint arXiv:2605.07355*, 2026.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Jiameng Li, Minye Wu, Jiezhang Cao, Aleksei Tiulpin, and Matthew B Blaschko. Echoprune: Interpreting redundancy as temporal echoes for efficient videollms. *arXiv preprint arXiv:2605.10050*, 2026.

Huanxuan Liao, Zhongtao Jiang, Yupu Hao, Yuqiao Tan, Shizhu He, Ben Wang, Jun Zhao, Kun Xu, and Kang Liu. Resadapt: Adaptive resolution for efficient multimodal reasoning. *arXiv preprint arXiv:2603.28610*, 2026.

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

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. In *International Conference on Learning Representations*, volume 2025, pp. 85485–85507, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *Advances in Neural Information Processing Systems*, 38:135547–135570, 2026.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, pengzhang liu, Sicheng Zhao, and Guiguang Ding. FastVID: Dynamic density pruning for fast video large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2xS4VtpApy.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From frames to clips: Training-free adaptive key clip selection for long-form video understanding. *arXiv preprint arXiv:2510.02262*, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001, June 2025.

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

Yahong Wang, Juncheng Wu, Zhangkai Ni, Longzhen Yang, Yihang Liu, Chengmei Yang, Ying Wen, Lianghua He, Xianfeng Tang, Hui Liu, and Yuyin Zhou. When token pruning is worse than random: Understanding visual token information in vllms, 2026. URL https://arxiv.org/abs/2512.07580.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.802. URL https://aclanthology.org/2025.findings-acl.802/.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9961–9980, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.505. URL https://aclanthology.org/2025.emnlp-main.505/.

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

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

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025b.