Sharp Eyes and Memory for VideoLLMs: Information-Aware Visual Token Pruning for Efficient and Reliable VideoLLM Reasoning

Jialong Qin[1,2,3], **Xin Zou**[1,2], **Di Lu**[1], **Yibo Yan**[1,2], **Xuming Hu**[1,2*]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]Beijing Institute of Technology
qinjialong5@gmail.com, xuminghu97@gmail.com

Abstract

Current Video Large Language Models (VideoLLMs) suffer from quadratic computational complexity and key-value cache scaling, due to their reliance on processing excessive redundant visual tokens. To address this problem, we propose SharpV, a minimalist and efficient method for adaptive pruning of visual tokens and KV cache. Different from most uniform compression approaches, SharpV dynamically adjusts pruning ratios based on spatial-temporal information. Remarkably, this adaptive mechanism occasionally achieves performance gains over dense models, offering a novel paradigm for adaptive pruning. During the KV cache pruning stage, based on observations of visual information degradation, SharpV prunes degraded visual features via a self-calibration manner, guided by similarity to original visual features. In this way, SharpV achieves hierarchical cache pruning from the perspective of information bottleneck, offering a new insight into VideoLLMs' information flow. Experiments on multiple public benchmarks demonstrate the superiority of SharpV. Moreover, to the best of our knowledge, SharpV is notably the first two-stage pruning framework that operates without requiring access to exposed attention scores, ensuring full compatibility with hardware acceleration techniques like Flash Attention.

Introduction

VideoLLMs have demonstrated remarkable capabilities in video understanding and reasoning tasks (Zhang et al. 2024b). However, the extensive spatiotemporal information in long videos consumes a significant portion of LLM input tokens (Li et al. 2024a; Xu et al. 2024a), leading to excessive computational overhead and memory usage (Xu et al. 2024b). As shown in Table 1, existing pre-LLM stage approaches have explored visual token pruning at various granularity levels to reduce token count and memory consumption. However, these methods (Chen et al. 2024; Lin et al. 2025; Hyun et al. 2025; Tao et al. 2025; Fu et al. 2024) uniformly apply fixed pruning ratios, which typically represent local optima tuned for specific datasets, lacking adaptive pruning based on video information content. This limitation affects their generalization and robustness. Moreover, since the primary objective of visual pruning is computational efficiency, the pruning process itself should maintain minimal complexity to avoid additional overhead. Existing approaches that rely on computationally intensive clustering algorithms (Huang, Zhou, and Han 2024) or mathematically sophisticated planning techniques (Sun et al. 2025; Alvar et al. 2025) may inadvertently introduce non-trivial overhead to the pruning process shown in Table 1, potentially limiting their practical applicability in real-world deployment scenarios where computational efficiency is paramount. To address these challenges, in pre-LLM stage, we propose a novel adaptive pruning method (Visual SharpV) that computes spatiotemporal importance for each token and determines frame-specific pruning ratios based on information volume, achieving fine-grained information preservation while maintaining low complexity.

In the intra-LLM stage, as demonstrated in Table 1, most existing works reduce cache by analyzing exposed attention scores. For instance, (Chen et al. 2024; Lin et al. 2025) retain visual tokens with high attention scores in specific layers based on inefficient visual attention phenomena, while (Huang, Zhou, and Han 2024) preserves tokens with high cross-attention scores. Although attention-based anal-

Methods *Pre-LLM Stage*	**Granularity Level**	**CLS Independent**	**Adaptive Ratio**	**Time Complexity**
DivPrune (CVPR25)	**Image**	✓	✗	$O(n^2 \cdot d + n^3)$
PruMerage (ICCV25)	**Image**	✗	✗	$O(n \cdot d)$
PruneVid* (ACL25)	**Segment**	✓	✗	$O(n^2 \cdot d + n \cdot d)$
DyCoke* (CVPR25)	**Frame**	✓	✗	$O(n \cdot d)$
VidCom^2 (2025.05)	**Frame**	✓	✗	$O(n \cdot d)$
SharpV* (ours)	**Frame**	✓	✓	$O(n \cdot d)$
Methods *Intra-LLM Stage*	**Flash Attention**	**Adaptive Layer**	**Cache Reduction**	**Space Complexity**
FastV (ECCV24)	✗	✗	✓	$O(n^2 \cdot d)$
VTW (CVPR25)	✗	✓	✓	$O(n^2 \cdot d)$
PruneVid† (ACL25)	✗	✗	✓	$O(n^2 \cdot d)$
DyCoke† (CVPR25)	✗	✗	✓	$O(n^2 \cdot d)$
FrameFusion (ICCV25)	✗	✗	✓	$O(n^2 \cdot d)$
SharpV† (ours)	✓	✓	✓	$O(n \cdot d)$

Table 1: Why SharpV Outperforms: A Systematic Comparison, where * denotes *Pre-LLM* pruning, † indicates *Intra-LLM* pruning.

*Corresponding Author

ysis provides reasonable motivation and valuable insights into attention mechanism dynamics, the emergence of efficient attention computation methods like flash attention (Dao et al. 2022; Dao 2023), which reduces the quadratic complexity of native attention to linear complexity while not exposing attention scores, renders such approaches incompatible. Therefore, we propose an information-theoretic perspective for intra-LLM stage pruning by measuring similarity between decoder layer inputs and original visual features to dynamically retain visual tokens for subsequent generation steps. Our method reduces complexity from $O(n^2 \cdot d)$ to $O(n \cdot d)$, fulfilling the lightweight design principle and the core objective of pruning.

We integrate SharpV into two widely adopted VideoLLMs with varying capabilities: PLLaVA (Xu et al. 2024a), LLaVA-OneVision (Li et al. 2024a). Extensive evaluations are conducted on four public datasets covering various video types (long, medium, short), including MVBench (Li et al. 2024b), VideoMME (Fu et al. 2025), NExTQA(Xiao et al. 2021), ActNet-QA(Yu et al. 2019), with comparisons against recent SOTA methods and baselines. Our experiments demonstrate the effectiveness of information-aware adaptive pruning ratios: at similar or lower retention ratios, SharpV achieves higher accuracy than SOTA methods while delivering more significant speedups due to its extremely low computational overhead.

Our main contributions are threefold:

- For the pre-LLM stage, we propose information-aware visual token pruning that maximize useful information retention at both frame and token levels, achieving video-specific adaptive pruning while maintaining lightweight $O(n \cdot d)$ complexity to accelerate VideoLLM inference.
- Departing from conventional attention-based analysis, we introduce an information-theoretic perspective for intra-LLM stage pruning. From the perspective of mutual information, our method is the first framework totally compatible with flash attention, reducing memory overhead, and offering new insights into LLM information flow.
- Comprehensive experiments on public datasets demonstrate SharpV's consistent performance and efficiency advantages. Our method consistently achieves a competitive trade-off between simplicity, efficiency, and accuracy compared to other SOTA approaches.

Related Work

The Intrinsic Redundancy in Videos

Videos inherently exhibit spatio-temporal locality, which has been exploited in deep learning approaches. Early works like 3D CNNs (Carreira and Zisserman 2017; Tran et al. 2015) and VideoTransformers (Tong et al. 2022; Feichtenhofer et al. 2022) achieved promising results in feature learning by leveraging this redundancy. Similarly, in multimodal LLM inference scenarios, visual token compression methods have enabled efficient model reasoning. Some works employ various strategies to identify spatially important tokens, typically by filtering tokens based on features from visual encoders (Rao et al. 2021; Vasu et al. 2023; Bolya et al. 2022; Kong et al. 2022). For instance, PruneMerge (Shang et al. 2024) selects crucial tokens based on inherent attention characteristics (Radford et al. 2021) between [CLS] token and other visual token. Other works rely on similarity-based methods to identify important tokens. DivPrune (Alvar et al. 2025) reduces spatial redundancy by solving for the maximally diverse set that maintains minimum pairwise distances. Cross-modal interaction methods like SimIgnore (Zhang et al. 2025) filter instruction-relevant visual tokens by comparing their similarity with prompt tokens. However, these approaches only consider spatial redundancy while neglecting temporal redundancy in videos.

For spatio-temporal video compression, LLaVA-Scissor (Sun et al. 2025) preserves semantic information through two-stage connectivity analysis, but suffers from high computational costs due to graph construction and Union-Find algorithms. PruneVid (Huang, Zhou, and Han 2024) focuses on compressing static backgrounds but requires video segmentation via clustering, which is computationally intensive and struggles with dynamic parameter settings. VidCom^2 (Liu et al. 2025) quantifies frame uniqueness through global token representations, which produces excessively coarse-grained features and proves inadequate for real-time inference scenarios. Dynamic methods like DyCoKe (Tao et al. 2025) and Frame Fusion (Fu et al. 2024) efficiently compress inter-frame tokens through similarity measures, yet face challenges in setting adaptive thresholds for diverse video types.

In contrast, SharpV introduces frame-specific dynamic pruning without manual intervention, maintaining low latency while avoiding additional computational overhead.

Intra-LLM Visual Token Pruning

During prefilling and decoding phases of multimodal large models, most approaches reduce computational costs and KV cache by analyzing attention distributions. FastV (Chen et al. 2024) and VTW (Lin et al. 2025) discovered that visual tokens receive significantly diminished attention in deeper layers, enabling pruning or removal in late stages. FitPrune (Ye et al. 2025) formulates pruning as a statistical distribution fitting problem, minimizing attention distribution discrepancies pre- and post-pruning. PrunerVid (Huang, Zhou, and Han 2024) performs uniform layer-wise pruning based on cross-attention scores between prompts and visual tokens, effectively reducing KV cache. DyCoKe (Tao et al. 2025) designs a special DP Cache to retain potentially useful tokens according to attention scores during subsequent decoding stages, optimizing matrix operations. Frame Fusion (Fu et al. 2024) accumulates cross-layer attention scores to measure token importance.

While attention-based pruning is theoretically reasonable, modern hardware acceleration strategies like flash attention (Dao et al. 2022; Dao 2023) have reduced quadratic complexity to linear complexity at the cost of hiding attention scores, rendering many existing pruning methods obsolete. Addressing this challenge, SharpV pioneers a novel pruning paradigm that operates independently of exposed attention scores, providing fresh insights for visual token compression.

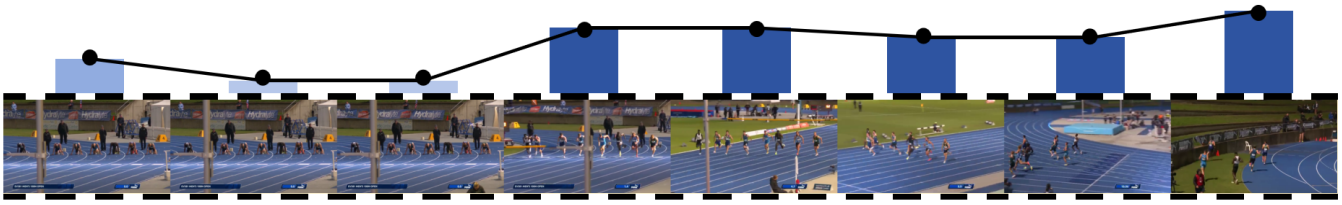


Figure 1: VideoLLM Response Demonstration Using Training-Free SharpV Pruning, Blue bars denote the per-frame token retention ratios dynamically computed by Visual SharpV, reflecting information-aware pruning.

Methodology

Information-Aware Visual Space Pruning

Our visual space pruning framework (Visual SharpV) operates through two stages: For a video sampling n frames, first, we evaluate the intra-frame token importance to obtain the spatial dimension scores of n frames and determine the pruning ratio for the first frame. Then, we analyze motion patterns within the temporal sequence to assess inter-frame information variation, computing temporal dimension scores and pruning ratios for the remaining $n-1$ frames. Finally, by aggregating each token's spatial and temporal scores, we retain high-scoring tokens according to the dynamically calculated per-frame ratios, resulting in an information-aware visual space representation.

Dissimilarity Computation Module As is well-known, cosine similarity effectively measures directional alignment between vectors. In this work, we aim to quantify token dissimilarity as a positive measure. While some approaches simply rescale cosine similarity using (1 - cosine similarity), this becomes problematic in high-dimensional spaces (Tessari, Yao, and Hogan 2024; Yamagiwa, Oyama, and Shimodaira 2024), where random vectors have expected cosine similarity approaching zero (Steck, Ekanadham, and Kallus 2024).

We therefore propose SharpV Dissimilarity, which measures directional differences using Euclidean distance between unit vectors. This metric amplifies subtle directional discrepancies between nearly-aligned vectors:

$$\text{Dissim}(\mathbf{v}_1, \mathbf{v}_2) = \|\hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_2\|_2 = \sqrt{2 - 2\cos(\theta)}, \quad (1)$$

where $\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ denotes L2-normalized vectors and θ is their angular separation. This paper employs Dissim to measure all visual token dissimilarities. We offer this module to prepare for a self-calibrated adaptive pruning ratio. Notably, replacing Dissim with (1 - cosine similarity) and a manually set fixed similarity threshold K is also compatiable with SharpV framework.

Spatio-Temporal Token Importance While some approaches compute pairwise token similarity to preserve critical semantic information (Alvar et al. 2025; Sun et al. 2025), this inevitably leads to quadratic computational complexity. Unlike existing methods (Liu et al. 2025) that coarsely estimate frame-level uniqueness, we propose a novel token-level spatio-temporal importance assessment mechanism. Assisted by our designed Dissimilarity Computation Module, each token's importance is comprehensively evaluated along dual dimensions.

For VideoLLMs, videos are sampled into n frames, each frame is mapped into f visual tokens. VideoLLMs take $n \times f$ visual tokens as complete visual input. For the tokens within each individual frame, only spatial information exists. In the spatial domain, the dissimilarity between each visual token and the overall spatial representation constitutes its Spatial Importance. Therefore, in a straightforward approach, we quantify the uniqueness of each token by measuring its dissimilarity with the frame's average representation. Given the set of tokens $F_t \in R^{fd}$ for frame t, where the global average representation is denoted as $\overline{F_t} \in \mathbb{R}^d$, the Spatial Importance $\mathcal{S}$ for each token is defined as:

$$\mathcal{S} = \textbf{Dissim}(F_t, \overline{F_t}) \in \mathbb{R}^f, \quad (2)$$

where $\overline{F_t}$ is unsqueezed along the $0th$ dimension before being input into the dissimilarity module.

The temporal importance we defined for visual tokens is calculated through the following steps: Firstly, we reshape the token matrix $V \in \mathbb{R}^{N_v \times d}$ into a three-dimensional matrix $\mathbb{R}^{n \times f \times d}$ where d is the token dimension and $n \times f = N_v$. Subsequently, temporal importance $\mathcal{T}$ is calculated by dissimilarity between corresponding tokens in adjacent frames:

$$\mathcal{T} = \textbf{Dissim}(\mathbf{F}_t, \mathbf{F}_{t-1}) \in \mathbb{R}^f \quad t \in \mathbb{Z} \cap [1, n], \quad (3)$$

where $\mathbf{F}_t, \mathbf{F}_{t-1} \in \mathbb{R}^{f \cdot d}$ are the token set of frame t and $t-1$.

The final importance score $\mathcal{I}$ of each token is computed by jointly considering its spatial and temporal importance:

$$\mathcal{I} = \mathcal{T} + w \cdot \mathcal{S}, \quad (4)$$

Since the distributions of variables $\mathcal{T}$ and $\mathcal{S}$ are different, we introduce a hyperparameter w to control the incorporation ratio of $\mathcal{S}$. By jointly modeling and aggregating temporal and spatial token importance, Visual SharpV provides more comprehensive token evaluation than single-dimension approaches.

Information-Aware Threshold Pruning In this section, we offer a lightweight adaptive threshold strategy that dynamically adjusts token retention rates frame-by-frame. This approach computes inter-frame variation without additional parameters and preserves only regions exhibiting meaningful changes, achieving fine-grained pruning at frame-level granularity. The adaptive mechanism automatically increases token retention for high-motion sequences while aggressively pruning static frames.

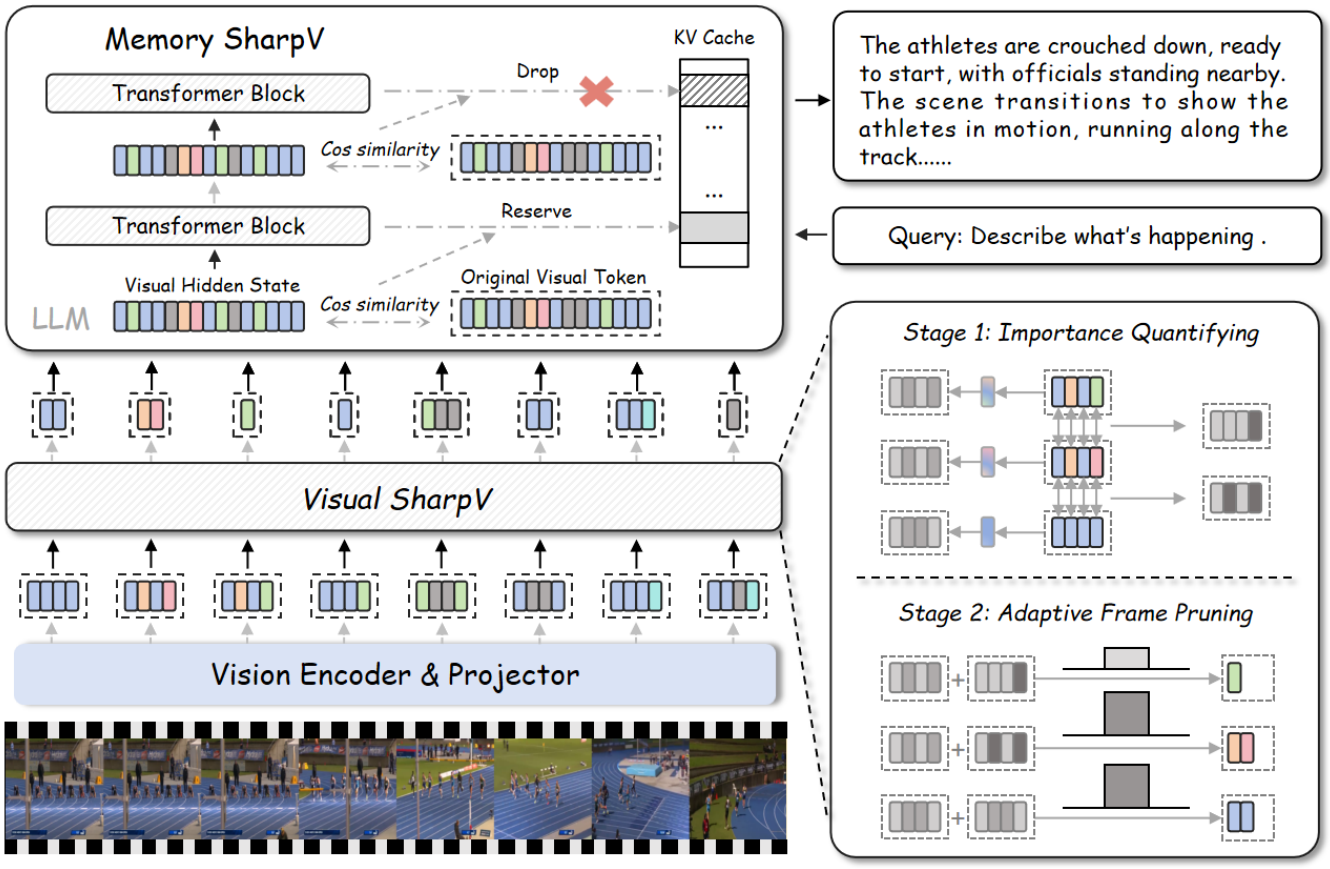


Figure 2: The Detailed Overview of SharpV. SharpV is a two-stage training-free plug-and-play framework for video LLM pruning. In the pre-LLM stage, Visual SharpV selects important visual tokens based on spatio-temporal scores, with an adaptive pruning ratio determined by L2 norm and a dissimilarity module. In the intra-LLM stage, Memory SharpV dynamically discards key-value cache by evaluating layer-wise visual information degradation.

Since the Euclidean norm of a frame t's temporal importance $\|\mathcal{T}_t\|_2 \in [0, 2\sqrt{f}]$, the Euclidean distance of a frame's temporal importance quantifies its visual variation relative to the preceding frame. The upper bound $2\sqrt{f}$ represents maximal inter-frame variation, while zero temporal importance indicates identical frames, corresponding to static visual content. Hence, we calculated threshold_t for frame $t \in \mathbb{Z} \cap [2, n]$ as follow:

$$\text{threshold}_t = \frac{\|\mathcal{T}_t\|_2}{2\sqrt{f}} \in [0, 1]. \tag{5}$$

For the first frame, since temporal importance is not applicable as no preceding frames exist, we maintain consistency by computing its threshold in a manner analogous to the subsequent $n-1$ frames. Specifically:

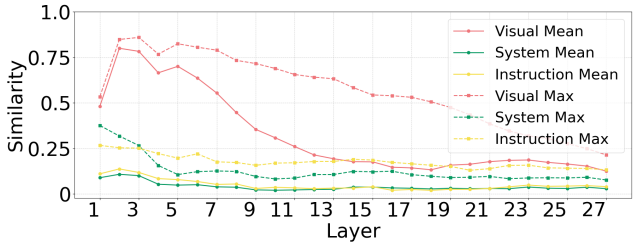
$$\text{threshold}_1 = \frac{\|\mathcal{S}_1\|_2}{2\sqrt{f}} \in [0, 1]. \tag{6}$$

This is equivalent to virtually supplementing a complete reference frame where all tokens are represented by $\overline{F_1}$ as the preceding frame for temporal threshold computation of the first frame.

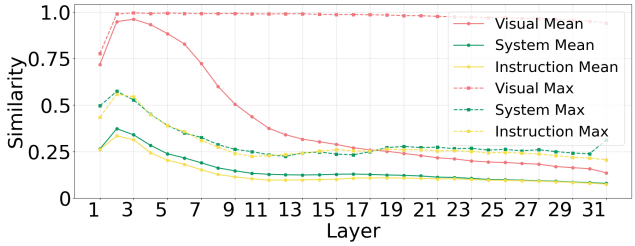
As illustrated in Fig. 1, our Information-Aware Threshold Strategy dynamically adjusts pruning thresholds by monitoring token norm variations, enabling frame-wise adaptive pruning through top-K selection of token importance scores $\mathcal{I}$. Detailed overview is shown in Fig. 2.

Information-Aware Memory Space Pruning

Visual Information Degradation Hypothesis As shown in Fig. 3, we comprehensively compare the cosine similarity between hidden states $G_l = (S_l, V_l, I_l) \in \mathbb{R}^{(N_s+N_v+N_p)\times d}$ at different layers l and their original positions in $G = (S, V, I) \in \mathbb{R}^{(N_s+N_v+N_p)\times d}$ for both LLaVA-OneVision-7B and PLLaVA, visualizing their average values. A striking observation is that in both models, the visual tokens V_l exhibit high similarity on average in shallow layers and demonstrate a clear decreasing trend to their original counterparts V as layer depth increases. In contrast, system and instruction tokens rapidly converge to near-zero similarity (0-0.2 range). Mathematically, the similarity of random high-dimensional vectors approaches zero, suggesting that system and instruction tokens undergo more complex unknown transformations through Attention and MLP oper-



(a) Llava-OneVision-7B



(b) PLLaVA-7B

Figure 3: Similarity across different layers

ations. Meanwhile, visual tokens show measurable degradation of original visual information via cosine similarity, which we term *Visual Information Degradation*. Although the exact relationships between transformed token embeddings remain unclear (Zhang et al. 2025), the similarity patterns reveal that visual information undergoes rapid decay in shallow layers resembling human memory curves before stabilizing at low similarity levels (¡0.2) after layer 14. This suggests LLMs primarily process visual information in early layers, with diminishing importance in deeper layers.

Insights Research on visual attention reveals a consistent pattern: Inefficient Visual Attention phenomenon (Chen et al. 2024; Lin et al. 2025) validate that visual tokens receive drastically diminished attention scores in deeper layers of multimodal LLMs. We further quantify this via cosine similarity, a robust metric for high-dimensional feature decay (Kalhori, Alamdari, and Ye 2018; Luo et al. 2018; Dehak et al. 2010; Lahitani, Permanasari, and Setiawan 2016).

From the perspective of information theory, this degradation pattern aligns with the concept of progressive information compression in hierarchical systems. The rapid similarity decay in early layers suggests intensive feature extraction, while the stable low similarity in deeper layers indicates visual information has been sufficiently encoded into subsequent representations. This mirrors the information bottleneck principle (Tishby, Pereira, and Bialek 2000; Hsu, Kennedy, and Chang 2006), where networks discard irrelevant details while preserving task-relevant features through successive transformations (Tishby and Zaslavsky 2015). The distinct behavior of visual versus textual tokens further implies modality-specific information (Yang et al. 2022; Barsalou et al. 2003) processing pathways in multimodal LLMs, analogous to attentional load theory in human perception (Chun and Wolfe 2005).

Model	D	L	T	F
PLLaVA-7B	4096	32	144	16
LLaVAOV-0.5B	896	24	196	32
LLaVAOV-7B	3584	28	196	32

Table 2: Model Specifications: D: dimensions, L: layers, T: tokens, F: frames.

Degradation-Aware Pruning Building upon this observation, we propose a simple yet effective strategy of discarding degraded visual information. Specifically, when the cosine similarity between a layer's visual tokens and the original visual information falls below a threshold M, we discard the corresponding layer's KV Cache. Formally,

$$\text{Discard}(l) = \begin{cases} \text{True}, & \text{if } \cos(\mathbf{V}_l, \mathbf{V}) < M \\ \text{False}, & \text{otherwise} \end{cases}, \quad (7)$$

where $\mathbf{V}_l$ represents the visual tokens at layer l, $\mathbf{V}$ denotes the original visual information and M is defined as the degradation threshold. When $\text{Discard}(l) = \text{True}$, the KV cache of layer l is discarded to prevent degraded features from propagating further. This ensures that only high-quality visual information is retained for subsequent decoding.

Experiment

Evaluation Setting

Benchmarks We employ lmms-eval (Zhang et al. 2024a) as our primary evaluation framework. For PLLaVA (Xu et al. 2024a) specifically, we use its official implementation to ensure fair comparison. Our comprehensive evaluation covers diverse video benchmarks, including MVBench (Li et al. 2024b) and VideoMME (Fu et al. 2025) for multi-dimensional video understanding and NExT-QA (Xiao et al. 2021) and ActNet-QA (Yu et al. 2019) for video question answering.

Baselines We implement FastV (Chen et al. 2024), PruMerge (Shang et al. 2024) and DyCoKe (Tao et al. 2025) on two widely used LLMs for video understanding: PLLaVA(Xu et al. 2024a) and LLaVA-OneVision (Li et al. 2024a) with details shown in Tab.2. To demonstrate scalability, we evaluate 0.5B and 7B versions of LLaVA-OneVision.

Metrics The term Token Budget combines the visual pruning ratio (VR) from the pre-LLM stage and the memory pruning ratio (MR) from the intra-LLM stage, defined as $VR \times MR$ following (Tao et al. 2025; Fu et al. 2024). VR measures the ratio of retained visual tokens fed into LLMs (Liu et al. 2025; Alvar et al. 2025), while MR estimates KV-Cache utilization during decoding (Fu et al. 2024). For efficiency comparison, we evaluate FLOPs, GPU memory and acceleration metrics: Time To First Token (TTFT) and Time Per Output Token (TPOT).

Implementation Details All experiments are conducted on NVIDIA A800 GPUs with 80GB memory. Since SharpV provides adaptive pruning strategies, we compare

Method	Token Budget	MVBench	VideoMME	NextQA	ActNet-QA		Avg.	Efficiency Analysis				
		Acc. ↑	wo ↑	mc ↑	Acc. ↑	Sco. ↑	Acc. ↑	Memory	TTFT ↓	VR ↓	MR ↓	TFlops
LLaVA-OneVision 0.5B	100%	46.6	45.9	57.5	47.9	2.66	49.5	18G	0.61s (1.00×)	100%	100%	3.4
+ FastV	30%	44.7	40.9	55.8	46.3	2.53	46.9	14G	0.49s (1.24×)	100%	30%	1.4
+ PruMerge	55%	42.5	38.8	54.5	41.7	2.35	44.4	13G	1.38s (0.44×)	55%	100%	1.5
+ DyCoke	19%	46.3	45.1	57.2	47.8	2.65	49.1	12G	0.46s (1.33×)	50%	38%	1.8
+ DyCoke	15%	46.5	45.2	57.5	47.7	2.65	49.2	10G	0.41s (1.49×)	30%	48%	1.2
+ SharpV (Manual)	19%	**46.7**	46.7	**57.8**	**47.9**	**2.66**	**50.4**	10G	0.39s (1.56×)	49%	39%	1.7
+ SharpV (Adaptive)	12%	46.6	**46.9**	57.7	47.8	2.65	49.8	**9G**	**0.37s** (1.65×)	32%	39%	**1.1**
LLaVA-OneVision 7B	100%	57.7	58.7	79.1	51.9	2.86	61.9	32G	1.05s (1.00×)	100%	100%	41.4
+ FastV	30%	56.3	56.4	76.5	50.7	2.80	60.0	30G	0.81s (1.30×)	100%	30%	17.9
+ PruMerge	55%	52.2	52.9	75.0	50.5	2.78	57.7	28G	1.73s (0.61×)	55%	100%	21.1
+ DyCoke	19%	57.7	59.3	78.4	52.1	2.88	61.9	28G	0.77s (1.36×)	50%	37.5%	24.1
+ DyCoke	15%	57.2	58.3	78.1	51.8	2.85	61.4	24G	0.73s (1.44×)	30%	47.5%	17.9
+ SharpV (Manual)	19%	57.8	59.5	**79.0**	**52.0**	**2.87**	62.1	23G	0.67s (1.57×)	49%	39%	23.9
+ SharpV (Adaptive)	12%	**58.2**	**60.0**	78.8	51.9	2.86	**62.2**	**22G**	**0.64s** (1.64×)	32%	39%	**16.7**

Table 3: Comparison of training-free visual token pruning methods using different VideoLLMs according to adaptive threshold pre-filling token budgets. Performance is measured by Accuracy (Acc), Computational Flops, and Time To First Token (TTFT).

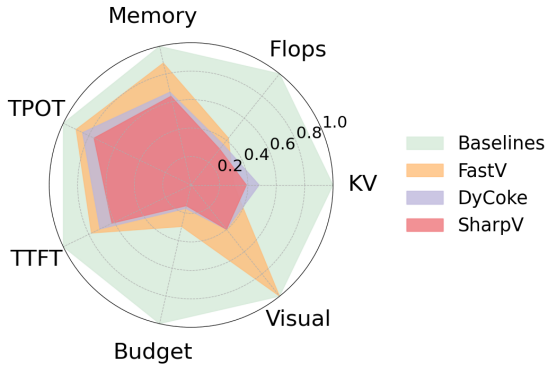


Figure 4: Average Efficiency

information-aware adaptive threshold ($M = 0.2, w = 1$) with a semi-adaptive strategy using ($K = 1.6, M = 0.2, w = 1$). Token budgets are averaged across all datasets. Positional embedding undergoes re-encoding during intra-LLM pruning in Memory SharpV. For FastV, we set the selected layer to 3 and the pruning ratio to 0.3, following the official implementation. For DyCoKe, we adopt the hyper-parameters (K, L, P) as (0.5, 3, 0.7) and (0.7, 3, 0.7), consistent with the original paper.

Performance and Efficiency

(1) As demonstrated in Table 3, SharpV consistently outperforms state-of-the-art methods across various benchmarks while maintaining scalability on both LLaVA-OV-7B and LLaVA-OV-0.5B architectures. Overall, SharpV achieves performance comparable to the dense model with pruning rates between 10% and 20% across all four benchmarks, surpassing DyCoke, FastV, and PruMerge while maintaining a lower token budget.

(2) Notably, as shown in Tables 3 and 4, SharpV occasionally exceeds the performance of the dense model baseline by 1% to 2% on VideoMME, NExTQA, and PLLaVA (VideoMME, MVBench). This improvement can be attributed to the fact that appropriate pruning helps reduce video noise and enables the model to better focus on key visual elements. In contrast, FastV suffers from performance degradation due to its fixed, coarse-grained one-time pruning approach. PruMerge underperforms because it preserves visual encoder attention information that may not align with the LLM's internal attention mechanisms. DyCoke's uniform pruning rate in the first stage results in insufficient preservation of information-rich videos while potentially introducing more noise from information-sparse videos.

(3) Moreover, during the pre-LLM stage, SharpV's adaptive threshold mechanism dynamically retains approximately 32% of video tokens on average without requiring manual parameter tuning, demonstrating both robustness and generality in achieving optimal performance.

(4) We further compare the multi-dimensional efficiency of different methods in Fig. 4 and Table 3. While maintaining comparable or superior performance, SharpV achieves significantly lower resource consumption in both GPU memory and FLOPs, attributable to its low computational complexity. Notably, our method demonstrates: 1) reduced TTFT through Visual SharpV's initial pruning stage, and 2) accelerated TPOT enabled by Memory SharpV's subsequent KV cache pruning.

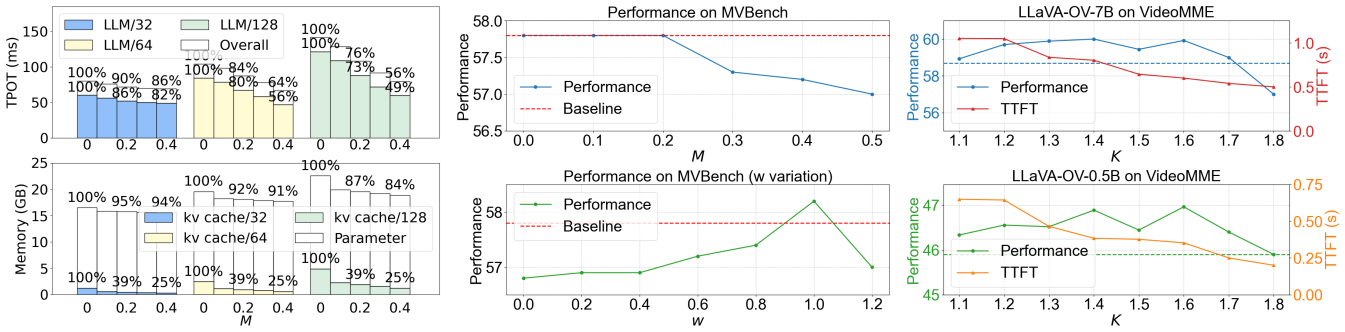
Ablation Study

Visual SharpV In pre-LLM evaluation, we compare our spatial-temporal scoring with two random strategies: 1) V-Random† replaces topK selection with random selection while keeping Visual SharpV's per-frame pruning ratios, comparing score-based versus random selection under identical pruning budgets. 2) V-Random* randomly selects tokens video-wide without frame-wise ratio control, maintaining only the total pruning threshold. Table 4 shows Visual SharpV consistently outperforms V-Random†, proving spatio-temporal selection's effectiveness. V-Random* suffers significant degradation versus V-Random†, demonstrating the importance of frame-adaptive pruning ratios. Our approach dynamically allocates more tokens to information-rich frames through fine-grained ratio control, better capturing video semantics while reducing redundancy and noise.

Memory SharpV During the intra-LLM stage, we systematically evaluate the KV cache and GPU memory con-

Method	MVBench																				VideoMME			Avg.
	AA	AC	AL	AP	AS	CO	CI	EN	ER	FA	FP	MA	MC	MD	OE	OI	OS	ST	SC	UA	Short	Medium	Long	ACC.
PLLaVA 7B	55.5	39.5	26.0	49.0	58.0	53.5	31.0	30.5	48.0	41.0	42.0	52.0	45.0	23.5	56.0	61.0	36.0	82.0	42.0	61.0	51.4	43.4	35.4	46.6/43.4
Visual SharpV	**62.0**	**41.0**	**27.5**	**50.5**	**56.5**	**53.5**	32.0	**32.0**	**47.5**	**42.5**	**49.5**	**56.0**	**39.5**	**20.5**	**57.0**	**64.0**	**37.0**	**82.5**	**43.5**	**65.0**	**52.1**	**42.2**	**36.5**	**48.0/43.6**
V-Random†	**62.0**	38.5	26.0	50.0	56.0	47.0	**37.0**	31.5	**47.5**	41.5	40.5	52.0	39.0	20.0	53.0	60.5	34.0	82.0	**43.5**	63.5	50.3	40.4	35.2	46.3/42.0
V-Random*	61.0	38.0	24.0	49.5	55.0	45.0	29.0	31.5	47.0	39.0	42.0	51.0	37.5	18.0	54.0	59.5	34.5	81.5	42.0	60.5	49.4	40.2	34.3	45.0/41.3
LLaVA-OV 7B	78.0	47.5	59.0	58.0	71.0	72.0	50.0	33.0	55.0	47.0	54.0	70.0	48.5	21.0	47.5	82.5	35.5	92.5	52.0	80.0	71.0	57.1	48.1	57.7/58.7
Visual SharpV	**81.0**	**46.5**	**58.0**	**59.0**	75.0	68.0	**48.0**	**34.0**	**53.5**	**45.0**	**58.0**	62.5	46.0	**28.0**	57.0	**83.0**	36.0	**93.5**	**52.0**	**79.0**	70.8	**57.4**	**50.2**	**58.2/59.5**
V-Random†	79.5	46.0	47.5	57.5	**76.0**	**69.0**	39.5	33.0	**52.0**	44.5	54.5	**68.0**	46.5	23.0	**60.5**	82.0	**37.5**	93.0	50.0	78.5	**71.0**	55.7	49.0	56.9/58.6
V-Random*	63.0	37.5	33.5	32.0	50.5	48.0	34.0	31.0	45.0	43.5	20.0	57.0	**47.0**	30.5	51.5	49.5	37.0	74.0	35.0	64.5	49.2	43.4	39.3	44.2/44.0
LLaVA-OV 0.5B	54.0	39.0	35.0	43.5	54.0	55.0	37.5	31.0	45.0	39.5	49.0	53.0	33.5	21.0	47.5	70.5	32.0	84.0	37.5	70.0	55.7	43.7	38.4	46.6/45.9
Visual SharpV	53.5	39.5	30.0	**42.0**	52.5	52.0	39.5	30.5	44.0	**37.5**	**46.5**	**52.5**	30.5	20.5	**48.5**	69.0	**33.5**	**82.5**	**39.5**	**68.0**	**57.3**	43.3	**39.4**	**46.3/46.7**
V-Random†	**55.5**	38.0	30.0	40.5	**55.0**	**51.5**	40.0	**32.0**	**45.5**	**37.5**	45.0	48.5	**31.0**	**22.0**	48.0	**69.5**	31.0	**82.5**	**39.5**	67.5	55.0	**44.3**	39.2	45.5/46.2
V-Random*	53.5	**40.0**	**31.5**	24.0	36.0	34.0	**40.5**	27.0	37.5	37.0	25.5	**52.5**	28.0	20.5	47.0	45.5	34.0	64.0	38.0	58.0	38.8	36.4	32.3	39.1/35.9

Table 4: Ablation Study of Information-aware Threshold in Pre-LLM Stage. † donates random selection according to information-aware threshold each frame, * indicates random pruning with same threshold of each video sample.



(a) Memory and TPOT Across Frame Sizes (b) Ablation Study of Parameter M and w (c) Ablation Study of Parameter K

Figure 5: Ablation Study of Parameters: M, w, K

sumption across different frame sizes, as illustrated in Fig. 5a. Our experiments reveal a sharp drop in KV cache usage between 0 and 0.1, where approximately 54% of visual token layers exhibit similarity scores below 0.1 compared to the original information, while showing stable variations in the $[0.1, 0.4]$ range, which align with the analysis presented in Fig. 3. Consequently, TPOT achieves computational acceleration through optimized KV cache reduction.

Hyper-parameters The parameter w controls the degree of spatial information incorporation, accounting for potential distribution differences between intra-frame and inter-frame dissimilarities. As shown in Fig. 5b, optimal performance is achieved at w=1, even surpassing the MVBench baseline. The threshold parameter m determines the degradation boundary. Analysis of similarity degradation in Fig. 3 reveals that the decline plateaus when similarity falls within $[0, 0.2]$. This observation is empirically validated in Fig. 5b, where performance remains stable for $m <= 0.2$ but deteriorates beyond this threshold. Additionally, threshold K of SharpV(Manual) in Tab. 3 is decided according to Fig. 5c. A larger K reduces VR and TTFT, with $K = 1.6$ offering a balanced efficiency-performance trade-off.

Discussion and Limitations

SharpV pioneers token-similarity-based dynamic pruning for efficient VLLM inference, yet challenges persist. First, while Visual SharpV's spatiotemporal compression effectively reduces token redundancy, subtle visual details (e.g., micro-expressions) may still suffer minor information loss. Future work could explore finer-grained token pruning for fine-detail-sensitive scenes. Second, although visual degradation is empirically validated through cosine similarity assisted with information theory and attention shift analysis, a rigorous theoretical framework for deep-layer token transformation mechanisms remains elusive. Further studies could bridge this gap.

Conclusion

SharpV introduces a novel approach to intra-LLM visual token pruning that operates without reliance on exposed attention scores, achieving linear complexity while maintaining compatibility with Flash Attention. Furthermore, we propose a simple yet effective pre-LLM token selection method, offering a self-calibrated, adaptive thresholding scheme with strong generality and robustness based on information measurement. These components synergize to reduce the final pruning rate around 12%, while frequently improving accuracy across multiple benchmarks due to reduced visual noise, enhancing both efficiency and reliability in Video-LLM reasoning.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality

References

Alvar, S. R.; Singh, G.; Akbari, M.; and Zhang, Y. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9392–9401.

Barsalou, L. W.; Simmons, W. K.; Barbey, A. K.; and Wilson, C. D. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2): 84–91.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Chun, M. M.; and Wolfe, J. M. 2005. Visual attention. *Blackwell handbook of sensation and perception*, 272–310.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359.

Dehak, N.; Dehak, R.; Glass, J. R.; Reynolds, D. A.; Kenny, P.; et al. 2010. Cosine similarity scoring without score normalization techniques. In *Odyssey*, volume 15.

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35: 35946–35958.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Fu, T.; Liu, T.; Han, Q.; Dai, G.; Yan, S.; Yang, H.; Ning, X.; and Wang, Y. 2024. Framefusion: Combining similarity and importance for video token reduction on large visual language models. *arXiv preprint arXiv:2501.01986*.

Hsu, W. H.; Kennedy, L. S.; and Chang, S.-F. 2006. Video search reranking via information bottleneck principle. In *Proceedings of the 14th ACM international conference on multimedia*, 35–44.

Huang, X.; Zhou, H.; and Han, K. 2024. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*.

Hyun, J.; Hwang, S.; Han, S. H.; Kim, T.; Lee, I.; Wee, D.; Lee, J.-Y.; Kim, S. J.; and Shim, M. 2025. Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs. *arXiv preprint arXiv:2507.07990*.

Kalhori, H.; Alamdari, M. M.; and Ye, L. 2018. Automated algorithm for impact force identification using cosine similarity searching. *Measurement*, 122: 648–657.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Shen, X.; Yuan, G.; Ren, B.; Tang, H.; et al. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, 620–640. Springer.

Lahitani, A. R.; Permanasari, A. E.; and Setiawan, N. A. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, 1–6.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lin, Z.; Lin, M.; Lin, L.; and Ji, R. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5334–5342.

Liu, X.; Wang, Y.; Ma, J.; and Zhang, L. 2025. Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models. *arXiv preprint arXiv:2505.14454*.

Luo, C.; Zhan, J.; Xue, X.; Wang, L.; Ren, R.; and Yang, Q. 2018. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In *International conference on artificial neural networks*, 382–391. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34: 13937–13949.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

Steck, H.; Ekanadham, C.; and Kallus, N. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, 887–890.

Sun, B.; Zhao, J.; Wei, X.; and Hou, Q. 2025. LLaVA-Scissor: Token Compression with Semantic Connected Components for Video LLMs. *arXiv preprint arXiv:2506.21862*.

Tao, K.; Qin, C.; You, H.; Sui, Y.; and Wang, H. 2025. DyCoke: Dynamic Compression of Tokens for Fast Video Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18992–19001.

Tessari, F.; Yao, K.; and Hogan, N. 2024. Surpassing Cosine Similarity for Multidimensional Comparisons: Dimension Insensitive Euclidean Metric. *arXiv preprint arXiv:2407.08623*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. Ieee.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vasu, P. K. A.; Gabriel, J.; Zhu, J.; Tuzel, O.; and Ranjan, A. 2023. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5785–5795.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024a. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Xu, M.; Gao, M.; Gan, Z.; Chen, H.-Y.; Lai, Z.; Gang, H.; Kang, K.; and Dehghan, A. 2024b. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*.

Yamagiwa, H.; Oyama, M.; and Shimodaira, H. 2024. Revisiting Cosine Similarity via Normalized ICA-transformed Embeddings. *arXiv preprint arXiv:2406.10984*.

Yang, X.; Xiong, B.; Huang, Y.; and Xu, C. 2022. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 3063–3071.

Ye, W.; Wu, Q.; Lin, W.; and Zhou, Y. 2025. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22128–22136.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; et al. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.

Zhang, X.; Zeng, F.; Quan, Y.; Hui, Z.; and Yao, J. 2025. Enhancing multimodal large language models complex reason via similarity computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10203–10211.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024b. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.