Vista-LLM: Decoupled Query-Guided Visual Token Pruning for Efficient Long-Video Large Language Models

Zhenyu Li[1], **Zuchao Li**[2*], **Ping Wang**[3], **Lefei Zhang**[1], **Haojun Ai**[4]
[1]School of Computer Science ,Wuhan University
[2]School of Artificial Intelligence ,Wuhan University
[3]School of Information Management ,Wuhan University
[4]School of Cyber Science and Engineering ,Wuhan University, Wuhan, China
{zhenyu-li, zcli-charlie, wangping, zhanglefei, aihj}@whu.edu.cn

Abstract

Long-video understanding is bottlenecked by the high cost of processing massive visual tokens. Current reduction strategies often rely on static allocation or inefficient in-network selection that disrupts optimized attention kernels. In this paper, we introduce Vista-LLM, a decoupled framework for query-guided visual token pruning. By filtering redundancy prior to inference with minimal overhead, Vista-LLM ensures full compatibility with Flash Attention. Our method employs a coarse-to-fine pipeline: (1) Query-Guided Dynamic Budgeting for adaptive temporal allocation; (2) a lightweight Semantic Scout for fine-grained, query-specific selection; and (3) Structure-Aware Compensation to preserve global context. Extensive experiments on benchmarks like Video-MME and MLVU demonstrate a significantly improved Pareto frontier. Notably, on LLaVA-OneVision, Vista-LLM reduces visual tokens by 90% and accelerates inference while retaining over 98% of baseline performance on average, effectively filtering visual noise. Our code is available at https://github.com/lizhenyu-123/Vista-LLM.

1 Introduction

The transition from static image understanding (Radford et al., 2021; Li et al., 2023; Liu et al., 2023, 2024a; Chen et al., 2023) to long-video comprehension (Zhang et al., 2023; Maaz et al., 2024; Li et al., 2025; Zhang et al., 2025b; Hurst et al., 2024; Team, 2025) is hindered by the low signal-to-noise ratio inherent in visual streams. Videos are characterized by high spatiotemporal redundancy (Sullivan and Wiegand, 2005), where a significant portion of frames typically consists of repetitive backgrounds or static scenes that contribute little to answering a specific user query. However, current Video-LLMs largely treat visual tokens equally, forcing the model to process a massive number of redundant tokens to locate sparse relevant information. This inefficient brute-force approach not only exceeds the context window limits of LLMs (Liu et al., 2024b) but also distracts the model with irrelevant visual noise, often leading to hallucinations (Bai et al., 2024) and degraded reasoning capabilities in long-context scenarios (Guo et al., 2025).

To mitigate this computational bottleneck, prior research has explored various visual token pruning techniques (Shao et al., 2025b), yet they suffer from three critical limitations. First, a prevalent flaw across most paradigms is static budget allocation, where a fixed token count is enforced per frame regardless of content (Bolya et al., 2023; Yue et al., 2021). This temporal rigidity fails to adapt to the varying information density of videos, wasting resources on static backgrounds while truncating complex motion segments. Second, regarding token selection, many methods operate in a task-agnostic manner, relying solely on visual statistics (Yang et al., 2025; Fu et al., 2025b; Shao et al., 2025a; Shen et al., 2025; Qu et al., 2024). By neglecting the user's query, they inherently risk discarding visually inconspicuous but semantically pivotal details. Third, even methods that incorporate language guidance (Zhang et al., 2025a; Tao et al., 2025) are often systemically inefficient. Relying on in-network attention for selection necessitates processing the entire sequence through shallow layers, introducing unavoidable overhead. Furthermore, extracting intermediate attention weights is fundamentally incompatible with Flash Attention (Dao et al., 2022), as avoiding matrix materialization is central to its I/O optimization. This forces a fallback to memory-intensive implementations, often negating the theoretical speedup.

In response to these challenges, we introduce

*Corresponding author.

This work was supported by the National Natural Science Foundation of China (No. 62306216), the Technology Innovation Program of Hubei Province (No. 2024BAB043) and the National Social Science Foundation of China (No. 24&ZD186).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13171–13187
July 2-7, 2026

Vista-LLM, a resource-efficient visual tokenization framework designed to resolve the tension between efficiency and comprehension. The framework consists of three synergistic components. First, to break temporal rigidity, we propose a Query-Guided Dynamic Budgeting strategy. Second, to bridge the semantic gap, we employ a BLIP-2 Q-Former (Li et al., 2023) as a semantic scout. By conditioning the Q-Former on the user's question, we identify and retain the most salient visual tokens based on cross-modal attention weights. Finally, to prevent semantic tunnel vision, we incorporate a Structure-Aware Compensation mechanism utilizing a Density Peak Clustering (DPC) (Du et al., 2016; Rodriguez and Laio, 2014) strategy. We explicitly identify representative structural anchors by selecting tokens characterized by high local density and distinct spatial separation, ensuring comprehensive coverage of the frame's geometric layout. Furthermore, to mitigate information loss, we implement a soft-pruning technique where unselected background features are aggregated into their nearest structural anchors, thereby preserving the integrity of environmental contexts without incurring computational redundancy.

We validate the robustness and universality of Vista-LLM by seamlessly integrating it with state-of-the-art models such as LLaVA-OneVision (Li et al., 2025) , LLaVA-Video (Zhang et al., 2025b) and Qwen2.5-VL (Bai et al., 2025). Experiments across diverse benchmarks—including Video-MME (Fu et al., 2025a) for general understanding, MLVU and LongVideoBench (Zhou et al., 2025; Wu et al., 2024) for long-context reasoning, and MVBench (Li et al., 2024) for temporal perception—demonstrate that our method achieves an optimal trade-off between efficiency and performance, reducing visual token consumption by up to 90% while consistently maintaining or even surpassing the accuracy of full-token baselines.

2 Related Work

2.1 Efficient Cross-Modal Alignment

Aligning visual features with linguistic semantics is fundamental to MLLMs. While dual-encoder models like CLIP (Radford et al., 2021) provide effective global alignment, they lack the granularity required for fine-grained token selection. To address this limitation, BLIP-2 (Li et al., 2023) introduced the Querying Transformer (Q-Former), a lightweight bottleneck module designed to extract query-driven visual features via cross-attention, with optional support for text-conditioned querying in vision-language tasks. Although originally developed for representation learning and generation, the Q-Former's architecture inherently supports efficient, query-aware visual filtering. Our work repurposes this capability, treating the Q-Former from a feature extractor as a semantic scout to identify informative regions prior to LLM inference.

2.2 Large Multimodal Models for Video

Recent advancements in Multimodal Large Language Models (MLLMs) generally follow two paradigms: projection-based adaptation and unified modeling. The former, representing the dominant open-source approach, aligns pre-trained visual encoders with frozen LLMs via learnable interfaces. Models such as Video-LLaMA (Zhang et al., 2023) and LLaVA-Video (Zhang et al., 2025b) treat video as dense frame sequences, leveraging adapters to map temporal features into the text space. In parallel, native multimodal models like Chameleon (Team, 2024) and the proprietary GPT/Gemini families (Hurst et al., 2024; Team, 2025) explore processing interleaved inputs in a unified architecture. While native models define the state-of-the-art, they demand prohibitive pre-training resources. Consequently, projection-based architectures remain the practical standard for academic research, though they face a critical scalability issue: linearly growing frame tokens trigger quadratic computational complexity in self-attention (Vaswani et al., 2017). This memory bottleneck, which severely limits KV-cache capacity during inference (Shi et al., 2024; Tang et al., 2025b; Zhao et al., 2025), creates an urgent need for efficient visual token reduction in long-video understanding, as discussed below.

2.3 Visual Token Compression

Strategies for reducing visual redundancy in Large Multimodal Models (LMMs) can be broadly categorized based on the information source used for selection: visual-centric reduction and query-aware selection.

Visual-Centric Reduction. These methods compress visual representations by exploiting intrinsic spatial-temporal correlations, independent of user inputs. A primary direction involves aggregating similar tokens to preserve context; ToMe (Bolya et al., 2023) applies a bipartite soft matching algorithm to progressively merge visually sim-

ilar tokens within transformer layers. Adopting a hybrid strategy, VisionZip (Yang et al., 2025) retains dominant tokens with high intrinsic attention while merging the remaining contextual tokens based on semantic similarity. Addressing the unique challenges of video data, FrameFusion (Fu et al., 2025b) integrates similarity-based token merging with importance-based pruning, specifically targeting spatially corresponding tokens across adjacent frames to reduce temporal redundancy. Subsequently, FastVID (Shen et al., 2025) introduces dynamic density pruning to partition videos into temporally ordered segments, effectively identifying and removing spatiotemporal redundancy. HoliTom (Shao et al., 2025a) further proposes a holistic framework that synergizes outer-LLM global temporal segmentation with inner-LLM similarity-based merging. Efficiency is also addressed at the input level, where TS-LLaVA (Qu et al., 2024) synthesizes a global thumbnail with sparse temporal sampling to construct compact inputs directly at the source.

Query-Aware Selection. To align visual retention with user intent, recent research integrates language guidance to filter information based on relevance to the textual instruction. FastV (Chen et al., 2024) identifies redundancy in deep layers and executes pruning in early blocks by ranking visual tokens via text-image attention scores. To mitigate pruning-induced information loss, SparseVLM (Zhang et al., 2025a) employs designated text tokens as significance evaluators and incorporates a feature recycling mechanism to preserve context. Distinct from these static approaches, DyCoke (Tao et al., 2025) implements dynamic token maintenance during the decoding phase, monitoring attention weights to adaptively discard or recall visual tokens throughout the generation process. Beyond these training-free methods, learning-based approaches like CoViPAL (Tang et al., 2025a) train a lightweight classifier that jointly processes visual and textual tokens to predict and prune redundancy.

Despite these advancements, existing paradigms face inherent trade-offs. Visual-centric methods risk discarding semantically important details due to their task-agnostic nature. In contrast, most query-aware approaches rely on in-network selection within the LLM, requiring additional shallow-layer computation to obtain token importance and depending on attention statistics that are incompatible with modern optimized attention kernels such as Flash Attention (Dao et al., 2022). We address these limitations with a decoupled, pre-inference framework that performs hybrid selection before the LLM stage, preserving semantic relevance while fully enabling efficient attention implementations.

3 Methodology

3.1 Overview

Formal Definition. Given a video sequence $V = \{v_1, v_2, \ldots, v_T\}$ consisting of T frames and a user textual query Q, a standard Video-LLM projects V into a sequence of visual tokens $X \in \mathbb{R}^{N_T \times D}$, where N_T is typically massive for long-form videos. Our goal is to derive a compressed token sequence $\hat{X}$ with length $N_K \ll N_T$, such that the likelihood of generating the correct response $P(A|\hat{X}, Q)$ is maximized under a strict computation budget.

The Coarse-to-Fine Framework. As illustrated in Figure 1, we propose a decoupled, training-efficient framework that selects visual tokens through a three-stage synergistic pipeline. (i) **Query-Guided Dynamic Budgeting:** Instead of treating all frames uniformly, we first segment the video based on visual coherence. We then compute the relevance between each segment and the query Q, adaptively allocating a token budget B_t to each frame v_t. This ensures that critical moments receive higher resolution while redundant frames are heavily compressed. (ii) **Hybrid Spatial Selection:** Within each frame, we employ a dual-stream strategy to fill the assigned budget B_t. We utilize a fine-tuned BLIP-2 Q-Former (Li et al., 2023) as a semantic scout to identify query-relevant patches serving as the foveal focus, while simultaneously selecting representative tokens via density peak clustering to preserve structural context. (iii) **Aggregation and Contiguous Assembly:** To prevent information loss, unselected tokens are aggregated into their nearest selected anchors. Finally, the selected foveal and peripheral tokens are concatenated into a single tensor $\hat{X}$ for efficient inference.

System Efficiency. All token selection is performed prior to Video-LLM inference, decoupling query-aware filtering from the LLM. This design avoids additional shallow-layer computation and preserves full compatibility with optimized attention kernels such as Flash Attention.

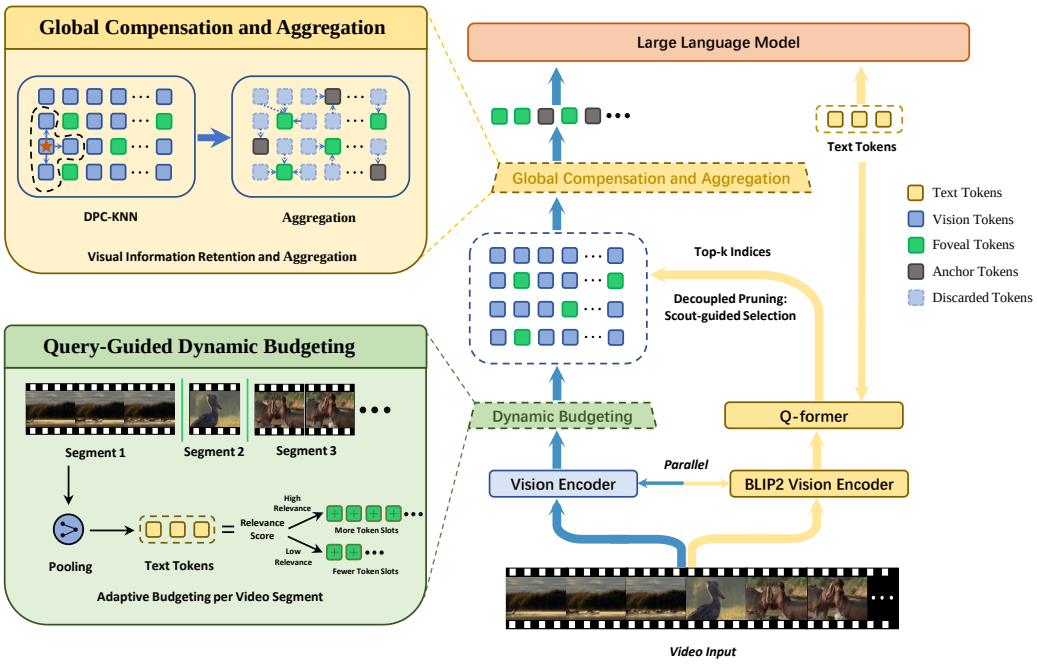


Figure 1: The Vista-LLM architecture. The framework consists of three key components: A decoupled pruning pipeline that identifies query-salient Foveal Tokens; A Query-Guided Dynamic Budgeting module for adaptive token allocation; and A Visual Information Retention and Aggregation mechanism that maintains global context via Anchor Tokens and feature aggregation.

3.2 Query-Guided Dynamic Budgeting

To address the inefficiency of static allocation, we propose a Query-Guided Dynamic Budgeting mechanism that redistributes computational resources based on semantic density. First, to partition the video into coherent events, we derive a global feature vector $\mathbf{f}_t$ for each frame v_t by performing average pooling over its spatial tokens. We then compute the inter-frame cosine similarity s_t between adjacent feature vectors $\mathbf{f}_t$ and $\mathbf{f}_{t+1}$. A segment boundary is determined whenever the visual coherence drops below a threshold τ:

$$\mathbf{f}_t = \frac{1}{L}\sum_{k=1}^{L} v_{t,k} \tag{1}$$

$$s_t = \frac{\mathbf{f}_t \cdot \mathbf{f}_{t+1}}{\|\mathbf{f}_t\|\|\mathbf{f}_{t+1}\|}, \quad \text{Boundary if } s_t < \tau \tag{2}$$

where L is the number of tokens per frame. This yields a set of variable-length segments $\mathcal{S}$. For each segment S_m, we derive a global prototype $\mathbf{h}_m$ via average pooling and measure its semantic relevance r_m to the query embedding $\mathbf{q}$ as follows:

$$\mathbf{h}_m = \frac{1}{|S_m|}\sum_{v_k \in S_m} v_k \tag{3}$$

$$r_m = \frac{\mathbf{h}_m \cdot \mathbf{q}}{\|\mathbf{h}_m\|\|\mathbf{q}\|} \tag{4}$$

The raw scores r_m are normalized into importance weights α_m via a Softmax function. Crucially, to ensure robustness against alignment errors, we employ a base-plus-bonus allocation strategy: we guarantee a minimum token count b_{min} for every frame to preserve temporal continuity, while distributing the remaining budget proportional to α_m. Given a total budget B_{total}, the specific budget allocated to segment S_m is defined as:

$$B(S_m) = |S_m| \cdot b_{min} + \alpha_m \left(B_{total} - T \cdot b_{min}\right) \tag{5}$$

This mechanism ensures that query-relevant moments receive high-resolution attention while redundant segments retain only essential structural sketches.

3.3 The Lightweight Semantic Scout

To efficiently bridge the semantic gap, we employ a BLIP-2 Q-Former (Li et al., 2023) as a parameter-efficient Semantic Scout. Unlike standard pooling, this module actively identifies regions pertinent to the user's intent. Specifically, we feed the user's textual instruction Q and a set of learnable query embeddings $\mathbf{Q}_l$ into the Q-Former. Through internal self-attention layers, $\mathbf{Q}_l$ interacts with Q, transforming generic queries into text-aware extractors tailored to the specific question. These conditioned

queries then attend to the input frame features $\mathbf{V}_t$ via cross-attention, generating the attention map $\mathcal{A} \in \mathbb{R}^{N_q \times N_h \times N_p}$, where N_q, N_h, N_p denote the length of query tokens, heads, and patches, respectively.

To derive a fine-grained importance score $s_{t,i}$ for the i-th patch, we aggregate $\mathcal{A}$ to capture the strongest alignment signals. Mirroring our implementation, we perform a max-pooling operation across attention heads to preserve peak confidence, followed by a summation across the query tokens. Guided by a semantic sampling ratio ρ, the patch importance and the foveal budget $K_{t,f}$ are formulated as:

$$s_{t,i} = \sum_{j=1}^{N_q} \max_{h=1}^{N_h} \mathcal{A}_{t,h,j,i} \tag{6}$$

$$K_{t,f} = \lfloor \rho \cdot B_t \rfloor \tag{7}$$

where B_t is the dynamic budget for frame t. We retain the top-$K_{t,f}$ tokens based on $s_{t,i}$ to form the foveal token set $\mathcal{V}_{t,f}$. This mechanism ensures that the model's computational focus is precisely directed by the textual query toward the most semantically relevant visual evidence.

3.4 Global Compensation and Aggregation

Exclusive reliance on the Semantic Scout may neglect structural context, creating a fragmented view of the scene. To prevent this, we allocate the remaining token budget $K_{t,r} = B_t - K_{t,f}$ to a structure-aware compensation mechanism. Instead of relying on a simplistic global average, we employ a Density Peak Clustering (DPC) strategy to identify representative anchors that capture the diverse geometric structures of the frame.

Let $\mathcal{C}_t$ be the set of candidate tokens not selected by the Scout. For each candidate $\mathbf{v}_{t,i} \in \mathcal{C}_t$, we first compute its local density ρ_i based on the average squared distance to its K-nearest neighbors (KNN). Subsequently, we calculate δ_i, the minimum distance to any token with higher density. Crucially, for the token with the highest global density, δ_i is defined as the maximum distance to any other token:

$$\rho_i = \exp\left(-\frac{1}{K} \sum_{k \in \text{KNN}(i)} \|\mathbf{v}_{t,i} - \mathbf{v}_{t,k}\|_2^2 \right) \tag{8}$$

$$\delta_i = \begin{cases} \min_{j:\rho_j > \rho_i} \|\mathbf{v}_{t,i} - \mathbf{v}_{t,j}\|_2 & \text{if } \exists j, \rho_j > \rho_i \\ \max_j \|\mathbf{v}_{t,i} - \mathbf{v}_{t,j}\|_2 & \text{otherwise} \end{cases} \tag{9}$$

We then define the structural importance score as $\gamma_i = \rho_i \cdot \delta_i$. This metric favors tokens that are both locally dense (cluster centers) and distant from other peaks (distinct structures). The peripheral set $\mathcal{V}_{t,r}$ is derived by selecting the top-$K_{t,r}$ candidates based on γ_i:

$$\mathcal{V}_{t,r} = \arg\operatorname*{topk}_{i \in \mathcal{C}_t}(\gamma_i, K_{t,r}) \tag{10}$$

The final preserved set is $\mathcal{S}_t = \mathcal{V}_{t,f} \cup \mathcal{V}_{t,r}$. To mitigate information loss, we apply a soft-pruning aggregation. We partition the discarded tokens $\mathcal{D}_t$ by assigning each $\mathbf{v}_{t,j}$ to its nearest spatial-semantic anchor in $\mathcal{S}_t$. The local cluster $\mathcal{N}_{t,i}$ for anchor $\mathbf{v}_{t,i}$ is:

$$\mathcal{N}_{t,i} = \left\{ \mathbf{v}_{t,j} \in \mathcal{D}_t \mid i = \arg\min_{k:\mathbf{v}_{t,k} \in \mathcal{S}_t} \|\mathbf{v}_{t,j} - \mathbf{v}_{t,k}\|_2 \right\} \tag{11}$$

We update the anchor via weighted fusion to absorb local texture details. The updated feature $\hat{\mathbf{v}}_{t,i}$ is computed as:

$$\hat{\mathbf{v}}_{t,i} = (1-\alpha)\mathbf{v}_{t,i} + \alpha \cdot \frac{1}{|\mathcal{N}_{t,i}|} \sum_{\mathbf{v}_{t,j} \in \mathcal{N}_{t,i}} \mathbf{v}_{t,j} \tag{12}$$

where α is a balancing hyperparameter. Finally, we concatenate the optimized tokens across all frames to form the global compressed sequence $\hat{X}$ for the LLM.

4 Experiments

4.1 Experimental Settings

Benchmarks. To assess diverse video understanding capabilities, we conduct comprehensive evaluations across four benchmarks. Video-MME (Fu et al., 2025a) serves as a large-scale evaluation suite for general-purpose analysis across broad domains. MLVU (Zhou et al., 2025) and LongVideoBench (Wu et al., 2024) assess long-context reasoning over extended video sequences requiring multi-hop comprehension. MVBench (Li et al., 2024) explicitly evaluates fine-grained temporal perception and dynamic action recognition.

Table 1: Comparison with state-of-the-art token reduction methods on LLaVA-OneVision.

Method	Prefill FLOPs Ratio	Retention Ratio	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
						Short	Med.	Long	Score	Score	%
LLaVA-OneVision	100%	100%	58.56	58.26	68.11	72.00	56.33	47.89	58.74	60.92	100.0
DyCoke	27.2%	26.6%	60.75	57.97	62.58	71.67	55.11	49.67	58.81	60.03	98.5
FrameFusion	25.5%	25%	58.61	58.61	65.39	**71.56**	57.33	**48.56**	59.15	60.30	99.0
VisionZip	25.6%	25%	**59.04**	57.37	67.05	70.78	55.33	47.22	57.78	60.31	99.0
FastVID	25.6%	25%	58.50	57.07	67.14	70.22	56.22	47.33	57.93	60.16	98.8
Vista-LLM	25.6%	25%	58.80	**58.71**	**68.11**	**71.56**	**57.67**	**48.56**	**59.26**	**61.22**	**100.5**
FrameFusion	20.5%	20%	58.77	57.67	64.65	69.56	56.00	48.00	57.85	59.74	98.1
VisionZip	20.6%	20%	58.50	56.99	66.64	**71.78**	**57.11**	47.22	**58.70**	60.21	98.8
FastVID	20.6%	20%	58.48	57.37	66.50	70.78	56.44	46.67	57.96	60.08	98.6
Vista-LLM	20.6%	20%	**58.93**	**58.64**	**67.79**	70.22	**57.11**	**48.22**	58.52	**60.97**	**100.1**
FrameFusion	15.6%	15%	58.34	55.57	63.13	68.33	53.56	46.33	56.07	58.28	95.7
VisionZip	15.6%	15%	57.99	55.35	65.07	68.89	55.56	48.22	57.56	58.99	96.8
FastVID	15.6%	15%	**58.90**	56.32	63.82	69.11	55.78	**49.33**	58.07	59.28	97.3
Vista-LLM	15.6%	15%	58.23	**57.74**	**67.70**	**69.89**	**56.78**	47.78	**58.15**	**60.45**	**99.2**
FrameFusion	10.6%	10%	57.29	53.55	59.59	66.78	51.67	44.89	54.44	56.22	92.3
VisionZip	10.6%	10%	55.12	51.16	61.98	62.44	54.44	47.89	54.93	55.80	91.6
FastVID	10.6%	10%	57.88	55.42	64.41	68.00	55.11	**47.89**	57.00	58.68	96.3
Vista-LLM	10.6%	10%	**57.96**	**56.62**	**67.24**	**68.44**	**56.22**	47.56	**57.41**	**59.81**	**98.2**

Compared Baselines. We compare our framework against state-of-the-art visual token compression methods. VisionZip (Yang et al., 2025) reduces spatial redundancy using attention distributions. DyCoke (Tao et al., 2025) merges redundant temporal tokens and progressively prunes the KV cache. FrameFusion (Fu et al., 2025b) integrates similarity-based merging with importance-based pruning. HoliTom (Shao et al., 2025a) performs holistic token reduction; however, we evaluate solely its input-level spatiotemporal merging module to ensure fair structural comparison. FastVID (Shen et al., 2025) employs dynamic density pruning to exploit spatiotemporal dependencies. We utilize the official implementations for all baselines and strictly align their computational budgets.

Implementation Details. We implement our proposed framework upon three representative backbones: LLaVA-Video-7B, LLaVA-OneVision-7B, and Qwen2.5-VL-7B. For the LLaVA models, we adhere to the standard configuration of sampling 64 frames, resulting in a fixed $N_v = 196$ tokens per frame. For Qwen2.5-VL, we maintain the 64-frame sampling protocol but accommodate its native dynamic resolution, which yields variable token lengths across frames. Regarding hyperparameters, the semantic coherence threshold τ is set to 0.80 for retention ratios of 20% and 25%, and 0.95 for ratios of 10% and 15%, while the hybrid sampling ratio ρ is fixed at 0.85. For the Semantic Scout, we utilize the pre-trained visual encoder and Q-Former from the BLIP-2 Image-Text Matching model, discarding the LLM decoder to minimize memory overhead. The Q-Former undergoes a rapid parameter-efficient adaptation to align with the binary selection objective. Specifically, training is conducted using the PyTorch framework with the AdamW optimizer, an initial learning rate of 5×10^{-5}, and a batch size of 16 utilizing gradient accumulation over 20 steps. For the alignment loss, the weight α is set to 2.0, the Top-K value K to 64, and the temperature T to 0.1. Further fine-tuning details are provided in Appendix A. All inference experiments and latency measurements are conducted on NVIDIA A40 GPUs.

4.2 Main Results

Performance on LLaVA-OneVision. The evaluation on LLaVA-OneVision-7B (Table 1) further validates our framework's generalization. Notably, Vista-LLM outperforms the full-token baseline at 20% and 25% retention ratios. This counter-intuitive gain indicates that our query-guided pruning acts as an effective denoising filter, discarding redundant background details that otherwise distract the LLM, thereby enhancing reasoning precision. Compared to FastVID, which suffers a distinct drop at the 10% budget, our method demonstrates superior resilience, sustaining 98.2% relative performance. This consistency confirms

Table 2: Comparison with state-of-the-art token reduction methods on LLaVA-Video. We also adapt our framework to the structure-free setting (Newline Mode: None) for a strictly fair comparison with baselines like FrameFusion; detailed results are provided in Appendix B.3.

Method	Newline Mode	Prefill FLOPs Ratio	Retention Ratio	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
							Short	Med.	Long	Score	Score	%
LLaVA-Video	Grid	100%	100%	62.02	60.66	71.43	76.11	62.22	54.11	64.15	64.56	100.0
DyCoke	None	25.4%	26.6%	60.75	57.97	62.58	71.67	55.11	49.67	58.81	60.03	93.0
FrameFusion	None	26.6%	25%	60.85	58.49	66.59	73.56	60.78	51.44	61.93	61.96	96.0
VisionZip	Grid	30.6%	25%	60.99	60.13	69.22	73.44	61.22	52.33	62.33	63.17	97.8
HoliTom	Grid	30.2%	25%	**61.50**	60.03	68.94	74.19	61.62	**53.62**	**63.14**	63.40	98.2
Vista-LLM	Grid	30.3%	25%	**61.50**	**60.96**	**70.92**	**74.89**	**61.67**	51.56	62.70	**64.02**	**99.2**
FrameFusion	None	20.5%	20%	60.42	57.44	65.07	72.11	59.11	50.11	60.44	60.84	94.2
VisionZip	Grid	25.6%	20%	60.99	60.13	69.22	73.22	60.89	**52.33**	62.15	63.12	97.8
HoliTom	Grid	25.0%	20%	**61.39**	59.81	69.39	73.75	**62.40**	52.28	**62.81**	**63.35**	**98.1**
Vista-LLM	Grid	25.2%	20%	61.02	**61.02**	**70.37**	**74.00**	60.22	50.78	61.67	63.33	98.1
FrameFusion	None	15.5%	15%	60.18	57.07	63.13	70.41	58.67	49.56	59.54	59.98	92.9
VisionZip	Grid	21.1%	15%	60.37	56.47	67.19	71.00	**60.33**	52.22	61.19	61.30	95.0
HoliTom	Grid	20.8%	15%	60.38	59.66	67.64	72.75	60.18	**52.28**	**61.74**	62.35	96.6
Vista-LLM	Grid	20.4%	15%	**60.83**	**60.36**	**69.26**	72.67	59.67	50.78	61.04	**62.87**	**97.4**
FrameFusion	None	10.7%	10%	58.80	54.67	60.51	65.33	54.56	47.00	55.63	57.40	88.9
VisionZip	Grid	16.4%	10%	58.02	53.93	63.92	66.22	56.22	50.11	57.52	58.34	90.4
HoliTom	Grid	15.1%	10%	**60.49**	57.86	65.74	70.75	**58.95**	**51.50**	60.40	61.12	94.7
Vista-LLM	Grid	15.1%	10%	59.75	**59.16**	**67.37**	**73.56**	58.00	50.78	**60.78**	**61.76**	**95.7**

Table 3: Comparison with state-of-the-art token reduction methods on Qwen2.5-VL.

Method	Prefill FLOPs Ratio	Retention Ratio	MLVU	MVBench
Qwen2.5-VL	100%	100%	64.82	68.51
FastVID	21.6%	25%	62.66	65.13
Vista-LLM	21.5%	25%	**63.10**	**66.67**
FastVID	13.2%	15%	62.03	64.96
Vista-LLM	13.0%	15%	**62.76**	**66.32**

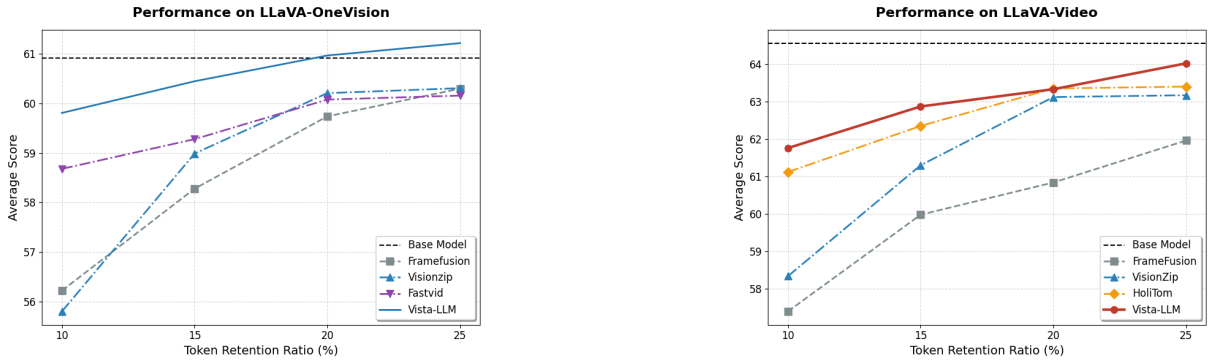
our ability to isolate discriminative visual evidence across all compression levels, offering a robust solution for resource-constrained scenarios.

Performance on LLaVA-Video. Table 2 presents the comparative analysis on the LLaVA-Video-7B backbone. In the grid-enhanced category, Vista-LLM demonstrates superior resilience compared to VisionZip and HoliTom. While HoliTom remains competitive at the 20% retention ratio, our approach exhibits greater stability as compression intensifies. Specifically, at the challenging 15% and 10% levels, Vista-LLM consistently surpasses HoliTom (e.g., maintaining 97.4% vs. 96.6% relative performance at 15%), highlighting the robustness of query-guided selection against the rapid degradation observed in visual-centric baselines.

Performance on Qwen2.5-VL. To demonstrate that our framework is entirely decoupled from specific tokenization paradigms, we evaluate Vista-LLM on the Qwen2.5-VL architecture, which inherently produces variable token lengths per frame. As detailed in Table 3, Vista-LLM consistently outperforms the FastVID baseline at both the 15% and 25% retention levels. Notably, at the challenging 15% setting, our method achieves a 0.73 score improvement on MLVU and a 1.36 improvement on MVBench. This resilience confirms that our Query-Guided Dynamic Budgeting and Semantic Scout modules effectively generalize to diverse architectures, successfully isolating query-relevant visual information completely independent of the underlying visual encoding strategy.

Table 4: **Generation Efficiency vs. Performance.** Generate Time denotes the end-to-end inference latency, encompassing both visual token processing and text decoding phases. Peak Memory reflects the maximum GPU memory allocated during the entire forward pass.

Retention Ratio	Method	Peak Memory	Generate Time	Avg. Score
100%	LLaVA-Video	29.94 GB	4.07s (1.0×)	64.56
25%	VisionZip	24.16 GB	2.61s (1.56×)	63.17
	HoliTom	52.68 GB	2.75s (1.48×)	63.40
	Vista-LLM	**28.42 GB**	**2.27s (1.79×)**	**64.02**
15%	VisionZip	24.15 GB	2.44s (1.67×)	61.30
	HoliTom	52.27 GB	2.47s (1.65×)	62.35
	Vista-LLM	**28.39 GB**	**2.02s (2.01×)**	**62.87**



(a) **LLaVA-OneVision-7B** (b) **LLaVA-Video-7B**

Figure 2: **Performance consistency across varying token retention ratios.** Vista-LLM (solid lines) establishes a superior trade-off compared to state-of-the-art baselines. **(a)** On LLaVA-OneVision, our method achieves higher accuracy than the full-token baseline at 20% and 25% retention ratios. **(b)** On LLaVA-Video, the framework exhibits exceptional stability at the challenging 10% retention level, significantly outperforming visual-centric approaches.

Efficiency Evaluation. We evaluate computational efficiency through prefilling FLOPs, generation latency, and peak memory allocation. As evidenced in Table 1, our method retains 98.2% of LLaVA-OneVision performance using only 10.6% of the original FLOPs. Regarding wall-clock inference on LLaVA-Video in Table 4, Vista-LLM demonstrates superior throughput, achieving a 2.01× speedup at a 15% retention ratio. Furthermore, our decoupled architecture strictly preserves FlashAttention compatibility for the primary LLM decoding stage and avoids LLM in-network attention materialization. This prevents the severe memory overhead observed in baselines like HoliTom, where explicit dense attention computation causes peak memory to exceed 52 GB. Vista-LLM reduces memory consumption below the uncompressed baseline to approximately 28.4 GB. Although VisionZip exhibits a marginally lower memory footprint, it suffers from significantly slower inference and degraded accuracy, confirming that our method achieves an optimal system-level equilibrium.

This comprehensive efficiency stems from the optimized design of our three selection components. First, Dynamic Budgeting relies on simple pooling operations that incur negligible latency. Second, we implement a parallel execution strategy for the Semantic Scout to maintain strict hardware efficiency. Both the primary Video-LLM and the BLIP-2 Vision Encoder remain fully compatible with FlashAttention and avoid attention matrix materialization. While the Q-Former structurally requires materializing attention matrices to compute cross-modal importance scores, this mathematical requirement does not introduce wall-clock latency overhead. We launch the lightweight Scout asynchronously alongside the primary visual backbone. Because the computational cost of the Q-Former is minimal relative to the dense Vision Encoder, the Q-Former completes its execution entirely within the primary feature extraction window. Consequently, the attention materialization cost is effectively masked, ensuring the large language model receives a pre-filtered sequence without computational delays. Finally, the Structure-Aware Compensation module remains lightweight as it operates solely on the residual candidate set and targets a limited token budget. This minimal selection overhead allows the inference pipeline to fully benefit from the reduced token count during the LLM decoding stage. Crucially, since our selection cost scales linearly with frame count while the LLM self-attention scales quadratically, the efficiency gains of Vista-LLM become increasingly significant as video duration extends.

4.3 Ablation Studies

To validate the individual contribution of each component within our framework, we conduct comprehensive ablation studies across the full suite of benchmarks, including VideoMME, MVBench, LongVideoBench, and MLVU. Unless otherwise stated, all ablation experiments are performed using the LLaVA-Video-7B backbone with a 25% retention ratio. Regarding hyperparameters, we fix the semantic coherence threshold τ at 0.95 and the hybrid sampling ratio ρ at 0.85. Notably, the specific evaluation for Dynamic Budgeting is conducted on the LLaVA-OneVision-7B backbone.

Table 5: **Ablation Studies on Component Efficacy.** Systematic evaluation of each module within the Vista-LLM framework. **Bold** denotes the best performance in each block.

Method	MVBench	LongVideo Bench	MLVU	VideoMME
(a) Impact of Semantic Scout				
Intrinsic Attn.	61.03	59.73	69.67	62.11
BLIP-2 (Ours)	**61.12**	**60.66**	**70.41**	**63.15**
(b) Dynamic Budgeting (LLaVA-OneVision)				
Static Alloc.	**58.77**	56.92	67.33	**59.48**
Dynamic (Ours)	58.72	**58.64**	**68.11**	58.85
(c) Context Compensation Strategy				
Random Supp.	60.45	60.21	70.41	**63.37**
Rep. Supp.	**61.15**	60.51	70.41	62.70
DPC-KNN (Ours)	61.12	**60.66**	**70.41**	63.15
(d) Feature Aggregation				
Hard Pruning	60.67	59.31	70.05	62.19
Soft Agg. (Ours)	**61.12**	**60.66**	**70.41**	**63.15**
(e) Task-Specific Adaptation				
Frozen Q-Former	60.67	59.84	68.39	61.89
Fine-tuned (Ours)	**61.12**	**60.66**	**70.41**	**63.15**

Impact of the Semantic Scout. Block (a) compares our query-guided design against an Intrinsic Attention baseline, a task-agnostic approach that ranks visual tokens based solely on inherent visual saliency derived from the self-attention weights of the frozen vision encoder. Our method achieves a consistent improvement across all benchmarks, raising the average relative performance from 97.8% to 98.9%. This verifies that pure visual saliency is insufficient for complex reasoning, whereas our query-conditional guidance effectively aligns token selection with specific semantic requirements.

Efficacy of Dynamic Budgeting. Block (b) examines temporal resource allocation using the LLaVA-OneVision backbone. While Static Allocation remains competitive on short-term tasks, the Dynamic strategy yields substantial gains on long-context benchmarks (e.g., +1.72 on LongVideoBench). Consequently, the overall average relative accuracy improves from 99.5% to 100.3%, confirming that prioritizing information-dense segments captures narrative evolution more effectively.

Strategy for Context Compensation. Block (c) evaluates peripheral token selection. We compare our approach against Representation Supplement, which prioritizes candidate tokens exhibiting the highest similarity to the global frame average. While this strategy achieves 98.6% relative accuracy by approximating the statistical center, our DPC-KNN approach attains 98.9%. This confirms that density-based clustering, by identifying distinct structural anchors, provides a more informative context than simply retaining tokens closest to the mean.

Impact of Feature Aggregation. Block (d) assesses the benefit of fusing unselected tokens. While Hard Pruning yields 97.7% relative accuracy, our Soft Aggregation strategy recovers performance to 98.9%, indicating that aggregating unselected regions into anchors effectively mitigates information loss.

Efficacy of Task-Specific Adaptation. Block (e) compares our fine-tuned module against the frozen Q-Former. The results show that fine-tuning improves relative accuracy from 97.1% to 98.9%, confirming that task-specific alignment is necessary to accurately locate query-relevant regions.

5 Conclusion

We proposed Vista-LLM, a framework that efficiently prunes visual tokens in long videos via Query-Guided Dynamic Budgeting and Structure-Aware Compensation. Unlike prior arts, Vista-LLM ensures compatibility with standard attention kernels while significantly reducing redundancy. Experiments demonstrate a state-of-the-art trade-off: on LLaVA-OneVision, we achieve a $\sim$90% reduction in computation with 98.2% performance retention, and a 2.01$\times$ speedup on LLaVA-Video. These results highlight Vista-LLM as a practical solution for deploying long-context MLLMs under strict resource constraints.

Limitations

The main constraint of Vista-LLM stems from the representational gap between the lightweight Semantic Scout and the LLM. Although the Scout is effective at filtering redundancy, its smaller scale implies that extremely subtle cues could be missed, leading to information loss before the reasoning stage. This limitation is particularly relevant for tasks requiring high-precision visual grounding. Future work could address this by exploring iterative scouting mechanisms or distilling stronger visual priors into the Scout to bridge this capacity gap.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pages 19–35. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Mingjing Du, Shifei Ding, and Hongjie Jia. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. Based Syst.*, 99:135–145.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025a. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. 2025b. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22654–22663.

Jiani Guo, Zuchao Li, Jie Wu, Qianren Wang, Yun Li, Lefei Zhang, Hai Zhao, and Yu-Jiu Yang. 2025. Tom: Leveraging tree-oriented mapreduce for long-context reasoning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 17793–17812. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Tingyu Qu, Mingxiao Li, Tinne Tuytelaars, and Marie-Francine Moens. 2024. Ts-llava: Constructing visual tokens through thumbnail-and-sampling for training-free video large language models. *CoRR*, abs/2411.11066.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025a. Holitom: Holistic token merging for fast video large language models. In *NeurIPS*.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. 2025b. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *CoRR*, abs/2507.20198.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, pengzhang liu, Sicheng Zhao, and Guiguang Ding. 2025. FastVID: Dynamic density pruning for fast video large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Luohe Shi, Hongyi Zhang, Yao Yao, Zuchao Li, and Hai Zhao. 2024. Keep the cost down: A review on methods to optimize llm' s kv-cache consumption. *CoRR*, abs/2407.18003.

Gary J. Sullivan and Thomas Wiegand. 2005. Video compression - from concepts to the H.264/AVC standard. *Proc. IEEE*, 93(1):18–31.

Zicong Tang, Ziyang Ma, Suqing Wang, Zuchao Li, Lefei Zhang, Hai Zhao, Yun Li, and Qianren Wang. 2025a. Covipal: Layer-wise contextualized visual token pruning for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20701–20714. Association for Computational Linguistics.

Zicong Tang, Luohe Shi, Zuchao Li, Baoyuan Qi, Guoming Liu, Lefei Zhang, and Ping Wang. 2025b. SpindleKV: A novel KV cache reduction method balancing both shallow and deep layers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycoke: Dynamic compression of tokens for fast video large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 18992–19001. Computer Vision Foundation / IEEE.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *CoRR*, abs/2405.09818.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. Visionzip: Longer is better but not necessary in vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19792–19802. Computer Vision Foundation / IEEE.

Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip H. S. Torr, Wayne Zhang, and Dahua Lin. 2021. Vision transformer with progressive sampling. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 377–386. IEEE.

Hang Zhang, Xin Li, and Lidong Bing. 2023. 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*, pages 543–553, Singapore. Association for Computational Linguistics.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025a. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025b. LLaVA-Video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025.

Yi Zhao, Yajuan Peng, Cam-Tu Nguyen, Zuchao Li, Xiaoliang Wang, Hai Zhao, and Xiaoming Fu. 2025. Smallkv: Small model assisted compensation of KV cache compression for efficient LLM inference. *CoRR*, abs/2508.02751.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13691–13701. Computer Vision Foundation / IEEE.

A Fine-tuning Details for Key Token Extraction

A.1 Dataset and Sample Construction

We utilize the region descriptions from the Visual Genome dataset for fine-tuning. To ensure semantic density, we exclude phrases containing fewer than three words, thereby removing overly generic descriptions. Since the raw dataset consists solely of positive pairs, we implement an online hard negative mining strategy to enhance the discriminative capability of the model. Specifically, we compute similarities between all images and texts within a batch and sample negative counterparts based on a multinomial distribution derived from these scores. This process prioritizes hard negatives that are semantically similar but factually incorrect. These dynamically mined negative samples are then combined with the positive pairs to construct the final training batch for the Image-Text Matching objective.

A.2 Model Architecture and Adaptations

Our method builds upon the pre-trained BLIP-2 architecture. To efficiently leverage its powerful pre-trained visual representations and reduce computational demands, we freeze the parameters of the Vision Encoder. The core components that undergo fine-tuning are the Q-Former and a newly introduced lightweight projection module. We introduce a Key Token Projector to bridge the raw visual patch embeddings and the text embedding space. This module is implemented as a sequential network consisting of Linear, LayerNorm, GELU, Dropout, and a final Linear layer. Its purpose is to project the weighted visual features from the vision dimension of 1408 into the Image-Text Contrastive embedding space of dimension 256. This projection facilitates direct alignment between the most relevant visual features and the text query tokens.

A.3 Training Objectives

The overall training objective is a weighted sum of three distinct loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{ITM} + \mathcal{L}_{ITC} + \alpha\mathcal{L}_{Align} \tag{13}$$

While our primary contribution is the alignment loss, the retention of the standard Image-Text Matching ($\mathcal{L}_{ITM}$) and Image-Text Contrastive ($\mathcal{L}_{ITC}$) losses is critical. These objectives act as essential regularizers that maintain the foundational global semantic alignment between modalities. Without these constraints, optimizing solely for local token alignment would likely lead the model to collapse into trivial solutions or overfit to local noise, effectively losing its ability to perform meaningful multimodal reasoning.

To explicitly guide the model towards focusing on key visual regions, we introduce the Soft Top-K Alignment Loss. Let $\mathbf{S}$ denote the patch importance score vector derived by averaging the cross-attention maps from the Q-Former. Instead of a hard selection, we employ a soft relaxation for the Top-K operation. We first determine a dynamic threshold value τ which corresponds to the K-th largest score in $\mathbf{S}$. A soft mask $\mathbf{m}$ is then generated using a sigmoid function with a temperature parameter T:

$$m_i = \sigma\left(\frac{S_i - \tau}{T}\right) \tag{14}$$

This soft mask is applied to the raw image patch embeddings. The key visual feature vector $\mathbf{v}_{key}$ is obtained by aggregating the image embeddings weighted by the normalized masked scores. Finally, we minimize the cosine distance between the projected key visual feature via the projector Φ and the corresponding text embedding $\mathbf{t}_{feat}$:

$$\mathcal{L}_{Align} = 1 - \cos(\Phi(\mathbf{v}_{key}), \mathbf{t}_{feat}) \tag{15}$$

A.4 Training Dynamics

We monitor the training progression by tracking the total loss and its individual components: $\mathcal{L}_{ITM}$, $\mathcal{L}_{ITC}$, and $\mathcal{L}_{Align}$. As illustrated in Figure 3, the model exhibits rapid convergence. Specifically, the Alignment Loss drops significantly within the initial training phase, indicating that the lightweight Key Token Projector efficiently learns to map visual features to the textual space without requiring extensive training epochs. The ITM and ITC losses maintain stable fluctuations, confirming their role in regularizing the model and preventing catastrophic forgetting of the pre-trained knowledge. Empirical evaluation on the validation set identifies the optimal checkpoint at step 625, highlighting the efficiency of our fine-tuning strategy in adapting the pre-trained BLIP-2 to the key token extraction task.

B Supplemental Experimental Results

B.1 Ablation Study on Sampling Ratio (ρ).

Table 6 examines the trade-off between semantic focus and structural preservation using the LLaVA-

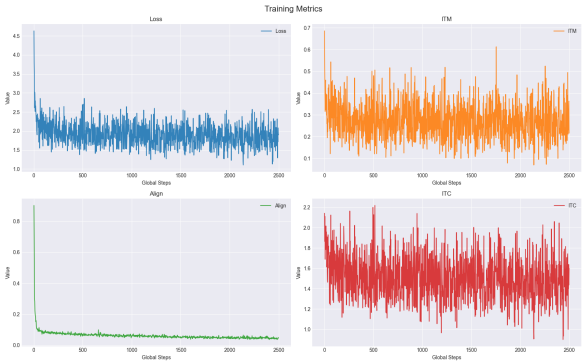


Figure 3: Training curves showing the evolution of Total Loss, ITM Loss, Alignment Loss, and ITC Loss. The distinct drop in Alignment Loss demonstrates rapid adaptation, while ITM and ITC losses provide consistent regularization.

Table 6: **Ablation Study on Sampling Ratio (ρ).** This parameter regulates the budget allocation between semantic-aware selection and structure-aware compensation.

Ratio (ρ)	MVBench	LongVideo Bench	MLVU	VideoMME	Avg. Acc.
0.0	60.48	60.36	69.40	62.56	97.9
0.2	60.29	60.43	70.00	62.70	98.1
0.4	60.75	**61.11**	70.18	62.40	98.5
0.6	**61.50**	60.96	70.37	62.33	98.8
0.8	61.02	60.21	**70.46**	62.67	98.5
0.85	61.12	60.66	70.41	**63.15**	**98.9**
1.0	60.80	60.06	70.32	62.70	98.3

Video backbone (25% retention, $\tau = 0.95$). At the lower bound where $\rho = 0$, the framework relies entirely on Structure-Aware Compensation, completely disregarding the textual instruction. This configuration yields the lowest performance, confirming that task-agnostic selection is insufficient for complex reasoning. Conversely, setting $\rho = 1$ allocates the full budget to the Semantic Scout. The subsequent decline in accuracy indicates that the total exclusion of background context leads to semantic isolation, hindering the model from grounding objects within the global scene. The peak performance at $\rho = 0.85$ validates that retaining a small proportion of structural anchors to complement high-relevance tokens ensures the most robust representation.

B.2 Ablation Study on Semantic Coherence Threshold (τ).

Table 7 investigates the impact of the Semantic Coherence Threshold τ under the LLaVA-Video setting (25% retention, $\rho = 0.85$). The empirical results exhibit a distinct performance peak at

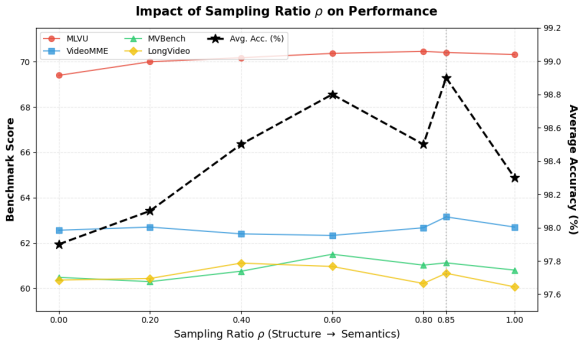


Figure 4: The dashed vertical line indicates the optimal configuration at 0.85, outperforming both the structure-only ($\rho = 0$) and semantics-only ($\rho = 1$) baselines.

$\tau = 0.80$, yielding an average relative accuracy of 99.2%. Deviating from this optimum leads to performance degradation: lower thresholds ($\tau < 0.80$) risk obscuring rapid visual changes by merging distinct events, whereas higher thresholds ($\tau > 0.80$) cause over-segmentation that fragments narrative continuity. Consequently, we adopt $\tau = 0.80$ as the default configuration to ensure a balanced temporal partition.

Table 7: **Ablation Study on Semantic Coherence Threshold (τ).** Validating that $\tau = 0.80$ strikes the optimal balance for dynamic segment generation.

Threshold (τ)	MVBench	LongVideo Bench	MLVU	VideoMME	Avg. Acc.
0.65	61.37	**61.26**	70.65	62.48	99.0
0.70	61.21	**61.26**	70.74	62.52	99.0
0.75	60.99	60.81	70.83	62.85	98.9
0.80	**61.50**	60.96	**70.92**	62.70	**99.2**
0.85	61.39	61.03	70.28	62.22	98.7
0.90	61.21	60.81	70.46	62.44	98.7
0.95	61.12	60.66	70.41	**63.15**	98.9

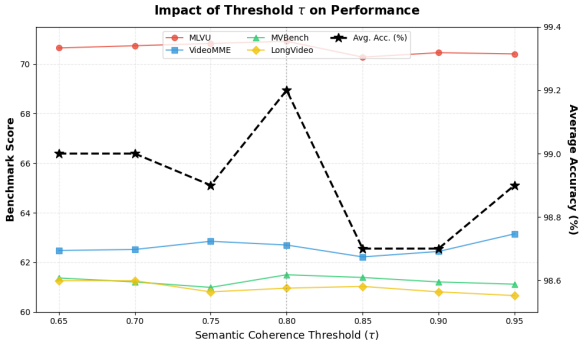


Figure 5: Performance trend across different Semantic Coherence Thresholds (τ). The curve indicates that $\tau = 0.80$ provides the most robust feature segmentation across diverse benchmarks.

Table 8: Comparison of methods using the None Newline Mode on LLaVA-Video. This setting strictly excludes newline tokens, allowing for a direct comparison of token selection capabilities between Vista-LLM* and structure-free baselines.

Method	Newline Mode	FLOPs Ratio	Retention Ratio	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
							Short	Med.	Long	Score	Score	%
DyCoke	None	25.4%	26.6%	60.75	57.97	62.58	71.67	55.11	49.67	58.81	60.03	93.0
FrameFusion	None	26.6%	25%	60.85	58.49	66.59	**73.56**	**60.78**	**51.44**	**61.93**	61.96	96.0
Vista-LLM*	None	24.0%	25%	**61.39**	**60.06**	**68.16**	73.00	60.33	50.67	61.33	**62.74**	**97.2**
FrameFusion	None	20.5%	20%	60.42	57.44	65.07	**72.11**	**59.11**	50.11	60.44	60.84	94.2
Vista-LLM*	None	19.2%	20%	**60.58**	**59.09**	**67.97**	72.00	58.78	**51.00**	**60.59**	**62.06**	**96.1**
FrameFusion	None	15.5%	15%	**60.18**	57.07	63.13	70.41	**58.67**	49.56	**59.54**	59.98	92.9
Vista-LLM*	None	14.6%	15%	59.99	**58.86**	**66.31**	**70.67**	57.44	**49.89**	59.33	**61.12**	**94.7**
FrameFusion	None	10.7%	10%	58.80	54.67	60.51	65.33	54.56	47.00	55.63	57.40	88.9
Vista-LLM*	None	10.1%	10%	**58.88**	**58.86**	**64.98**	**69.22**	**56.22**	**50.00**	**58.48**	**60.30**	**93.4**

B.3 Evaluation on Structure-Free Settings

In the main paper, we demonstrated the performance of Vista-LLM using the standard grid structure of LLaVA-Video. However, some baselines such as DyCoke and FrameFusion operate in a structure-free manner, discarding the newline tokens typically used to represent spatial layouts. To ensure a strictly fair comparison, we introduce a variant, denoted as **Vista-LLM***, which removes newline tokens to match the structural assumptions of these baselines.

Performance Analysis. Table 8 presents the comparison between Vista-LLM* and structure-free baselines. Compared to DyCoke at the 25% retention level, Vista-LLM* achieves a significantly higher average score (62.74 vs. 60.03) with a lower FLOPs ratio. Notably, when compared to FrameFusion, our method demonstrates superior robustness as the compression rate increases. At the challenging 10% retention ratio, Vista-LLM* outperforms FrameFusion by 2.9 points (60.30 vs. 57.40) and maintains a 93.4% relative performance to the full model, whereas FrameFusion drops to 88.9%.

B.4 Scalability with Video Length

To evaluate the scalability of the proposed framework across varying video durations, we conduct a stratified performance analysis utilizing a combined dataset comprising VideoMME, MLVU, and LongVideoBench. This aggregation yields a total of 6207 evaluation samples. Table 9 delineates the performance variations when integrating the proposed framework into two distinct backbone models at a 25% token retention ratio. The numerical values presented in the table represent the relative accuracy changes compared to the uncompressed full token baselines of each respective model. Positive percentages indicate performance improvements achieved by our method, whereas negative percentages reflect accuracy degradation compared to the original architectures.

Table 9: Relative performance changes of the proposed framework compared to the uncompressed baselines across different video durations at a 25% token retention ratio.

Duration	Count	LLaVA-OneVision	LLaVA-Video
< 3m	1261	-0.39%	-1.90%
3 - 10m	2696	+1.83%	+0.10%
10 - 20m	938	-2.03%	-1.49%
20 - 30m	274	-1.46%	+3.28%
> 30m	1038	+0.58%	-1.93%
Overall	6207	+0.21%	-1.20%

The empirical results indicate that the framework demonstrates consistent efficacy in medium length videos ranging from 3 to 10 minutes, which constitute the majority of the evaluation samples. Within this duration bracket, the LLaVA-OneVision backbone achieves an accuracy improvement of 1.83 percent over its full token counterpart. This observation supports the hypothesis that medium length videos frequently contain substantial temporal redundancy, including static backgrounds and repetitive scenes. The dynamic budgeting module effectively filters this redundancy without compromising semantic integrity. Conversely, a marginal performance decline is observed in short videos under 3 minutes. These brief clips typically possess high information density and minimal visual repetition, rendering them less amenable to aggressive

compression. However, the substantial efficiency gains achieved in longer contexts justify this minor degradation for long context applications.

For videos exceeding 10 minutes, the evaluation exhibits increased performance variance, with divergent behaviors between the two backbone models. Most notably, the LLaVA-Video model experiences a 1.93 percent performance decrease in videos longer than 30 minutes. This volatility can be attributed to the inherent characteristics of standard video inference protocols. Conventional temporal sampling strategies extract frames sparsely across the entire video sequence to meet context limits. Consequently, the semantic weight carried by each individual token becomes exceptionally high. Removing 75 percent of the visual tokens from an already sparse information pool naturally introduces stochasticity. Discarding any discriminative visual feature in such sparse settings can significantly impact the reasoning accuracy of the model. Despite this inherent instability, the ability of the framework to achieve positive gains in several extended video brackets validates its potential to mitigate visual noise in long context understanding.

C More Visualizations

Question: Which object was taken by the person?
Answer: The blanket.

Figure 6: Visualization of the token selection strategy on a sample video clip. The red patches indicate high-relevance regions identified by the Semantic Scout, which aligns closely with the textual query regarding the person and the blanket. The green patches denote structural anchors retained by the Structure-Aware Compensation module to preserve global context. This hybrid approach effectively captures the core interaction while maintaining scene coherence.