Stop Looking for "Important Tokens" in Multimodal Language Models: Duplication Matters More

Anonymous ACL submission

Abstract

Vision tokens in multimodal large language models often dominate huge computational overhead due to their excessive length compared to linguistic modality. Abundant recent methods aim to solve this problem with token pruning, which first defines an importance criterion for tokens and then prunes the unimportant vision tokens during inference. However, in this paper, we show that the importance is not an ideal indicator to decide whether a token should be pruned. Surprisingly, it usually results in inferior performance than random token pruning and leading to incompatibility to efficient attention computation operators. Instead, we propose **DART** (**D**uplication-**A**ware **R**eduction of **T**okens), which prunes tokens based on its duplication with other tokens, leading to significant and training-free acceleration. Concretely, DART selects a small subset of pivot tokens and then retains the tokens with low duplication to the pivots, ensuring minimal information loss during token pruning. Experiments demonstrate that DART can prune **88.9%** vision tokens while maintaining comparable performance, leading to a **1.99**× and **2.99**× speed-up in total time and prefilling stage, respectively, with good compatibility to efficient attention operators.

1 Introduction

Multimodal large language models (MLLMs) exhibit remarkable capabilities across a diverse range of multimodal tasks, including image captioning, visual question answering (VQA), video understanding (Wang et al., 2024b), and multimodal reasoning (Wang et al., 2024c). However, such impressive performance is always accompanied by huge computation costs, which are mainly caused by massive vision tokens in the input data, especially for high-resolution images (Li et al., 2024b) and multi-frame video (Tang et al., 2023), leading to challenges in their applications.

Figure 1: **Comparison between DART and FastV.** Red text indicates hallucination from vanilla LLaVA-1.5-7B, green text represents hallucination from DART, and blue text represents hallucination from FastV.

To solve this problem, abundant recent methods introduce *token pruning* to remove the vision tokens in a training-free manner, which usually first defines the importance score of each token, and then prunes the most unimportant tokens during the inference phrase (Chen et al., 2024; Zhang et al., 2024b; Liu et al., 2024e). The key to a token pruning method is the definition of the importance of vision tokens, where most existing methods are based on the attention scores between vision-only tokens and vision-language tokens. However, this paper argues that these importance-based methods have several serious problems.

(I) Ignoring interactions between tokens during pruning: Although the interaction between different tokens is considered in attention scores, however, importance-based methods directly remove the most unimportant tokens, ignoring the truth that the importance of each token should be adjusted when other tokens are pruned or preserved. For

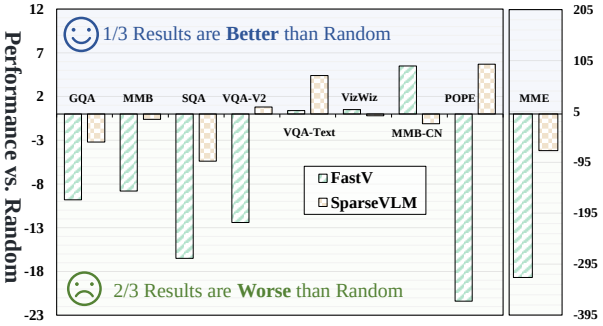


Figure 2: **Performance of FastV and SparseVLM compared with random token pruning** on the LLaVA-1.5-7B, with a **88.9%** token reduction ratio.

instance, for two similar tokens, if one of both is determined to be pruned, then the importance of the other token should be improved and vice versa. Unfortunately, previous importance-based token pruning methods fail to model such interaction.

(II) Incompatibility to efficient attention: Efficient attention operators such as FlashAttention (Dao et al., 2022) have become the default configure in neural networks, which accelerates attention computation by around $2\times$ and reduce the memory costs from $O(N^2)$ to $O(N)$. However, these efficient attention operators make attention scores not accessible during computation, indicating conflicts with most previous importance-based token pruning methods. Disabling FlashAttention for accessing attention scores significantly improves the overall latency and memory footprint.

(III) Bias in token positions: As claimed by abundant recent works (Endo et al., 2024; Zhang et al., 2024a) and shown in Figure 1, attention scores have position bias, where the tokens are positionally close to the last token tend to have a higher attention score, making attention score does not truly reveal the value of this token.

(IV) Significant accuracy drop: Although the aforementioned three problems have reminded us the ineffectiveness of importance-based token pruning, however, it is still extremely surprising to find that ***some influential importance-based token pruning methods show inferior accuracy than random token pruning***, (*i.e.*, randomly selecting the tokens for pruning), as shown in Figure 2.

The above observations demonstrates the disadvantages of importance-based token pruning methods, while also introducing the expectation for the ideal alternative: The expected method should consider both the individual value of a token and its interaction to other tokens. It should be cheap in computation and friendly to hardware, and shows no bias in the positions of tokens.

These insights inspire us to incorporate token duplication into the token reduction. Intuitively, when multiple tokens exhibit identical or highly similar representations, it is natural to retain only one of them for the following computation, thereby maintaining efficiency without harming accuracy. Building upon this idea, we introduce a simple but effective token pruning pipeline referred to as **DART** (**D**uplication-**A**ware **R**eduction of **T**okens) with the following two steps.

Firstly, we begin by selecting a small subset of tokens as pivot tokens, which comprise no more than 2% of the total tokens. Such pivot tokens can be selected based on the norm of tokens or even randomly selected, which does not introduce notable computations. Secondly, we then calculate the cosine similarity between pivot tokens and the remaining image tokens. Since the pivot tokens are fewer than 2%, such computation is efficient in both computing and memory. With a desired token reduction ratio, we retain only those vision tokens with the lowest cosine similarity to pivot tokens and remove the similar ones. The entire process is simple and highly efficient, completing in no more than **0.08** seconds, friendly to efficient attention operators, and leading to significantly higher accuracy than previous methods.

In summary, our contributions are three-fold:

- **Rethink Token Importance.** Through empirical analysis, we demonstrate the suboptimality of relying on attention scores to measure token importance to guide the token reduction paradigm.
- **Token Duplication as a Key Factor.** Building on token duplication, we introduce a training-free, plug-and-play token reduction method that seamlessly integrates with Flash Attention.
- **Superior Performance with Extreme Compression.** Extensive experiments across four diverse MLLMs and over 10 benchmarks demonstrate the clear superiority of DART. For instance, our method outperforms the second-best method by 2.2% (93.7% vs. 91.5%) on LLaVA-1.5-7B with an 88.9% reduction ratio.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) (Liu et al., 2024b; Li et al., 2023a; Zhu et al., 2023; Liu et al., 2024d) achieve remarkable performance in tasks involving images, videos, and multimodal reasoning by integrating vision and text process-

(a) Pivot Token Selection

(b) Calculate ϵ-Duplicate Score Between Pivot Tokens and the Remainder

(c) Token Reduction to Keep Tokens With Least Duplication

Figure 3: **The overview of DART.** The process includes (a) selecting pivot tokens, (b) calculating ϵ-Duplicate scores between pivot tokens and other tokens, and (c) reducing tokens to retain those with the least duplication.

ing. However, processing visual data poses computational challenges due to the redundancy and low information density of high-resolution tokens (Liang et al., 2022) and the quadratic scaling of attention mechanisms (Vaswani et al., 2017). For instance, models like LLaVA (Liu et al., 2023) and mini-Gemini-HD (Li et al., 2024b) encode high-resolution images into thousands of tokens, while video-based models such as VideoLLaVA (Lin et al., 2023) and VideoPoet (Kondratyuk et al., 2023) require even more tokens to process multiple frames. These challenges highlight the need for efficient token representations and longer context lengths to enable scalability. Recent advancements, such as Gemini (Team et al., 2023) and LWM (Liu et al., 2024a), address these issues by optimizing token efficiency and extending context length, paving the way for more scalable and effective MLLMs.

2.2 Visual Token Compression

Visual tokens often exceed text tokens by tens to hundreds of times, with visual signals being more spatially redundant compared to information dense text (Marr, 2010). Various methods have been proposed to address this issue. For instance, LLaMA-VID (Li et al., 2024a) uses a Q-Former with context tokens, and DeCo (Yao et al., 2024a) applies adaptive pooling to downsample visual tokens at the patch level. However, these approaches require modifying model components and additional training, increasing computational and training costs. ToMe (Bolya et al., 2023) reduces tokens without training by adding a token merge module to ViTs (Alexey, 2020), but this disrupts early cross-modal interactions in language models (Xing et al., 2024). FastV (Chen et al., 2024) selects important visual tokens using attention scores, while SparseVLM (Zhang et al., 2024b) incorporates text guidance via cross-modal attention. However, these methods forgo Flash-Attention (Dao et al., 2022; Dao, 2024) and primarily focus on token importance, overlooking the impact of token duplication. In our work, we maintain hardware acceleration compatibility, including Flash Attention, while focusing on another critical factor, token duplication, for efficient token reduction.

3 Methodology

3.1 Preliminary

Architecture of MLLM. The architecture of Multimodal Large Language Models (MLLMs) typically comprises three core components: a visual encoder, a modality projector, and a language model (LLM). Given an image I, the visual encoder and a subsequent learnable MLP are used to encode I into a set of visual tokens e_v. These visual tokens e_v are then concatenated with text tokens e_t encoded from text prompt p_t, forming the input for the LLM. The LLM decodes the output tokens y sequentially, which can be formulated as: $y_i = f(I, p_t, y_0, y_1, \cdots, y_{i-1})$.

3.2 Beyond Token Importance: Questioning the Status Quo

Given the computational burden associated with the length of visual tokens in MLLMs, numerous studies have embraced a paradigm that utilizes attention scores to evaluate the significance of visual tokens, thereby facilitating token reduction. Specifically, in transformer-based MLLMs, each layer performs attention computation as illustrated below:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_k}}\right) \cdot \mathbf{V}, \tag{1}$$

where d_k is the dimension of $\mathbf{K}$. The result of $\text{Softmax}(\mathbf{Q} \cdot \mathbf{K}^\top / \sqrt{d_k})$ is a square matrix known as the attention map. Existing methods extract the corresponding attention maps from one or multiple layers and compute the average attention score for each visual token based on these attention maps:

$$\phi_{\text{attn}}(x_i) = \frac{1}{N} \sum_{j=1}^{N} \text{Attention}(x_i, x_j), \tag{2}$$

where $\text{Attention}(x_i, x_j)$ denotes the attention score between token x_i and token x_j, $\phi_{\text{attn}}(x_i)$ is re-

garded as the importance score of the token x_i, N represents the number of visual tokens. Finally, based on the importance score of each token and the predefined reduction ratio, the most important visual tokens are selectively retained:

$$\mathcal{R} = \{x_i \mid (\phi_{\text{attn}}(x_i) \geq \tau)\}, \tag{3}$$

where $\mathcal{R}$ represents the set of retained visual tokens, and τ is a threshold determined by the predefined reduction ratio.

Problems: Although this paradigm has demonstrated initial success in enhancing the efficiency of MLLMs, it is accompanied by several inherent limitations that are challenging to overcome.

One key limitation is disregarding the dynamic nature of token importance during pruning. For a token sequence $\{x_1, \ldots, x_n\}$, importance-based methods compute static token importance via a scoring function $s_i = \mathcal{F}(x_i|X)$, where X is the full token set. The strategy retains Top-k tokens:

$$X_{\text{pruned}} = \arg \max_{X' \subseteq X, |X'|=k} \sum_{x_j \in X'} s_j \tag{4}$$

This implies an **independence assumption**: the score s_j remains unchanged for any subset $X' \subset X$, ignoring dynamic token interactions. For example, if two similar tokens x_p, x_q have $s_p \approx s_q$, removing x_q should recalibrate s_p as:

$$s'_p = \mathcal{F}(x_p|X' \setminus \{x_q\}) > s_p, \tag{5}$$

which leads to a bias in importance estimation $\Delta = s'_p - s_p$. This contradiction between static scoring and dynamic interaction can be quantified as:

$$\mathbb{E}_{X' \subset X} \left[\sum_{x_i \in X'} \left(\mathcal{F}(x_i|X') - \mathcal{F}(x_i|X) \right) \right] \tag{6}$$

Additionally, Figure 1 visualizes the results of token reduction, revealing that selecting visual tokens based on attention scores introduces a noticeable bias toward tokens in the lower-right region of the image, those appearing later in the visual token sequence. However, this region is not always the most significant in every image. Further, we present the outputs of various methods. Notably, FastV generates more hallucinations than the vanilla model, while DART effectively reduces them. We attribute this to the inherent bias of attention-based methods, which tend to retain tokens concentrated in specific regions, often neglecting the broader context of the image. In contrast, DART removes highly duplication tokens and preserves a more balanced distribution across the image, enabling more accurate and consistent outputs.

Furthermore, methods relying on attention scores for token importance are incompatible with Flash Attention, compromising speed, and sometimes even underperforming random token reduction in effectiveness (See Fig. 2).

3.3 Token Duplication: Rethinking Reduction

Given the numerous drawbacks associated with the paradigm of using attention scores to evaluate token importance for token reduction, *what additional factors should we consider beyond token importance in the process of token reduction?* Inspired by the intuitive ideas mentioned in §1 and the phenomenon of tokens in transformers tending toward uniformity (*i.e.*, over-smoothing) (Nguyen et al., 2023; Gong et al., 2021), we propose that token duplication should be a critical focus.

Due to the prohibitively high computational cost of directly measuring duplication among all tokens, we adopt a paradigm that involves selecting a minimal number of pivot tokens.

Definition 1 (Pivot Tokens)**.** *Let $\mathcal{P} = \{p_1, p_2, \ldots, p_k\} \subseteq \mathcal{X}$ denote the pivot tokens, where $k \ll n$ and n is the total length of the tokens $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. The pivot tokens $\mathcal{P}$ are a subset of $\mathcal{X}$, selected for their representativeness of the entire set.*

Given the pivot tokens, we can define the duplication score based on it.

Definition 2 (ϵ-duplicate Score)**.** *The token duplication score between a pivot token p_i and a visual token x_j is defined as:*

$$dup(p_i, x_j) = \frac{p_i^\top x_j}{\|p_i\| \|x_j\|}, \tag{7}$$

*where $\|\cdot\|$ denotes the Euclidean norm. Two tokens p_i, x_j are **ϵ-duplicates** if*

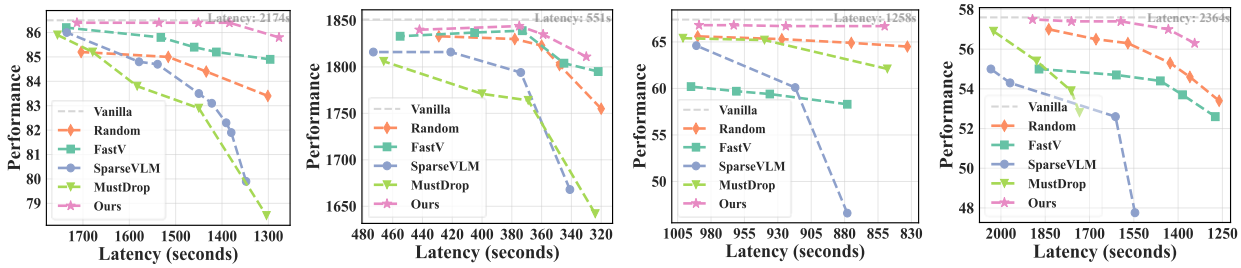
$$dup(p_i, x_j) > \epsilon. \tag{8}$$

With the ϵ-duplicate score, for each pivot p_i, the associated retained token set is defined as:

$$\mathcal{R}_i = \{x_j \mid dup(p_i, x_j) \leq \epsilon\} \tag{9}$$

The final retained set is:

$$\mathcal{R} = \mathcal{P} \cup \left(\bigcup_{p_i \in \mathcal{P}} \mathcal{R}_i \right) \tag{10}$$



(a) POPE **(b) MME** **(c) MMBench** **(d) VizWiz**

Figure 4: **Performance-Latency trade-off comparisons** across different datasets on LLaVA-Next-7B. DART consistently achieves better performance under varying latency constraints compared to other approaches.

where ϵ is the threshold dynamically determined for each pivot p_i based on reduction ratio. This ensures that only tokens that are sufficiently different from the pivot tokens are kept.

Our method is orthogonal to the paradigm of using attention scores to measure token importance, meaning it is compatible with existing approaches. Specifically, we can leverage attention scores to select pivot tokens, and subsequently incorporate token duplication into the process.

However, this still does not fully achieve compatibility with Flash Attention. Therefore, we explored alternative strategies for selecting pivot tokens, such as using K-norm, V-norm[1], or even random selection. Surprisingly, all these strategies achieve competitive performance across multiple benchmarks. This indicates that our token reduction paradigm based on token duplication is not highly sensitive to the choice of pivot tokens. Moreover, it suggests that removing duplicate tokens may be more critical than identifying "important tokens", highlighting token duplication as a more significant factor in token reduction. Detailed discussion on pivot token selection is provided in §5.2.

3.4 Theoretical Analysis

To further justify trustworthiness of our proposed method, we provide a theoretical analysis of it.

Assumption 1 (Transformer Property)**.** *For transformer property, we assume the following:*
(A1). *(Lipschitz continuity under Hausdorff distance). The model f is Lipschitz continuous with respect to the Hausdorff distance between token sets. Formally, there exists $K > 0$ such that for any two token sets $\mathcal{X}_1, \mathcal{X}_2 \subseteq \mathbb{R}^d$:*

$$\|f(\mathcal{X}_1) - f(\mathcal{X}_2)\| \leq K \cdot d_H(\mathcal{X}_1, \mathcal{X}_2),$$

where $d_H(\mathcal{X}_1, \mathcal{X}_2) \triangleq \max$

$$\left\{ \sup_{x_1 \in \mathcal{X}_1} \inf_{x_2 \in \mathcal{X}_2} \|x_1 - x_2\|, \sup_{x_2 \in \mathcal{X}_2} \inf_{x_1 \in \mathcal{X}_1} \|x_1 - x_2\| \right\}.$$

(A2). *(Bounded embedding). All tokens have bounded Euclidean norms:*

$$\|x\| \leq B, \quad \forall x \in \mathcal{X},$$

where $B > 0$ is a constant.

Lemma 1 (Bounded Distance)**.** $\min_{p_i \in \mathcal{P}} |p_i - x_j| \leq (2(1-\epsilon))^{1/2} B, \quad \forall x_j \in \mathcal{X} \setminus \mathcal{R}.$

Proof. Using A2 and Definition 2, we obtain:

$$\min_{p_i \in \mathcal{P}} |p_i - x_j|^2 = \min_{p_i \in \mathcal{P}} (|p_i|^2 + |x_j|^2 - 2p_i^\top x_j)$$
$$\leq \min_{p_i \in \mathcal{P}} (B^2 + B^2 - 2\epsilon \cdot B \cdot B) \leq 2(1-\epsilon)B^2$$

Therefore, the duplication distance bound is given by: $\min_{p_i \in \mathcal{P}} |p_i - x_j|^2 \leq (2(1-\epsilon))^{1/2} B$ □

Lemma 2 (Bounded Approximation Error)**.** *Under Assumption 1, the Hausdorff distance between original and retained tokens satisfies:*

$$d_H(\mathcal{X}, \mathcal{R}) \leq \sqrt{2(1-\epsilon)}B.$$

Proof. For any $x \in \mathcal{X}$:

- If $x \in \mathcal{R}$, then $\inf_{r \in \mathcal{R}} \|x - r\| = 0$
- If $x \notin \mathcal{R}$, by definition and Lemma 1 there exists $p_i \in \mathcal{P} \subseteq \mathcal{R}$ with $\|x - p_i\| \leq \sqrt{2(1-\epsilon)}B$

Thus:

$$\sup_{x \in \mathcal{X}} \inf_{r \in \mathcal{R}} \|x - r\| \leq \sqrt{2(1-\epsilon)}B.$$

Since $\mathcal{R} \subseteq \mathcal{X}$, Hausdorff distance simplifies to: $d_H(\mathcal{X}, \mathcal{R}) = \sup_{x \in \mathcal{X}} \inf_{r \in \mathcal{R}} \|x - r\| \leq \sqrt{2(1-\epsilon)}B.$ □

Theorem 1 (Performance Guarantee)**.** *Under Assumptions 1, the output difference between original and pruned token sets is bounded by:*

$$\|f(\mathcal{X}) - f(\mathcal{R})\| \leq K\sqrt{2(1-\epsilon)}B.$$

Proof. Direct application of Lipschitz continuity (A1) with Lemma 2: $\|f(\mathcal{X}) - f(\mathcal{R})\| \leq K \cdot d_H(\mathcal{X}, \mathcal{R}) \leq K\sqrt{2(1-\epsilon)}B.$ □

[1]Here, the K-norm and V-norm refer to the L1-norm of K matrix and V matrix in attention computing, respectively.

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	VizWiz	OCRBench	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**										
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	78.5	58.2	50.0	297	100%
LLaVA-1.5-7B	*Retain 192 Tokens* (↓ **66.7%**)										
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	-	88.5%
FastV (ECCV24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	291	91.2%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	74.9	47.4	50.1	190	91.5%
LLaVA-PruMerge (2024.05)	54.3	59.6	52.9	1632	71.3	67.9	70.6	54.3	50.1	253	90.8%
SparseVLM (2024.10)	57.6	62.5	53.7	1721	**83.6**	69.1	75.6	56.1	50.5	292	96.3%
PDrop (2024.10)	57.1	63.2	56.8	1766	82.3	68.8	75.1	56.1	51.1	290	96.7%
FiCoCo-V (2024.11)	58.5	62.3	55.3	1732	82.5	67.8	74.4	55.7	51.0	-	96.1%
MustDrop (2024.11)	58.2	62.3	55.8	1787	82.6	69.2	76.0	56.5	**51.4**	289	97.2%
DART (Ours)	**60.0**	**63.6**	**57.0**	**1856**	82.8	**69.8**	**76.7**	**57.4**	51.2	**296**	**98.8%**
LLaVA-1.5-7B	*Retain 128 Tokens* (↓ **77.8%**)										
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	-	80.4%
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	51.3	285	86.4%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	73.4	46.1	51.3	191	90.2%
LLaVA-PruMerge (2024.05)	53.3	58.1	51.7	1554	67.2	67.1	68.8	54.3	50.3	248	88.8%
SparseVLM (2024.10)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	280	93.8%
PDrop (2024.10)	56.0	61.1	56.6	1644	**82.3**	68.3	72.9	55.1	51.0	287	95.1%
FiCoCo-V (2024.11)	57.6	61.1	54.3	1711	82.2	68.3	73.1	55.6	49.4	-	94.9%
MustDrop (2024.11)	56.9	61.1	55.2	1745	78.7	68.5	74.6	56.3	**52.1**	281	95.6%
DART (Ours)	**58.7**	**63.2**	**57.5**	**1840**	80.1	**69.1**	**75.9**	**56.4**	51.7	**296**	**98.0%**
LLaVA-1.5-7B	*Retain 64 Tokens* (↓ **88.9%**)										
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	-	70.1%
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	245	77.3%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	69.7	44.2	50.2	191	87.0%
LLaVA-PruMerge (2024.05)	51.9	55.3	49.1	1549	65.3	68.1	67.4	54.0	50.1	250	87.4%
SparseVLM (2024.10)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	180	84.6%
PDrop (2024.10)	41.9	33.3	50.5	1092	55.9	68.6	69.2	45.9	50.7	250	78.1%
FiCoCo-V (2024.11)	52.4	60.3	53.0	1591	**76.0**	68.1	71.3	53.6	49.8	-	91.5%
MustDrop (2024.11)	53.1	60.0	53.1	1612	68.0	63.4	69.3	54.2	51.2	267	90.1%
DART (Ours)	**55.9**	**60.6**	**53.2**	**1765**	73.9	**69.8**	**72.4**	**54.4**	**51.6**	**270**	**93.7%**
LLaVA-Next-7B	*Upper Bound, 2880 Tokens* **(100%)**										
Vanilla	64.2	67.4	60.6	1851	86.5	70.1	81.8	64.9	57.6	517	100%
LLaVA-Next-7B	*Retain 320 Tokens* (↓ **88.9%**)										
FastV (ECCV24)	55.9	61.6	51.9	1661	71.7	62.8	71.9	55.7	53.1	374	86.4%
HiRED (AAAI25)	59.3	64.2	55.9	1690	83.3	66.7	75.7	58.8	54.2	404	91.8%
LLaVA-PruMerge (2024.05)	53.6	61.3	55.3	1534	60.8	66.4	69.7	50.6	54.0	146	79.9%
SparseVLM (2024.10)	56.1	60.6	54.5	1533	82.4	66.1	71.5	58.4	52.0	270	85.9%
PDrop (2024.10)	56.4	63.4	56.2	1663	77.6	67.5	73.5	54.4	54.1	259	86.8%
MustDrop (2024.11)	57.3	62.8	55.1	1641	82.1	68.0	73.7	**59.9**	54.0	382	90.4%
FasterVLM (2024.12)	56.9	61.6	53.5	1701	83.6	66.5	74.0	56.5	52.6	401	89.8%
GlobalCom^2 (2025.01)	57.1	61.8	53.4	1698	83.8	67.4	76.7	57.2	54.6	375	90.3%
DART (Ours)	**61.7**	**65.3**	**58.2**	**1710**	**84.1**	**68.4**	**79.1**	58.7	**56.1**	**406**	**93.9%**

Table 1: Comparative experiments on image understanding. In all experiments for DART, tokens are pruned after the second layer with 8 pivot tokens. The pivot tokens are selected based on the maximum K-norm.

4 Experiments

Experiment Setting. We conduct experiments on over four MLLMs across ten image-based and four video-based benchmarks. For details on implementation, please refer to Appendix B.

4.1 Main Results

Image understanding task. The results presented in Table 1 highlight the exceptional performance of **DART**, across a diverse range of image understanding tasks under varying vision token configurations. We can observe that **(i)** with only 192 tokens retained (↓ 66.7%), DART achieves an impressive average performance of 98.8%, substantially outperforming the second-best method, MustDrop by **1.6**%. **(ii)** This trend becomes even more pronounced under more aggressive reduction ratios, with DART outperforming by **2.2**% while retaining only 64 tokens. **(iii)** Moreover, DART scales seamlessly to more advanced models, as evidenced by its performance on the LLaVA-Next-7B backbone, where it achieves an average score of **93.9**% with only 11.1% of visual tokens retained, outperforming all competing methods by a significant margin. These results highlight DART's exceptional efficiency in leveraging limited visual tokens while preserving critical information, showcasing its robust performance across diverse tasks and seamless adaptability to various model architectures.

Video Understanding Task. To assess DART's capabilities in video understanding, we integrate it with Video-LLaVA (Lin et al., 2023) and benchmark it against state-of-the-art methods, including FastV (Chen et al., 2024). Following established protocols, Video-LLaVA processes videos by sampling 8 frames and extracting 2048 vision tokens, with 50% retained for evaluation. As demonstrated in Table 5, DART surpasses FastV across all bench-

Methods	Tokens ↓	Total Time ↓ (Min:Sec)	Prefilling Time ↓ (Min:Sec)	FLOPs ↓	KV Cache ↓ (MB)	POPE ↑ (F1-Score)	Speedup ↑ (Total)	(Prefilling)
Vanilla LLaVA-Next-7B	2880	36:16	22:51	100%	1512.1	86.5	1.00×	1.00×
+ FastV	320	18:17	7:41	**12.8%**	168.0	78.3	1.98×	2.97×
+ SparseVLM	320	23:11	-	15.6%	168.0	82.3	1.56×	-
+ DART	320	**18:13**	**7:38**	**12.8%**	168.0	**84.1**	**1.99**×	**2.99**×

Table 2: Inference costs of the number of tokens, Total-Time, Prefilling-Time, FLOPs, and KV Cache Memory.

marks, achieving a notable 4.0 score on MSVD, 46.3% accuracy on TGIF, and 56.7% accuracy on MSRVT. With an average accuracy of 58.0% and an evaluation score of 3.7, DART demonstrates superior reasoning over complex multimodal data.

5 Analysis and Discussion

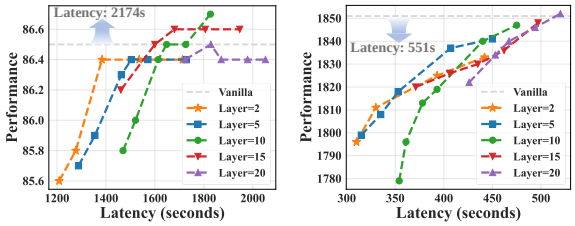
5.1 Efficiency Analysis

As shown in Table 2, we compare the total inference time, prefill time, FLOPs, and KV cache memory of multiple methods. **(i)** DART achieves a **2.99× speedup** in the prefill phase and a **1.99× speedup** in the entire inference, while its performance on POPE degrades by less than **3%** compared to the vanilla model. **(ii)** Further analysis reveals that *although the FLOPs reduction of different methods is similar, their actual inference speeds vary significantly.* For instance, SparseVLM increases FLOPs by only **2.8%** compared to DART, but its speedup drops by **21.6%**, demonstrating that relying solely on FLOPs to measure acceleration effectiveness lacks practical significance. **(iii)** We further evaluate the performance-latency trade-off of each method based on actual latency. As illustrated in Figure 4, *when considering the balance between latency and performance, some existing methods even underperform random visual token retention*. We argue that SparseVLM and MustDrop suffer from significant speed degradation due to their sequential multi-stage token processing. For FastV, its reliance on biased attention scores for "important" token selection results in slightly inferior performance. In contrast, DART seamlessly integrates Flash Attention and incurs less than **0.08s** overhead for token reduction, achieving a better trade-off between performance and speed.

5.2 Influence from Selection of Pivot Tokens

In this section, we investigate whether pivot token selection in DART significantly affects its performance. Table 6 in Appendix A.1 evaluates pivot tokens based on criteria such as maximum (♠), minimum (♡) attention scores, K-norm, V-norm, and random selection. Results show that various strategies achieve over 94.9% of the vanilla model's performance across benchmarks. *Even DART with randomly selected pivot tokens incurs only a* 1.2% *performance drop compared to the best strategy and outperforms the previous importance-based methods by* 2.1%. This observation shows the robustness in the selection of pivot tokens in DART, and highlights the crucial role of duplication in token reduction, as *selecting "important" pivot tokens based on attention scores is only 0.2% better than selecting "unimportant" ones as pivot tokens*.

Furthermore, on the MME benchmark, we analyze the visual tokens retained by selecting pivot tokens based on K-norm♠ and K-norm♡. Interestingly, statistical analysis shows that the overlap between tokens preserved by these two strategies is, on average, less than **50%**. Despite this low overlap, both strategies achieve highly effective results, *indicating the existence of multiple distinct groups of tokens which should not be pruned*. This finding challenges the conventional notion of a single critical token set defined by importance scores, demonstrating that diverse token subsets with minimal overlap can yield comparable performance.



(a) POPE (b) MME

Figure 5: Influence from the layer for token pruning.

5.3 Influence from Choice of the Pruned Layer and the Number of Pivot Tokens

We explore the impact of layer on model performance. As expected, pruning deeper layers yields performance closer to the vanilla model but increases latency, as shown in Figure 5. However, we observe two intriguing findings: **(i)** Pruning at layers 10, 15, and 20 surprisingly outperforms the vanilla model (Fig. 5(a)), consistent with Fig. 1, suggesting that removing duplicate tokens may reduce hallucinations in MLLMs on the POPE. **(ii)** At deeper layers (*e.g.*, 15, 20), the latency-minimizing points correspond to pruning all vision tokens, yet performance drops only by **0.1%∼1.6%**. This highlights a modality imbalance in MLLMs, in-

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	Avg.
Qwen2-VL-7B	*Upper Bound, All Tokens* **(100%)**							
Vanilla	62.2	80.5	81.2	2317	86.1	84.7	82.1	100%
Qwen2-VL-7B	*Token Reduction* (↓ **66.7%**)							
+ FastV (ECCV24)	58.0	76.1	75.5	2130	82.1	80.0	77.3	94.0%
+ DART (Ours)	60.2	78.9	78.0	2245	83.9	81.4	80.5	97.0%
Qwen2-VL-7B	*Token Reduction* (↓ **77.8%**)							
+ FastV (ECCV24)	56.7	74.1	73.9	2031	79.2	78.3	72.0	91.0%
+ DART (Ours)	58.5	77.3	77.1	2175	82.1	79.6	75.3	94.3%
Qwen2-VL-7B	*Token Reduction* (↓ **88.9%**)							
+ FastV (ECCV24)	51.9	70.1	65.2	1962	76.1	75.8	60.3	84.0%
+ DART (Ours)	55.5	72.0	71.7	2052	77.9	77.6	61.8	87.5%

Table 3: Comparative Experiments on Qwen2-VL-7B.

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	Avg.
MiniCPM-V2.6	*Upper Bound, All Tokens* **(100%)**							
Vanilla	51.5	79.7	77.9	2267	83.2	95.6	78.5	100%
MiniCPM-V2.6	*Token Reduction* (↓ **66.7%**)							
+ FastV (ECCV24)	43.2	74.9	73.1	1895	75.4	89.8	67.1	89.0%
+ DART (Ours)	47.8	76.5	74.8	1951	77.4	91.8	70.9	92.9%
MiniCPM-V2.6	*Token Reduction* (↓ **77.8%**)							
+ FastV (ECCV24)	41.3	72.9	70.4	1807	70.2	86.5	54.9	83.4%
+ DART (Ours)	47.8	73.8	71.4	1821	71.6	88.9	65.7	88.6%
MiniCPM-V2.6	*Token Reduction* (↓ **88.9%**)							
+ FastV (ECCV24)	35.5	61.4	60.8	1376	56.9	80.4	33.4	68.4%
+ DART (Ours)	42.5	66.2	64.0	1405	58.0	83.5	51.9	76.1%

Table 4: Comparative Experiments on MiniCPM-V2.6.

Methods	TGIF		MSVD		MSRVT		Avg.	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM-1B	41.9	-	32.2	-	16.8	-	30.3	-
VideoChat-7B	34.4	2.3	56.3	2.8	45.0	2.5	45.1	2.5
LLaMA-Adapter-7B	-	-	54.9	3.1	43.8	2.7	-	-
Video-LLaMA-7B	-	-	51.6	2.5	29.6	1.8	-	-
Video-ChatGPT-7B	51.4	3.0	64.9	3.3	49.3	2.8	55.2	3.0
Video-LLaVA-7B	47.0	3.4	70.2	3.9	57.3	3.5	58.2	3.6
+ FastV-7B	45.2	3.1	**71.0**	3.9	55	3.5	57.1	3.5
+ DART-7B (Ours)	**46.3**	**3.4**	**71.0**	**4.0**	**56.7**	**3.6**	**58.0**	**3.7**

Table 5: **Comparing MLLMs on Video Understanding tasks** with 50% visual tokens retained.

dicating underutilization of the visual modality.

Furthermore, we delved into the impact of the number of pivot tokens on performance. As depicted in Figure 6, choosing either an insufficient or an excessive number of pivot tokens leads to suboptimal outcomes. When a limited number of pivot tokens (*e.g.*, one or two), the lack of diversity among these tokens may impede their ability to comprehensively represent the entire feature space. In contrast, when an overly large number of pivot tokens, for example, 20 or more, are chosen, the majority of retained visual tokens tend to be pivot tokens. In extreme cases, our approach starts to resemble the importance-based method, where pivot tokens essentially transform into important tokens, overlooking the impact of duplication factors.

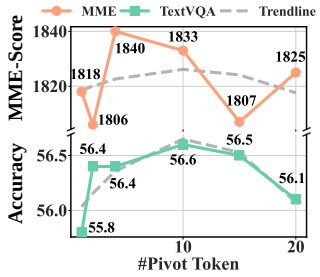
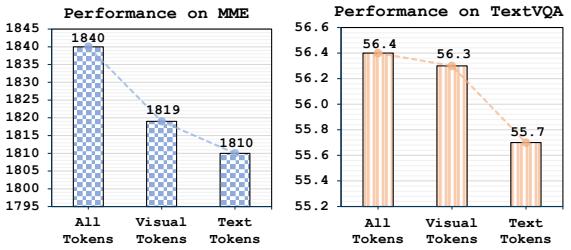


Figure 6: Impact of the number of pivot tokens.

5.4 Influence from Modalities of Pivot Tokens

We further analyze the impact of the source of pivot tokens on the overall performance of DART, with a particular focus on understanding whether guidance from the language modality is essential for effective token reduction. We evaluate the performance implications of selecting pivot tokens exclusively from either the visual or text modality, aiming to quantify the influence of each modality. As illustrated in Figure 7, the absence of pivot tokens from either modality leads to a noticeable decline in performance. This demonstrates that information from both modalities contributes to the token reduction process to varying degrees. Moreover, it highlights that we provide an effective method for incorporating textual guidance without the need to explicitly compute cross-modal attention scores while remaining compatible with Flash Attention.



(a) MME (b) TextVQA

Figure 7: **Analysis of pivot token sources**: "ALL Tokens" selects from both visual and textual modalities, while "Visual Tokens" and "Text Tokens" select exclusively from visual or textual modalities, respectively.

6 Conclusion

The pursuit of efficient token reduction in MLLMs has traditionally centered on token "importance" often measured by attention scores, sometimes suffering from worse performance than random token pruning. This study introduces DART, which focuses on the duplication of tokens, aiming to remove the tokens that are similar to others, leading to a superior balance between performance and latency in 13 benchmarks and 5 MLLMs (Tab. 1, 2, 3, 4, and Fig. 4). Our exploration yields surprising insights: multiple distinct sets of retained tokens, sharing less than 50% overlap, can deliver comparable and good performance (§5.2). Additionally, token pruning shows potential to mitigate hallucinations (§5.3). These observations highlight the limitations of importance-based methods and may offer valuable insights into the influence of vision tokens in MLLMs.

7 Limitations

One of the limitations of our work is that it cannot be applied to black-box models like the GPT (*e.g.* GPT 3.5 and more advanced versions) and Claude series, as we are unable to access their encoded tokens during the inference process. Additionally, our proposed DART exhibits slightly degraded performance when dealing with models that encode tokens with high information density. For instance, Qwen2-VL employs token merging during training, and MiniCPM-V-2.6 utilizes learnable queries in its Resampler module to map variable-length patch features into shorter feature representations. In such cases, removing a single visual token results in a greater loss of information, thereby impacting the effectiveness of our approach. Despite this, DART is still able to maintain 97.0% and 92.9% of its performance while reducing visual tokens by 66.7%, as demonstrated in Tables 3, 4.

References

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2024. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models in resource-constrained environments. *arXiv preprint arXiv:2408.10945*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your ViT but faster. In *International Conference on Learning Representations*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*.

Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. 2021. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv:2312.14125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024a. LLaMA-VID: An image is worth 2 tokens in large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*.

Y Liang, C Ge, Z Tong, Y Song, P Xie, et al. 2022. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *ICLR*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2024a. World model on million-length video and language with ringattention. *Preprint*, arXiv:2402.08268.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. 2024e. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*.

Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, Siteng Huang, and Honggang Chen. 2025a. Compression with global guidance: Towards training-free high-resolution mllms acceleration. *arXiv preprint arXiv:2501.05179*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025b. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024f. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

David Marr. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Tam Nguyen, Tan Nguyen, and Richard Baraniuk. 2023. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36:80233–80256.

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

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv:1706.03762*.

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

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024b. Internvideo2: Scaling video foundation models for multimodal video understanding. *Arxiv e-prints*, pages arXiv–2403.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024c. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM international conference on Multimedia*, pages 1645–1653.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024a. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv:2405.20985*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024b. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. 2024b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Additional Experiments

A.1 Supplementary Results on Pivot Token Selection

This section presents comprehensive experimental results conducted on the LLaVA-1.5-7B model, supporting the analysis of pivot token selection strategies within DART. Table 6 details performance metrics across multiple benchmarks, including GQA, MMB, MME, POPE, SQA, and VQA, with all experiments retaining 128 vision tokens. These findings further validate the robustness of DART under various pivot token selection criteria, ranging from random selection to methods based on attention scores and norm-based approaches. The table also includes comparisons with baseline methods (*e.g.*, SparseVLM and FastV), highlighting the consistent superiority of DART across different configurations. For additional insights, refer to the main discussion in §5.2.

B Detailed Experiment Settings

B.1 Datasets

Our experiments are conducted on a suite of widely recognized benchmarks, each designed to evaluate distinct aspects of multimodal intelligence. For image understanding task, we performed experiments on ten widely used benchmarks, including GQA (Hudson and Manning, 2019), MMBench (MMB) and MMB-CN (Liu et al., 2025b), MME (Fu et al., 2023), POPE (Li et al., 2023b), VizWiz (Bigham et al., 2010), SQA (Lu et al., 2022), VQA^{V2} (VQA V2) (Goyal et al., 2017), VQA^{Text} (TextVQA) (Singh et al., 2019), and OCRBench (Liu et al., 2024f). For video understanding task, we evaluated our method on three video-based benchmarks: TGIF-QA (Jang et al., 2017), MSVD-QA (Xu et al., 2017), and MSRVTT-QA (Xu et al., 2017).

GQA. GQA is structured around three core components: scene graphs, questions, and images. It includes not only the images themselves but also detailed spatial features and object-level attributes. The questions are crafted to assess a model's ability to comprehend visual scenes and perform reasoning tasks based on the image content.

MMBench. MMBench offers a hierarchical evaluation framework, categorizing model capabilities into three levels. The first level (L-1) focuses on perception and reasoning. The second level (L-2) expands this to six sub-abilities, while the third level (L-3) further refines these into 20 specific dimensions. This structured approach allows for a nuanced and comprehensive assessment of a model's multifaceted abilities. MMBench-CN is the Chinese version of the dataset.

MME. The MME benchmark is designed to rigorously evaluate a model's perceptual and cognitive abilities through 14 subtasks. It employs carefully constructed instruction-answer pairs and concise instructions to minimize data leakage and ensure fair evaluation. This setup provides a robust measure of a model's performance across various tasks.

POPE. POPE is tailored to assess object hallucination in models. It presents a series of binary questions about the presence of objects in images, using accuracy, recall, precision, and F1 score as metrics. This approach offers a precise evaluation of hallucination levels under different sampling strategies.

ScienceQA. ScienceQA spans a wide array of domains, including natural, language, and social sciences. Questions are hierarchically categorized into 26 topics, 127 categories, and 379 skills, providing a diverse and comprehensive testbed for evaluating multimodal understanding, multi-step reasoning, and interpretability.

VQA V2. VQA V2 challenges models with open-ended questions based on 265,016 images depicting a variety of real-world scenes. Each question is accompanied by 10 human-annotated answers, enabling a thorough assessment of a model's ability to accurately interpret and respond to visual queries.

TextVQA. TextVQA emphasizes the integration of textual information within images. It evaluates a model's proficiency in reading and reasoning about text embedded in visual content, requiring both visual and textual comprehension to answer questions accurately.

OCRBench. OCRBench is a comprehensive benchmark for evaluating the OCR capabilities of multi-modal language models across five key tasks: text recognition, scene text-centric and document-oriented VQA, key information extraction, and handwritten mathematical expression recognition.

TGIF-QA. TGIF-QA extends the image question-answering task to videos, featuring 165,000 question-answer pairs. It introduces tasks that require spatio-temporal reasoning, such as repetition count and state transition, as well as frame-based questions, promoting advancements in video question answering.

MSVD-QA. Based on the MSVD dataset,

Benchmark	Vanilla	Pivot Token Selection							Other Methods	
		Random	A-Score♠	A-Score♡	K-norm♠	K-norm♡	V-norm♠	V-norm♡	SparseVLM	FastV
GQA	61.9	59.0	59.2	58.4	58.7	59.1	57.3	59.4	56.0	49.6
MMB	64.7	63.2	63.1	62.9	63.2	64.0	62.5	64.3	60.0	56.1
MME	1862	1772	1826	1830	1840	1820	1760	1825	1745	1490
POPE	85.9	80.6	81.1	81.0	80.1	80.2	76.8	81.6	80.5	59.6
SQA	69.5	69.0	69.9	68.9	69.1	68.7	69.2	68.9	68.5	60.2
VQA^{V2}	78.5	75.2	75.9	76.0	75.9	75.6	75.4	76.1	73.8	61.8
VQA^{Text}	58.2	56.0	55.7	56.5	56.4	55.4	55.5	56.0	54.9	50.6
Avg.	100%	96.0%	96.9%	96.7%	96.8%	96.8%	94.9%	97.2%	93.9%	81.5%

Table 6: **Analysis on how to select the pivot token.** This study evaluates pivot tokens, comprising a fixed set of 4 visual and 4 text tokens, using various criteria with 128 retained tokens. **A-Score** denotes the Attention Score. ♠ represents selecting token with the highest value as the pivot token. ♡ represents selecting the token with the smallest value as the pivot token. For instance, **A-Score♠** means selecting the token with the highest value of Attention Score as the pivot token.

MSVD-QA includes 1970 video clips and approximately 50.5K QA pairs. The questions cover a broad spectrum of topics and are open-ended, categorized into what, who, how, when, and where types, making it a versatile tool for video understanding tasks.

MSRVTT-QA. MSRVTT-QA comprises 10K video clips and 243K QA pairs. It addresses the challenge of integrating visual and temporal information in videos, requiring models to effectively process both to answer questions accurately. Similar to MSVD-QA, it includes five types of questions, further enriching the evaluation landscape.

B.2 Models

We evaluate DART using various open-source MLLMs. For image understanding tasks, experiments are conducted on the LLaVA family, including LLaVA-1.5-7B[2] (Liu et al., 2024d) and LLaVA-Next-7B[3] (Liu et al., 2024c), with the latter used to validate performance on high-resolution images. Furthermore, we validate our method on more advanced models, including Qwen2-VL-7B[4] (Wang et al., 2024a) and MiniCPM-V-2.6[5] (Yao et al., 2024b). For video understanding tasks, we use Video-LLaVA (Lin et al., 2023) as the baseline model. following the settings reported in their paper to ensure a fair comparison.

[2] https://huggingface.co/liuhaotian/llava-v1.5-7b

[3] https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b

[4] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[5] https://huggingface.co/openbmb/MiniCPM-V-2_6

B.3 Baselines

We analyze multiple representative methods for accelerating multi-modal language models (MLLMs) through token reduction. These methods share the goal of improving efficiency by reducing redundant tokens, yet differ in their strategies, such as token merging, pruning, or adaptive allocation.

ToMe (Bolya et al., 2023) merges similar tokens in visual transformer layers through lightweight matching techniques, achieving acceleration without requiring additional training.

FastV (Chen et al., 2024) focuses on early-stage token pruning by leveraging attention maps, effectively reducing computational overhead in the initial layers.

SparseVLM (Zhang et al., 2024b) ranks token importance using cross-modal attention and introduces adaptive sparsity ratios, complemented by a novel token recycling mechanism.

HiRED (Arif et al., 2024) allocates token budgets across image partitions based on `CLS` token attention, followed by the selection of the most informative tokens within each partition, ensuring spatially aware token reduction.

LLaVA-PruMerge (Shang et al., 2024) combines pruning and merging strategies by dynamically removing less important tokens using sparse `CLS`-visual attention and clustering retained tokens based on key similarity.

PDrop (Xing et al., 2024) adopts a progressive token-dropping strategy across model stages, forming a pyramid-like token structure that balances efficiency and performance.

MustDrop (Liu et al., 2024e) integrates multiple strategies, including spatial merging, text-guided pruning, and output-aware cache policies, to reduce

tokens across various stages.

FasterVLM (Zhang et al., 2024a) evaluates token importance via CLS attention in the encoder and performs pruning before interaction with the language model, streamlining the overall process.

GlobalCom^2 (Liu et al., 2025a) introduces a hierarchical approach by coordinating thumbnail tokens to allocate retention ratios for high-resolution crops while preserving local details.

These methods collectively highlight diverse approaches to token reduction, ranging from attention-based pruning to adaptive merging, offering complementary solutions for accelerating MLLMs.

B.4 Implementation Details

All of our experiments are conducted on Nvidia A100-80G GPU. The implementation was carried out in Python 3.10, utilizing PyTorch 2.1.2, and CUDA 11.8. All baseline settings follow the original paper.

C Computational Complexity.

To evaluate the computational complexity of MLLMs, it is essential to analyze their core components, including the self-attention mechanism and the feed-forward network (FFN). The total floating-point operations (FLOPs) required can be expressed as:

$$\text{Total FLOPs} = T \times (4nd^2 + 2n^2d + 2ndm), \quad (11)$$

where T denotes the number of transformer layers, n is the sequence length, d represents the hidden dimension size, and m is the intermediate size of the FFN. This equation highlights the significant impact of sequence length n on computational complexity. Notable, we follow FastV (Chen et al., 2024) to roughly estimate various token reduction baseline FLOPs. The FLOPs after token pruning can be represented as:

$$\begin{aligned}&\text{Post-Pruning FLOPs} \\ &= L \times (4nd^2 + 2n^2d + 2ndm) + \\ &\quad (T - L) \times (4\hat{n}d^2 + 2\hat{n}^2d + 2\hat{n}dm),\end{aligned} \quad (12)$$

where L denotes the pruned layer, $\hat{n}$ represents token sequence length after pruning. The theoretical FLOPs reduction ratio related to visual tokens is computed as:

$$1 - \frac{\text{Post-Pruning FLOPs}}{\text{Total FLOPs}}. \quad (13)$$

D Future Works

As can be observed from Figure 1 and Figure 5(a), in certain cases, token pruning contributes to the reduction of hallucinations. Our method achieved better results than the vanilla model on the POPE benchmark, which is specifically designed for evaluating the hallucination issues of multimodal large language models. Therefore, we believe that it is worth exploring in the future why token pruning is beneficial for reducing hallucinations and how we can better utilize efficient techniques (*e.g.*, token pruning, and token merge) to reduce hallucinations while achieving acceleration benefits.

E Sparsification Visualization on Different Pivot Token Selection Strategy

Figure 8 showcases a diverse array of sparsification visualization examples on different pivot token selection strategy, including K-norm♠, K-norm♡, V-norm♠, V-norm♡, Attention Score♠, Attention Score♡, and Random. Here, we can observe two interesting points: (i) The commonality is that DART employs different pivot token selection strategies for token reduction, and the retained tokens are distributed in a relatively scattered manner without obvious bias, *i.e.*, spatial uniformity, which contributes to a more accurate understanding of the entire image and consistent responses. (ii) The difference lies in the fact that although each strategy achieves comparable performance, it is noticeable that the final set of retained tokens varies significantly across strategies, indicating the existence of multiple token sets that can deliver satisfactory results. This further corroborates the limitation of selecting a unique set of tokens based solely on importance scores.

Figure 8: **Sparsification Visualization examples of DART on different Pivot Token Selection Strategy.**