MMG-Vid: Maximizing Marginal Gains at Segment-level and Token-level for Efficient Video LLMs

Junpeng Ma[1,2], **Qizhe Zhang**[1], **Ming Lu**[1†], **Zhibin Wang**[3†], **Qiang Zhou**[3], **Jun Song**[3], **Shanghang Zhang**[1 ✉]

[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2] Institute of Science and Technology for Brain-inspired Intelligence, Fudan University
[3]Taobao & Tmall Group of Alibaba

Abstract

Video Large Language Models (VLLMs) excel in video understanding, but their excessive visual tokens pose a significant computational challenge for real-world applications. Current methods aim to enhance inference efficiency by pruning redundant visual tokens. However, they do not consider the dynamic characteristics and temporal dependencies of video frames, perceiving video understanding as a multi-frame task. To address these challenges, we propose **MMG-Vid**, a novel training-free visual token pruning framework that removes redundancy by **M**aximizing **M**arginal **G**ains at both segment-level and token-level. Specifically, we first divide the video into segments based on frame similarity, and then dynamically allocate the token budget for each segment to maximize the marginal gain of each segment. Subsequently, we propose a temporal-guided DPC algorithm that jointly models inter-frame uniqueness and intra-frame diversity, thereby maximizing the marginal gain of each token. By combining both stages, MMG-Vid can maximize the utilization of the limited token budget, significantly improving efficiency while maintaining strong performance. Extensive experiments on multiple benchmarks demonstrate that MMG-Vid can maintain over **99.5%** of the original performance, while effectively reducing **75%** visual tokens and accelerating the prefilling stage by **3.9x** on LLaVA-OneVision-7B.

Introduction

Recently, Video Large Language Models (VLLMs) have shown impressive abilities in understanding video content (Li et al. 2024a; Zhang et al. 2024e), which typically rely on densely sampling video frames, resulting in a substantial volume of visual tokens (Zhang et al. 2024b; Chen et al. 2024b). However, processing these visual tokens is computationally demanding due to the quadratic complexity of the self-attention mechanism (Vaswani et al. 2017; Liu et al. 2025c). This bottleneck leads to substantial inference latency and memory overhead, severely hindering their deployment in resource-constrained scenarios. Consequently, developing token compression algorithms to effectively reduce redundancy while preserving critical semantic information has become a primary focus of current research (Tao et al. 2025; Huang, Zhou, and Han 2024; Fu et al. 2024b).

Current methods do not effectively consider the dynamic characteristics and temporal dependencies of video frames, as they view video understanding as a multi-frame task, leading to suboptimal performance. The shortcomings of previous works can be summarized from two perspectives: 1) Static token budget: As shown in Figure 1, previous approaches typically assign a static token budget to all video segments, disregarding the dynamic characteristics (Tao et al. 2025; Shen et al. 2025; Sun et al. 2025). For example, a complex segment with multiple human interactions carries significantly richer semantic information than a relatively static landscape shot, and therefore should be allocated more computational resources. While recent works like VidCom^2 (Liu et al. 2025a) attempt to allocate token budgets based on the importance, their static allocation strategy fails to reflect the inherently dynamic characteristics. The importance of a segment should not be considered static but rather dynamically adjusted based on information already retained. We argue that the true value of a segment lies in **maximizing segment-level marginal gain**. 2) Disjoint pruning strategy: Most prior works treat diversity and importance as two separate metrics (Huang, Zhou, and Han 2024; Fu et al. 2024b; Shao et al. 2025). A typical pruning pipeline involves two stages: temporal redundancy removal and spatial importance selection, executed in a strictly sequential manner. However, a token that is spatially important may be discarded too early in the first stage, while a token kept in the second stage may not be globally optimal. We contend that token selection should also consider the temporal dependencies of video frames, aiming for globally optimal pruning by **maximizing token-level marginal gain**.

To address these limitations in a unified manner, we propose MMG-Vid, a novel training-free video token pruning framework that **M**aximizes **M**arginal **G**ains at segment-level and token-level to preserve critical information while reducing computational cost. MMG-Vid consists of three stages: **(i) Similarity-based Frame Segmentation:** Long videos often contain multiple discontinuous segments, and simply pruning individual frames does not take advantage of the temporal structure inherent in videos. To address this, we first divide the video into semantically coherent segments by analyzing the similarities between adjacent frames. This approach allows us to prune more effectively by leveraging the temporal relationships within each segment. **(ii)**

[†]Project leader.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

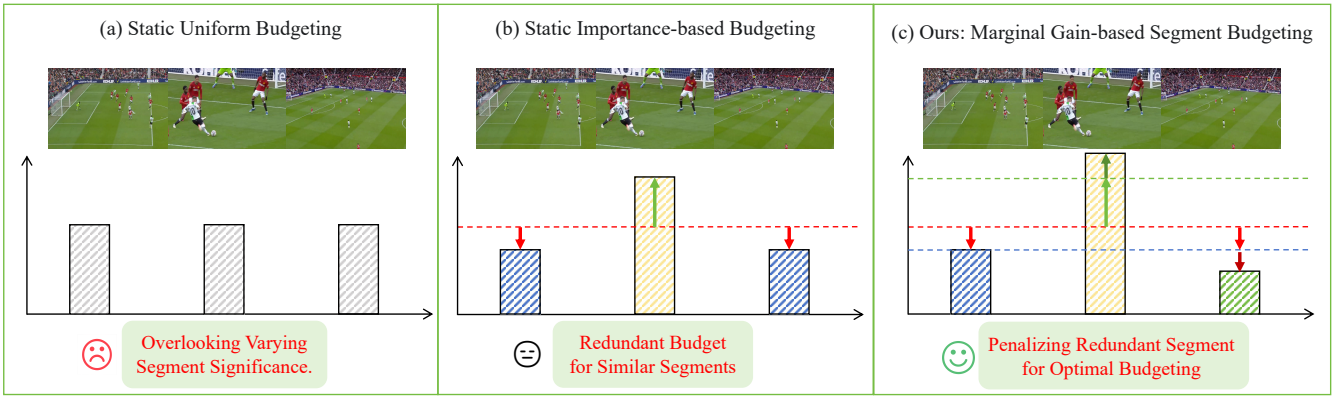

Figure 1: **Comparison of different budget allocation strategies.** (a) Static uniform budgeting overlooks varying segment significance. (b) Static importance-based budgeting acknowledges the significance of segments but wastes resources by allocating redundant budgets to two visually similar static segments (the first and third). (c) Our marginal gain-based segment budgeting further reduces redundancy by penalizing a segment's budget if its information is already included in previously selected segments. This results in an optimal allocation of the budget that takes into account the dynamic characteristics.

Marginal Gain-based Segment Budgeting: As previously mentioned, segments differ in their dynamic characteristics, and distributing the budget evenly can lead to the loss of important information in high-density segments. To tackle this issue, we propose an iterative dynamic budget allocation process. At each step, it evaluates the marginal gain of the current segment, ensuring that we make the best use of limited computational resources at the segment level. **(iii) Marginal Gain-based Token Pruning:** To address the limitations of disjoint token pruning, we propose a new approach called temporal-guided density peak clustering (TG-DPC). This algorithm models both inter-frame distinctiveness and intra-frame diversity, allowing it to effectively reduce spatiotemporal redundancy in a progressive manner. Ultimately, TG-DPC aims to maximize marginal gains at the token level.

We apply our MMG-Vid to the most widely used VLLMs, LLaVA-Video and LLaVA-OneVision, and evaluate on multiple video question answering benchmarks. Empirical results demonstrate that MMG-Vid substantially reduces the computational overhead of VLLMs while robustly preserving their performance across a wide range of retention ratios. For instance, when applied to LLaVA-OneVision-7B, MMG-Vid prunes **75%** of video tokens, achieving a **3.9x** speedup in the prefilling stage, all while maintaining **99.5%** of the model's original performance on average.

The key contributions are summarized as follows:

- We reformulate video token pruning as a problem of maximizing marginal gain, addressing the limitations of previous methods at both segment-level and token-level.
- We present MMG-Vid, a novel framework for training-free video token pruning that utilizes marginal gain-based segment budgeting and token pruning to efficiently manage limited computational resources.
- Our MMG-Vid achieves state-of-the-art performance on both LLaVA-Video and LLaVA-OneVision accross various benchmarks, substantially improving inference efficiency while largely maintaining original performance.

Related Work

Video Large Language Models

Typical Video Large Language Models (VLLMs) employ visual encoders and projectors to independently encode each frame into visual tokens, which are then concatenated with the user query and fed into the LLM. Recent studies have explored diverse methods for advancing video understanding (Song et al. 2024; Zhang et al. 2025a). LongVA (Zhang et al. 2024b) enhances long-video understanding by extending the LLM's context length. Qwen2-VL (Wang et al. 2024) improves temporal awareness using M-RoPE. LLaVA-OneVision (Li et al. 2024a) unifies image and video tasks and efficiently compresses tokens by bilinear interpolation. LLaVA-Video (Zhang et al. 2024e) uses newline tokens for spatial-temporal grounding. However, excessive video frames yield a vast quantity of visual tokens, limiting the practical application of VLLMs due to the quadratic computational complexity (Vaswani et al. 2017).

Token Compression for VLLMs

Token compression directly prunes visual tokens to improve inference efficiency in VLLMs. Although training-aware paradigms (Li, Wang, and Jia 2024; Ye et al. 2025; Zhang et al. 2025c; Yang et al. 2025b) effectively reduce sequence length while maintaining model performance, they require modifications during training and are thus computationally expensive. Consequently, an increasing number of works have begun exploring training-free approaches for enhancing inference efficiency (Zhang et al. 2024d; Xing et al. 2024; Shang et al. 2024; Han et al. 2024; Liu et al. 2025b; Zhang et al. 2024c, 2025b). FastV (Chen et al. 2024a) first identifies inefficient cross-modal attention within the language model and evaluates the importance of visual tokens based on the attention received from text tokens. VisionZip (Yang et al. 2025a) selects important tokens using attention from the [CLS] token in the visual encoder

and then merges the remaining ones. DivPrune (Alvar et al. 2025) formulates the token pruning problem as a Max-Min Diversity Problem and iteratively retains visual tokens using a greedy algorithm. However, none of the aforementioned methods are designed specifically for video understanding and ignore the temporal connections between different video frames. To address this, PruneVid (Huang, Zhou, and Han 2024) clusters frames into segments and classifies tokens within each segment as either static or dynamic, applying different pruning strategies accordingly. FrameFusion (Fu et al. 2024b) first merges tokens across frames and then selects important tokens. Inspired by DivPrune, we propose a training-free video pruning strategy that maximizes the marginal gains at both segment-level and token-level, enabling more effective use of limited computational resources while maintaining overall performance.

Methodology

In this section, we formulate the token pruning task as a constrained optimal subset selection problem, and present our MMG-Vid framework to derive a feasible solution. MMG-Vid employs the principles of Maximal Marginal Gains, primarily comprising three components: 1) Similarity-based Frame Segmentation, 2) Marginal Gain-based Segment Budgeting, and 3) Marginal Gain-based Token Pruning.

Problem Formulation

First, we claim that the essence of visual token pruning is to identify and select an optimal subset $\mathcal{S}_{\text{sub}}$ from a given visual token set $\mathcal{S}_{\text{v}}$ that maximizes a predefined quality function, under the constraint of a specific retention ratio R. We formally model this as a constrained subset selection problem.

Let V denote a video, represented by a set of tokens $\mathcal{T}_v = \{t_1, t_2, \ldots, t_N\}$, where N is the total number of tokens across all frames. We first define a quality function for a token subset that jointly quantifies the representativeness and diversity of each token within it, as expressed in Eq (1):

$$Q(\mathcal{T}_{\text{sub}}) = \underbrace{\sum_{t_i \in \mathcal{T}_{\text{sub}}} I(t_i)}_{\text{Total Representativeness}} - \beta \underbrace{\sum_{\substack{t_i, t_j \in \mathcal{T}_{\text{sub}} \\ i \neq j}} D(t_i, t_j)}_{\text{Total Redundancy}} \quad (1)$$

Our optimization objective is to find an optimal subset $\mathcal{T}^*_{\text{sub}}$ that maximizes the quality function, subject to a constraint determined by the pre-defined retention ratio R:

$$\begin{aligned} \mathcal{T}^*_{\text{sub}} = \underset{\mathcal{T}_{\text{sub}} \subseteq \mathcal{T}_{\text{v}}}{\text{argmax}} \quad & Q(\mathcal{T}_{\text{sub}}) \\ \text{subject to} \quad & |\mathcal{T}_{\text{sub}}| \leq R \cdot N \end{aligned} \quad (2)$$

This is a variant of the Max-Min Diversity Problem, which is known to be NP-hard (Porumbel, Hao, and Glover 2011; Parreño, Álvarez-Valdés, and Martí 2021). A brute-force search over all possible subsets is computationally intractable. Given the infeasibility of direct global optimization, we resort to a three-stage framework under the guidance of the MMG principle to find an approximate solution.

MMG-Vid

We propose a novel three-stage framework, MMG-Vid, an efficient approximation approach by decomposing the complex global token selection problem into a series of more tractable sub-problems, which are in turn solved via our proposed three-stage approach.

Stage-1: Similarity-based Frame Segmentation

We begin by reformulating the problem space. Motivated by the observation that video content is non-uniformly distributed across the temporal dimension, processing the entire video as a monolithic, unstructured token set fails to exploit its inherent semantic structure. To align our approach with this intrinsic structure, we first segment the video into K semantically coherent segments based on temporal dynamics.

Let $f_t \in \mathbb{R}^d$ be the feature embedding of frame F_t obtained by average pooling its corresponding visual tokens. We compute the cosine similarity between adjacent frames:

$$\begin{aligned} \text{sim}(F_t, F_{t+1}) &= \frac{f_t \cdot f_{t+1}}{\|f_t\| \|f_{t+1}\|} \\ B &= \{i | \text{sim}(F_t, F_{t+1}) < \tau\} \end{aligned} \quad (3)$$

We define an initial set of segment boundaries B by identifying frames where the similarity to the next frame drops below a threshold τ. An initial set of segments $\{S_1, S_2, ...\}$ is thus formed. To ensure temporal coherence and avoid fragmented, single-frame segments, we enforce a minimum segment length. Any segment S_k with $|S_k| = 1$ is merged into its most similar adjacent segment:

$$\begin{aligned} S_{k^*} &\leftarrow S_{k^*} \cup S_k, \\ \text{where } k^* &= \underset{j \in \{k-1, k+1\}}{\arg\max} \ \text{sim}(S_k, S_j) \end{aligned} \quad (4)$$

This process yields a final set of K segments, $\{S_1, S_2, ..., S_K\}$, where each segment represents a temporally and visually consistent segment of the video.

Stage-2: Marginal Gain-based Segment Budgeting

After segmenting the video into clips in Stage-1, the problem is thus reformulated as selecting optimal token subsets from each segment, thereby decomposing the overall task into a set of independent sub-problems. Prior works typically assign a uniform token budget to each segment and then perform intra-segment pruning, overlooking the inherent heterogeneity and inter-dependencies among them. Instead, to avoid losing critical information in content-dense segments while wasting the computational budget on content-sparse ones, we propose a budget allocation strategy based on dynamic marginal value, aiming to devise an optimal token allocation strategy within a constrained total budget, such that the resulting subset of tokens can maximize the semantic diversity and coverage of the entire video.

First, to prevent critical information loss, we establish a universal minimum retention ratio R_{min} for all segments. This guarantees a baseline representation for every segment, and the remaining discretionary token budget is calculated as $R_{extra} = R - R_{min}$, which is allocated iteratively based on the marginal gain of each segment.

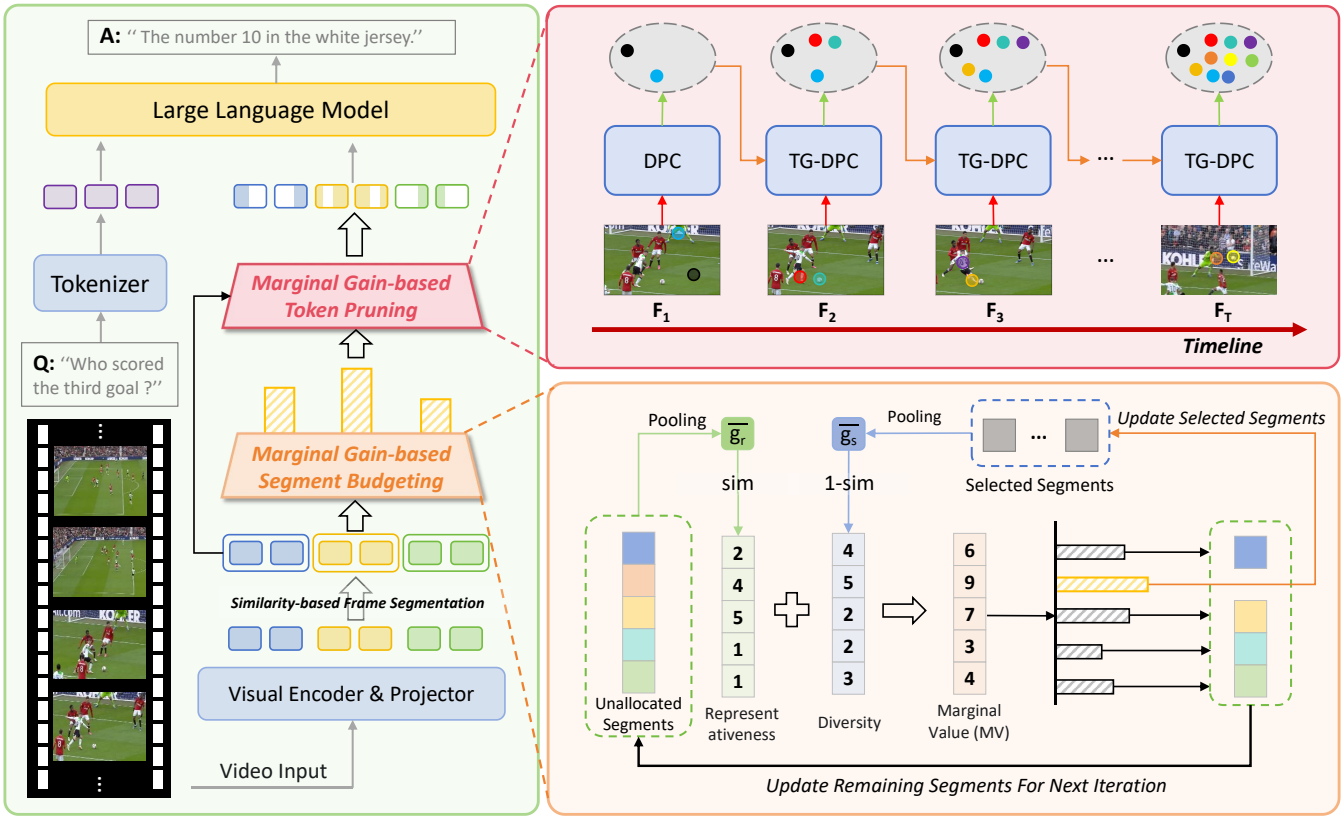

Figure 2: **Overall framework. Segment-level (Bottom-Right):** We iteratively calculate the marginal gain (a combination of representativeness and diversity) for each segment to dynamically allocate budget, prioritizing more informative segments. **Token-level (Top-Right):** Our proposed TG-DPC progressively prunes each frame by selecting tokens that are both salient within the frame and novel across the temporal dimension, guided by the set of previously selected tokens

Then, we argue that a segment's value is not absolute but is relative to the information already processed. Therefore, we sequentially and iteratively determine an optimal ordering and importance score for each segment. We begin with an empty set of selected segments. At each step, for all remaining unselected segments, we compute a Marginal Value (MV) to represent the optimal trade-off between the segment's intrinsic relevance and its diversity with respect to the set of segments already selected.

Let $\mathbf{g}_k$ be the representation of segment S_k, obtained by averaging all tokens within the segment. Additionally, let $\mathbb{S}_{\text{sel}}$ be the set of segments for which the budget has been allocated, and $\mathbb{S}_{\text{rem}}$ be the set of segments that remain unselected. We posit that the Marginal Value (MV) of a segment is co-determined by its representativeness, defined as its thematic relevance to the remaining segments, and its diversity, which reflects its capacity to introduce novel information:

$$\text{MV}(S_k|\mathbb{S}_{\text{sel}}) = \lambda \underbrace{\text{sim}(\mathbf{g}_k, \overline{\mathbf{g}_r})}_{\text{Representativeness}} + (1-\lambda) \underbrace{(1 - \text{sim}(\mathbf{g}_k, \overline{\mathbf{g}_s}))}_{\text{Diversity}}$$

$$\text{where} \quad \overline{\mathbf{g}_r} = \frac{1}{|\mathbb{S}_{\text{rem}}|} \sum_{S_i \in \mathbb{S}_{\text{rem}}} \mathbf{g}_i, \quad \overline{\mathbf{g}_s} = \frac{1}{|\mathbb{S}_{\text{sel}}|} \sum_{S_j \in \mathbb{S}_{\text{sel}}} \mathbf{g}_j. \tag{5}$$

λ is a parameter to balance representativeness and diversity. The segment with the highest MV is then chosen to compute its token budget via Z-score (illustrated in eq (6)).

$$R_k = R_{min} + R_{extra} \cdot \frac{\text{MV}(S_k) - mean(\text{MV})}{std(\text{MV})} \tag{6}$$

This process is repeated until all segments are allocated. Crucially, the marginal value $\text{MV}(S_k)$ utilized in the allocation is not a pre-computed, static metric but a dynamic score determined at the precise selection moment of segment S_k. This ensures that the extra budget R_{extra} is allocated based on a segment's actual marginal contribution to the video's overall information content, thereby maximizing information density under a given computational constraint.

Stage-3: Marginal Gain-based Token Pruning

After determining the token budget, we proceed to the fine-grained token pruning within each segment. Simply applying a pruning algorithm to each frame independently disregards the inherent temporal coherence of videos, leading to two primary issues: **1) Information Redundancy:** Static regions, such as backgrounds, may be repeatedly selected across consecutive frames, yielding highly similar tokens. **2) Loss of Dynamic Information:** Subtle yet crucial changes,

such as object movements or pose alterations, might be overlooked if are not sufficiently salient within a single frame.

To overcome these limitations, we propose a dynamic selection strategy based on temporal information gain that treats the initial frame as a visual anchor and progressively enriches the visual information in subsequent frames.

First Frame Token Pruning: For the first frame, we employ the standard DPC-KNN algorithm to select an optimal token subset. For each token t_i, we compute its local density:

$$\rho_i = \exp\left(-\frac{1}{n}\sum_{t_j \in \mathrm{kNN}(t_i)} \|t_i - t_j\|_2^2,\right), \tag{7}$$

where $\mathrm{kNN}(t_i)$ denotes the k-nearest neighbors of token t_i. Then, for each token, we compute its minimum distance δ_i to any other token with higher density :

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} & \|t_i - t_j\|_2^2, \quad \text{if } \exists j \text{ s.t. } \rho_j > \rho_i, \\ \max\limits_{j} & \|t_i - t_j\|_2^2, \quad \text{otherwise,} \end{cases} \tag{8}$$

where each token score is $\gamma_i = \rho_i \cdot \delta_i$. Tokens with higher scores are selected as initial cluster centers, which are both representative (high ρ) and unique (high δ). We select $R_k \cdot M$ tokens with the highest score (M is the token number of each frame) and add them to the selected set $\mathcal{T}_k^{guide}$.

Subsequent Frame Token Pruning: Starting from the second frame, we introduce the **Temporal-Guided-DPC (TG-DPC)** algorithm. For each candidate token in the current frame F_i, we redefine its density and distance properties using the set of previously selected tokens $\mathcal{T}_k^{guide}$.

• **Temporal Relevance Density (ρ^t):** We replace the conventional local density with temporal relevance density. The ρ_i^t of token i is no longer determined by its neighbors in the current frame but by the historically selected tokens.

$$\rho_i^t = 1 - \exp\left(-\frac{1}{k}\sum_{j=1}^{k}\left\|t_i - t_g^{(j)}\right\|_2^2\right) \tag{9}$$

where $t_g^{(j)}$ is the j-th nearest token to t_i in $\mathcal{T}_k^{guide}$. It suggests that if a token differs from the selected tokens, the marginal gain is significant and should be chosen.

• **Intra-Frame Separation (δ^t):** We retain the concept in standard DPC-KNN, but its calculation is now based on our new temporal relevance density ρ^t. For a token t_i, it intra-frame separation δ_i^t is defined as the minimum distance to any other token t_m within the same frame that has a higher ρ^t (i.e., $\rho_m^t > \rho_i^t$) :

$$\delta_i^t = \begin{cases} \min\limits_{m:\rho_m^t > \rho_i^t} & \|t_i - t_m\|_2^2, \quad \text{if } \exists m \text{ s.t. } \rho_m^t > \rho_i^t, \\ \max\limits_{m} & \|t_i - t_m\|_2^2, \quad \text{otherwise.} \end{cases} \tag{10}$$

Therefore, the final score of each token is defined as $\gamma_i^t = \rho_i^t \cdot \delta_i^t$. A higher score indicates that a token is both novel (dissimilar to past information) and core (distinct within the current frame). We then also select the top $R_k \cdot M$ tokens based on the new score γ_i^t and upadte the selected set $\mathcal{T}_k^{guide} \leftarrow \mathcal{T}_k^{guide} \bigcup \{\mathcal{T}_i^*\}$. This temporally-aware selection process greedily builds a diverse set of tokens across the entire segment and yields two key benefits:

1. Capturing Dynamics: When an object in the segment moves or deforms (e.g., a person waving hands), the tokens with significant marginal gains will be preserved due to their novelty compared to those from the previous static state.

2. Background Completion: If the tokens chosen from the initial frame focus on a foreground object, the algorithm will give a higher ρ_i^t value to background tokens in later frames due to their significant dissimilarity.

Experiments

Experimental Settings

Benchmarks and Baselines: We employ lmms-eval (Zhang et al. 2024a) to evaluate our method on four widely-used video understanding benchmarks: MVBench (Li et al. 2024b), LongVideoBench (Wu et al. 2024), MLVU (Zhou et al. 2024), and VideoMME (Fu et al. 2024a). For comparison, we compare our method against several representative open-source methods, including FastV (Chen et al. 2024a), VisionZip (Yang et al. 2025a), PruneVid (Huang, Zhou, and Han 2024), and FrameFusion (Fu et al. 2024b).

Implementation Details: MMG-Vid is implemented into two VLLMs: LLaVA-OneVision-7B (Li et al. 2024a) and LLaVA-Video-7B (Zhang et al. 2024e). All experiments are conducted on NVIDIA H100-SXM5-80GB GPUs. Consistent with the official settings, we set the number of input frames to 32 for LLaVA-OneVision and 64 for LLaVA-Video. Furthermore, we set the similarity threshold τ to 0.95 and λ to 0.5 for all experiments. For FastV, we prune tokens at the 2nd layer. For all intra-LLM methods, to ensure a fair comparison, we use the *equivalent retention ratio* for token number calculation, which is defined as the average percentage of tokens processed across all layers of the LLM. For PruneVid, which compresses both visual tokens and the KV Cache, we evaluate only on its token compression strategy.

Main Results

To comprehensively evaluate the performance of our proposed MMG-Vid method, we conducted an extensive comparison against four advanced token compression methods on the LLaVA-Video and LLaVA-OneVision models.

(i) State-of-the-art Performance: As detailed in Table 1, we established four distinct token Retention Ratios—25%, 20%, 15%, 10%—to thoroughly assess the performance of MMG-Vid and evaluate its robustness under varying compression intensity. The experimental results demonstrate that MMG-Vid consistently achieves optimal performance across all configurations, significantly outperforming all other baseline methods. Specifically, with 75% of tokens compressed, MMG-Vid achieves a total average accuracy of **98.1%**, whereas the best-performing baseline reached 96.8%. As the compression ratio increases, the performance degradation of our method is substantially less pronounced

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Overall	Short	Medium	Long		
Duration	16s	1∼60min	3∼120min	1∼60min	1∼3min	3∼30min	30∼60min	Score	%
LLaVA-Video	59.2	58.9	67.4	64.3	77.3	62.3	53.2	62.4	100.0
Retention Ratio: 25%									
FastV (ECCV24)	52.1	54.8	57.8	58.6	68.7	58.4	48.7	55.8	89.4
VisionZip (CVPR2025)	56.0	58.0	64.4	61.9	73.4	**60.6**	51.6	60.1	96.3
PruneVid (ACL2025)	55.0	57.9	64.1	60.5	72.2	58.6	50.7	59.4	95.2
FrameFusion (ICCV2025)	56.5	57.7	65.9	61.3	72.8	59.3	51.9	60.4	96.8
MMG-Vid	**57.6**	**58.8**	**66.2**	**62.3**	**74.7**	60.0	**52.1**	**61.2**	**98.1**
Retention Ratio: 20%									
FastV (ECCV24)	50.8	52.4	55.2	57.3	65.8	57.2	48.8	53.9	86.4
VisionZip (CVPR2025)	55.7	56.3	63.3	60.9	71.1	59.4	**52.2**	59.1	94.7
PruneVid (ACL2025)	54.6	57.3	63.3	60.2	73.0	58.2	49.2	58.9	94.4
FrameFusion (ICCV2025)	55.3	57.4	63.9	60.8	72.2	59.8	50.3	59.4	95.2
MMG-Vid	**57.2**	**58.6**	**64.9**	**61.4**	**73.4**	**60.2**	50.6	**60.5**	**97.0**
Retention Ratio: 15%									
FastV (ECCV24)	46.9	49.8	53.4	54.0	60.9	54.4	46.7	51.0	81.7
VisionZip (CVPR2025)	55.2	56.3	62.2	60.3	70.3	58.8	**51.7**	58.5	93.8
PruneVid (ACL2025)	54.0	56.2	63.3	59.4	71.2	57.7	49.3	58.2	93.3
FrameFusion (ICCV2025)	54.3	54.9	61.3	59.3	69.9	58.3	49.7	57.5	92.1
MMG-Vid	**56.1**	**57.7**	**64.8**	**61.1**	**72.3**	**60.1**	50.8	**59.9**	**96.0**
Retention Ratio: 10%									
FastV (ECCV24)	43.2	46.5	53.1	49.6	54.0	50.3	44.6	48.1	77.1
VisionZip (CVPR2025)	53.8	52.9	60.3	58.7	67.4	57.7	**51.1**	56.4	90.4
PruneVid (ACL2025)	53.0	55.7	61.0	58.0	69.3	55.4	49.3	56.9	91.2
FrameFusion (ICCV2025)	52.8	53.0	58.0	56.7	66.3	55.4	48.2	55.1	88.3
MMG-Vid	**54.9**	**56.3**	**63.4**	**59.4**	**71.0**	**57.9**	49.2	**58.5**	**93.8**

Table 1: Comparison of state-of-the-art methods across video understanding benchmarks on LLaVA-Video-7B.

Retention Ratio	Time (ms)		Acc%
	Prefill	Generate	
LLaVA-OV	207.5 (1.0x)	329.7 (1.0x)	100
25%	52.8 (3.9x)	107.2 (3.1x)	99.5
15%	34.6 (6.0x)	78.3 (4.2x)	98.3
LLaVA-Video	277.5 (1.0x)	411.6 (1.0x)	100
25%	65.0 (4.3x)	142.9 (2.9x)	98.1
15%	42.5 (6.5x)	109.6 (3.8x)	96.0

Table 2: **Efficiency comparison of different Retention Ratios.** "Prefill Time": Time for model to generate first token; "Generate Time": Time for model to generate response.

than that of other methods, which demonstrates the efficacy of our marginal gain-based segment budgeting algorithm in dynamically adjusting the token budget to preserve the maximum visual information, especially under extreme compression intensity. When only 10% of tokens are retained, our method still maintains an accuracy of **93.8%**, a significant **16.7%** higher than the 77.1% of FastV, and outperforms the best baseline by **2.6%**. Notably, the superiority of MMG-Vid is particularly evident on short videos across all compression levels (e.g., on MVBench, MMG-Vid is **2.53%** and **1.86%** higher than the best baseline at 20% and 10% retention ratio, respectively). We attribute this to the stronger temporal correlation between sampled frames in shorter videos. Our proposed MMG-Vid is designed to effectively leverage this temporal information, enabling it to achieve more comprehensive coverage of visual content and thus demonstrate superior performance. Besides, to assess the cross-model robustness of MMG-Vid, we extended our evaluation to the LLaVA-OneVision, as detailed in Table 3. The results reveal that MMG-Vid still maintains a consistent superiority over all competing methods at various compression rates. Impressively, it exhibits negligible performance degradation at retention ratios of 25% and 15%, preserving 99.5% and 98.3% of the baseline performance, respectively.

(ii) Superior Inference Efficiency: Beyond Performace, Table 2 presents comprehensive real-world inference latency. MMG-Vid achieves a **3.9x** acceleration in the prefill phase and a **3.1x** acceleration in the generation phase, while maintaining **99.5%** of the model performance on LLaVA-OneVision, significantly reducing computational costs.

Ablation study

Ablation study on different modules To validate the effectiveness of our two strategies, we perform ablation experiments on the LLaVA-Video model with a 25% retention ratio. The results are presented in Figure 3. Replacing

Method	MVBench	LongVideo Bench	MLVU	VideoMME				Average	
				Overall	Short	Medium	Long		
Duration	16s	1∼60min	3∼120min	1∼60min	1∼3min	3∼30min	30∼60min	Score	%
LLaVA-OneVision	57.6	56.6	63.1	58.5	70.1	56.6	48.8	59.0	100.0
Retention Ratio: 25%									
FastV (ECCV24)	54.8	**56.8**	59.3	55.9	66.0	54.6	47.2	56.7	96.1
VisionZip (CVPR2025)	**56.9**	56.0	62.9	58.0	68.9	**57.4**	47.6	58.5	99.2
PruneVid (ACL2025)	55.7	55.1	**63.4**	57.0	68.8	54.4	47.7	57.8	98.0
FrameFusion (ICCV2025)	56.0	54.8	61.7	57.5	68.2	55.7	**48.6**	57.5	97.5
MMG-Vid	56.7	56.6	62.9	**58.6**	**71.2**	56.6	48.1	**58.7**	**99.5**
Retention Ratio: 15%									
FastV (ECCV24)	52.3	51.5	56.3	51.9	58.4	51.7	45.4	53.0	89.8
VisionZip (CVPR2025)	55.7	54.2	60.0	55.5	63.8	54.4	**48.3**	56.4	95.6
PruneVid (ACL2025)	55.0	55.6	**61.9**	56.8	67.9	54.3	48.1	57.3	97.1
FrameFusion (ICCV2025)	55.1	53.0	58.3	55.5	65.8	54.1	46.7	55.5	94.1
MMG-Vid	**56.5**	**55.9**	61.6	**57.9**	**69.6**	**56.0**	48.2	**58.0**	**98.3**

Table 3: Comparison of state-of-the-art methods across video understanding benchmarks on LLaVA-OneVision-7B.

λ in Budgeting	MLVU	VideoMME				Acc %
		Overall	Short	Medium	Long	
Vanilla	67.4	64.3	77.3	62.3	53.2	100.0
1.0	65.2	62.0	74.3	60.2	51.3	96.6
0.8	65.5	61.9	74.2	60.2	51.1	96.7
0.6	**66.2**	62.0	74.3	59.7	51.9	97.3
0.5	66.2	**62.3**	**74.7**	60.0	**52.1**	**97.6**
0.4	65.9	61.8	74.3	59.8	51.3	97.0
0.2	65.4	61.6	74.1	59.6	51.0	96.4
0	66.0	62.0	74.2	**60.3**	51.3	97.2

Table 4: **Ablation study for marginal gain-based segment budgeting on LLaVA-Video (Retention Ratio: 25%).** Our method consists of Representiveness and Diversity, and λ determines the proportion between them.

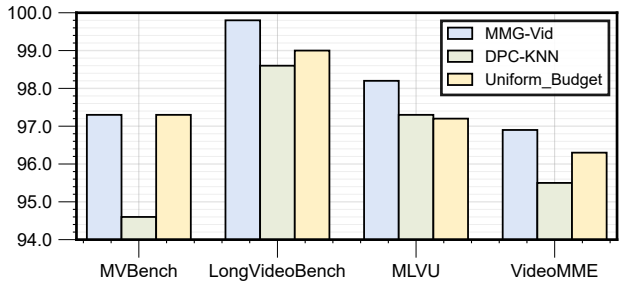

Figure 3: **Ablation study of MMG-Vid's modules on LLaVA-Video (Retention Ratio: 25%).** "DPC-KNN" refers to using the standard DPC-KNN algorithm instead of our proposed TG-DPC. "Uniform Budget" refers to the conventional method of assigning a fixed budget to each frame.

our TG-DPC with a standard DPC-KNN algorithm leads to a significant performance degradation across all applicable benchmarks. This is particularly evident on short-video benchmarks like MVBench, where performance drops from 97.3% to 94.6%, underscoring the superiority of our TG-DPC in capturing complex temporal dynamics compared to conventional methods. Furthermore, compared to a uniform static budget allocation, our dynamic approach boosts performance by 0.8% and 1.0% on long-video benchmarks such as LongVideoBench and MLVU, respectively. This demonstrates that our method can judiciously allocate the budget to maximize information density coverage in long videos, which typically feature more diverse segments.

Ablation study on budgeting Our marginal gain-based segment budgeting module balances two components: Representativeness (selecting core content) and Diversity (reducing redundancy), controlled by the hyperparameter λ. As shown in Table 4, the Diversity-only setting outperforms the Representativeness-only setting by 0.6%. The peak scores on MLVU and VideoMME are obtained with a set to 0.6 and 0.5, respectively. This demonstrates that a balanced integration of both modules is crucial for preserving the most comprehensive video information.

Conclusion

In this paper, we introduce MMG-Vid, a novel training-free visual token pruning framework that significantly reduces the computational cost of VLLMs. Unlike prior methods that relied on static uniform budgeting and disjoint pruning, our MMG-Vid formulates the token pruning task as a constrained subset selection problem and optimizes the selected token subset by consistently maximizing marginal information gain. We achieve more comprehensive coverage of visual representation by first segmenting videos into semantically coherent segments, then dynamically allocating budgets based on contextual relevance, and finally performing unified spatiotemporal pruning via our proposed TG-DPC algorithm. Extensive experiments across multiple VLLMs and benchmarks demonstrate that MMG-Vid effectively reduce the inference latency while maintaining superior performance, enabling the practical deployment of VLLMs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62476011), and by the Beijing Natural Science Foundation (L252060).

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.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024a. 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.

Chen, Y.; Xue, F.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; et al. 2024b. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

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

Han, Y.; Liu, X.; Zhang, Z.; Ding, P.; Wang, D.; Chen, H.; Yan, Q.; and Huang, S. 2024. Filter, correlate, compress: Training-free token reduction for mllm acceleration. *arXiv preprint arXiv:2411.17686*.

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

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.

Li, Y.; Wang, C.; and Jia, J. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 323–340. Springer.

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

Liu, X.; Wang, Z.; Han, Y.; Wang, Y.; Yuan, J.; Song, J.; Zheng, B.; Zhang, L.; Huang, S.; and Chen, H. 2025b. Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models. *arXiv preprint arXiv:2501.05179*.

Liu, X.; Wen, Z.; Wang, S.; Chen, J.; Tao, Z.; Wang, Y.; Jin, X.; Zou, C.; Wang, Y.; Liao, C.; et al. 2025c. Shifting ai efficiency from model-centric to data-centric compression. *arXiv preprint arXiv:2505.19147*.

Parreño, F.; Álvarez-Valdés, R.; and Martí, R. 2021. Measuring diversity. A review and an empirical analysis. *European Journal of Operational Research*, 289(2): 515–532.

Porumbel, D. C.; Hao, J.-K.; and Glover, F. 2011. A simple and effective algorithm for the MaxMin diversity problem. *Annals of Operations Research*, 186: 275–293.

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*.

Shao, K.; Tao, K.; Qin, C.; You, H.; Sui, Y.; and Wang, H. 2025. HoliTom: Holistic Token Merging for Fast Video Large Language Models. *arXiv preprint arXiv:2505.21334*.

Shen, L.; Gong, G.; He, T.; Zhang, Y.; Liu, P.; Zhao, S.; and Ding, G. 2025. Fastvid: Dynamic density pruning for fast video large language models. *arXiv preprint arXiv:2503.11187*.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2024. Pyramidrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025a. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19792–19802.

Yang, S.; Li, J.; Lai, X.; Yu, B.; Zhao, H.; and Jia, J. 2025b. VisionThink: Smart and Efficient Vision Language Model via Reinforcement Learning. *arXiv preprint arXiv:2507.13348*.

Ye, X.; Gan, Y.; Huang, X.; Ge, Y.; and Tang, Y. 2025. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29836–29846.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

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, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024b. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024c. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *arXiv preprint arXiv:2412.01818*.

Zhang, Q.; Liu, M.; Li, L.; Lu, M.; Zhang, Y.; Pan, J.; She, Q.; and Zhang, S. 2025b. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. *arXiv preprint arXiv:2506.10967*.

Zhang, S.; Fang, Q.; Yang, Z.; and Feng, Y. 2025c. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2024d. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

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

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.