KTV: Keyframes and Key Tokens Selection for Efficient Training-Free Video LLMs

Baiyang Song[1*], Jun Peng[2*†], Yuxin Zhang[1], Guangyao Chen[3], Feidiao Yang[2], Jianyuan Guo[4†]

[1]Department of Artificial Intelligence, School of Informatics, Xiamen University
[2]Peng Cheng Laboratory
[3]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[4]City University of Hong Kong, HKSAR, China
{songbaiyang, yuxinzhang}@stu.xmu.edu.cn, {pengj01, yangfd}@pcl.ac.cn, gy.chen@pku.edu.cn, jianyguo@cityu.edu.hk

Abstract

Training-free video understanding leverages the strong image comprehension capabilities of pre-trained vision language models (VLMs) by treating a video as a sequence of static frames, thus obviating the need for costly video-specific training. However, this paradigm often suffers from severe visual redundancy and high computational overhead, especially when processing long videos. Crucially, existing keyframe selection strategies, especially those based on CLIP similarity, are prone to biases and may inadvertently overlook critical frames, resulting in suboptimal video comprehension. To address these significant challenges, we propose **KTV**, a novel two-stage framework for efficient and effective training-free video understanding. In the first stage, KTV performs question-agnostic keyframe selection by clustering frame-level visual features, yielding a compact, diverse, and representative subset of frames that mitigates temporal redundancy. In the second stage, KTV applies key visual token selection, pruning redundant or less informative tokens from each selected keyframe based on token importance and redundancy, which significantly reduces the number of tokens fed into the LLM. Extensive experiments on the Multiple-Choice VideoQA task demonstrate that KTV outperforms state-of-the-art training-free baselines while using significantly fewer visual tokens, *e.g.*, only 504 visual tokens for a 60-min video with 10800 frames, achieving 44.8% accuracy on the MLVU-Test benchmark. In particular, KTV also exceeds several training-based approaches on certain benchmarks.

Code — https://github.com/songbaiyang07-star/KTV

Introduction

Vision-language models (VLMs) (Liu et al. 2023; Li et al. 2023a) have achieved remarkable advancements, enabling impressive multimodal understanding and reasoning over the static image. A natural next step is to extend these capabilities to the more complex domain of video, where temporal dynamics and richer contextual information present both opportunities and challenges. While recent video LLMs (Zhang, Li, and Bing 2023; Lin et al. 2024) combine specialized video encoders with pre-trained LLMs (Brown et al. 2020; Touvron et al. 2023), their development faces significant challenges due to the scarcity of large-scale, high-quality video-text data needed for effective training or fine-tuning such models.

To circumvent the demand for massive training data, the idea of training-free methods is compelling. It treats a video as a sequence of individual frames and ingeniously leverages the strong image understanding and reasoning capabilities of pre-trained VLMs for analyzing these frames.

Despite the progress achieved by training-free methods (Kim et al. 2024; Wu 2024), two critical challenges persist. First, videos, particularly long-form ones, exhibit severe temporal redundancy, where consecutive frames often contain nearly identical visual content. This redundancy leads to inefficient computation when processed by VLMs. Second, even with frame sampling, the resulting number of image tokens, *e.g.*, 576 per frame, is still prohibitive for LLMs: aggregating tokens across hundreds of frames either exceeds LLM's context window or drastically slows down inference.

To address these challenges, we propose a novel two-stage framework for training-free video understanding, termed **KTV**. In the first stage, KTV reduces temporal redundancy by selecting a compact yet representative subset of keyframes from the entire video. Unlike prior methods (Gorti et al. 2022; Han et al. 2024; Zhang et al. 2025), which rely on CLIP-processed features for keyframe selection, we observe that these methods can introduce inherent biases and result in the omission of crucial visual cues, as shown in Fig. 1. To address this issue, we introduce a sampling strategy that does not depend on textual queries, ensuring more reliable and comprehensive keyframe selection. In practice, visual features are extracted from all frames using a pre-trained off-the-shelf visual encoder. K-means clustering is then applied to group similar frames, with the frame closest to each cluster's centroid being selected as the representative keyframe for that cluster.

In the second stage, KTV further reduces the number of visual tokens by selecting key tokens from each keyframe. Our goal is to retain essential semantic information while preserving visual features diversity. To achieve this, KTV evaluates the importance of each token by measuring its sim-

[*]These authors contributed equally.
[†]Corresponding authors.

Figure 1: The pipeline of our two-stage training-free **K**eyframes and **K**ey **T**okens selection for empowering LLMs in **V**ideo understanding, termed **KTV**. Given a video clip (1^{st} row), KTV first performs question-agnostic keyframe selection (3^{rd} row), identifying representative keyframes that provide a diverse and unbiased summary of the video content. This contrasts with question-relevant selection methods (2^{nd} row), which may overlook general scene information. Then, KTV selects only the most informative and non-redundant visual tokens from each keyframe (4^{th} row), significantly reducing spatial redundancy.

ilarity not only with the [CLS] token (Shang et al. 2024; Zhang et al. 2024a), but also with other visual tokens within the same frame, ensuring a rich and comprehensive representation of the keyframe. Additionally, we leverage the well-trained CLIP text encoder to perform question-aware vision token pruning, thereby enhancing the relevance of the retained tokens with respect to the query, ensuring that they are highly aligned with the question at hand. By jointly performing keyframe selection and fine-grained token pruning, KTV produces a concise yet semantically rich visual representation of the video. This enables efficient and effective multimodal reasoning within a pre-trained VLM, without requiring any video-specific training or fine-tuning.

We evaluate KTV on Multiple-Choice VideoQA benchmarks using pre-trained LLaVA-v1.6 (Liu et al. 2024) as the VLM. Three KTV variants, *i.e.*, *sparse*, *normal*, and *dense*, are assessed, corresponding to different key visual token sequence lengths. Experimental results show that KTV significantly reduces visual tokens while outperforming or matching state-of-the-art training-free methods. For example, a 60-minute video with 10800 frames can be processed with only 504 visual tokens, fewer than typically extracted from a single image. Moreover, KTV achieves superior performance with smaller VLMs compared to the larger models used in other methods, and even outperforms training-based approaches, underscoring its efficiency and effectiveness.

In conclusion, our contributions are as follows: (i) We propose KTV, a two-stage training-free framework for video understanding. It first performs question-agnostic keyframe selection, followed by fine-grained key visual token selection to reduce both temporal and spatial redundancy. (ii) Extensive evaluations on standard VideoQA benchmarks demonstrate our KTV's robustness and generalizability.

Related Work

Vision Language Models (VLMs). The evolution of VLMs focuses on bridging powerful vision encoders (Dosovitskiy et al. 2021) with LLMs (Chiang et al. 2023; Touvron et al. 2023). Early works like BLIP-2 (Li et al. 2023a) proposed complex bridging modules, *i.e.*, Q-Former, to distill visual information. In contrast, LLaVA (Liu et al. 2023) introduced a streamlined architecture using a simple MLP to project visual features into the LLM's embedding space, achieving strong performance with minimal overhead. However, the challenge with this simplicity is that projecting a full grid of visual tokens significantly increases sequence length and inference cost. To address this, follow-up works such as LLaVA-PruMerge (Shang et al. 2024) introduced token merging and pruning strategies, effectively reducing the visual input length without substantial accuracy loss. Never-

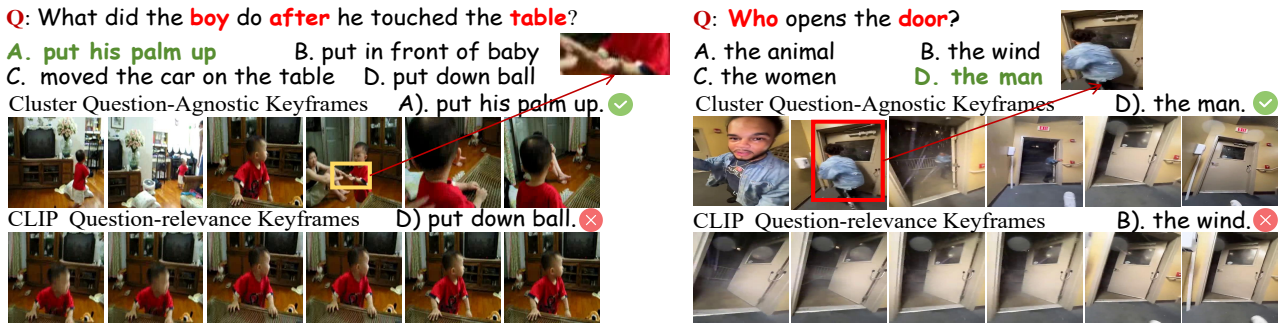

Figure 2: Keyframes selection compared between DINO Cluster and CLIP Text-Frame similarity. CLIP selects frames that are overly dependent on the word in question, making it easy to get into "semantic traps". In contrast, DINO Cluster captures the video's maximal visual diversity, yielding a holistic and unbiased summary of its core scenes.

theless, merging similar tokens without model fine-tuning may distort feature distributions and compromise compatibility with the pretrained LLM.

Video Large Language Models (Video LLMs). The success of image-based VLMs has catalyzed the development of video LLMs, which aim to incorporate temporal context for complex video reasoning. A typical pipeline involves extracting frame-level features via pre-trained visual encoders and feeding them into LLMs through an interface module. These modules vary from simple spatial-temporal pooling, *e.g.*, VideoChatGPT (Maazi et al. 2024), to transformer-based bridges like Q-Former in VideoChat (KunChang Li and Qiao 2023). Efforts such as Video-LLaVA (Lin et al. 2024) propose shared projectors to align image and video encoders for joint training. Another trend focuses on instruction tuning with large-scale video-text datasets, as seen in VideoChat2 (Li et al. 2024), and applying alignment techniques like Direct Preference Optimization (DPO) (Rafailov et al. 2023), as in LLaVA-NeXT-Video (Liu et al. 2024), to improve response quality and task adherence.

Training-free Video LLMs. Several recent works reframe video understanding as a multi-image task, enabling the use of pre-trained image-based VLMs without extra video-specific training. FreeVA (Wu 2024) explores temporal-spatial pooling across sampled frames to build compact representations. IG-VLM (Kim et al. 2024) arranges uniformly sampled frames into one image for inference, but loses detail due to limited resolution. SF-LLaVA (Xu et al. 2024) adopts a dual-stream design to balance spatial resolution and temporal coverage, yet still yields long visual sequences. DYTO (Zhang et al. 2024b) selects keyframes using clustering on [CLS] tokens and prunes tokens within each frame based on semantic similarity, but relies on a large number of tokens for downstream reasoning.

While these methods reduce spatial and temporal redundancy to varying degrees, they often overlook frame content diversity, *e.g.*, uniform sampling in FreeVA and IG-VLM, or incur high token counts that increase inference latency and memory usage, *e.g.*, 3680 used by SF-LLaVA and DYTO.

To address these limitations, our proposed method KTV, introduces a two-stage strategy that jointly reduces temporal and spatial redundancy. First, it clusters frame-level features to select a compact subset of semantically diverse keyframes. Then, it performs fine-grained token pruning within each keyframe based on relevance to the [CLS] token and intra-frame similarity. This enables precise control over the visual token budget, *e.g.*, 504 tokens (KTV-sparse), 936 tokens (KTV-normal), and 1872 tokens (KTV-dense), while maintaining strong performance.

Method

In this paper, we propose KTV, a novel two-stage method for training-free video understanding that effectively mitigates both temporal and spatial redundancy. KTV first selects keyframes by clustering frame-wise visual features, then prunes each keyframe's visual tokens based on their importance and redundancy. This strategy preserves both essential content and overall scene information. The overall framework of KTV is illustrated in Fig.3.

Question-Agnostic Keyframe Selection

While question-relevance methods, which select frames based on full sentence (Gorti et al. 2022; Han et al. 2024) or extracted keywords (Zhang et al. 2025), appear efficient, they are prone to "semantic traps", *i.e.*, selecting frames that superficially align with question keywords while overlooking crucial contextual information that lacks explicit textual correspondence. This issue is especially detrimental for complex reasoning tasks involving causality or temporal progression, where the "cause" of an event may be visually distinct from the "effect" referenced in the question.

For instance, as shown on the left of Fig.2, given the question *"What did the boy do after he touched the table"*, a question-relevance method, *e.g.*, selecting the frames most similar to the question via CLIP (Radford et al. 2021), tends to focus on visual cues tied to keywords like *"boy"*, *"touched"* and *"table"*. As a result, it selects frames showing the boy touching the table, but misses the temporal cue *"after"*, leading to an incorrect answer.

In contrast, our question-agnostic clustering approach captures maximal visual diversity, offering a holistic and unbiased summary of the video. This allows it to select key moments, including the boy's action after touching the table, enabling correct reasoning and answering.

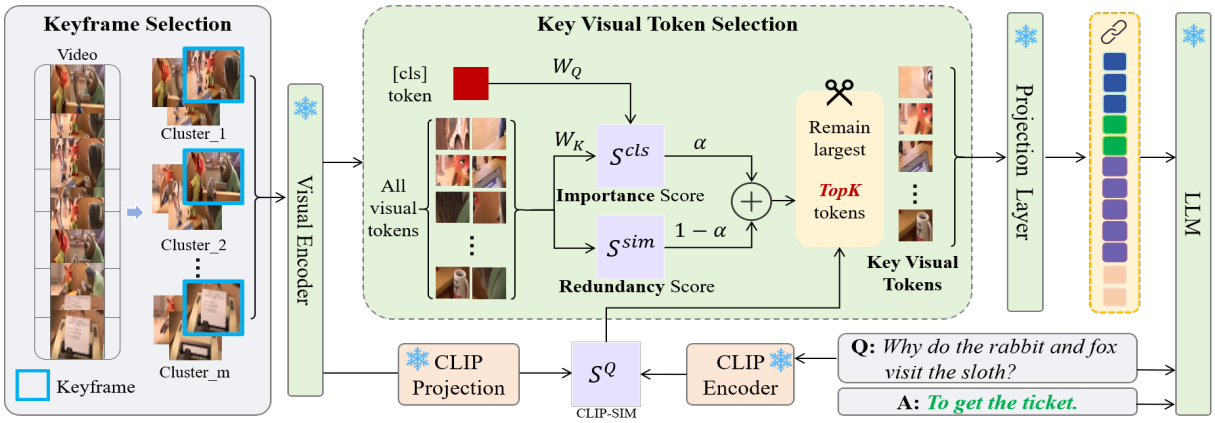

Figure 3: Framework of KTV, which is a two-stage method of training-free video understanding built upon LLaVA-v1.6. First, we extract the visual features of video frames and cluster them into m clusters, whose centroids are selected as keyframes to mitigate the temporal redundancy. Second, for each frame, we select top-$k = \beta \cdot L$ key visual tokens based on their importance and redundancy, which will be fed to the LLM and prune other visual tokens to mitigate the spatial redundancy. At last, we concatenate all the remaining visual tokens and text tokens and feed them to the LLM to generate the answer.

Specifically, given a video $\mathcal{V}$ consisting of T frames, $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$. KTV first utilizes DINOv2 (Oquab et al. 2024) to extract frame-level feature vector:

$$\{f_1, f_2, ..., f_T\} = \mathcal{E}_{dinov2}(\mathcal{V}) \tag{1}$$

Subsequently, the extracted frame features $F_\mathcal{V}$ are grouped into m clusters using the K-means algorithm. To mitigate temporal redundancy, KTV selects the most representative frame from each cluster, *i.e.*, the one whose feature vector is closest to the corresponding centroid:

$$j^* = \arg\min_{f_j \in C_i} \|f_j - c_i\|, \; i \in [1, m] \tag{2}$$

in which C_i and c_i denote the i-th cluster and corresponding centroid, respectively. Reorganize the selected keyframes by time order, denoted as $\mathcal{K} = \{I'_1, I'_2, \cdots, I'_m\}$. This keyframe selection effectively reduces temporal redundancy while preserving diverse and representative visual content throughout the video.

Key Visual Tokens Selection

After selecting the keyframes, KTV further mitigates spatial redundancy within each selected frame to reduce the total number of visual tokens input to the LLM. Specifically, each keyframe is encoded using the frozen visual encoder of the VLM, typically an image encoder such as CLIP-L, to extract token-level features. Denote the resulting representation as $F_i \in \mathbb{R}^{L \times d}$, where L is the number of visual tokens and d is the dimensionality of each token. This encoding process is formulated as follows:

$$F_i = \mathcal{E}_{vis}(I'_i) = \{t_{i,1}, ..., t_{i,L}\}, i \in [1, m] \tag{3}$$

where $t_{i,j}$ is the j-th token in the i-th keyframe I'_i. For each keyframe, we retain only the most informative and non-redundant visual tokens, yielding a compact representation $F'_i \in \mathbb{R}^{topk \times d}$ to be fed into the LLM. We aim to preserve essential and globally relevant information within each frame. To this end, we assess the importance of each visual token $t_{i,j}$ by computing its attention score to the corresponding [CLS] token t_i^{cls}:

$$S_{i,j}^{cls} = S^{cls}(t_{i,j}) = \text{softmax}\left(\frac{W_Q \cdot t_i^{cls} \cdot (W_K \cdot t_{i,j})^T}{\sqrt{d}}\right) \tag{4}$$

where W_Q and W_K are the projection weight matrices.

To measure redundancy, we compute a redundancy score by calculating the average cosine similarity between each token and the other ones within the same keyframe.

$$S_{i,j}^{sim} = S^{sim}(t_{i,j}) = \frac{1}{L-1} \sum_{k=1, k \neq j}^{L} \frac{f_{i,j} \cdot f_{i,k}}{\|f_{i,j}\| \cdot \|f_{i,k}\|} \tag{5}$$

Selecting tokens based solely on importance may lead to redundancy, while relying solely on low redundancy could result in the omission of critical information. To strike a balance, KTV integrates both aspects by assigning each token a final score, calculated as a weighted combination of its importance and redundancy scores.

$$\text{Norm}(S_{i,j}^{cls}) = \frac{S_{i,j}^{cls} - \min(S_i^{cls})}{\max(S_i^{cls}) - \min(S_i^{cls})} \tag{6}$$

$$\text{Norm}(S_{i,j}^{sim}) = \frac{S_{i,j}^{sim} - \min(S_i^{sim})}{\max(S_i^{sim}) - \min(S_i^{sim})} \tag{7}$$

$$\begin{aligned} S_{i,j} = &\text{Norm}(S_{i,j}^{cls}) \times \alpha \\ &+ (1 - \text{Norm}(S_{i,j}^{sim})) \times (1 - \alpha) \end{aligned} \tag{8}$$

where the hyperparameter $\alpha \in [0, 1]$ balances importance and redundancy, helping the model retain informative yet diverse tokens.

The top-k tokens with the highest scores are then retained as the key visual tokens for each keyframe. To determine the value of top-k, we use a set of descending hyperparameters, $\beta = \{\beta_1, \cdots, \beta_m\}$, where each $\beta_i \in [0, 1]$ specifies the proportion of tokens to keep. Then we compute the CLIP

Method	LLM Size	Vis Encoder	NExTQA	Ego Schema	Intent QA	STAR	Video MME	MVBench
Training-Based Methods								
Video-LLaVA (Lin et al. 2024)	7B	ViT-L	-	-	-	-	39.9	41.0
VideoLLaMA2 (Cheng et al. 2024)	46.7B	CLIP-L	-	53.3	-	-	-	53.9
Training-Free Methods								
IG-VLM (Kim et al. 2024)	7B	CLIP-L	63.1	35.8	60.3	48.6	39.7	42.9
SF-LLaVA-7B (Xu et al. 2024)	7B	CLIP-L	64.0	44.2	60.5	48.8	39.4	43.3
DYTO (Zhang et al. 2024b)	7B	CLIP-L	65.7	48.6	61.6	50.7	41.2	-
KTV-7B-sparse	7B	CLIP-L	64.5	49.6	61.2	52.3	43.6	46.2
KTV-7B-normal	7B	CLIP-L	65.1	50.4	61.3	52.5	43.7	**46.4**
KTV-7B-dense	7B	CLIP-L	**65.8**	**51.0**	**62.0**	**52.7**	**44.0**	46.0
IG-VLM (Kim et al. 2024)	34B	CLIP-L	70.7	53.4	64.5	50.5	50.3	49.0
SF-LLaVA-7B (Xu et al. 2024)	34B	CLIP-L	70.9	55.0	66.1	51.3	51.9	49.6
DYTO (Zhang et al. 2024b)	34B	CLIP-L	72.9	56.8	66.4	51.1	53.4	-
KTV-34B-sparse	34B	CLIP-L	71.2	55.6	65.9	54.2	52.2	51.9
KTV-34B-normal	34B	CLIP-L	72.3	55.6	66.6	54.6	53.0	**52.1**
KTV-34B-dense	34B	CLIP-L	72.7	57.0	68.0	54.7	53.2	51.5

Table 1: KTV performance compared to existing models on Multiple-Choice VideoQA benchmarks. Models are grouped by LLM size (7B or 34B) and whether they are training-based or training-free. **Bold** indicates best performance among 7B models. Underline indicates best performance among 34B models.

similarity between each keyframe and the question Q and rank the keyframes I'_i by relevance:

$$S_i^Q = \text{CLIP-SIM}(I'_i, Q), i \in [1, m] \tag{9}$$

The i-th most relevant keyframe is then assigned β_i, and the number of retained tokens is computed as $\beta_i \times L$, where L is the number of tokens per frame. This strategy allocates more tokens to frames more relevant to the question, and fewer to less relevant ones, balancing efficiency and informativeness.

Subsequently, the retained tokens from each keyframe are passed through a projection layer to align with the text embedding space. These projected visual tokens are then concatenated with text tokens, which typically consist of a structured prompt and the question. Finally, the combined token sequence is fed into the LLM to produce the answer.

Experiments

Benchmarks and Metrics

We evaluate our proposed KTV on seven Multiple-Choice VideoQA benchmarks, *i.e.*, NExT-QA (Xiao et al. 2021), EgoSchema (Mangalam, Akshulakov, and Malik 2023), IntentQA (Li et al. 2023b), STAR (Wu et al. 2024), VideoMME (Fu et al. 2025), MVBench (Li et al. 2024), and MLVU-Test (Zhou et al. 2025). Notably, MVBench and MLVU-Test are multi-task benchmarks that include a broad range of question types and video content.

These benchmarks are designed to assess complex video understanding and reasoning abilities, covering diverse question types, *e.g.*, causal, temporal, and spatial reasoning. We report accuracy as the metric for selecting the correct answer from the candidates. For MVBench, the breakdown is included in the appendix or supplementary material.

To rigorously evaluate visual understanding, we exclude any extra reference data, *e.g.*, subtitles, ensuring the model relies solely on visual content, question, and answer choices.

Implementation Details

Following IG-VLM and SF-LLaVA, we adopt LLaVA-v1.6 as our VLM. All experiments are conducted on two Huawei Ascend 910C NPUs, each has 64 GB of memory. To ensure fair comparison with baselines, we cluster video frames into 6 clusters, resulting in $m = 6$ keyframes selected by KTV. Moreover, we define three configurations, KTV-sparse, KTV-normal, and KTV-dense, by assigning different values to the hyperparameter β^{sparse}, β^{normal}, and β^{dense}, respectively. Detailed settings can be found in the appendix.

Experimental Results

Multiple Choice VideoQA In Tab. 1, we report the overall accuracy of KTV under three configurations, sparse, normal, and dense, using both LLaVA-v1.6-7B and LLaVA-v1.6-34B. As can be seen, KTV consistently outperforms or matches other training-free baselines while using significantly fewer visual tokens. Specifically, KTV-7B-sparse outperforms IG-VLM and SF-LLaVA across all benchmarks, and outperforms DYTO on EgoSchema, STAR, and VideoMME. Notably, it uses only 504 visual tokens, fewer than the 576 tokens encoded from a single image in LLaVA-v1.6, and achieves an inference time of 1.19s, just 36.4% of that required by SF-LLaVA-7B. KTV-7B-normal further improves accuracy across all benchmarks, and KTV-7B-dense achieves the best performance among all training-free methods, using only 1872 visual tokens with an inference time of 1.35s (41.3% of SF-LLaVA-7B).

As for KTV-34B, the sparse variant also outperforms IG-VLM-34B and SF-LLaVA-34B on most benchmarks, using 504 visual tokens and achieving an inference time of 1.23s, which is just 28.0% of SF-LLaVA-34B. KTV-34B-normal

Method	LLM Size	Vision Encoder	Holistic LVU		Single-Detail LVU			Multi-Detail LVU			M-Avg
			TR	AR	NQA	ER	PQA	SQA	AO	AC	TQA	
Training-Based Methods												
Video-LLaVA (Lin et al. 2024)	7B	CLIP-L	70.3	38.5	13.3	26.4	26.0	38.9	20.0	21.7	20.9	30.7
Video-LLaMA2 (Cheng et al. 2024)	13B	CLIP-L	52.7	12.8	13.3	17.0	12.0	19.4	15.7	8.3	18.6	18.9
Video-LLaMA2 (Cheng et al. 2024)	72B	CLIP-L	80.2	53.8	36.7	54.7	54.0	38.9	42.9	16.7	32.6	45.6
InternVL2 (Chen et al. 2024)	76B	InternViT-6B	85.7	51.3	48.3	47.2	52.0	44.4	32.9	15.0	34.9	45.7
Training-Free Methods												
IG-VLM-7B (Kim et al. 2024)	7B	CLIP-L	74.7	33.3	26.7	18.9	24.0	33.3	17.1	15.0	27.9	32.1
SF-LLaVA-7B (Xu et al. 2024)	7B	CLIP-L	68.1	23.1	25.0	34.0	24.0	33.3	22.9	16.7	16.3	32.7
KTV-7B-sparse	7B	CLIP-L	73.6	43.6	35.0	41.5	34.0	36.1	27.1	18.3	20.9	36.5
KTV-7B-normal	7B	CLIP-L	72.5	51.3	35.0	41.5	34.0	38.9	24.3	21.7	25.6	36.1
KTV-7B-dense	7B	CLIP-L	69.2	48.9	33.3	39.6	34.0	38.9	27.1	**23.3**	27.9	36.9
IG-VLM-34B (Kim et al. 2024)	34B	CLIP-L	68.1	35.9	21.7	30.1	34.0	50.0	20.0	6.7	20.1	33.3
SF-LLaVA-34B (Xu et al. 2024)	34B	CLIP-L	76.9	43.6	36.7	39.6	44.0	47.2	27.1	8.3	32.6	43.6
KTV-34B-sparse	34B	CLIP-L	81.3	51.3	**53.3**	**47.2**	**50.0**	52.8	**37.1**	11.7	34.9	44.8
KTV-34B-normal	34B	CLIP-L	81.3	**56.4**	46.7	45.3	46.0	52.8	32.9	8.3	**41.9**	44.2
KTV-34B-dense	34B	CLIP-L	**85.7**	51.3	48.7	**47.2**	48.0	**58.3**	35.7	10.0	37.2	**45.0**

Table 2: Performance of each subcategory in MLVU-Test, compared with the same or similar size of LLMs.

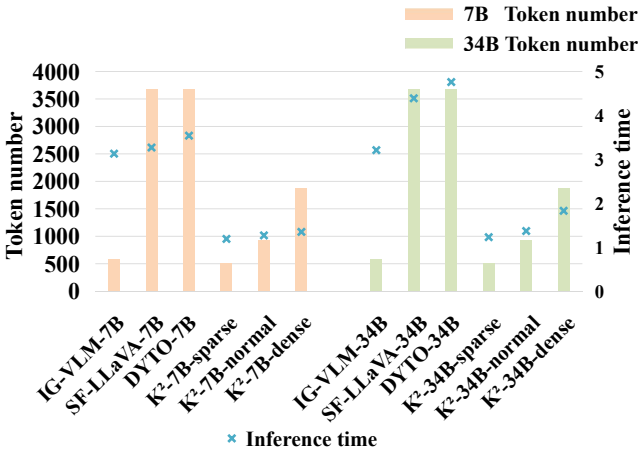

Figure 4: Number of visual tokens fed to the LLM and the average inference time per sample on NExTQA.

further improves accuracy and consistently surpasses IG-VLM-34B and SF-LLaVA-34B. KTV-34B-dense attains the highest accuracy on EgoSchema, IntentQA, and STAR. In addition, not only does it perform comparably to DYTO on NExT-QA and VideoMME, but it also requires only 1872 visual tokens, almost half of DYTO's 3680 tokens. Also, we compare the number of tokens and the average VLM inference time in Fig.4. KTV uses fewer tokens and less inference time than other methods.

The above results demonstrate that KTV outperforms other training-free methods with a smaller LLM on the STAR benchmark and even surpasses some training-based methods on NExTQA, EgoSchema and VideoMME, which well validates the effectiveness and efficiency of KTV for training-free video understanding.

MLVU-Test We report the mean average accuracy (M-Avg) and subcategory-wise accuracy on MLVU-Test in Tab.2. KTV-7B-dense achieves an M-Avg of 36.9, exceeding IG-VLM-7B and SF-LLaVA-7B by $+4.8\%$ and $+4.2\%$, respectively. KTV-34B-dense further raises M-Avg to 45.0, which not only surpasses all training-free baselines but also outperforms several training-based methods with comparable or even larger LLMs. For instance, it exceeds VILA-1.5-40B by $+0.8\%$, and is only marginally lower than Video-LLaMA and InternVL2 by -0.6% and -0.7%, despite their use of twice the LLM size. In addition, across all subcategories, KTV not only consistently outperforms other training-free approaches but also surpasses many training-based methods in specific tasks. For example, KTV-34B-sparse achieves the highest M-Avg of 53.3 in Needle QA among all methods. KTV-34B-dense ranks first in Sports QA with 58.3 M-Avg. Besides, KTV-7B-dense achieves the highest M-Avg of 23.3, among all methods with the smallest 7B LLM.

This comprehensive strength underscores that our keyframe and key token selection strategy effectively preserves both global and fine-grained visual cues, supporting strong reasoning across diverse video understanding tasks.

Ablation Study We conduct an ablation study to assess the key components of KTV in Tab. 3. When visual tokens are pruned based solely on either importance or redundancy, performance consistently degrades. For instance, on EgoSchema, removing RS and IS leads to drops of -1.0% and -1.8%, respectively. This demonstrates the complementarity and effectiveness of combining the two pruning strategies. Moreover, we observe that CK+IS consistently outperforms CK+RS across all benchmarks, which suggests that semantic relevance plays a more crucial role than visual diversity in guiding effective token selection for VideoQA.

In addition to visual token pruning, we also evaluate

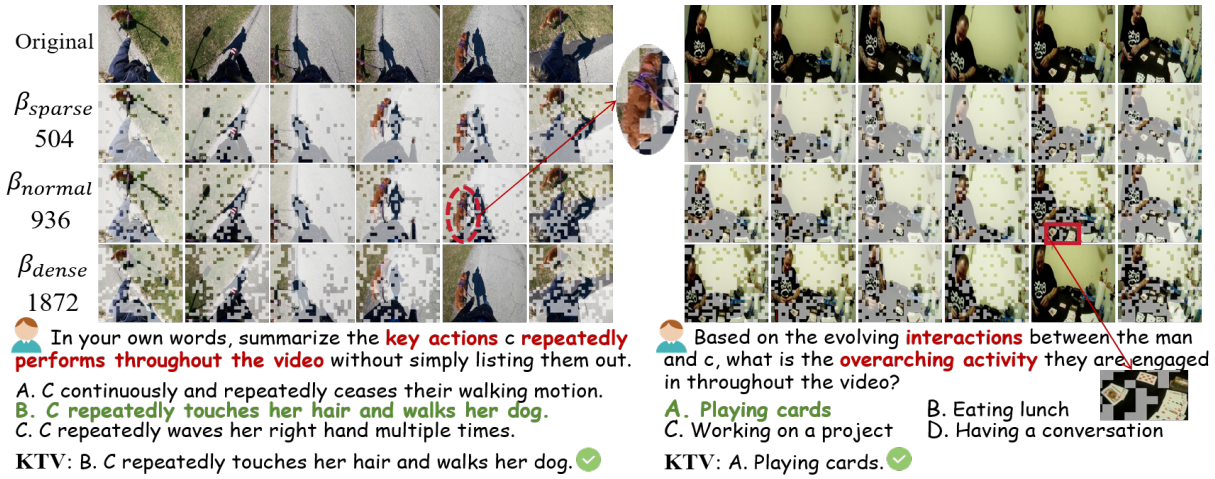

Figure 5: Visualization of KTV using different pruning rate β. The translucent patches are the pruned visual tokens.

Components	NExTQA	EgoSchema	IntentQA
CK+IS+RS	**65.1**	**50.4**	**61.3**
QK+IS+RS	63.5	46.1	60.3
CK+IS	64.8	49.4	61.0
CK+RS	64.6	48.6	60.9
RS+IS	64.1	48.0	60.7

Table 3: Ablation of KTV-7B-normal on NExTQA, Egoschema, and IntentQA. CK: Cluster Keyframes; QK: Question-relevant Keyframes; IS: Importance Score; RS: Redundancy Score;

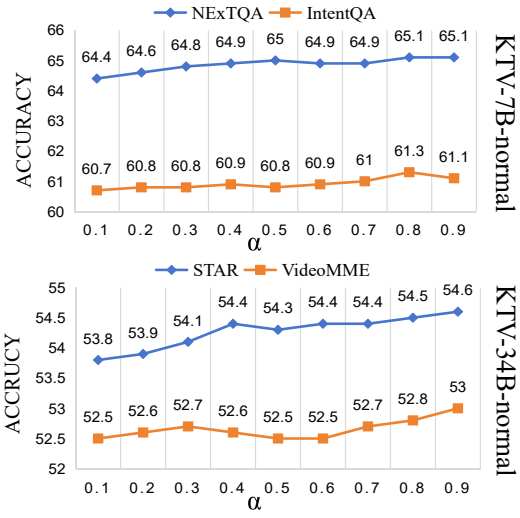

Figure 6: Accuracy under different setting of α.

the efficacy of Cluster Keyframes(CK) selection. Compared with selecting keyframes based on temporal order or question, which respectively correspond to IS+RS setting and QK+IS+RS setting, our method of clustering frame-level visual features yields notable improvements, especially on long videos, *e.g.*, on EgoSchema, CK achieves a significant improvement of $+2.6\%$ and $+4.3\%$ respectively. This gap reveals the limitation of uniform sampling, which may either missing brief yet crucial moments or redundantly sample static scenes, and question-relevant selection, which selects keyframes excessively related to the question, losing the global information. In contrast, our clustering strategy yields visually diverse and representative frames, enabling a more comprehensive understanding of video content.

As for α, we report accuracy under different values in Fig.6. Accuracy generally increases with higher α, peaking at values of 0.8 or 0.9. In addition, our method consistently outperforms the baseline and single-criterion pruning (*i.e.*, importance-only or redundancy-only), highlighting the complementary strengths of the two scores.

Visualization of KTV Fig. 5 shows the visualization of KTV under different pruning rates β. Despite significantly reducing the number of visual tokens, our key token selection strategy effectively retains the most salient visual cues, ensuring that the LLM receives sufficient contextual information to produce accurate answers.

Conclusion

In this paper, we propose **KTV**, a novel training-free, two-stage framework that addresses the critical challenges of temporal and spatial redundancy in video understanding for LLMs. KTV first clusters frame features to select a concise set of keyframes, then selects key visual tokens using a balanced combination of their importance and redundancy scores, which yields a concise yet highly informative representation of video content. Extensive experiments show that KTV consistently achieves state-of-the-art or competitive performance across diverse VideoQA benchmarks, while significantly reducing the number of visual tokens and VLM inference time. These results highlight KTV's potential not only for training-free approaches but also as a strong foundation for training-based video understanding methods.

Acknowledgments

This work was supported by the Guangdong Province's Key Research and Development Project (2024B0101010003), the National Natural Science Foundation of China (624B2119, 62402015), the Postdocotoral Fellowship Program of CPSF (GZB20230024), the China Postdoctoral Science Foundation (2024M750100), and the Start-up Grant (No.9382010) of the City University of Hong Kong.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

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.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5006–5015.

Han, K.; Guo, J.; Tang, Y.; He, W.; Wu, E.; and Wang, Y. 2024. Free video-llm: Prompt-guided visual perception for efficient training-free video llms. *arXiv preprint arXiv:2410.10441*.

Kim, W.; Choi, C.; Lee, W.; and Rhee, W. 2024. An Image Grid Can Be Worth a Video: Zero-shot Video Question Answering Using a VLM. *arXiv preprint arXiv:2403.18406*.

KunChang Li, Y. W. Y. L. W. W. P. L. Y. W. L. W., Yinan He; and Qiao, Y. 2023. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Wei, P.; Han, W.; and Fan, L. 2023b. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11963–11974.

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

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5971–5984.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. llava. *Advances in neural information processing systems*, 36: 34892–34916.

Maazi, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *62nd Annual Meeting of the Association-for-Computational-Linguistics (ACL)/Student Research Workshop (SRW), Bangkok, THAILAND, aug 11-16, 2024*, 12585–12602. ASSOC COMPUTATIONAL LINGUISTICS-ACL.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*.

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.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

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*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Wu, W. 2024. Freeva: Offline mllm as training-free video assistant. *arXiv preprint arXiv:2405.07798*.

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

Zhang, H.; Li, X.; and Bing, L. 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*, 543–553.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024a. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. *arXiv e-prints*, arXiv–2412.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2025. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *International Conference on Machine Learning*.

Zhang, Y.; Zhao, Z.; Chen, Z.; Ding, Z.; Yang, X.; and Sun, Y. 2024b. Beyond training: Dynamic token merging for zero-shot video understanding. *arXiv preprint arXiv:2411.14401*.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Liang, Z.; Xiao, S.; Qin, M.; Yang, X.; Xiong, Y.; Zhang, B.; et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13691–13701.