DToMA: Training-free Dynamic Token MAnipulation for Long Video Understanding

Bowen Yuan[1] , **Sisi You**[1,3*] , **Bing-Kun Bao**[1,2]
[1]Nanjing University of Posts and Telecommunications,
[2]Pengcheng Laboratory,
[3]State Key Laboratory of Tibetan Intelligence.
yuanbw0925@gmail.com, {ssyou, bingkunbao}@njupt.edu.cn

Abstract

Video Large Language Models (VideoLLMs) often require thousands of visual tokens to process long videos, leading to substantial computational costs, further exacerbated by visual token inefficiency. Existing token reduction and alternative video representation methods improve efficiency but often compromise comprehension abilities. In this work, we analyze the reasoning processes of VideoLLMs in multi-choice VideoQA task, identifying three reasoning stages—shallow, intermediate, and deep stages—that closely mimic human cognitive processing. Our analysis reveals specific inefficiencies at each stage: in shallow layers, VideoLLMs attempt to memorize all video details without prioritizing relevant content; in intermediate layers, models fail to re-examine uncertain content dynamically; and in deep layers, they continue processing video even when sufficiently confident. To bridge this gap, we propose DToMA, a training-free Dynamic Token MAnipulation method inspired by human adjustment mechanisms in three aspects: 1) Text-guided keyframe-aware reorganization to prioritize keyframes and reduce redundancy, 2) Uncertainty-based visual injection to revisit content dynamically, and 3) Early-exit pruning to halt visual tokens when confident. Experiments on 6 long video understanding benchmarks show that DToMA enhances both efficiency and comprehension, outperforming state-of-the-art methods and generalizing well across 3 VideoLLM architectures and sizes. Code is available at https://github.com/yuanrr/DToMA.

1 Introduction

Long video understanding requires models to process and reason over video content spanning several minutes to hours. Large language models (LLMs) [Yang *et al.*, 2024] have shown exceptional comprehension of textual information. Multimodal LLMs (MLLMs) [Li *et al.*, 2024; Wang *et al.*, 2024] further equip LLMs with multimodal perception ability, greatly extending their comprehension of images and short videos. Those advances motivate researchers to investigate methods for effectively understanding long videos.

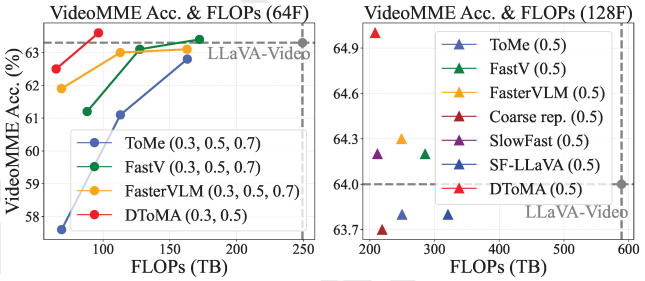

Figure 1: Comparison of the accuracy and FLOPs on VideoMME based on LLaVA-Video between DToMA and existing methods, including token reduction methods at various compression ratios (0.3, 0.5, 0.7) for a 64-frame input, and other efficient video representations at a fixed 0.5 compression ratio with a 128-frame input.

Existing VideoLLMs typically process videos by converting each frame into hundreds of visual tokens, leading to a rapid token increase as video length grows. For instance, LLaVA-Video [Zhang *et al.*, 2024d] generates over 11k tokens for merely 64 frames. However, these visual tokens often have heavy redundancy [Chen *et al.*, 2025], exacerbating the computational challenges. To address this, prior works focus on token reduction and alternative video representation strategies. The former includes pruning visual content outside LLMs [Shang *et al.*, 2024; Shen *et al.*, 2024] and compression mechanisms within LLMs [Chen *et al.*, 2025; Ye *et al.*, 2024]. The latter methods, such as image grids [Kim *et al.*, 2024], slow-fast processing [Xu *et al.*, 2024], or query-based methods [Li *et al.*, 2025], design alternative ways to effectively represent videos. While improving efficiency, they are often at the expense of visual details or temporal dependencies, limiting their comprehension ability. Effective visual token processing remains to be explored for simultaneously improving efficiency and video comprehension.

To enhance both efficiency and comprehension abilities, we explore the internal reasoning mechanisms of VideoLLMs, inspired by human reasoning processes. Human reasoning consists of two key aspects: a general reason-

*Corresponding author

ing process and intuitive adjustment mechanisms [Dundas and Chik, 2011]. Humans generally approach video understanding tasks by first perceiving video content, then analyzing based on questions, and finally inferring based on prior knowledge. For quicker and better comprehension, they employ strategies such as conducting an initial coarse scan to adjust focus, rewatching video to clarify uncertainties or refine their understanding, and ceasing examining visual content once confident in their conclusions. We delve into the internal reasoning mechanisms of VideoLLMs to analyze whether they exhibit reasoning processes similar to humans. Inspired by recent studies [Fang *et al.*, 2024] and [Chen *et al.*, 2025], we analyze reasoning patterns in multi-choice VideoQA tasks through entropy dynamics and cross-attention. Specifically, we identify three distinct reasoning stages in VideoLLMs: 1) **Shallow Layers**: Text strongly attends to visual tokens, facilitating rapid multimodal information exchange. Entropy remains high, suggesting an information-gathering phase [Chen *et al.*, 2025], akin to human initial perception of video content. 2) **Intermediate Layers**: Attention declines and narrows its focus, and entropy fluctuates as the model explores and analyzes visual content, similar to humans refining and analyzing visual content given questions. 3) **Deep Layers**: The model greatly reduces attention on visual tokens and entropy steadily declines, indicating a shift to knowledge-driven reasoning, resembling how humans draw conclusions after gathering sufficient information. Our analysis reveals an important insight: while VideoLLMs align with the general reasoning process observed in humans, they overlook the intuitive adjustment mechanisms crucial for human cognition. Unlike humans, VideoLLMs do not prioritize question-relevant content but instead attempt to memorize all video details. They cannot re-examine the visual content when uncertain and cease processing visual content once sufficiently confident in their conclusions. This limits leads to suboptimal efficiency and comprehension abilities in VideoLLMs.

Based on these observations, we propose **DToMA**, a training-free Dynamic Token MAipulation method to improve efficiency and comprehension abilities of VideoLLMs in three aspects: text-guided keyframe-aware reorganization, uncertainty-based visual injection, and early-exit pruning. Firstly, text-guided keyframe-aware reorganization imitates humans to selectively focus after an initial scan. It uses cross-attention scores from shallow layers to identify keyframes, allocating more tokens to keyframes while reducing non-keyframe tokens. Secondly, uncertainty-based visual injection emulates humans to revisit related content when uncertainty arises. For samples with high entropy, this process dynamically reintroduces visual tokens in the model's feed-forward layers, enhancing alignment with visual content and reducing uncertainty. Thirdly, early-exit pruning mirrors humans to stop examining visual content once they are confident about their conclusions. This process prunes visual tokens before the model reaches its deep layers, reducing redundant computations without sacrificing reasoning accuracy. As shown in Fig.1, by systematically emulating these human-inspired cognitive strategies, DToMA achieves gains in both efficiency and comprehension capability.

Our contributions are summarized as follows:

- We analyze VideoLLMs' reasoning processes through entropy and attention dynamics, revealing that while these models partially emulate human general reasoning processes, they lack key intuitive adjustments critical for efficiency and comprehension capability.
- We propose DToMA, a training-free dynamic visual token manipulation method that improves efficiency and reasoning capability by imitating human intuitive adjustment mechanisms, including text-guided keyframe-aware reorganization, uncertainty-based visual injection, and early-exit pruning.
- Experiments on 6 long video understanding benchmarks demonstrate that DToMA outperforms state-of-the-art methods and generalizes across diverse LLM architectures and sizes.

2 Related Works

2.1 Video Large Language Models

The demand for enhanced understanding capabilities in MLLMs has led to a substantial increase in visual tokens. Early works like LLaVA [Liu *et al.*, 2024a] encode a 336px image into 576 tokens, while further efforts [Li *et al.*, 2024; Liu *et al.*, 2024b] splitting higher resolution images into multiple sub-images, leading to more extended tokens, *e.g.*, 2,306 tokens for a 672px image. Video understanding tasks exacerbate this challenge due to sequential frames, *e.g.*, up to 4k tokens for a 16-frame video, straining LLM context limits [Liu *et al.*, 2024b]. To process longer videos, efforts like Gemini-1.5 [Team *et al.*, 2024] and LongVA [Zhang *et al.*, 2024b] extend context windows but suffer from prohibitive computational costs, with Gemini-1.5 encoding one-hour videos into 920k tokens. Compression strategies, such as LLaMA-VID [Li *et al.*, 2025] and Vid-Compress [Lan *et al.*, 2024], reduce the token numbers but often sacrifice visual detail. Alternative approaches like image grids [Kim *et al.*, 2024] and slow-fast processing [Xu *et al.*, 2024] improve efficiency but struggle with long videos.

The key challenge remains in the excessive number of tokens to represent video content. To address this, we propose DToMA, a training-free dynamic token manipulation method that enhances efficiency and understanding of VideoLLMs.

2.2 Visual Token Reduction

Extensive research has focused on reducing visual tokens to enhance efficiency without obviously compromising visual understanding. Methods like LLaVA-PruMerge [Shang *et al.*, 2024], FasterVLM [Zhang *et al.*, 2024c], and FoPru [Jiang *et al.*, 2024a] drop redundant tokens using attention guidance in visual encoder but lack textual adaptability. Methods like TRIM [Song *et al.*, 2024] and [Chen *et al.*, 2024b] employ paired text encoders, *i.e.*, CLIP, to assist token filtering but are incompatible with fine-tuned image encoders. Recent works consider compression within LLMs, such as FastV [Chen *et al.*, 2025] and Feather [Endo *et al.*, 2024], retaining tokens with high visual-text attention scores. FitPrune [Ye *et al.*, 2024] and G-Search [Zhao *et al.*, 2024] further progressively prune tokens within LLM layers, thereby enhancing computational efficiency.

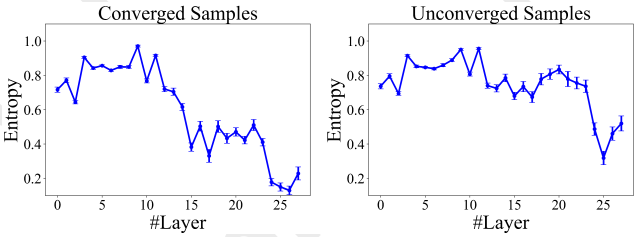

Figure 2: Visualization of entropy dynamics in Video-MME. According to whether the final convergence is distinguished, we divided all the samples into converged (left) and unconverged (right) samples. It illustrates the uncertain reasoning processes in intermediate layers, which also propagate to deeper layers.

Existing methods do not adequately balance efficiency with performance, often compromising video understanding. To address this gap, DToMA 1) mitigates inefficiency by pruning non-keyframe tokens and exiting all visual tokens in LLM deep layers, and 2) improves comprehension by exerting emphasis on keyframes and reintroducing visual information when uncertainty arises. Thus, DToMA improves both efficiency and comprehension without additional training.

2.3 Interpreting MLLM Internal Mechanism

Understanding the inner workings of MLLMs can gain insights into their reasoning processes, identify inefficiencies, and develop targeted strategies for improvement. Recent interpretability studies using attention scores [Chen *et al.*, 2025; Li *et al.*, 2022], intermediate representations [Jiang *et al.*, 2024b], and activation patterns [Chen *et al.*, 2024a] mitigate token redundancy and image hallucination issues. For instance, analyzing attention scores provides insight into the model's focus at each layer. Recent studies leverage this method to reveal that LLMs often emphasize textual information over visual content [Liu *et al.*, 2025; Fu *et al.*, 2024b]. By better balancing the emphasis on visual tokens, these methods effectively mitigate image hallucinations.

Building on these insights, we analyze VideoLLMs through the lenses of attention and entropy dynamics to compare their reasoning processes to human cognition. We find that while VideoLLMs can imitate general human reasoning phases, they lack key intuitive strategies for efficient and accurate reasoning. DToMA bridges this gap by incorporating human-inspired strategies to enhance reasoning efficiency and comprehension.

3 Method

3.1 Preliminary

VideoLLMs generate text responses based on input video and text queries. The core components include a visual encoder $E(\cdot)$, a projector for modality alignment, an LLM with L layers, and an LM head $\mathrm{Vocab}(\cdot)$ which predicts the vocabulary distribution of the next token. Given a video with N frames, the visual encoder extracts N frame embeddings. The projector maps frame embeddings as visual tokens $T_v \in \mathbb{R}^{N\times P\times d}$ to enable processing by the LLM, where P, d are the number of patches and model dimension. Then, visual tokens T_v and

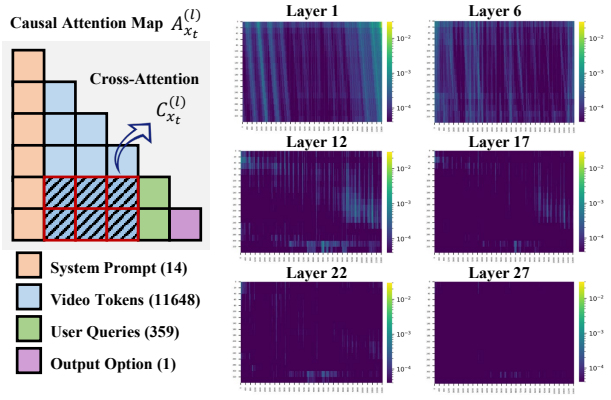

Figure 3: Visualization of cross-attention $C_{x_t}^{(l)}$ across LLM layers, where $C_{x_t}^{(l)}$ are derived from the causal attention maps $A_{x_t}^{(l)}$. It shows how cross-attention evolves through LLM layers: in shallow layers, there is extensive interaction between text and visual tokens; in intermediate layers, the model narrows its focus with reduced attention weights; in deep layers, cross-attention diminishes almost entirely.

text query tokens $T_t \in \mathbb{R}^{N_t\times d}$ are concatenated as the input sequence, *i.e.*, $\{x_i\}_1^t, t = N \times P + N_t$, and are processed through L LLM layers. The LLM final layer output $h_{x_t}^{(L)}$ is used by $\mathrm{Vocab}(\cdot)$ to predict the next token x_{t+1} as:

$$P(x_{t+1}|x_{1:t}) = \mathrm{Softmax}(\mathrm{Vocab}(h_{x_t}^{(L)})) \quad (1)$$

With greedy decoding, the most probable word for output would be selected.

3.2 Observations of LLM Internal Behaviors

In this section, we aim to explore the inefficiencies in the reasoning processes of VideoLLMs by analyzing their internal behaviors. Through this analysis, we identify key areas where the models struggle and propose potential improvements.

Specifically, to gain insights into these behaviors, we employ two critical metrics: entropy and cross-attention dynamics. These metrics are chosen because they provide a comprehensive view of the model's confidence levels and its prioritization of multimodal information, respectively. Entropy helps track uncertainty and confidence across layers [Fang *et al.*, 2024], while cross-attention reveals how the model integrates information from different modalities [Chen *et al.*, 2025]. Detailed observation implementations are as follows:

- Entropy dynamics are observed by tracking the evolution of answer probabilities across layers, inspired by recent interpretability research [nostalgebraist, 2020] which finds intermediate outputs $h_{x_t}^{(l)}$ in each l-th layer can be interpreted by $\mathrm{Vocab}(\cdot)$. By prompting the model to answer directly with the letter option from given choices, we can decode the first token output at each layer l to track answer probabilities $P(x = o|x_{1:t})$ following Eq. (1), denoted as P_o, where $o \in \{A, B, C, D\}$. Then, the normalized entropy is computed as $\mathcal{H} = \sum -P_o \log P_o / \log m$, where $m = 4$ is the number of options. $\mathcal{H}$ is the model's uncertainty at each layer.

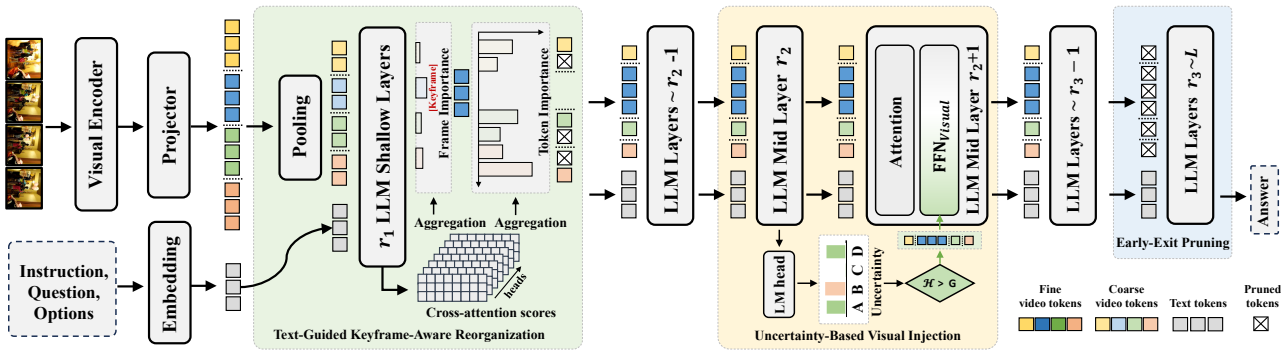

Figure 4: We propose DToMA, a training-free Dynamic visual Token MAnipulation method for efficient long video understanding.

- Cross-attention dynamics are analyzed through attention maps from each layer. Specifically, for attention maps $A_{x_t}^{(l)} \in \mathbb{R}^{n\times t\times t}$ in each layer, where n is the number of heads, we extract the cross-attention scores between text query T_t and visual tokens T_v, *i.e.*, $C_{x_t}^{(l)} \in \mathbb{R}^{n\times N_t\times NP}$, from $A_{x_t}^{(l)}$. Aggregating $C_{x_t}^{(l)}$ across all attention heads yields: $\overline{C}_{x_t}^{(l)} = \sum^n C_{x_t}^{(l)}$. Measuring $\overline{C}_{x_t}^{(l)}$ allows us to analyze how text token aggregates visual information.

Our analysis reveals distinct stages in VideoLLMs' reasoning process, as in Fig.2, 3. In shallow layers, entropy is high and P_o is nearly random, yet text tokens strongly attend to visual tokens, indicating an initial information-gathering phase [Chen *et al.*, 2025]. In intermediate layers, entropy exhibits dynamic fluctuations. For simple questions, it decreases steadily as the model converges on the correct answer. However, for complex ones, prolonged fluctuations suggest ongoing uncertainty. Attention declines and narrows its focus, reflecting information refinement. In deep layers, entropy rapidly converges for simple questions but remains higher and fluctuates for complex ones. Attention to visual tokens diminishes, indicating internal textual reasoning.

The above observations highlight several inefficiencies: in shallow layers, models attempt to perceive all visual content. Methods like FastV [Chen *et al.*, 2025] reduce the influence of redundancy but do not adequately emphasize key information, leading to unnecessary computational overhead and suboptimal performance. Though uncertainty occurs in intermediate layers, the diminishing cross-attention from text tokens makes it challenging for the model to independently and adequately reassess necessary visual evidence. In deep layers, reduced attention to visual tokens suggests inefficiency in continuously computing visual context for final reasoning.

3.3 DToMA

To this end, we propose DToMA, a training-free Dynamic visual Token MAnipulation method to address those inefficiencies and improve reasoning capabilities, as in Fig. 4. DToMA has 3 token processing strategies: text-guided keyframe-aware reorganization, uncertainty-based visual injection, and early-exit pruning, detailed in the following sections.

Text-Guided Keyframe-Aware Reorganization (TKR). To solve the inefficiency that the model attempts to memorize all video details in shallow layers, we propose TKR that prioritizes keyframes while reducing non-keyframe tokens for both efficiency and understanding ability improvement. The method involves a two-pass process. The first pass is executed only in LLM shallow layers to obtain question-related guidance for keyframe selection and non-keyframe tokens reduction. Then, the second pass is executed formally through all layers with our reorganized visual tokens. Details of TKR are shown in Algorithm 1. Specifically, the LLM first performs an r_1-layer forward pass using coarse representations $\hat{T}_v$ to get cross-attention $\overline{C}_{\hat{x}_t}^{(r)}$, which is then averaged to obtain frame importance $F^{(r)}$ for keyframe selection. Note that $F^{(r)}$ would be divided into S non-overlapping segments, *i.e.*, $F_1^{(r)}, ..., F_S^{(r)}$, and the top$k$ frames are selected from each segment to avoid concentrating keyframes in a local video clip. For non-keyframes, we drop the lowest attended tokens in $\hat{T}_v$, ensuring the pruning ratio respects the token budget. Finally, keyframes retain their fine representations and non-keyframes are with reduced coarse representations, both are reorganized in chronological order for full propagation through the LLM as: $\mathrm{LLM}_{1:L}(\mathrm{Concat}(\mathrm{T_{final}}, \mathrm{T_t}))$. This method balances computational efficiency with representational fidelity, highly adaptable to various computational budgets and tasks.

Uncertainty-Based Visual Injection (V-Inj). Intermediate layers often exhibit uncertainty for complex questions. As attention to visual tokens diminishes over layers, it is challenging for the model to incorporate necessary visual evidence independently, leading to propagated confusion. To address this, we propose uncertainty-based visual injection, which identifies uncertain layers and reintroduces visual evidence to support more accurate reasoning. Inspired by studies showing that feed-forward networks (FFNs) can act as key-value memories for factual knowledge retrieval [Geva *et al.*, 2021] and that visual information can be effectively integrated into FFNs for hallucination mitigation [Zou *et al.*, 2025], we inject visual tokens into the FFN during uncertain scenarios.

The standard FFN is defined as: $\mathrm{FFN}(x) = \sigma(x \cdot \boldsymbol{W}_1)\boldsymbol{W}_2$, where $\boldsymbol{W}_1 \in \mathbb{R}^{d\times D}$, $\boldsymbol{W}_2 \in \mathbb{R}^{D\times d}$, σ is activation function, D is the intermediate dimension. Furthermore, by represent weights as $\boldsymbol{W}_1 = (\boldsymbol{k}_1, \boldsymbol{k}_2, ..., \boldsymbol{k}_D)$ and $\boldsymbol{W}_2 = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_D)^T$, $\boldsymbol{k}_i, \boldsymbol{v}_i \in \mathbb{R}^d$, the FFN can be rewritten in

Algorithm 1: Token Reorganization Strategy

Input: visual tokens $T_v \in \{T_v^1, ..., T_v^N\}$, $T_v^i \in \mathbb{R}^{P \times d}$. Text tokens T_t, selected shallow layer r_1, number of keyframes m, number of segments S, token budget B.
Output: Reorganized tokens T_{final}

First Pass: Early Layer Execution

1. Obtain coarse frame representations (e.g., pooling):

$$\hat{T}_v = \text{Pool}(T_v), \quad \hat{T}_v \in \mathbb{R}^{P_c \times N \times d}, \quad P_c < P$$

2. Run r_1-layer forward pass to get $\overline{C}_{\hat{x}_t}^{(r)}$:

$$\overline{C}_{\hat{x}_t}^{(r_1)} \leftarrow \text{LLM}_{1:r_1}(\text{Concat}(\hat{T}_v, T_t))$$

3. Aggregate $\overline{C}_{\hat{x}_t}^{(r_1)}$ to get frame importance:

$$F^{(r_1)} = \text{Mean}(\overline{C1}_{\hat{x}_t}^{(r_1)}), \quad F^{(r_1)} \in \mathbb{R}^N$$

4. Identify top-k frames in each segment:

$$N_{\text{key}} = \bigcup_{i=1}^{S} \text{Topk\_id}(F_i^{(r_1)}, m/S)$$

Second Pass: Token Reorganization

5. Compute pruning ratio:

$$\text{ratio} = \min\left(\frac{N_{\text{key}} P + (N - N_{\text{key}}) P_c}{B}, 1\right)$$

6. Reorganize visual tokens T_v:

$$T_{\text{final}}^i = \begin{cases} T_v^i, & i \in N_{\text{key}} \\ \text{Topk}(\hat{T_v}^i, P_c \cdot \text{ratio}), & i \notin N_{\text{key}} \end{cases}$$

return $T_{\text{final}} = \bigcup_{i=1}^{N} T_{\text{final}}^i$

key-value form:

$$\text{FFN}(x) = \sum_{i=1}^{D} \sigma(\langle x \cdot \boldsymbol{k}_i \rangle) \cdot \boldsymbol{v}_i \tag{2}$$

where FFN can be interpreted as using input x as queries to compute similarity with keys $\boldsymbol{k}_i$ and gather values v_i.

To inject visual evidence, we convert our reorganized visual tokens $T_{\text{final}} \in \mathbb{R}^{B \times d}$ into keys and values and perform similar retrieval from the visual key-value memory as:

$$\text{VisInj}(x) = \sum_{i=1}^{B} \sigma(\langle x \cdot \mathcal{K}(T_i) \rangle) \cdot \mathcal{V}(T_i) \tag{3}$$

where $\mathcal{K}(T_i), \mathcal{V}(T_i) \in \mathbb{R}^d$ are key and value corresponding to visual token $T_i \in T_{\text{final}}$. We employ identical mappings for $\mathcal{K}(\cdot)$ and $\mathcal{V}(\cdot)$ to maintain its training-free nature. This visual memory is integrated into FFN as:

$$\text{FFN}_{\text{visual}}(x) = (1 - \alpha) \cdot \text{FFN}(x) + \alpha\beta \cdot \text{VisInj}(x) \tag{4}$$

where $\beta = \text{Norm}(\text{FFN}(x))/\text{Norm}(\text{VisInj}(x))$ is a balance factor and α is injection strength. For modern Gated Linear Units (GLU) instead of FFN in recent LLMs, the visual injection bypasses the gating mechanism for simplicity.

We dynamically activate visual injection based on the entropy $\mathcal{H}$ of output probabilities in intermediate layers $r_s - r_e$. If entropy in layer $r_2 \in [r_s, r_e]$ exceeds a threshold G, visual evidence is injected at the next layer $r_2 + 1$ by replacing $\text{FFN}^{(r_2+1)}$ as $\text{FFN}_{visual}^{(r_2+1)}$. This injection is performed only once in uncertain scenarios to preserve feature stability. This method enriches the visual cues tied to the question and enhances reasoning consistency with visual content, thus improving the model's capability to integrate multimodal information for accurate predictions.

Early-Exit Pruning (EP). To reduce inefficiency in deep layers, we introduce early-exit pruning by omitting overall visual token processing. Initially, we explored a dynamic early exit based on entropy convergence, enabling the model to bypass unnecessary computations for high-confidence samples. However, experiments showed that samples with prolonged uncertainty paid minimal attention to image tokens in deep layers, indicating that delayed exits do not enhance understanding and instead increase computing costs. Therefore, we adopted a fixed-layer exit strategy, which simplifies implementation and ensures consistent efficiency.

Specifically, for layers $r > r_3$, all visual token computations are omitted. The attention mechanism is simplified to only focus on text tokens T_t and the FFN computations for visual tokens are bypassed. This pruning strategy significantly reduces computational costs while ensuring the model maintains reasoning accuracy.

4 Experiments

4.1 Benchmarks and Metrics

We conducted evaluations of our method on 6 long video understanding benchmarks, including VideoMME [Fu *et al.*, 2024a], LongVideoBench [Wu *et al.*, 2024], EgoSchema [Mangalam *et al.*, 2023], MLVU [Zhou *et al.*, 2024], NExT-QA [Xiao *et al.*, 2021], and PerceptionTest [Patraucean *et al.*, 2024]. Following evaluation tool LMMs-Eval [Zhang *et al.*, 2024a], we perform standardized evaluation settings and metrics, *i.e.*, accuracy, on each benchmark. We evaluate the efficiency by computing FLOPs and prefill time of LLMs using the library from LLM-Viewer [Yuan *et al.*, 2024], and assume 1000 text tokens for LLaVA-Video due to their time prompt, while 100 for the others.

4.2 Implementation Details

Our implementation mainly follows LLaVA-Video-7B [Zhang *et al.*, 2024d]. We adopt SigLIP [Zhai *et al.*, 2023] as the vision encoder and Qwen2 [Yang *et al.*, 2024] as the LLM. For DToMA, the selected layer $r_1 = 3, r_2 \in [12, 18], r_3 = 21$. For TKR, following optimal design [Du *et al.*, 2024] for SigLIP, we use 2×2 pooling for keyframes, while 3×3 for coarse non-keyframes. Token budget B is pre-defined according to experimental requirements, and token compression ratio is self-adaptive according to B. With no specifically stated, we set $m = S = N/4$. For V-Inj, we set threshold $G = 0.75$, factor $\alpha = 0.25$. We also adapt DToMA in LLaVA-OV-0.5B [Li *et al.*, 2024], Qwen2-VL-2B [Wang *et al.*, 2024] to evaluate generalizing across VideoLLM architectures and sizes.

4.3 Comparative Evaluation

Comparison with SOTA Methods. We compare DToMA with state-of-the-art (SOTA) methods, focusing on long video comprehension under identical token budgets. Specifically,

Models	#Tokens Budget	VideoMME	LongVideoBench	MLVU	EgoSchema	PerceptionTest	NExT-QA
		w/o sub.	Val	M-Avg	Test	Val	Test
Proprietary Models							
GPT4-V	-	59.9	59.1	49.2	-	-	-
GPT4-o	-	71.9	66.7	64.6	-	-	-
Gemini-1.5-Pro	-	75.0	64.0	-	72.2	-	-
Open-Source Video MLLMs							
LLaVA-OV-0.5B	6k	44.0	43.4	50.3	26.8	49.2	57.2
w. DToMA	6k	45.8	44.9	52.4	27.2	49.8	59.7
Qwen2-VL-2B	16k	55.6	-	-	54.9	53.9	-
Qwen2-VL-2B*	8k	54.2	48.1	57.2	54.3	53.2	75.2
w. DToMA	8k	55.1	48.9	59.1	56.5	53.9	76.8
LongVA-7B	18k	52.6	-	56.3	-	-	68.3
LLaVA-OV-7B	6k	58.6	56.5	64.7	60.1	57.1	79.4
LongVU-7B	8k	60.6	-	65.4	**67.6**	-	-
Qwen2-VL-7B	16k	63.3	-	-	66.7	62.3	-
LLaVA-Video-7B	12k	63.3	58.2	70.8	57.3	67.9	83.2
w. DToMA	12k	**65.0**	**59.6**	**71.7**	59.3	**68.9**	**83.8**

Table 1: Performance on 6 long video benchmarks. * represent our implementation results with 64 frame inputs.

Methods	#Tokens	#Frames	FLOPs (TB)	Prefill Time (ms)	VideoMME w/o subs
LLaVA-Video	11,648	64	249.3	1,214	63.3
ToMe	70%	64	163.1	734.5	62.8
FastV	70%	64	172.2	789.3	63.4
FasterVLM	70%	64	162.9	734.2	63.1
ToMe	50%	64	113.0	483.7	61.1
FastV	50%	64	127.5	565.8	63.1
FasterVLM	50%	64	112.9	483.5	63.0
DToMA	50%	64	**96.6**	**410.2**	**63.6**
ToMe	30%	64	69.0	276.6	57.6
FastV	30%	64	88.1	374.8	61.2
FasterVLM	30%	64	68.9	276.5	61.9
DToMA†	30%	64	**64.9**	**262.2**	**62.5**
LLaVA-Video	23,296	128	589.0	3,224	64.0
Corase rep.	11,520	128	219.1	1,056	63.7
SlowFast	11,200	128	211.8	990.6	64.2
SF-LLaVA	15,232	128	320.8	1,643	63.8
ToMe	50%	128	249.8	1,231	63.8
FastV	50%	128	285.4	1,429	64.2
FasterVLM	50%	128	249.0	1,217	64.3
DToMA	50%	128	**208.3**	**988.6**	**65.0**

Table 2: Comparison results on VideoMME based on LLaVA-Video. As DToMA uses coarse representations for non-keyframes, reducing tokens to less than 70% without token-level compression, we report results only at 50% and 30% ratios. † At 30% ratio, we use the visual encoder's attention to compress keyframe token counts.

we evaluate DToMA across three different architectures and sizes of video models, *i.e.*, LLaVA-OV-0.5B, Qwen2-VL-2B, and LLaVA-Video-7B. By compressing visual tokens and using coarse representations for non-keyframes, DToMA enables more frame inputs within the same token budget. We set a double frame input for DToMA, with an 80.5% non-keyframes compression ratio. Table 1 summarizes the results on 6 video understanding benchmarks. DToMA consistently improves performance across all 3 tested models, with the averaged improvement of 1.48%, 1.35%, and 1.43% on the 6 benchmarks, respectively. Specifically, when applied to LLaVA-Video-7B, it achieves SOTA results on 5 out of 6 benchmarks. Moreover, under the same token budget, the efficiency of DToMA also shows an improvement, with 83.6% FLOPs and 79.0% prefill time, which will be discussed in the ablation. This demonstrates the effectiveness of DToMA in enhancing both efficiency and understanding capabilities, and the generalization ability across diverse architectures and sizes of VideoLLMs.

Comparison with token reduction and efficient video representation Methods. To assess performance and efficiency (FLOPs and prefill time) of DToMA, we compared it with existing token reduction techniques, *i.e.*, ToMe [Bolya *et al.*, 2023], FastV [Chen *et al.*, 2025], FasterVLM [Zhang *et al.*, 2024c], and efficient video representation strategies, *i.e.*, coarse representation $\hat{T}_v$, SlowFast [Zhang *et al.*, 2024d], SF-LLaVA [Xu *et al.*, 2024], on VideoMME using LLaVA-Video with 64 and 128 frame inputs. The results are shown in Table 2. It reveals that DToMA achieves superior performance in terms of FLOPs and prefill time while improving accuracy on VideoMME. Specifically, With 64-frame videos, DToMA at a 50% compression ratio surpasses baseline (+0.4%) whereas other methods perform below baseline. For 128-frame videos, DToMA's advantages are more highlighted. This is especially notable as 128-frame inputs increase redundancy challenges for LLaVA-Video which is trained on 64-frame inputs. DToMA consistently outperforms the baseline (+1.0%) and other methods (+0.7%) with fewer FLOPs and prefill time, proving the effectiveness of DToMA in improving efficiency and understanding abilities.

4.4 Ablation Study

Effectiveness of Each Component. We ablate the three key components of DToMA: text-guided keyframe-aware reorganization (TKR), uncertainty-based visual injection (V-Inj), and early-exit pruning (EP), in VideoMME dataset based on LLaVA-Video. The results are shown in Table 4. TKR reduces the token count by 50% and thus achieves marked efficiency gains, using only 49.6% of baseline FLOPs and 43.9% of prefill time while improving performance. It demonstrates that TKR effectively emphasizes the related visual context and reduces redundancy, thereby enhancing both efficiency and accuracy. V-Inj reintroduces visual information when uncertainty is detected, adding less than 0.5% FLOPs but boosting accuracy by 0.4% with baseline tokens and 0.7% with reorganized tokens in 128 frames. The results prove its effectiveness in mitigating uncertainty and aligning reasoning with visual context with only a small computing increase. EP uses 72.9% FLOPs and 71.7% prefill time of baseline to pro-

TKR	V-Inj	EP	#Token	#Frame	FLOPs	Prefill Time	VideoMME
LLaVA-Video			11,648	64	249.3	1,214	63.3
✓			5,824	64	123.7	533.2	63.4
	✓		11,648	64	250.2	1,215	63.7
		✓	11,648	64	181.9	870.7	63.1
✓	✓		5,824	64	124.4	533.9	63.8
✓	✓	✓	5,824	64	**96.6**	**410.2**	63.6
LLaVA-Video			23,296	128	589.0	3,224	64.0
✓			11,648	128	275.2	1,317	64.5
	✓		23,296	128	590.4	3,225	64.4
		✓	23,296	128	424.5	2,336	63.8
✓	✓		11,648	128	276.4	1,318	**65.2**
✓	✓	✓	11,648	128	208.3	988.6	65.0

Table 3: Ablation Studies of each component in DToMA based on LLaVA-Video in VideoMME dataset.

r_1 in TKR			r_2 in V-Inj			r_3 in EP		
Layer	**Acc.**	**FLOPs**	**Layer**	**Acc.**	**FLOPs**	**Layer**	**Acc.**	**FLOPs**
1	64.4	**190.5**	13	64.6	**207.4**	19	63.9	**191.5**
2	64.9	199.2	17	64.7	**207.4**	20	64.8	199.8
3	**65.0**	208.3	12-18	64.6	208.6	**21**	**65.0**	208.3
4	**65.0**	216.9	**13-18**	**65.0**	208.3	22	64.7	216.7
5	64.8	225.6	13-19	64.9	208.4	Dynamic	**65.0**	226.3

Table 4: Deployment layer analysis of each component in DToMA based on LLaVA-Video in VideoMME dataset.

cess the same visual tokens by pruning all visual tokens in deep layers, improving model efficiency without compromising performance. Overall, these strategies enable DToMA to consistently improve both performance and computing efficiency. Notably, with 128-frame inputs, DToMA achieves a 1.7% improvement over the 64-frame baseline, using just 83.5% of the FLOPs and 81.4% of the prefill time, proving its excellent accuracy-efficiency trade-off.

Layer selection analysis. In DToMA, the methods TKR, V-Inj, and EP target inefficiencies in shallow, intermediate, and deep LLM layers, respectively. We select target layers based on attention and entropy dynamics. TKR is applied at the first layer when cross-attention shifts from inclined to vertical patterns. V-Inj starts when entropy diverges between easy and hard questions and ends when entropy rises sharply again, with visual attention drops. EP is applied when attention nearly vanishes. We evaluated various layer indices to find the best balance between accuracy and efficiency. Results in Table 4 show the optimal setup is $r_1 = 3, r_2 \in [13-18], r_3 = 21$, maximizing accuracy while minimizing resource usage. We also tested dynamic r_3 based on model uncertainty, but found no clear performance benefits and delayed the early exit due to diminished attention focus. Thus, we opt for a fixed early exit layer.

Uncertainty Decline Analysis. To investigate the impact of V-Inj on uncertainty, we visualize the uncertainty $\mathcal{H}$ across LLM layers in two datasets: VideoMME and LongVideoBench. The visualization results in Fig. 5 reveal that injecting visual content at intermediate layers greatly reduces model uncertainty. This reduction is not only immediate but also propagates to deeper layers, thus ensuring a sustained decrease in uncertainty throughout VideoLLM. V-Inj also lowers the uncertainty peaks in intermediate and deep layers, enabling the model to make more informed responses based on visual context in complex scenarios. The improved consistency and confidence underscore the value of V-Inj in enhancing the performance and reliability of VideoLLMs.

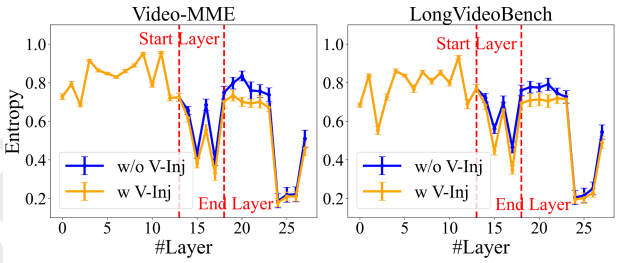

Figure 5: Visualization of uncertainty w. and w/o. V-Inj on the VideoMME and LongVideoBench datasets based on LLaVA-Video.

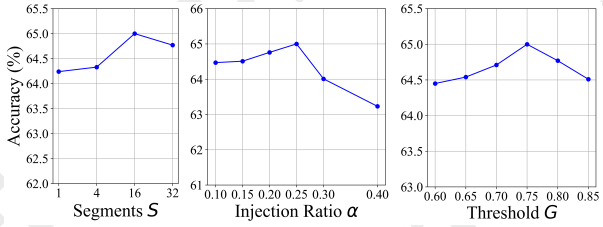

Figure 6: Hyper-parameter analysis on VideoMME dataset based on LLaVA-Video.

Hyper-parameter Analysis. To investigate the impact of hyper-parameters on DToMA, we conduct experiments by varying segment S in TKR, injection ratio α, and threshold G in V-Inj. More specifically, we consider the following settings: $S \in \{1, 4, 16, 32\}$, $\alpha \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.4\}$, and $G \in \{0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}$. The results in Fig. 6 show that the optimal performance is achieved with $S = 16$, $\alpha = 0.25$, and $G = 0.75$.

5 Conclusion

This paper addresses the challenge of efficiently processing long videos in VideoLLMs while enhancing comprehension. By analyzing reasoning patterns of VideoLLMs in multi-choice VideoQA task through entropy and cross-attention dynamics, we identify three distinct reasoning stages—shallow, intermediate, and deep stages—and the inefficiencies existing in these stages. To address these, we proposed DToMA, a training-free Dynamic Token Manipulation method in three aspects: text-guided keyframe-aware reorganization, uncertainty-based visual injection, and early-exit pruning. Experiments on 6 long video benchmarks show that DToMA boosts both performance and efficiency, outperforms SOTA methods, and generalizes well across various VideoLLMs. In sum, DToMA offers a promising alternative to directly using the baseline, enabling superior long video understanding within reduced computational resources.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62325206 and 62301276, the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, and the Opening Foundation of the State Key Laboratory of Tibetan Intelligence, Key Laboratory of Tibetan Information Processing, Ministry of Education (2024-2-003).

References

[Bolya *et al.*, 2023] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023.

[Chen *et al.*, 2024a] Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024.

[Chen *et al.*, 2024b] Yi Chen, Jian Xu, Xu-Yao Zhang, Wen-Zhuo Liu, Yang-Yang Liu, and Cheng-Lin Liu. Recoverable compression: A multimodal vision token recovery mechanism guided by text information. *arXiv preprint arXiv:2409.01179*, 2024.

[Chen *et al.*, 2025] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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*, pages 19–35. Springer, 2025.

[Du *et al.*, 2024] Yifan Du, Yuqi Huo, Kun Zhou, Zijia Zhao, Haoyu Lu, Han Huang, Wayne Xin Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. Exploring the design space of visual context representation in video mllms. *arXiv preprint arXiv:2410.13694*, 2024.

[Dundas and Chik, 2011] Jitesh Dundas and David Chik. Implementing human-like intuition mechanism in artificial intelligence. *arXiv preprint arXiv:1106.5917*, 2011.

[Endo *et al.*, 2024] Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. *arXiv preprint arXiv:2412.13180*, 2024.

[Fang *et al.*, 2024] Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. From uncertainty to trust: Enhancing reliability in vision-language models with uncertainty-guided dropout decoding. *arXiv preprint arXiv:2412.06474*, 2024.

[Fu *et al.*, 2024a] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[Fu *et al.*, 2024b] Yuhan Fu, Ruobing Xie, Jiazhen Liu, Bangxiang Lan, Xingwu Sun, Zhanhui Kang, and Xirong Li. Magnifier prompt: Tackling multimodal hallucination via extremely simple instructions. *arXiv preprint arXiv:2410.11701*, 2024.

[Geva *et al.*, 2021] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

[Jiang *et al.*, 2024a] Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. Fopru: Focal pruning for efficient large vision-language models. *arXiv preprint arXiv:2411.14164*, 2024.

[Jiang *et al.*, 2024b] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024.

[Kim *et al.*, 2024] Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *arXiv preprint arXiv:2403.18406*, 2024.

[Lan *et al.*, 2024] Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Vidcompress: Memory-enhanced temporal compression for video understanding in large language models. *arXiv preprint arXiv:2410.11417*, 2024.

[Li *et al.*, 2022] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13874–13883, 2022.

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

[Li *et al.*, 2025] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[Liu *et al.*, 2025] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2025.

[Mangalam *et al.*, 2023] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[nostalgebraist, 2020] nostalgebraist. Interpreting gpt: The logit lens. *LessWrong*, 2020.

[Patraucean *et al.*, 2024] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Shang *et al.*, 2024] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[Shen *et al.*, 2024] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[Song *et al.*, 2024] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. *arXiv preprint arXiv:2409.10994*, 2024.

[Team *et al.*, 2024] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[Wu *et al.*, 2024] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

[Xiao *et al.*, 2021] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[Xu *et al.*, 2024] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[Ye *et al.*, 2024] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024.

[Yuan *et al.*, 2024] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, et al. Llm inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*, 2024.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[Zhang *et al.*, 2024a] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[Zhang *et al.*, 2024b] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[Zhang *et al.*, 2024c] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024.

[Zhang *et al.*, 2024d] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[Zhao *et al.*, 2024] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. Accelerating multimodel large language models by searching optimal vision token reduction. *arXiv preprint arXiv:2412.00556*, 2024.

[Zhou *et al.*, 2024] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

[Zou *et al.*, 2025] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *The Forty-second International Conference on Machine Learning (ICML)*, 2025.