

Fit and Prune: Fast and Training-free Visual Token Pruning for Multi-modal Large Language Models

Weihao Ye^{1,2}, Qiong Wu^{1,2}, Wenhao Lin^{1,2}, Yiyi Zhou^{1,2*}

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China.

²Institute of Artificial Intelligence, Xiamen University, 361005, P.R. China.
{weihaoye, qiong, wenhaolin}@stu.xmu.edu.cn, zhoyiyi@xmu.edu.cn

Abstract

Recent progress in *Multimodal Large Language Models* (MLLMs) often use large image tokens to compensate the visual shortcoming of MLLMs, which not only exhibits obvious redundancy but also greatly exacerbates the already high computation. Token pruning is an effective solution for speeding up MLLMs, but when and how to drop tokens still remains a challenge. In this paper, we propose a novel and training-free approach for the effective visual token pruning of MLLMs, termed *FitPrune*, which can quickly produce a complete pruning recipe for MLLMs according to a pre-defined budget. Specifically, *FitPrune* considers token pruning as a statistical problem of MLLM and its objective is to find out an optimal pruning scheme that can minimize the divergence of the attention distributions before and after pruning. In practice, *FitPrune* can be quickly accomplished based on the attention statistics from a small batch of inference data, avoiding the expensive trials of MLLMs. According to the pruning recipe, an MLLM can directly remove the redundant visual tokens of different examples during inference. To validate *FitPrune*, we apply it to a set of recent MLLMs, including LLaVA-1.5, LLaVA-HR and LLaVA-NEXT, and conduct extensive experiments on a set of benchmarks. The experimental results show that our *FitPrune* can not only reduce the computational complexity to a large extent, while retaining high performance, *e.g.*, -54.9% FLOPs for LLaVA-NEXT with only 0.5% accuracy drop. Notably, the pruning recipe can be obtained in about 5 minutes.

Code — <https://github.com/ywh187/FitPrune>

Extended version — <https://arxiv.org/pdf/2409.10197>

Introduction

Recently, the great success of *large language models* (LLMs) (Zhang et al. 2022; OpenAI 2023; Touvron et al. 2023) also arouses an influx of interest in extending them to more modalities, *e.g.*, *vision and language* (VL) (Liu et al. 2023c; Zhu et al. 2023; Wang et al. 2023; Lin et al. 2023; Xu et al. 2023). For VL tasks, a simple yet intuitive solution for building *multimodal LLMs* (MLLMs) is to directly project the extracted image features onto the seman-

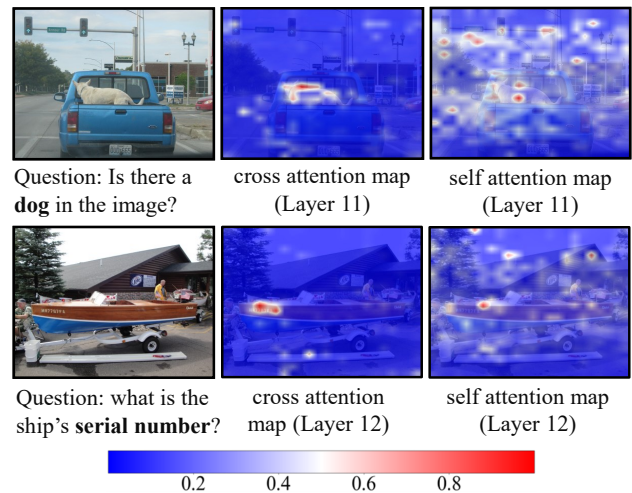


Figure 1: Visualization of the cross and self attention of image tokens of LLaVA-1.5 7B (Liu et al. 2023c). These tokens become less active at the higher layer of an MLLM.

tic space of LLMs as visual tokens, thus performing multimodal sequence modeling in one Transformer architecture (Liu et al. 2023c; Luo et al. 2024). Despite effectiveness, this paradigm often suffers from excessive computation. For instance, LLaVA (Liu et al. 2023c) uses 576 image patches as the visual tokens, which requires 6.2 times the computation compared to its text-only inference on ScienceQA (Lu et al. 2022). More recently, practitioners (Li et al. 2023c; Luo et al. 2024) resort to increasing image resolution to alleviate the widely criticized visual shortcoming of MLLMs (Tong et al. 2024), which also further exacerbates the already high computation of MLLMs.

In addition, the use of large visual tokens also involves obvious redundancy in MLLMs. Compared with previous Transformer-based VL models (Zhou et al. 2020, 2021), the multi-head attention of MLLMs is unidirectional rather than truly “global”. Simply put, MLLMs only propagate information from the previous tokens to the subsequent ones, and their visual tokens are usually placed in front of the text questions. In this case, they mainly serve to give visual semantics to the text tokens, but in fact not many of them are

*Corresponding Author.

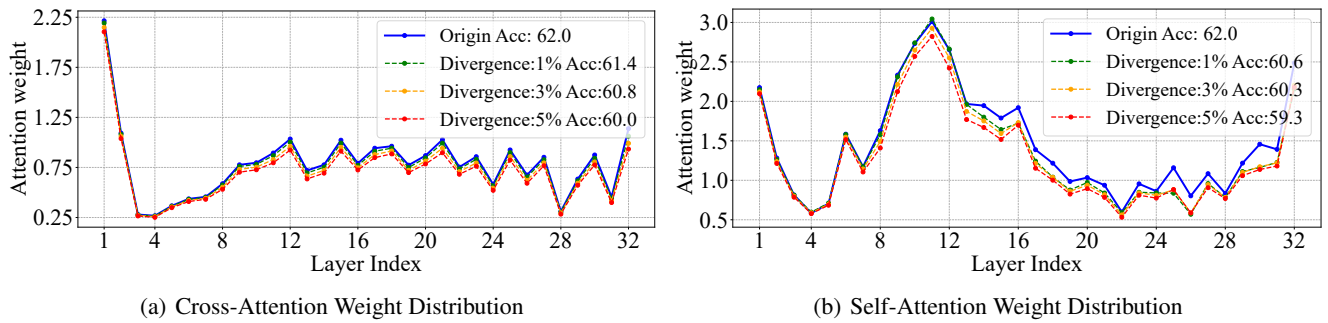


Figure 2: The impact of token pruning according to the fitting of cross- and self-attention distributions of visual tokens of LLaVA 1.5. The GQA accuracy (Hudson and Manning 2019) is reported. For pruning recipes, the better fitting of attention distribution can retain better performance. However, only considering a single distribution is hard to obtain the optimal pruning recipe. In this paper, our FitPrune will consider the fitting of both cross- and self-attentions.

active. As shown in Fig. 1, the attention from image to text at the 12-th layer (32 in total) of LLaVA becomes very focused, indicating that only a small number of visual tokens engage in multimodal reasoning. Theoretically, pruning the less active tokens would have limited impact on model performance. This assumption is also validated in our experiments, *i.e.* removing a set of visual tokens of MLLMs barely decline the accuracy on most benchmarks.

However, when and how to prune the visual tokens of MLLMs remains an open problem, especially according to a predefined computation budget. Although token pruning has been extensively studied in *natural language processing* (NLP) (Ye et al. 2021; Liu et al. 2023a; Anagnostidis et al. 2024) and *computer vision* (Kong et al. 2022; Dong et al. 2023; Wei et al. 2023), most pruning methods still need to be manually validated for different MLLMs, of which expenditure is also much more expensive. For instance, the researchers often need to define the optimal pruning ratios of an MLLM and its layers via numerous trials. The special properties of MLLMs, such as unidirectional self-attention and the involved visual modeling, also make some previous metrics (Fayyaz et al. 2022; Wang, Dedhia, and Jha 2023) not applicable or suboptimal for VL tasks. For instance, only considering the cross-attention between visual and text tokens will drop the tokens that are also important in visual modeling, and *vice versa*. To this end, we question that

“Can we find a solution that can directly determine the optimal pruning recipe for MLLMs?”

To approach this target, we propose an effective and training-free pruning method for MLLMs, termed *Fit and Prune* (FitPrune). In particular, we define token pruning as a task of distribution fitting, and adopt statistical principle to obtain the optimal pruning recipe. Concretely, it aims to minimize the divergence of attention distributions before and after pruning, thereby reducing the negative impact on performance. In practice, the attention statistics can be well represented by a small batch of data, as shown in Fig. 2. Afterwards, FitPrune searches the token slots whose removal would have minimal impact on the default distribution, based on which the pruning ratio of each layer can be directly set. To better meet the property of MLLMs, Fit-

Prune also considers the distributions of both cross- and intra-attentions of visual tokens, *i.e.*, the two distributions in Fig.2. With these innovative designs, FitPrune can directly produce a pruning recipe according to a predefined computation budget.

To validate FitPrune, we apply it to a set of advanced MLLMs, including LLaVA-1.5 (Liu et al. 2023b), LLaVA-HR (Luo et al. 2024) and LLaVA-NEXT (Liu et al. 2024), on a bunch of highly competitive VL benchmarks (Lu et al. 2022; Singh et al. 2019; Hudson and Manning 2019; Goyal et al. 2017; Gurari et al. 2018; Fu et al. 2023; Liu et al. 2023d; Yu et al. 2023; Li et al. 2023b). Experiment results show that FitPrune can significantly reduce the computational complexity of MLLMs while retaining high performance on all benchmarks. For instance, FitPrune can reduce 54.9% TFLOPs of LLaVA-NEXT with only 0.5% performance degradation. More importantly, its pruning recipe can be obtained in about 5 minutes for all VL tasks.

Conclusively, the contribution of this paper is three-fold:

- We reveal the heavy redundancy of visual tokens of MLLMs via investigating their attention patterns, indicating that a great number of visual tokens can be actually discarded during inference.
- To achieve fast and effective token pruning of MLLMs, we propose a novel and training-free method called *fit-and-prune* (FitPrune), which consider token pruning as a task of cross-distribution fitting.
- On ten VL benchmarks, FitPrune not only significantly reduces the computation overhead, *e.g.*, -54.1% TFLOPs of LLaVA-NEXT with marginal performance decline, but also achieves a better trade-off than existing visual token pruning methods.

Related Work

Vision-Language Models

Recently, the prevalence of *large language models* (LLMs) leads to a trend of extending these LLMs to VL tasks (Touvron et al. 2023; Le Scao et al. 2023), termed *multimodal LLMs* (MLLMs) or *vision-LLMs* (VLLMs). Compared with

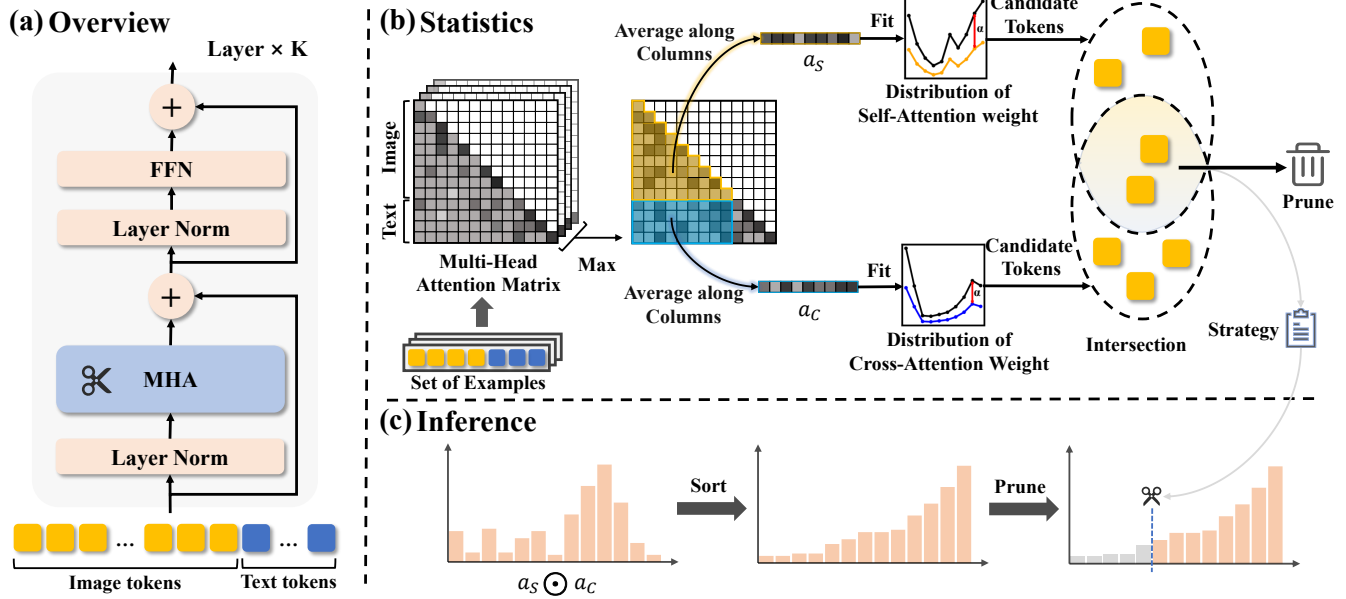


Figure 3: Illustration of our FitPrune. (a) FitPrune is used to reduce the length of visual tokens in the MHA of each layer. (b) The generated pruning recipe is obtained via binary search based on the attention statistics of a set of examples. Its principle is to find out the optimal pruning recipe that reduce the gap of distributions before and after pruning. (c) During inference, the MLLM can drop tokens according to the pruning recipe of FitPrune.

previous single-task or BERT-style VLMs (Radford et al. 2021; Bao et al. 2022), MLLMs (Liu et al. 2023c, 2024) differ in the direct use of language models for both language modeling and multimodal interactions, rather than building another deep multimodal branch. A main design paradigm of these MLLMs is to use a cross-modal projector to connect the visual encoder (Bai et al. 2023; Liu et al. 2023c) with LLMs (Touvron et al. 2023; Le Scao et al. 2023), which can be a deep Query-based branch (Bai et al. 2023; Li et al. 2023a) or a simple projection network (Luo et al. 2024; Liu et al. 2024). Despite effectiveness, existing MLLMs still suffer from visual shortcoming (Tong et al. 2024) or visual hallucination (Huang et al. 2024) mainly due to the inferior descriptive power of input visual tokens (Liu et al. 2023e). To this end, recent efforts (Li et al. 2023c; Luo et al. 2024) have been devoted to increasing the image resolution for MLLMs, which exhibit obvious improvements in fine-grained VL tasks, such as TextVQA (Singh et al. 2019). However, the increase of image resolution also greatly exacerbate the already high computation of MLLMs. Meanwhile, the large visual tokens also involve obvious redundancy as discussed in this paper. In this case, how to effectively and adaptively prune less important tokens of MLLM is a critical demand for its applications.

Token Pruning

Token pruning, as a subclass of *dynamic neural networks* (Han et al. 2021), has become a hot research topic with the prevalence of Transformer-based networks (Vaswani et al. 2017; Kenton and Toutanova 2019; Dosovitskiy et al. 2020), aiming to speed up inference by dynamically reducing less

important tokens based on input-dependent importance. This task has been actively investigated in either *natural language processing* (NLP) (Ye et al. 2021; Kim et al. 2022) or *computer vision* (CV) (Rao et al. 2021; Meng et al. 2022). Existing methods (Rao et al. 2021; Meng et al. 2022; Kong et al. 2022; Wei et al. 2023; Dong et al. 2023) mainly focus on the design of pruning metrics, *e.g.*, low-information, and often require post-pruning tuning. It is prohibitively expensive for LLMs and MLLMs. For example, DiffRate (Chen et al. 2023) relies on prediction loss to guide pruning and requires a large dataset for optimization, while FitPrune uses only a small batch of data, leveraging attention distribution fitting before and after pruning. Some recent approaches explore training-free token pruning (Fayyaz et al. 2022; Bolya et al. 2022; Wang, Dedhia, and Jha 2023; Chen et al. 2024). For instance, ATS (Fayyaz et al. 2022) adopts token sampling based on importance scores, ToMe (Bolya et al. 2022) gradually merges tokens by similarity, and Zero-TPrune (Wang, Dedhia, and Jha 2023) employs a weighted *PageRank* algorithm to identify less critical tokens. In the visual-language field, token pruning is nascent. MADTP (Cao et al. 2024) employs a novel multimodal alignment guidance module coupled with *dynamic token pruning* to efficiently compress the model. FastV (Chen et al. 2024) introduces a pruning strategy that determines token importance based on average attention received by each token. Despite effectiveness, existing methods still need to manually determine the pruning ratio of Transformer and its layers, which makes trials expensive for MLLMs. In this paper, we focus on a training-free and cost-effective method to quickly determine the pruning recipe based on the predefined target.

Method

In this paper, we propose a novel and training-free method towards fast and effective visual token pruning of *multi-modal large language models* (MLLMs), termed *fit-and-prune* (FitPrune). The main principle of FitPrune is to remove the redundant visual tokens based on the fit of attention distributions before and after pruning. In practice, FitPrune performs statistical analysis to derive pruning strategies, as illustrated in Fig. 3.

Concretely, given a model $G(\cdot)$ and its averaged attention distribution D on a set of examples, the objective of FitPrune is to find a strategy P that minimizes the divergence between D and the pruned attention distribution D' given a computation budget δ :

$$\arg \min_P d(D, D') \quad \text{s.t.} \quad \Phi(G, P) \leq \delta, \quad (1)$$

where $\Phi(G, P)$ denotes the computation overhead of $G(\cdot)$ under the pruning strategy P , and δ is the given computation budget. In this paper, we primarily focus on FLOPs.

To meet the property of MLLMs, FitPrune considers the self- and cross-attentions at the same time, denoted as $D_S = [\bar{a}_s^i]_{i=1}^K$ and $D_C = [\bar{a}_c^i]_{i=1}^K$, respectively. Here, K is the number of layers, and \bar{a}_s^i denotes the averaged self-attention value of visual tokens at the i -th layer:

$$\bar{a}_s^i = \frac{1}{N} \sum_{m=1}^N \sum_{n=1}^N A_{m,n}^i, \quad (2)$$

where $A^i \in \mathbb{R}^{(N+M) \times (N+M)}$ represents the attention matrix of the i -th layer, and N and M denote the numbers of visual and text tokens, respectively. In practice, A is the averaged statistics of a batch of examples, akin with D .

Similarly, \bar{a}_c^i is the averaged cross-attention value of visual tokens, obtained by

$$\bar{a}_c^i = \frac{1}{M} \sum_{m=N+1}^{N+M} \sum_{n=1}^N A_{m,n}^i. \quad (3)$$

Afterwards, we denote these two distributions after pruning as $D'_S = [\bar{a}'_s^i]_{i=1}^K$ and $D'_C = [\bar{a}'_c^i]_{i=1}^K$, respectively. Here, \bar{a}'_s^i and \bar{a}'_c^i are obtained by:

$$\bar{a}'_s^i = \frac{1}{N - t_i} \sum_{m=1}^{N-t_i} \sum_{n=1}^{N-t_i} A_{m,n}^i, \quad (4)$$

$$\bar{a}'_c^i = \frac{1}{M} \sum_{m=N-t_i+1}^{N+M-t_i} \sum_{n=1}^{N-t_i} A_{m,n}^i. \quad (5)$$

where $A^i \in \mathbb{R}^{(N+M-t_i) \times (N+M-t_i)}$ is the attention matrix after pruning. Since the number of removed tokens increases layer by layer, t_i is a cumulative value, defined by $t_i = \sum_{i=1}^{i-1} t_i + t_i^*$, and t_i^* is the number of new tokens removed at the i -th layer. In this case, the final pruning strategy P can be denoted by

$$P = [t_1^*, t_2^*, \dots, t_K^*]. \quad (6)$$

Algorithm 1: FitPrune

Require: model G , target FLOPs budget δ and binary search threshold ϵ

Ensure: Pruning strategy P

```

1: Initialize  $D_S, D_C$ 
2: Initialize  $\alpha_L = 0, \alpha_R = 1, P$ 
3: while  $\alpha_R - \alpha_L > \epsilon$  do
4:    $\alpha = (\alpha_L + \alpha_R)/2$ 
5:   for each layer  $i$  in model  $G$  do
6:     Initialize the candidate sets  $\mathcal{T}_S = \emptyset, \mathcal{T}_C = \emptyset$ 
7:     Compute  $a_s^i$  and  $a_c^i$  according to Eq.8
8:      $I_S = \text{argsort}(a_s^i), I_C = \text{argsort}(a_c^i)$ 
9:      $\mathcal{S} = \{(I_S, \bar{a}_s^i, a_s^i, \mathcal{T}_S), (I_C, \bar{a}_c^i, a_c^i, \mathcal{T}_C)\}$ 
10:    for each  $(I, \bar{a}^i, a^i, \mathcal{T})$  in  $\mathcal{S}$  do
11:      for each  $k$  in  $I$  do
12:         $\mathcal{T} = \mathcal{T} \cup \{k\}$ 
13:        if  $\left| \frac{\bar{a}^i - \sum_{k \in \mathcal{T}} a^{i,k}}{\bar{a}^i} \right| > \alpha$  then
14:           $\mathcal{T} = \mathcal{T} \setminus \{k\}$ 
15:        break
16:      end if
17:    end for
18:    end for
19:     $t_i^* = |\mathcal{T}_S \cap \mathcal{T}_C|$ 
20:  end for
21:  if  $\Phi(G(P, x)) \leq \delta$  then
22:     $\alpha_R = \alpha$ 
23:  else
24:     $\alpha_L = \alpha$ 
25:  end if
26: end while
27: return strategy  $P = [t_1^*, t_2^*, \dots, t_K^*]$ 

```

Lastly, the optimization of FitPrune in Eq.1 can be reformulated as

$$\arg \min_P [d(D_S, D'_S) + d(D_C, D'_C)], \quad (7)$$

s.t. $\Phi(G, P) \leq \delta$.

To accomplish the objective of Eq.7, we adopt the principle of *binary search* to obtain the optimal pruning recipe. The search algorithm is depicted in Algorithm 1. Concretely, we first obtain the default the self-attention distribution $D_S = [a_s^i]_{i=1}^K$ and cross-attention distribution $D_C = [a_c^i]_{i=1}^K$ from a set of examples. Then, we employ binary search to determine the smallest divergence upper bound α . The binary search is initialized with $\alpha_L = 0$ as the left boundary (no divergence) and $\alpha_R = 1$ as the right boundary (maximum divergence). In each iteration of the binary search, the midpoint $\alpha = (\alpha_L + \alpha_R)/2$ is evaluated. For each candidate value of α , we greedily remove the token that has the smallest impact on the attention distributions.

Concretely, for the i -th layer, we first evaluate the received attention weights for each visual token. The j -th token's attention weights in two distributions are denoted as $a_s^i \in \mathbb{R}^{N-t_{i-1}}$ and $a_c^i \in \mathbb{R}^{N-t_{i-1}}$, calculated by:

Method	Pruning Ratio	ScienceQA - IMG		GQA		TextVQA		Vizwiz		VQA _{v2}	
		Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*
LLaVA - 1.5 7B	0%	68.0	0%	62.0	0%	58.2	0%	50.0	0%	78.5	0%
+FitPrune	40%	67.7	-30.5%	61.9	-33.7%	58.1	-31.3%	50.2	-33.0%	78.5	-33.9%
+FitPrune	60%	67.8	-46.8%	61.5	-51.6%	58.2	-47.9%	50.4	-50.5%	78.3	-51.9%
LLaVA - HR 7B	0%	68.0	0%	64.2	0%	67.1	0%	48.7	0%	81.9	0%
+FitPrune	40%	68.0	-30.6%	64.2	-32.1%	67.2	-30.7%	48.5	-31.7%	81.9	-32.1%
+FitPrune	60%	67.7	-46.7%	64.1	-49.1%	66.7	-47.1%	48.5	-48.5%	81.8	-49.2%
LLaVA - NEXT 7B	0%	70.2	0%	64.2	0%	64.9	0%	57.7	0%	81.8	0%
+FitPrune	40%	70.1	-34.3%	64.2	-36.6%	64.9	-36.0%	57.4	-36.5%	81.7	-36.6%
+FitPrune	60%	70.1	-51.7%	64.0	-55.0%	64.2	-54.0%	57.3	-54.9%	81.5	-55.0%

Table 1: Results of MLLMs with FitPrune on 5 MLLM benchmarks. Pruning ratio denotes the target reduction of FLOPs by visual tokens. TFLOPs* reflect the computation and actual FLOPs reduction of the entire MLLM. Due to the different lengths of text decoding, the overall reduction varies for different benchmarks.

Method	Pruning Ratio	POPE		MM - Vet		MMB		MMB ^{CN}		MME	
		Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*	Accuracy	TFLOPs*
LLaVA - 1.5 7B	0%	85.9	0%	31.6	0%	64.3	0%	58.3	0%	1510.7	0%
+FitPrune	40%	86.9	-33.9%	31.5	-30.7%	64.9	-31.3%	58.2	-29.8%	1502.5	-33.5%
+FitPrune	60%	86.5	-51.8%	32.8	-47.1%	64.6	-47.9%	58.4	-45.7%	1507.9	-51.4%
LLaVA - HR 7B	0%	87.6	0%	31.2	0%	65.0	0%	60.6	0%	1554.9	0%
+FitPrune	40%	88.4	-32.2%	31.5	-31.1%	65.0	-30.7%	60.1	-30.1%	1543.9	-32.1%
+FitPrune	60%	88.7	-49.2%	31.0	-47.5%	64.9	-47.0%	60.2	-46.0%	1561.6	-49.0%
LLaVA - NEXT 7B	0%	86.5	0%	43.9	0%	67.4	0%	60.6	0%	1519.0	0%
+FitPrune	40%	87.6	-36.5%	44.2	-34.3%	68.0	-35.5%	60.3	-35.0%	1506.0	-36.5%
+FitPrune	60%	87.6	-54.8%	41.7	-52.2%	67.5	-53.4%	60.0	-52.6%	1486.0	-54.8%

Table 2: Performance with different pruning ratios on three MLLMs for 5 multimodal benchmarks for MLLMs. Pruning ratio denotes the target reduction of FLOPs by visual tokens. TFLOPs* reflect the computation and actual reduction of FLOPs of the entire MLLM.

$$a_s^{i,j} = \sum_{m=1}^{N-t_{i-1}} A_{m,j}^i, \quad a_c^{i,j} = \sum_{m=N-t_{i-1}+1}^{N+M-t_{i-1}} A_{m,j}^i. \quad (8)$$

Next, we select the sets of tokens \mathcal{T}_S and \mathcal{T}_C with the smallest attention weights in the self-attention and cross-attention distributions, so that their removal results in a distribution divergence close to α :

$$\mathcal{T}_S = \operatorname{argmax}_{\mathcal{T} \subseteq [N-t_{i-1}]} |\mathcal{T}| \text{ s.t. } \left| \frac{\bar{a}_s^i - \sum_{j \in \mathcal{T}} a_s^{i,j}}{\bar{a}_s^i} \right| \leq \alpha, \quad (9)$$

$$\mathcal{T}_C = \operatorname{argmax}_{\mathcal{T} \subseteq [N-t_{i-1}]} |\mathcal{T}| \text{ s.t. } \left| \frac{\bar{a}_c^i - \sum_{j \in \mathcal{T}} a_c^{i,j}}{\bar{a}_c^i} \right| \leq \alpha, \quad (10)$$

where \mathcal{T} represents a subset of the remaining tokens. The goal is to maximize the number of tokens $|\mathcal{T}|$ that can be removed while ensuring that the divergence between the original and pruned attention distributions does not exceed the divergence threshold α . Then, the number of pruned tokens in i -th layer can be calculated by:

$$t_i^* = |\mathcal{T}_i|, \quad \text{where } \mathcal{T}_i = \mathcal{T}_S \cap \mathcal{T}_C. \quad (11)$$

This iterative process continues by checking whether $\Phi(G, P)$ meets the target computation budget δ after each

iteration, and accordingly adjusting the binary search boundaries α_L and α_R . The search terminates when the difference between α_R and α_L is no more than the threshold ϵ .

During inference, tokens are ranked by the combined importance metric $a_u^{i,j} = a_s^{i,j} \cdot a_c^{i,j}$, where $a_s^{i,j}$ and $a_c^{i,j}$ are the attention weights received by the j -th image token in the i -th layer from other image tokens and the text tokens, respectively. Tokens with the lowest $a_u^{i,j}$ are pruned first. The tokens to be pruned in i -th layer can be defined as:

$$\mathcal{T}_i = \operatorname{argmin}_{\mathcal{T} \subseteq [N-t_{i-1}], |\mathcal{T}|=t_i^*} \sum_{j \in \mathcal{T}} a_u^{i,j} \quad (12)$$

where \mathcal{T} represents a subset of the remaining tokens, and t_i^* represents the number of tokens to be pruned from layer i as specified in the pruning strategy P . This ensures that the least important tokens are removed, preserving model performance while enhancing computational efficiency. FitPrune does not need any additional modules and also requires no gradient computations, thus ensuring its efficient deployment with minimal computational overhead.

Experiments

Datasets and Metrics

To validate FitPrune, we conduct extensive experiments on 10 benchmarks. The datasets include the common vision-language benchmarks such as VQA_{v2} (Goyal et al.

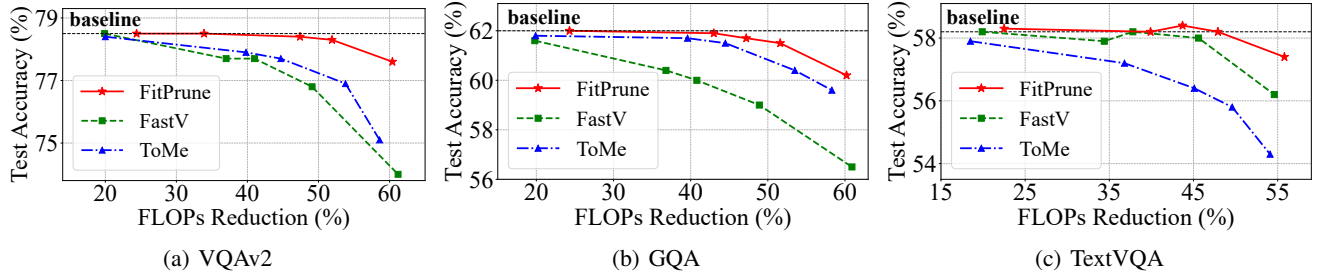


Figure 4: Performance comparison of FitPrune and other pruning methods on the LLaVA-1.5 7B *w.r.t* different pruning ratios.



Figure 5: The visualization of impact of our FitPrune. (a) Distribution of cross-attention weights. (b) Distribution of self-attention weights. These distributions are based on LLaVA-1.5 7B for the GQA benchmark. (c) The number of tokens across the 32 layers under the different settings.

2017) for visual question answering, GQA (Hudson and Manning 2019) for compositional reasoning, VizWiz (Gurari et al. 2018) for zero-shot generalization, ScienceQA-IMG (Lu et al. 2022) for scientific question answering, and TextVQA (Singh et al. 2019) for text understanding in images. The multimodal benchmarks for MLLMs are also used, including POPE (Li et al. 2023b) for object hallucination, MM-Vet (Yu et al. 2023) for integrated vision-language capabilities, MMBench (Liu et al. 2023d) for diverse perception and reasoning tasks, MMBench^{CN} for Chinese language evaluation, and MME (Fu et al. 2023) for systematic multimodal assessment. All experiments on these datasets follow the default setting and metric.

Implementation Details

To conduct the statistical analysis, we use 655 samples (0.1%) from the LLaVA-655k data (Liu et al. 2023b) to generate the pruning strategy. To validate FitPrune, we apply it to LLaVA (Liu et al. 2023c), LLaVA-HR (Luo et al. 2024), and LLaVA-NEXT (Liu et al. 2024). The base language models used in these MLLMs are all Vicuna (Chiang et al. 2023). Under the default setting, LLaVA-1.5 7B has 576 visual tokens, while LLaVA-HR 7B stands out with 1024 visual tokens, resulting in better performance and higher computational complexity. For a higher resolution, LLaVA-NEXT 7B dynamically introduces visual tokens up to 2880, improving the performance by 6.0% with 3.5 \times computation overhead compared to LLaVA. In our experiments, we randomly select 655 samples (0.1%) from LLaVA-655k

dataset (Liu et al. 2023b) to generate the pruning strategy. In our statistical process, the binary search tolerance parameter ϵ is set to 0.01. Our experiments are conducted on a single A100 40G GPU. More details can refer to our projects.

Experimental Results

Effects of FitPrune on different MLLMs. We report the effects of FitPrune on a set of MLLMs in Tab. 1 and 2, including LLaVA-1.5 7B (Liu et al. 2023c), LLaVA-HR 7B (Luo et al. 2024) and LLaVA-NEXT 7B (Liu et al. 2024). From Tab. 1 and Tab. 2, we can first observe that our FitPrune method significantly reduces computational overhead while maintaining competitive performance of these MLLMs. For instance, when decreasing 40% FLOPs from visual tokens in statistic, FitPrune decreases computation overhead by 33.2% on average without dropping performance. When defining a pruning ratio of 60% FLOPs, our FitPrune reduces the overall computation overhead of LLaVA-NEXT by 55%, with only 0.3% drop for VQAv2. Similar results can be also obtained on the MLLM benchmarks, in Tab. 2. With a 60% pruning ratio from visual tokens, FitPrune achieves 52.4% computation reduction, with only 0.1% performance drop on average. Overall, these experiments well validate the effectiveness of FitPrune in token pruning.

Comparison with SOTA pruning methods. We further compare our FitPrune with two representative pruning methods on LLaVA-1.5, *i.e.*, ToMe (Bolya et al. 2022) and FastV (Chen et al. 2024). From Fig. 4, we can observe

Num. of Sample	Search Time	Accuracy (%)
6.6K	199 min	61.6
2.0K	66 min	61.6
0.6K	19 min	61.5
65	2 min	61.6
10	0.5 min	61.2

Table 3: Ablation study on the scale of data for binary search of FitPrune for LLaVA-1.5 on GQA.

that our FitPrune method has obvious advantages over these methods. At lower pruning ratios, all methods maintain competitive performance, highlighting the presence of visual redundancy. For example, with a 20% pruning ratio on the GQA dataset, ToMe experiences only a 0.2% performance drop, and FastV shows a 0.4% reduction, while FitPrune maintains its performance without any degradation. However, as the pruning ratio increases, FitPrune’s advantages become more evident. With a 60% ratio, FitPrune achieves +2.5% and +4.5% performance gains compared to ToMe and FastV on the VQAv2 dataset. This demonstrates FitPrune’s ability to preserve model performance even under more aggressive pruning.

With a larger reduction target, FitPrune can better maintain the performance. For instance, when reducing FLOPs by 60%, FitPrune only drops 0.9% performance on VQAv2, whereas FastV loses 4.5%. In summary, compared with other methods, the proposed FitPrune method shows robustness as the pruning ratio increases, proving its effectiveness in the token pruning task.

Ablation study. Next, we ablate the key settings of FitPrune. We first compare the performance of different pruning ratios, and also visualize the attention distributions after pruning in Fig. 5-(a) and (b). From these figures, we can first see that FitPrune is robust and can well maintain the high performance when increasing the pruning ratio. Besides, we can also see that the fit of distributions also affect the final performance. For instance, when the pruning ratio is 80%, the average divergence of the self-attention and cross-attention distributions compared to the original distributions reaches approximately 27%, leading to a performance drop of 3.6%. Furthermore, the divergence caused by token pruning is relatively uniform across layers. However, compared to cross-attention, self-attention is more unstable and significantly affected by token pruning, indicating that MLLMs rely more heavily on text tokens to gather information from visual tokens. Overall, these results well validate our assumptions about the distribution fitting and the consideration of both self- and cross-attention modeling.

Additionally, we also visualize the pruning strategy in Fig. 5-(c). It can be seen that as the pruning ratios increase, FitPrune tends to focus more on pruning tokens in the shallower layers, leading to a significant reduction in computational overhead.

In Tab. 3, we ablate the number of examples for FitPrune to produce a pruning recipe. The first observation is that our FitPrune method maintains stable performance even

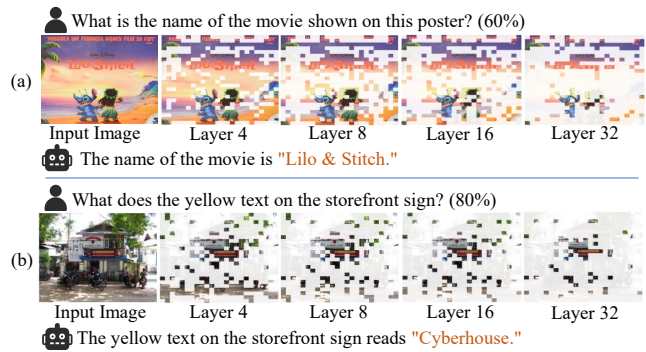


Figure 6: Visualization of token pruning process. The percentage in brackets indicates the pruning ratio.

with a small scale of statistical data. For instance, when using only 0.01% of the dataset (65 samples), the model achieves an accuracy of 61.6%, comparable to using 1% of the dataset (6.6K samples). This result indicates that the pattern of information exchange may depend more on the nature of the model itself, and our FitPrune method can capture this pattern efficiently. Furthermore, as the amount of data decreases, the statistical time is also reduced substantially. Overall, these experiment results well confirm the motivation and efficiency of FitPrune.

Visualization results. In Fig. 6, we visualize the token pruning process for 2 challenging samples on LLaVA-1.5 7B. From Fig. 6, it can be seen that our FitPrune retains key object tokens according to the question while pruning irrelevant tokens. For example, our FitPrune retains the poster title in Fig. 6-(a). Besides, as the number of deleted tokens increases, tokens from the primary object are gradually pruned. For the text-related tasks, we can notice that FitPrune retains all text-related tokens. For instance, in Fig. 6-(b), the yellow text is preserved intact until the end. These visualizations highlight FitPrune’s effectiveness in retaining key tokens essential for accurate question answering.

Conclusion

In this paper, we introduce FitPrune, a novel and training-free approach for visual token pruning of MLLMs. By framing token pruning as a statistical problem, FitPrune aims to minimize divergence in attention distributions, enabling the efficient pruning of redundant tokens. FitPrune can generate an optimal pruning strategy based on a small batch of data, avoiding costly manual trials. We validate FitPrune on various MLLMs, demonstrating its ability to reduce computational overhead while preserving performance. FitPrune offers a practical and efficient solution for optimizing MLLMs, and can be seamlessly integrated into existing models, paving the way for more scalable and efficient vision-language tasks. To further unlock its potential, future work will focus on expanding the application of FitPrune to other domains and tasks, while investigating the in-depth impact of token pruning.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No.2022J06001).

References

- Anagnostidis, S.; Pavlo, D.; Biggio, L.; Noci, L.; Lucchi, A.; and Hofmann, T. 2024. Dynamic context pruning for efficient and interpretable autoregressive transformers. *Advances in Neural Information Processing Systems*, 36.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlm0: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inform. Process. Syst.*, 35: 32897–32912.
- Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.
- Cao, J.; Ye, P.; Li, S.; Yu, C.; Tang, Y.; Lu, J.; and Chen, T. 2024. MADTP: Multimodal Alignment-Guided Dynamic Token Pruning for Accelerating Vision-Language Transformer. *arXiv preprint arXiv:2403.02991*.
- Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. *arXiv preprint arXiv:2403.06764*.
- Chen, M.; Shao, W.; Xu, P.; Lin, M.; Zhang, K.; Chao, F.; Ji, R.; Qiao, Y.; and Luo, P. 2023. Diffrate: Differentiable compression rate for efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17164–17174.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023).
- Dong, P.; Sun, M.; Lu, A.; Xie, Y.; Liu, K.; Kong, Z.; Meng, X.; Li, Z.; Lin, X.; Fang, Z.; et al. 2023. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 442–455. IEEE.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Fayyaz, M.; Koohpayegani, S. A.; Jafari, F. R.; Sengupta, S.; Joze, H. R. V.; Sommerlade, E.; Pirsiavash, H.; and Gall, J. 2022. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 396–414. Springer.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.
- Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6325–6334.
- Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.
- Han, Y.; Huang, G.; Song, S.; Yang, L.; Wang, H.; and Wang, Y. 2021. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7436–7456.
- Huang, W.; Liu, H.; Guo, M.; and Gong, N. Z. 2024. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.
- Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.
- Kim, S.; Shen, S.; Thorsley, D.; Gholami, A.; Kwon, W.; Hassoun, J.; and Keutzer, K. 2022. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 784–794.
- Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Shen, X.; Yuan, G.; Ren, B.; Tang, H.; et al. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, 620–640. Springer.
- Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2023c. Monkey: Image resolution and

- text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.
- Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.
- Liu, C.; Tao, C.; Liang, J.; Feng, J.; Shen, T.; Huang, Q.; and Zhao, D. 2023a. Length-Adaptive Distillation: Customizing Small Language Model for Dynamic Token Pruning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Liu, Y.; Li, Z.; Yang, B.; Li, C.; Yin, X.; Liu, C.-l.; Jin, L.; and Bai, X. 2023e. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.
- Luo, G.; Zhou, Y.; Zhang, Y.; Zheng, X.; Sun, X.; and Ji, R. 2024. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. *arXiv preprint arXiv:2403.03003*.
- Meng, L.; Li, H.; Chen, B.-C.; Lan, S.; Wu, Z.; Jiang, Y.-G.; and Lim, S.-N. 2022. Advait: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12309–12318.
- OpenAI. 2023. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>. Accessed: 2025-01-01.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C.-J. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34: 13937–13949.
- Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.
- Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, H.; Dedhia, B.; and Jha, N. K. 2023. Zero-TPrune: Zero-Shot Token Pruning through Leveraging of the Attention Graph in Pre-Trained Transformers. *arXiv preprint arXiv:2305.17328*.
- Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Wei, S.; Ye, T.; Zhang, S.; Tang, Y.; and Liang, J. 2023. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2092–2101.
- Xu, H.; Ye, Q.; Yan, M.; Shi, Y.; Ye, J.; Xu, Y.; Li, C.; Bi, B.; Qian, Q.; Wang, W.; et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, 38728–38748. PMLR.
- Ye, D.; Lin, Y.; Huang, Y.; and Sun, M. 2021. Tr-bert: Dynamic token reduction for accelerating bert inference. *arXiv preprint arXiv:2105.11618*.
- Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. URL <https://arxiv.org/abs/2205.01068>.
- Zhou, Y.; Ji, R.; Sun, X.; Luo, G.; Hong, X.; Su, J.; Ding, X.; and Shao, L. 2020. K-armed bandit based multi-modal network architecture search for visual question answering. In *Proceedings of the 28th ACM international conference on multimedia*, 1245–1254.
- Zhou, Y.; Ren, T.; Zhu, C.; Sun, X.; Liu, J.; Ding, X.; Xu, M.; and Ji, R. 2021. Trar: Routing the attention spans in transformer for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2074–2084.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.