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., 2024d) 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

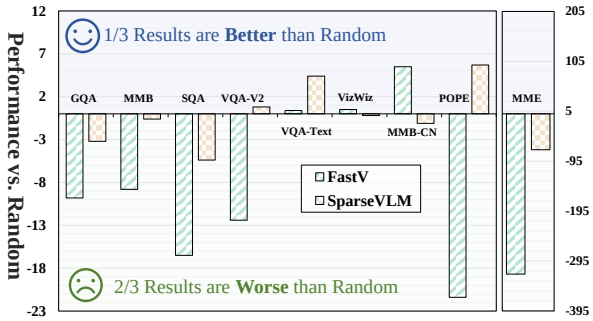


Figure 2: **Performance of FastV and SparseVLM compared with random token pruning** on the LLaVA-1.5-7B, with a **88.9%** token reduction ratio.

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 of 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

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) excel at image, video, and multimodal reasoning by integrating vision and text. However, visual data processing is costly due to redundancy, low information density (Liang et al.,

(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.

2022), and the quadratic cost of attention (Vaswani et al., 2017). For instance, models like LLaVA (Liu et al., 2023) and mini-Gemini-HD (Li et al., 2024d) encode high-resolution images into thousands of tokens, while video models like VideoLLaVA (Lin et al., 2023) and VideoPoet (Kondratyuk et al., 2023) handle even more tokens across frames. These challenges highlight the need for efficient token representations and longer context. Recent work like Gemini (Team et al., 2023) and LWM (Liu et al., 2024a) addresses this by improving token efficiency and extending context, enabling more scalable MLLMs.

Visual Token Compression Visual tokens often outnumber text tokens by tens to hundreds of times, as visual signals are more spatially redundant than information-dense text (Marr, 2010). LLaMA-VID (Li et al., 2024c) employs a Q-Former with context tokens, and DeCo (Yao et al., 2024a) uses adaptive pooling. DTMFormer (Wang et al., 2024d) improves ViTs' efficiency in medical image segmentation by merging redundant tokens during training. MADTP (Cao et al., 2024) reduces computation by aligning cross-modal features and pruning tokens. However, these require modifying components and additional training. ToMe (Bolya et al., 2023) merges tokens without training but disrupts cross-modal interactions (Xing et al., 2024). FastV (Chen et al., 2024) selects via attention scores, while SparseVLM (Zhang et al., 2024b) uses text guidance. Yet, these forgo Flash-Attention (Dao et al., 2022; Dao, 2024), neglecting token duplication. We preserve hardware acceleration (*i.e.*, Flash-Attention) and target 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 the 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 regarded 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:

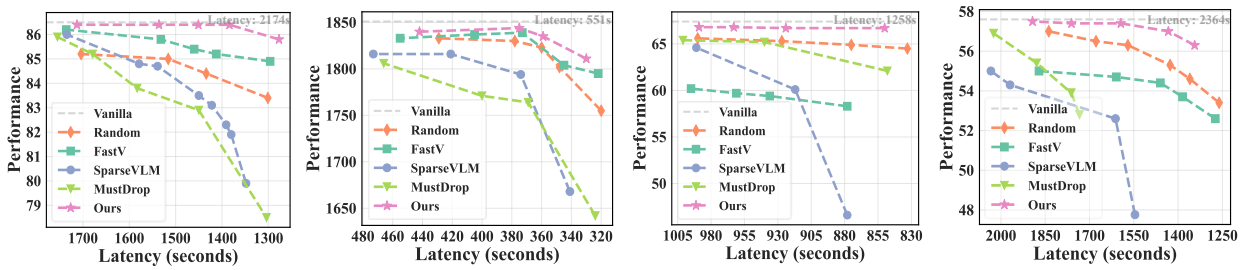
$$\mathcal{R}_i = \{x_j \mid dup(p_i, x_j) \leq \epsilon\} \tag{9}$$

The final retained set is:

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

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.



(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.

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.

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

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

This provides a theoretical guarantee that DART preserves model output within a controllable bound, thereby supporting the trustworthiness and robustness of our method.

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	100.0%
LLaVA-1.5-7B	*Retain 192 Tokens* (↓ **66.7%**)										
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	-	-
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%
FitPrune (AAAI25)	**60.4**	63.3	56.4	1831	83.4	67.8	-	57.4	50.9	-	-
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 (ICML25)	57.6	62.5	53.7	1721	**83.6**	69.1	75.6	56.1	50.5	292	96.3%
PDrop (CVPR25)	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%**
DART † (Ours)	60.9	66.3	59.5	1829	85.3	70.1	78.2	56.8	51.3	304	100.4%
LLaVA-1.5-7B	*Retain 128 Tokens* (↓ **77.8%**)										
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	-	-
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%
FitPrune (AAAI25)	58.5	62.7	56.2	1776	77.9	68.0	-	55.7	51.7	-	-
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 (ICML25)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	280	93.8%
PDrop (CVPR25)	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%**
DART † (Ours)	59.8	65.6	58.3	1849	84.4	70.7	77.5	56.4	52.6	299	99.9%
LLaVA-1.5-7B	*Retain 64 Tokens* (↓ **88.9%**)										
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	-	-
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%
FitPrune (AAAI25)	52.3	58.5	49.7	1556	60.9	68.0	-	51.2	51.1	-	-
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 (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	180	84.6%
PDrop (CVPR25)	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%**
DART † (Ours)	57.1	64.7	56.7	1823	79.3	71.1	74.6	54.7	52.1	286	97.2%

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. DART † indicates that DART is applied during the training stage of LLaVA-1.5-7B.

Methods	Tokens ↓	Total Time ↓ (Min:Sec)	Prefilling Time ↓ (Min:Sec)	FLOPs ↓	KV Cache ↓ (MB)	POPE ↑ (F1-Score)	Speedup ↑ (Total)	Speedup ↑ (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.

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 C.

4.1 Main Results

Image understanding task. The results presented in Tables 1 and 3 highlight **DART**'s exceptional performance across diverse image understanding tasks under varying token configurations. We observe that **(i)** with only 192 tokens, DART achieves an impressive 98.8% average performance, substantially outperforming second-best MustDrop by **1.6**%. **(ii)** This trend strengthens under aggressive reduction ratios, with DART leading by **2.2**% using just 64 tokens. **(iii)** Moreover, DART scales seamlessly to advanced and larger models like LLaVA-Next-7B and Qwen2-VL-72B (See Tab. 7), achieving **93.9**% with only 11.1% tokens, outperforming all competitors significantly. **(iv)** Inspired by (Wen et al., 2025), we apply DART during training. DART † in Table 1 shows better performance-efficiency trade-offs, maintaining full performance with just 192 visual tokens, highlighting the strong adaptability of our method. These results demonstrate DART's efficiency in leveraging limited tokens while preserving critical information, showcasing robust performance across tasks, model architectures, and model size. For more comparisons, please refer to Tables 4, 5, and Appendix A.3.

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

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	VizWiz	OCRBench	Avg.
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	100.0%
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 (ICML25)	56.1	60.6	54.5	1533	82.4	66.1	71.5	58.4	52.0	270	85.9%
PDrop (CVPR25)	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 3: Comparative experiments are performed on LLaVA-Next-7B using the same settings as LLaVA-1.5-7B.

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 6, DART surpasses FastV across all benchmarks, 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 prefill and **1.99× speedup** in inference, while its performance on POPE degrades by less than **3%** versus the vanilla model. **(ii)** Analysis reveals *although FLOPs reduction is similar across methods, their speeds vary significantly.* For instance, SparseVLM increases FLOPs by **2.8%** versus DART, but its speedup drops **21.6%**, showing FLOPs alone poorly measure acceleration. **(iii)** We evaluate performance-latency trade-offs using actual latency. Figure 4 shows *some methods underperform random token retention*. SparseVLM and MustDrop suffer speed degradation from sequential token processing. FastV's biased attention scores yield worse performance. In contrast, DART integrates Flash Attention with under **0.08s** overhead, achieving better performance-speed balance.

5.2 Influence from Selection of Pivot Tokens

In this section, we investigate whether pivot token selection in DART significantly affects its performance. Table 8 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.

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 6. However, we observe two intriguing findings: **(i)** Pruning at layers 10, 15, and 20 surprisingly outperforms the vanilla model (Fig. 6(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

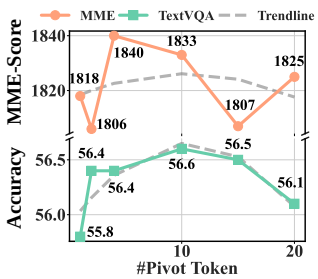


Figure 5: Impact of the number of pivot tokens.

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 4: Comparative Experiments on Qwen2-VL-7B.

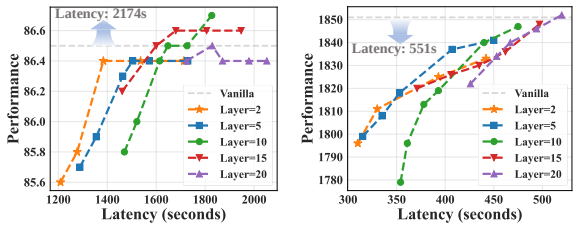
Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	Avg.
MiniCPM-V2.6	*Upper Bound, All Tokens* **(100%)**							
Vanilla	51.5	79.7	77.9	2267	83.2	95.6	78.5	100%
MiniCPM-V2.6	*Token Reduction* (↓ **66.7%**)							
+ FastV (ECCV24)	43.2	74.9	73.1	1895	75.4	89.8	67.1	89.0%
+ DART (Ours)	47.8	76.5	74.8	1951	77.4	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 5: 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 6: **Comparing MLLMs on Video Understanding tasks** with 50% visual tokens retained.

highlights a modality imbalance in MLLMs, indicating underutilization of the visual modality. Furthermore, we delved into the impact of the number of pivot tokens on performance. As depicted in Figure 5, 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.

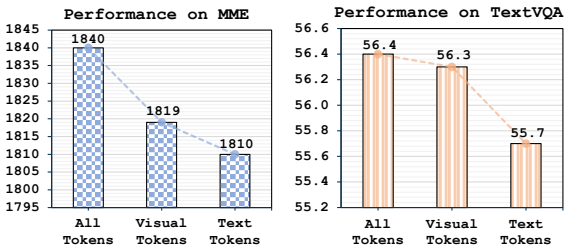


(a) POPE (b) MME

Figure 6: Influence from the layer for token pruning.

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 focused on token "importance", often measured by attention scores, but sometimes performs worse than random pruning. This study introduces DART, which targets token duplication, removing tokens similar to others and achieving better balance between performance and latency across multiple benchmarks and MLLMs (Tab. 1, 2, 3, 4, 5, 7, 9 and Fig. 4). Our exploration yields surprising insights: distinct retained token sets, with under 50% overlap, deliver similarly strong performance (§5.2). Moreover, token pruning may reduce hallucinations (§5.3). These findings expose limits of importance-based methods and offer insights into vision tokens in MLLMs.

7 Limitations

Similar to many other methods aimed at improving efficiency, such as network pruning, quantization, distillation, model merging, and speculative decoding, 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. Moreover, due to space limitations in the main text, we had to move some experimental results that we believe are particularly insightful and interesting to the appendix. These include, for example, our investigation of strategies for pivot token selection, a more detailed analysis of the impact of the number of pivot tokens, and validations of our method on larger-scale models, which may slightly affect the overall reading experience.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Ruichuan An, Sihan Yang, Ming Lu, Kai Zeng, Yulin Luo, Ying Chen, Jiajun Cao, Hao Liang, Qi She, Shanghang Zhang, et al. 2024. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*.

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*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.

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*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *Preprint*, arXiv:2307.15818.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15710–15719.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

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*.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. 2024. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE.

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.

Yuhang Han, Xuyang Liu, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. 2024. Rethinking token reduction in mllms: Towards a unified paradigm for training-free acceleration. *arXiv preprint arXiv:2411.17686*.

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.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. 2024a. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *Preprint*, arXiv:2411.19650.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. 2024b. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. 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. 2024d. 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.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. 2024. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer.

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.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.

Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan. 2024. Long-form speech generation with spoken language models. *arXiv preprint arXiv:2412.18603*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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*.

Kimi Team. 2024. Kimi-audio technical report. *Preprint*, arXiv:arXiv:placeholder.

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*.

Zhehao Wang, Xian Lin, Nannan Wu, Li Yu, Kwang-Ting Cheng, and Zengqiang Yan. 2024d. Dtmformer: Dynamic token merging for boosting transformer-based medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5814–5822.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*.

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.

Siyu Xu, Yunke Wang, Chenghao Xia, Dihao Zhu, Tao Huang, and Chang Xu. 2025. Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation. *arXiv preprint arXiv:2502.02175*.

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*.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22128–22136.

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*.

Appendix

A Additional Experiments **13**
- A.1 Supplementary Results on Pivot Token Selection 13
- A.2 Influence from the Number of Pivot Tokens 13
- A.3 More Experimental Results on Larger MLLMs 13

B Extensions to Other Scenarios **14**
- B.1 Exploring the Effectiveness of DART in Audio Modalities 14
- B.2 Enhancing VLA Efficiency with DART 15

C Detailed Experiment Settings **16**
- C.1 Datasets 16
  - C.1.1 Image Understanding . . . 16
  - C.1.2 Video Understanding . . . 17
  - C.1.3 Automatic Speech Recognition. 17
  - C.1.4 Vision-Language-Action Models Simulation Platform 17
- C.2 Models 17
- C.3 Baselines 17
- C.4 Implementation Details 18

D Computational Complexity. **18**

E Future Works **18**

F Sparsification Visualization on Different Pivot Token Selection Strategy **19**

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 8 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.

A.2 Influence from the Number of Pivot Tokens

Beyond the investigation of pivot token numbers on MME and TextVQA in §5.3, we conduct additional experiments on several representative visual benchmarks to further support our insight. Figure 8 illustrates that our observations on benchmarks such as POPE and SQA align with those in §5.3—namely, that both insufficient and excessive pivot tokens can lead to suboptimal performance. While an insufficient or excessive number of pivot tokens may result in suboptimal outcomes, our statistical analysis reveals that **even the worst-performing settings still match or surpass the performance of existing token pruning approaches.** This further demonstrates the superiority of DART.

A.3 More Experimental Results on Larger MLLMs

Method	MME	POPE	GQA	TextVQA	SQA	Avg.
Qwen2-VL-72B	*Upper Bound, Full Tokens* **(100%)**					
Vanilla	2521	87.4	65.3	82.8	91.6	100%
Qwen2-VL-72B	*Token Reduction* (↓ **66.7%**)					
FastV (ECCV24)	2376	83.8	62.5	81.5	87.6	96.0%
DART (Ours)	2511	85.7	64.2	82.1	90.9	98.9%
Qwen2-VL-72B	*Token Reduction* (↓ **77.8%**)					
FastV (ECCV24)	2219	81.1	59.2	79.6	85.1	92.1%
DART (Ours)	2496	83.8	62.5	80.4	88.1	96.8%
Qwen2-VL-72B	*Token Reduction* (↓ **88.9%**)					
FastV (ECCV24)	2089	78.7	55.7	75.4	83.3	88.0%
DART (Ours)	2350	79.3	59.2	76.6	86.0	92.2%

Table 7: Comparative experiments on Qwen2-VL-72B.

While prior experiments primarily focused on models with 7B parameters, we further validate the effectiveness and robustness of DART on substantially larger models, including LLaVA-v1.5-13B[2] and Qwen2-VL-72B[3]. Our results demonstrate that DART consistently outperforms prior token pruning methods such as FastV (Chen et al., 2024) and SparseVLM (Zhang et al., 2024b) across various pruning ratios and downstream tasks, while maintaining near-Vanilla performance.

[2] https://huggingface.co/liuhaotian/llava-v1.5-13b

[3] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct

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	$59.0_{\pm 0.3}$	59.2	58.4	58.7	59.1	57.3	59.4	56.0	49.6
MMB	64.7	$63.2_{\pm 0.7}$	63.1	62.9	63.2	64.0	62.5	64.3	60.0	56.1
MME	1862	$1772_{\pm 17.9}$	1826	1830	1840	1820	1760	1825	1745	1490
POPE	85.9	$80.6_{\pm 0.49}$	81.1	81.0	80.1	80.2	76.8	81.6	80.5	59.6
SQA	69.5	$69.0_{\pm 0.3}$	69.9	68.9	69.1	68.7	69.2	68.9	68.5	60.2
VQA^{V2}	78.5	$75.2_{\pm 0.2}$	75.9	76.0	75.9	75.6	75.4	76.1	73.8	61.8
VQA^{Text}	58.2	$56.0_{\pm 0.3}$	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 8: **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. For the **Random** pivot token selection strategy, we conducted experiments five times using five different random seeds, and report the corresponding standard deviation to reflect variability.

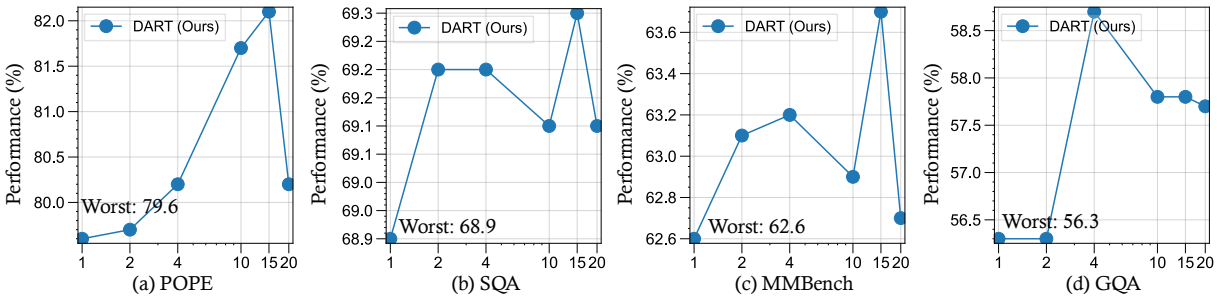


Figure 8: Impact of the number of pivot tokens on performance across additional visual benchmarks. All experiments are conducted with a token reduction ratio of 77.8%. It is noteworthy that even under relatively extreme numbers of pivot tokens, our worst-case performance still matches or surpasses that of existing token pruning methods.

As shown in Table 9, on LLaVA-1.5-13B with an 88.9% pruning ratio, DART achieves 94.7% average performance, significantly outperforming SparseVLM (79.7%) and FastV (81.0%). Similarly, on Qwen2-VL-72B, DART reaches 92.2% under the same pruning ratio, surpassing FastV (88.0%) (Table 7). At a moderate 66.7% pruning ratio, DART retains 99.5% and 98.9% accuracy on LLaVA-1.5-13B and Qwen2-VL-72B, respectively, with minimal degradation.

DART also excels on specific tasks, achieving 60.9 GQA on LLaVA-1.5-13B at 77.8% pruning and 90.9 ScienceQA on Qwen2-VL-72B at 66.7%, both outperforming FastV. These results demonstrate DART 's scalability and its ability to balance compression and performance in large MLLMs.

B Extensions to Other Scenarios

B.1 Exploring the Effectiveness of DART in Audio Modalities

In recent years, the integration of audio as a core modality (Abouelenin et al., 2025; Team, 2024; Chu et al., 2024) within Multimodal Large Language Models (MLLMs) has garnered increasing attention. As these models evolve to handle complex, real-world tasks that span language, vision, and sound, the ability to effectively process spoken language becomes crucial. Audio understanding, particularly in the form of automatic speech recognition (ASR), plays a foundational role in applications such as virtual assistants, transcription services, voice-controlled systems, and multimodal reasoning agents. Therefore, beyond the widely explored domains of image and video understanding in the visual modality, we further extend our investigation to evaluate the effectiveness of our proposed method on tasks within the audio modality. To conduct our study, we select Phi-4-Multimodal-Instruct[4], an MLLM with strong audio modality capabilities, and evaluate it on two representative speech benchmarks: FLEURs-en (Conneau et al., 2023) and LibriSpeech-long (Park et al., 2024). As demonstrated in Table 10, our proposed method DART consistently outperforms baseline

[4] https://huggingface.co/microsoft/Phi-4-multimodal-instruct

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	VizWiz	Avg.
LLaVA-1.5-13B	*Upper Bound, 576 Tokens* **(100%)**								
Vanilla	63.3	68.9	62.3	1818	85.9	72.8	61.3	56.6	100%
LLaVA-1.5-13B	*Retain 192 Tokens* (↓ **66.7%**)								
FastV (ECCV24)	59.1	54.0	51.2	1641	82.3	56.4	51.6	56.9	87.8%
SparseVLM (ICML25)	58.7	67.4	61.0	1768	82.2	73.1	45.4	56.5	94.5%
DART (Ours)	**62.1**	**68.2**	**61.4**	**1855**	**84.0**	**73.6**	**60.2**	**57.3**	**99.5%**
LLaVA-1.5-13B	*Retain 128 Tokens* (↓ **77.8%**)								
FastV (ECCV24)	57.7	57.9	48.8	1673	79.3	57.0	56.0	55.3	88.2%
SparseVLM (ICML25)	57.9	65.8	55.8	1774	81.1	69.9	49.9	56.3	93.2%
DART (Ours)	**60.9**	**67.4**	**60.7**	**1839**	**81.8**	**74.3**	**59.0**	**57.3**	**98.5%**
LLaVA-1.5-13B	*Retain 64 Tokens* (↓ **88.9%**)								
FastV (ECCV24)	53.7	50.9	42.1	1567	69.3	56.8	47.1	56.7	81.0%
SparseVLM (ICML25)	50.6	61.3	54.8	1402	65.0	69.0	22.7	54.5	79.7%
DART (Ours)	**57.1**	**65.4**	**59.3**	**1722**	**75.4**	**74.1**	**55.9**	**57.4**	**94.7%**

Table 9: Comparative experiments on LLaVA-1.5-13B. 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.

Method	FLEURs ↓	LibriSpeech ↓	Avg. ↓
Phi-4-Multimodal-Instruct	*Upper Bound, Full Audio Tokens* **(100%)**		
Vanilla	3.49	6.40	4.95
Phi-4-Multimodal-Instruct	*Token Reduction* (↓ **20%**)		
+ Random	8.15	25.23	16.69
+ FastV (ECCV24)	19.82	27.90	23.86
+ DART (Ours)	**5.05**	**6.95**	**6.00**
Phi-4-Multimodal-Instruct	*Token Reduction* (↓ **30%**)		
+ Random	13.18	39.42	26.3
+ FastV (ECCV24)	34.10	51.60	42.85
+ DART (Ours)	**5.84**	**11.64**	**8.74**
Phi-4-Multimodal-Instruct	*Token Reduction* (↓ **50%**)		
+ Random	37.57	76.85	57.21
+ FastV (ECCV24)	180.0	88.38	134.19
+ DART (Ours)	**18.93**	**49.13**	**34.03**

Table 10: Comparative experiments on Automatic Speech Recognition tasks. In all experiments for DART, tokens are pruned after the 2nd layer with 8 pivot tokens. The pivot tokens are selected based on the maximum K-norm. The evaluation metric is Word Error Rate (WER).

approaches under varying token reduction ratios on both FLEURs-en and LibriSpeech-long benchmarks. While random pruning and FastV result in substantial degradation in recognition performance, particularly under higher reduction rates, DART maintains significantly lower Word Error Rates (WER), showcasing its robustness and effectiveness in preserving critical audio information even with limited token usage.

B.2 Enhancing VLA Efficiency with DART

Building on recent progress in multimodal understanding from vision-language models (Awadalla et al., 2023; Li et al., 2022; Radford et al., 2021; An et al., 2024; Luo et al., 2024), Vision-Language-Action (VLA) models represent a significant step toward embodied intelligence. Systems such as OpenVLA (Kim et al., 2024), CogACT (Li et al., 2024a), pi_0(Black et al., 2024), and RT-2(Brohan et al., 2023) seamlessly translate multimodal inputs into executable actions. Leveraging large-scale datasets (Fang et al., 2024; O'Neill et al., 2024), these models have demonstrated impressive capabilities in complex robotic manipulation and reasoning tasks. As a potential pathway toward Artificial General Intelligence (AGI), we place great emphasis on improving the efficiency of VLA models through our approach.

To this end, we employ the SIMPLER environment (Li et al., 2024b), a simulation-based benchmark specifically designed for table-top manipulation to evaluate our method. SIMPLER aims to closely mirror real-world dynamics observed in robots such as the Google Robot and WidowX, exhibiting strong consistency between simulated and real-world performance. In this setup, the Vision-Language-Action (VLA) model receives 224×224 RGB image observations along with natural language task instructions (*e.g.*, "Pick coke can") and generates a sequence of actions in 7-DoF Cartesian space. SIMPLER supports two evaluation configurations: **Visual Matching**, which emphasizes visual fidelity to real-world scenes, and **Variant Aggregations**, which introduces variability through changes in lighting, background, and surface textures. For the Google Robot, both configurations include the same set of four tasks: Pick coke can; Move near; Open/close drawer and Open top drawer and place apple. Performance is assessed using success rate as the evaluation metric.

As shown in Table 11, DART demonstrates superior performance compared to other baseline methods in the SIMPLER environment. With only 56 retained visual tokens, DART achieves the highest average success rates of 75.2% and 64.4% in Visual Matching and Variant Aggregation, respectively, outperforming Random Dropping (Wen et al., 2025), FastV (Chen et al., 2024), VLA-Cache (Xu et al., 2025), and even vanilla

SIMPLER	Method	Retained Tokens	PickCan	MoveNear	Drawer	DrawerApple	Average	FLOPs ↓	Speedup ↑
Visual Matching	CogACT	256	91.3%	85.0%	71.8%	50.9%	74.8%	100.0%	1.00×
	Random Dropping	112	9.7%	20.4%	53.5%	0.0%	20.9%	58.5%	1.20×
	FastV	56	92.6%	81.4%	69.8%	52.4%	74.1%	42.0%	1.21×
	VLA-Cache	-	92.0%	83.3%	70.5%	51.6%	74.4%	80.1%	1.38×
	DART	56	**95.6%**	**85.8%**	69.9%	49.5%	**75.2%**	44.7%	1.25×
Variant Aggregation	CogACT	256	89.6%	80.8%	28.3%	46.6%	61.3%	100.0%	1.00×
	Random Dropping	112	4.0%	16.1%	15.6%	0.0%	8.9%	58.5%	1.20×
	FastV	56	91.4%	78.6%	27.6%	50.6%	62.1%	42.0%	1.19×
	VLA-Cache	-	91.7%	79.3%	32.5%	45.8%	62.3%	82.6%	1.37×
	DART	56	**92.4%**	77.0%	**35.9%**	**52.4%**	**64.4%**	44.7%	1.25×

Table 11: Performance of DART on the CogACT versus the other baselines in the SIMPLER environment. Random Dropping denotes a method involving the random retention of visual tokens.

CogACT (Li et al., 2024a). Moreover, DART significantly reduces computational cost, achieving the lower FLOPs (44.7%), which corresponds to a speedup of 1.25× compared to the CogACT. These results highlight DART 's efficiency in maintaining high task performance while substantially reducing computational demands.

C Detailed Experiment Settings

C.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). Furthermore, to validate the effectiveness and applicability of our approach, we extended the evaluation scenarios of DART. Specifically, we tested our token reduction method in both the speech modality—on automatic speech recognition (audio token reduction) (Conneau et al., 2023; Park et al., 2024), and on a vision-language-action model within a simulated environment (Li et al., 2024b).

C.1.1 Image Understanding

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. 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 interoperability.

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.

C.1.2 Video Understanding

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, 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.

C.1.3 Automatic Speech Recognition.

FLEURS. FLEURS is a benchmark for evaluating universal speech representations across 102 languages, built on top of the FLoRes-101 dataset. It contains 12 hours of speech data per language, with parallel speech and text for tasks like ASR, Speech LangID, and cross-modal retrieval.

LibriSpeech-Long. LibriSpeech-Long is a benchmark dataset for long-form speech generation, derived from the original LibriSpeech dataset. It provides 4-minute long continuous speech and corresponding transcripts, enabling the evaluation of long-form speech continuation. This benchmark supports reference-based evaluation for long-form speech tasks and facilitates research in generating coherent and contextually relevant speech over extended durations.

C.1.4 Vision-Language-Action Models Simulation Platform

SIMPLER. SIMPLER is a simulation platform for evaluating real-world robot manipulation policies. It features realistic simulated environments that match common real robot setups (*e.g.*, Google Robot and WidowX) and tasks (*e.g.*, picking and moving objects). By addressing control and visual disparities between simulation and reality, SIMPLER achieves strong correlation with real-world performance, providing a scalable and reproducible evaluation tool.

C.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[5] (Liu et al., 2024d) and LLaVA-Next-7B[6] (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[7] (Wang et al., 2024a) and MiniCPM-V-2.6[8] (Yao et al., 2024b). Moreover, to enhance the effectiveness of our proposed method, we also validate DART on larger MLLMs, such as Qwen2-VL-72B and LLaVA-1.5-13B. 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.

C.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.

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

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

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

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

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.
FiCoCo (Han et al., 2024) introduces a unified "filter-correlate-compress" paradigm to streamline training-free token reduction in Multimodal Large Language Models (MLLMs).
FitPrune (Ye et al., 2025) proposes a method that generates an efficient token pruning strategy for multi-modal large language models by removing redundant visual tokens. FitPrune is easy to deploy and is designed to meet a predefined computational budget while maintaining model performance.

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

C.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.

D 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:

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

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:

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

E Future Works

As can be observed from Figure 1 and Figure 6(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.

F Sparsification Visualization on Different Pivot Token Selection Strategy

Figure 9 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 9: **Sparsification Visualization examples of DART on different Pivot Token Selection Strategy.**