PAR: Prompt-Aware Token Reduction Method for Efficient Large Multimodal Models

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) demonstrate strong performance across visual tasks, but their efficiency is hindered by significant computational and memory demands from processing long contexts in multimodal inputs. To address this, we introduce PAR (Prompt-Aware Token Reduction), a novel and plug-and-play approach that reduces visual tokens efficiently without compromising model performance. Unlike previous methods that rely heavily on attention mechanisms and overlooking cross-modal interactions , we uses a prompt-aware strategy to adpative identify and cluster essential visual tokens. PAR categorizes visual context redundancy into two types: external and internal. External redundancy is minimized through semantic retrieval, while internal redundancy is addressed using a token routing mechanism. This method substantially reduces computational load without requiring additional training or complex architectural modifications. **Experimental results demonstrate that across various visual question answering tasks, PAR reduces FLOPs by 83% with a compression ratio of 89%, while retaining 97% of baseline accuracy.** The adaptive design of PAR achieves a 2x token reduction ratio compared to prior approaches, enabling a better balance between performance and efficiency.

1 Introduction

Thanks to advanced architectures and large-scale training, large language models (LLMs) have achieved remarkable success in natural language understanding and generation tasks in a range of NLP domains(Touvron et al.; Ouyang et al., 2022; Chowdhery et al., 2023). While traditional LLMs operate on textual inputs, real-world data spans multiple modalities, including images, audio, and point clouds. To bridge this gap, recent research has extended LLM capabilities to multimodal domains(Jin et al., 2024; Song et al., 2023), giving rise to multimodal large language models (MLLMs) such as GPT-4(OpenAI, 2024), Gemini(Team, 2024), LLaVA(Liu et al., 2024), and MiniGPT-4(Zhu et al., 2023). These models leverage pre-trained modality-specific encoders and alignment strategies to integrate information across modalities efficiently, enabling broader applicability without the cost of training from scratch.

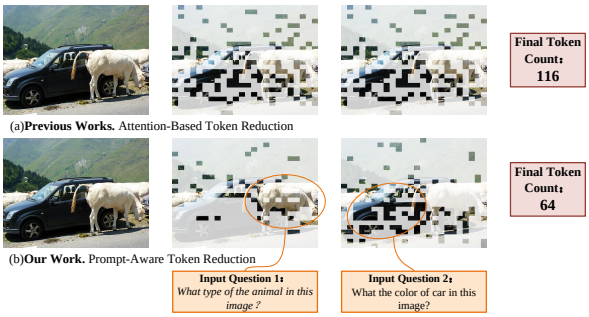

Figure 1: Previous works relying on the attention mechanism, focus on global visual tokens and cause unnecessary redundancy. In contrast, our method is guided by prompts and focuses more effectively on the task-relevant visual tokens. **Our approach achieves a token reduction ratio about 2x of previous methods.**

Despite their potential, MLLMs face significant computational and memory challenges, particularly in deployment and inference. MLLMs rely on visual encoders to convert image inputs into large sets of visual tokens, concatenated with prompt tokens, forming long input sequences. This increase in sequence length raises storage and computational demands. How to reduce the number of visual tokens without significantly impacting generation performance is a critical challenge that must be addressed.

Traditional token reduction methods (Chen et al., 2023; Shang et al., 2024; Chen et al., 2025) typically rely on attention scores to remove unimportant tokens. However, these methods struggle to adaptively balance efficiency and task relevance, and often overlook the unique redundancy patterns

of visual tokens. Some also require customized architectures (Chu et al., 2023; Cao et al., 2024; Haurum et al., 2023), adding extra complexity. In this work, we innovatively take an information-theoretic view and classify visual redundancy into two types: external redundancy, referring to task-irrelevant information, and internal redundancy, referring to semantically repetitive yet task-relevant tokens. This formulation guides our design of a more efficient reduction method.

Fully considering both external and internal redundancy observed in visual tokens, we introduce a simple yet effective training-free token reduction method called PAR (**P**rompt-**A**ware Token **R**eduction). In this approach, we develop a prompt-aware strategy to identify and retrieve important visual tokens within the given context, effectively optimizing the selection of tokens based on their semantic relevance.

Our method proceeds as follows. First, we apply predefined templates to rewrite the user prompt, enhancing its semantic representation in future steps. Next, we use a graph-based clustering algorithm to partition visual tokens into meaningful semantic clusters based on similarity distribution. We then perform prompt-guided semantic retrieval, matching the most prompt-relevant visual tokens to eliminate external redundancy. Finally, a token router simplifies the retained tokens by refining the final selection thus removing internal redundancy.

In the design of the experiments, we fully considered the balance between performance and efficiency. *Our goal was to achieve the optimal trade-off by minimizing the final number of visual tokens used while ensuring a minimal loss in accuracy.* Experimental results show that PAR achieves an 83% reduction in FLOPs and an 89% compression ratio across diverse visual question-answering tasks, all while preserving 97% of the baseline accuracy. Notably, in hallucination benchmarks(Li et al., 2023), PAR outperforms the original model under specific settings, indicating effective mitigation of hallucination phenomena in MLLMs through reduced external redundancy.

In summary, our work makes three main contributions:

- We conducted an in-depth analysis of the redundancy present in visual representations, categorizing it into external and internal redundancy. External redundancy refers to representations irrelevant to the task, while internal redundancy represents those that contribute overlapping semantic information.
- Inspired by human cognition, we propose **PAR**, a training-free method for reducing visual tokens. PAR leverages prompt semantics to eliminate external redundancy, while a token router adaptively filters internal redundancies, retaining only the most relevant tokens.
- Experimental results show that PAR not only surpasses previous methods in accuracy but also achieves a **2x** token reduction, effectively balancing efficiency and performance.

2 Related works

2.1 Multimodal Large Language Models

The development of Large Language Models (LLMs) such as GPT-3(Brown et al., 2020), LLaMA(Touvron et al.), and GPT-4(OpenAI, 2024) has seen substantial progress in recent years. These advancements have inspired the evolution of Multimodal Large Language Models (MLLMs), which extend the capabilities of LLMs to include images. Notable examples of this progress are LLava(Liu et al., 2024), MiniGPT-4(Zhu et al., 2023), InstructBLIP(Dai et al., 2023),Qwen(Bai et al., 2023),and Gemini(Team, 2024) .

These MLLMs primarily utilize a visual encoder(Radford et al., 2021) to process visual input. They then align this visual data with text through linear projection and concatenate the visual information with text tokens for generation by a pre-trained LLM. By integrating data from various modalities, MLLMs enhance contextual understanding, thereby improving the accuracy of information processing and generation.

Despite these advancements, MLLMs face significant computational costs during inference and deployment, highlighting the need for efficient token reduction techniques. Especially when dealing with videos or high-resolution images, processing thousands of tokens becomes necessary.

2.2 Visual token pruning

The quadratic complexity of Transformers (Vaswani, 2017) is a significant challenge, especially for MLLMs where image inputs are converted into numerous tokens. This process results in substantial computational costs and limits scalability due to the high memory demands of processing long sequences.

Recent research has sought to improve inference efficiency by pruning visual tokens. LLaVA-PruMerge (Shang et al., 2024) employs an adaptive visual token reduction strategy that takes advantage of the sparsity in visual encoders, selectively retaining essential tokens and enhancing their informational content through clustering and weighted averaging. FastV (Chen et al., 2025) reduces inference costs by learning adaptive attention patterns in early layers and pruning tokens in later layers. However, these approaches still rely heavily on attention mechanisms and often overlook the cross-modal relevance of tokens. Due to the inherent properties of the attention mechanism, they struggle to adaptively reduce tokens within a task-aware context, making it difficult to maintain a balance between performance and efficiency.

3 Methods

3.1 Preliminaries

Multimodal large language models(MLMMs) (Liu et al., 2024)(OpenAI, 2024) represent a significant advancement in artificial intelligence by integrating visual encoders with pre-trained large language models. This integration allows these models to jointly process and understand diverse modalities, such as images and text, enabling more comprehensive and contextually aware analyses.

For a given image I, the visual encoder f_{visual} transforms it into a series of token representations $\mathcal{V}$:

$$\mathcal{V} = f_{visual}(I) \tag{1}$$

Here, $\mathcal{V}$ denotes a set of visual tokens $\{v_i\}_{i=1}^{n} \subset R^d$ that capture essential visual features and semantic information from the image. This process allows image content to be encoded in a format compatible with the language model, facilitating the effective integration of visual and textual data.

3.2 Redundancy Analysis and Motivation

MLLMs typically handle lengthy sequences of visual tokens densely extracted from images. Due to the quadratic complexity of the attention mechanism(Vaswani, 2017), processing such long visual inputs imposes a significant computational burden, regardless of whether all tokens are necessary for the downstream task. While prior methods often rely on intrinsic model properties, such as attention mechanism for token selection or pruning, we take a different perspective. We propose an information-theoretic framework to formally characterize and eliminate redundancy in MLLMs' visual inputs.

In this section, we provide a conceptual and theoretical analysis of two prevalent forms of redundancy: *external redundancy*, where tokens are irrelevant to the task, and *internal redundancy*, where tokens carry overlapping information. This analysis serves as the motivation for our proposed method, which explicitly addresses both redundancy types to improve efficiency and preserve semantic integrity.

3.2.1 External Redundancy: Task-Irrelevant Tokens

External redundancy refers to visual tokens that are semantically unrelated to the task objective. These tokens often originate from uninformative background regions or task-irrelevant objects. This issue is particularly prominent in Visual Question Answering (VQA) tasks, where the model is expected to attend only to task-relevant regions, rather than uniformly processing all visual input.

From an information-theoretic perspective(Kraskov et al., 2004), the irrelevance of such tokens can be formalized as:

$$I(v_e;\, Y \mid T) \approx 0 \tag{2}$$

where T denotes the task instruction, Y is the model's output, and $v_e \subset V$ is the subset of visual tokens deemed externally redundant.

Tokens with near-zero conditional mutual information contribute minimally to the final output. Retaining such irrelevant content not only introduces unnecessary computational overhead but may also increase the risk of hallucination during generation, thereby compromising the faithfulness of the model's predictions.

3.2.2 Internal Redundancy: Repetitive or Overlapping Tokens

Internal redundancy arises when multiple tokens convey overlapping semantic content despite being individually task-relevant. This commonly occurs due to repeated textures or structures in the image. With fixed patch-based vision encoders, a single object may be split into multiple similar tokens, producing duplicated representations.

Let $\mathcal{V}' = \{v_1, v_2, \ldots, v_k\} \subset \mathcal{V}$ be a subset of visual tokens. We characterize internal redundancy

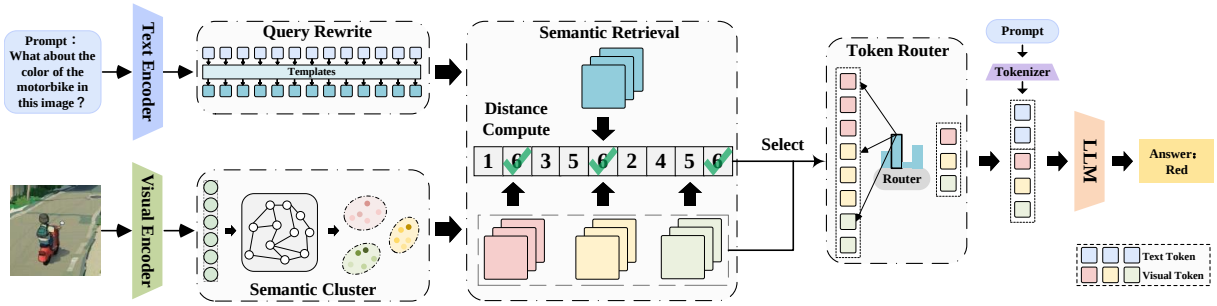

Figure 2: **The framework of our method.** Given an input of image and text, PAR processes each modality separately: the text is structured using predefined templates, and the image undergoes semantic clustering. Prompt tokens are then retrieval with visual tokens to select relevant ones, reducing external redundancy. Finally, the token router refines these selections, removing internal redundancy.Finally, the prompt token is concatenated with the refined visual tokens and fed into the large language model (LLM) to generate the final answer.

via mutual information as follows:

$$\exists\, v_i, v_j \in \mathcal{V}' : \tag{3}$$

$$I(v_i;\, v_j) \gg 0, \tag{4}$$

$$I(v_i;\, Y \mid T) \approx I(v_j;\, Y \mid T) \tag{5}$$

Equation (4) indicates a high degree of semantic similarity between tokens, suggesting duplication. Equation (5) implies that their influence on the final output Y, given the task instruction T, is nearly indistinguishable.

Together, these conditions identify token sets that are semantically redundant and contribute repetitive information. Such redundancy not only offers limited information but also undermines inference efficiency. Moreover, this type of redundancy is unique to the visual domain and remains inadequately addressed by existing compression or pruning methods, thus motivating a dedicated solution.

3.3 Semantic Retrieval for External Redundancy

Building upon the analysis of external redundancy in Section 3.2.1, we propose a novel token reduction designed for external redundancy.

We formulate external redundant token reduction as a mutual information maximization problem. Given a visual token set V and an task instruction T, the goal is to select a subset $V' \subset V$ of size k that retains the most informative content:

$$\max_{V' \subset V,\, |V'|=k} \sum_{v_i \in V'} I(v_i;\, T) \tag{6}$$

However, directly estimating mutual information $I(v_i;\, T)$ is intractable in practice. To avoid this, we leverage the shared semantic embedding space in MLLMs, where both visual and textual inputs are encoded into a unified representation. Within this space, semantic similarity naturally correlates with mutual information and can thus be used as a practical proxy.

We thus approximate $I(v_i;\, T)$ using the similarity between a visual token v_i and the task instruction T, denoted as $\text{sim}(v_i, T)$:

$$I(v_i;\, T) \propto \text{sim}(v_i, T) \tag{7}$$

In implementation, both v_i and T are encoded using the same multimodal encoder thus ensuring alignment in the shared embedding space.

Based on this approximation, we formulate the token reduction problem as a semantic retrieval task—identifying a subset of visual tokens that are most semantically aligned with the task prompt. To enhance retrieval efficiency and mitigate the effects of diverse visual appearances and prompt styles, we introduce a query rewrite mechanism and a graph-based semantic cluster algorithm.

Query Rewrite. It is not appropriate to use all prompt tokens as the query for retrieval. Original prompts may contain biases that inadequately represent certain semantic information in the embedding space, thus impacting retrieval accuracy.

To address the issue of information asymmetry between different modalities, we rewrite the query prompts ahead of retrieval. By intentionally controlling the structure and word choices of the query, we ensure that the text embeddings more accurately capture the semantic features relevant to the target image, thus enhancing the performance and retrieval precision.

We simply employ a text rewriting framework based on predefined templates, as illustrated below:

$$< Prefix >< Main >< Info > \tag{8}$$

This framework systematically organizes different forms of text descriptions to ensure that the prompts effectively convey semantic information aligned with visual features. Additionally, key elements can be precisely adjusted based on task requirements to optimize the retrieval performance of the text prompts.

Semantic Cluster. The semantic understanding of visual tokens is highly dependent on contextual information. If retrieval is conducted on a single-token level, it may result in the loss or blurring of higher-level semantic details.

To address this issue, we propose a graph-based semantic clustering strategy that supports prompt-aware retrieval. The key insight is that by grouping semantically coherent tokens before retrieval, we can (1) aggregate local semantic context to form robust cluster representations and (2) reduce the number of retrieval candidates without losing semantic coverage.

Formally, given a visual token set $\mathcal{V} = \{v_1, \ldots, v_n\} \subset R^d$, we construct a semantic similarity graph $G = (\mathcal{V}, E)$, where an undirected edge $(v_i, v_j) \in E$ exists if the semantic similarity between v_i and v_j exceeds a predefined threshold ϵ. We then extract connected components from this graph as semantic clusters, each representing a locally consistent region in the image. The detail algorithm is depicted in Algorithm 1.

Algorithm 1 Graph-Based Semantic Clustering

Require: Visual tokens: $\mathcal{V} = \{v_i\}_{i=1}^n \subset R^d$; similarity threshold ϵ

Ensure: Semantic clusters $\mathcal{C} = \{C_k\}_{k=1}^m$; cluster prototypes $\mathcal{P} = \{\mathbf{c}_k\}_{k=1}^m$

procedure ClusterTokens($\mathcal{V}, \epsilon$)
  Compute cosine similarity matrix $S \in R^{n \times n}$, where $S_{ij} = \cos(v_i, v_j)$
  Construct graph $G = (\mathcal{V}, E)$, with edges $E = \{(v_i, v_j) \mid S_{ij} \geq \epsilon\}$
  Identify connected components $\mathcal{C} = \{C_1, \ldots, C_m\}$ from G cluster $C_k \in \mathcal{C}$
  Compute prototype $\mathbf{c}_k = \frac{1}{|C_k|} \sum_{v_i \in C_k} v_i$
  return $\mathcal{C}, \mathcal{P}$
end procedure

This clustering process transforms a dense visual token sequence into a compact, semantically meaningful candidate set. The prototypes $\mathcal{P}$ are then used for prompt-aware semantic retrieval, ensuring that the model focuses on main visual concepts while discarding irrelevant or redundant background content.

To finalize the selection, we perform semantic retrieval using the rewritten instruction prompt T' against the set of semantic cluster prototypes $\{\mathbf{c}_k\}$. We compute the similarity between each cluster prototype and the rewritten prompt, and rank the clusters accordingly to select the top-k candidates:

$$\text{Top-}k = \arg\text{top}_k \ \text{sim}(\mathbf{c}_k, T') \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes the semantic similarity between the cluster representation $\mathbf{c}_k$ and the rewritten prompt embedding T'.

3.4 Token Router for Internal Redundancy

In the previous section, we addressed *external redundancy* by constructing a compact set of task-relevant tokens through semantic retrieval. However, the resulting token set $\mathcal{V}'$ may still contain *internal redundancy*—tokens that are individually relevant but semantically overlapping, thus contributing redundant information to the final representation.

To eliminate this redundancy, we propose a lightweight yet effective post-retrieval pruning mechanism called *token router*. The key idea is to iteratively retain the most informative and semantically distinct tokens while discarding those with high pairwise similarity to already selected ones. This ensures the final token set maintains diversity and avoids unnecessary duplication.The detail as illustrated in Algorithm 2

By routing semantically diverse tokens in this way, our method reduces internal redundancy while preserving the most informative content, ultimately enhancing the efficiency and effectiveness of downstream generation.

4 Experiments

4.1 Experimental Setting

Datasets.To accurately assess the practical effectiveness of our method, we selected four types of popular visual question-answering benchmarks, including (1)General visual question answering benchmark: GQA(Hudson and Manning, 2019) (2)Hallucination benchmark: POPE(Li et al., 2023) (3)Optical character-based visual question answering benchmark: TextVQA(Singh et al., 2019) (4)Comprehensive benchmark:MME(Fu et al., 2024),MMBench(Liu et al., 2025), MMVet(Yu et al., 2023).

Algorithm 2 Token Routing Algorithm

Require: Retrieved visual tokens $V = \{v_i\}_{i=1}^{k} \subset R^d$, threshold $\tau \in [0, 1]$
Ensure: Filtered token set V'

function TOKENROUTER(V, τ)
  Normalize all vectors: $\hat{v}_i \leftarrow v_i / \|v_i\|$
  Compute similarity matrix $S_{ij} \leftarrow \hat{v}_i \cdot \hat{v}_j$
  Set $S_{ii} \leftarrow 0$, and $S_{ij} \leftarrow 0$ if $S_{ij} \leq \tau$
  Initialize result set $\mathcal{R} \leftarrow \emptyset$, candidate index list $Q \leftarrow [1, ..., k]$
  while Q not empty **do**
    Pop t from Q, add to $\mathcal{R}$
    Remove all $j \in Q$ where $S_{tj} > 0$ (i.e., similar to t)
  end while
  return $V' \leftarrow \{v_i \mid i \in \mathcal{R}\}$
end function

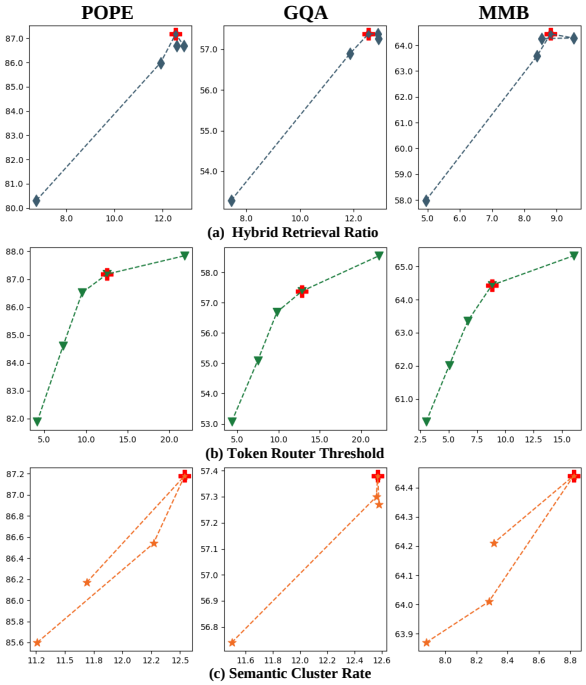

Figure 3: **Hyperparameters ablation results** about **Hybrid Retrieval Ratio** ,**Token Router Threshold** and **Semantic Cluster Rate** across three datasets. To illustrate the trade-off between performance and efficiency, we use Token Ratio as the x-axis and Accuracy as the y-axis. The red sign represents the selected parameters.

Model.We first applied our method to LLaVA-1.5(Liu et al., 2024), an open-sourced multimodal large model designed for tasks such as visual question answering and image captioning.LLaVA-1.5 uses CLIP(Radford et al., 2021) as the visual encoder and a LLaMA-based (Touvron et al.)backbone large language model. The encoder and model are connected with a linear projector. To ensure semantic consistency, we use the CLIP text encoder as a text feature extractor in our method.

Notably, our approach does not require additional training or fine-tuning, setting it apart from most mainstream methods. All experiments were conducted on an NVIDIA A100 80G GPU.

4.2 Performance evaluation

In Table 1, we present PAR's performance on several visual question-answering benchmarks.

Compared to the baseline LLaVA-1.5 model, our approach maintains approximately 97% of the original model's accuracy across diverse datasets, using only around 11% of the visual tokens. This yields a tenfold increase in efficiency. Notably, on the multimodal hallucination benchmark POPE, PAR surpasses the original model in accuracy, underscoring its effectiveness in handling hallucination issues.

To further validate PAR's efficiency, we benchmarked it against two state-of-the-art methods: LLaVA-PruMerge+ (Shang et al., 2024) and Fast-V (Chen et al., 2025). Using identical experimental conditions and reproduced open-source code and weights, PAR consistently outperformed these methods both in terms of accuracy and token pruning efficiency.

These results highlight that our method, by effectively reducing redundancy in visual tokens, achieves near-baseline performance with a fraction of the computational cost. Furthermore, its strong results on hallucination benchmarks illustrate PAR's capability to minimize irrelevant contextual interference, demonstrating its robustness and precision in multimodal inference.

4.3 Ablation Study

This section establishes the purpose and structure of the ablation study clearly, setting up an in-depth examination of each component's contributions.

Effectiveness of Each Modules. Table 2 show the effectiveness of each module in our method and attention based method. We first compare the effectiveness of using only semantic retrieval versus using only attention-based selection. Both approaches achieve comparable accuracy and speedup under similar conditions. Subsequently, we integrate the token router component with each of the two methods. In both cases, this addition leads to a notable reduction in the number of tokens.

Method	GQA	VQA^{text}	POPE	MME_{per}	MME_{cog}	MMBench	MMvet	Avg
LLaVA-v1.5-7B	61.23 / 100	60.12 / 100	85.3 / 100	1529.77 / 100	439.29 / 100	65.42 / 100	29.3 / 100	**100/100**
LLaVA-PruMerge+	54.41 / 19.82	56.94 / 20.49	82.33 / 19.96	1501.84 / 20.24	435.64 / 20.24	65.70 / 20.79	28.8 / 20.5	96.43 / 20.29
Fast-V	48.04 / 20.00	45.00 / 20.00	77.48 / 20.00	1424.70 / 20.00	405.00 / 20.00	60.64 / 20.00	24.3 / 20.00	86.29 / 20
PAR (Ours)	57.38 / 12.56	57.06 / 12.65	87.18 / 12.54	1527.71 / 10.57	432.86 / 10.57	64.33 / 8.95	27.2 / 13.44	97.03(↓2.97) / 11.60(↓88.4)

Table 1: **Model performance comparison across various datasets.** Each metric is shown as "Accuracy / Token ratio". Compared to the baseline, our method results in a 2.97% decrease in accuracy, but achieves an 88.4% reduction in token usage, striking a better balance between performance and efficiency.

Table 2: **Performance Comparison of Different Modules.** R is semantic retrieval, A is attention-based selection from LLaVA-PruMerge,T is token router.

	POPE		MMB	
Module	Acc	Ratio	Acc	Ratio
LLaVA	85.3	100	65.42	100
+A	82.33	19.96	63.70	20.79
+R	83.02	20.00	64.13	20.00
+A+T	84.48	13.44	64.12	11.43
+R+T(PAR)	87.18	12.54	64.33	8.95

However, our semantic retrieval method combined with token routing consistently outperforms the attention-based counterpart in terms of accuracy and overall performance. These results demonstrate the effectiveness of each individual component.

Token Router Threshold. Figure3 shows how different routing thresholds affect performance. The Token Router reduces internal redundancy by filtering out semantically similar tokens from clustering results. We test thresholds from 0.4 to 0.7 and observe that higher thresholds retain more tokens and generally yield better accuracy. However, beyond 0.6, the performance gain slows, and excessive tokens reduce efficiency. Thus, we select 0.6 as the optimal threshold to balance accuracy and compression.

Clustering Rate. Figure 3 demonstrates the effect of varying semantic clustering rates on experimental results. Selecting an appropriate clustering rate is crucial for optimizing retrieval efficiency. If the rate is set too high, it may fail to effectively aggregate nearby semantic information, while a rate set too low can introduce noise, diminishing clustering effectiveness. To assess how different clustering rates influence generation accuracy and token retention, we tested four rates: 0.6, 0.7, 0.8, and 0.9. The results show that for the first three rates, our method follows a pattern similar to direct retrieval, with accuracy increasing as more tokens are retained. Based on these findings, we selected 0.8 as the optimal semantic clustering rate.

Distance Metric. Table 3a presents our investigation into different distance-based retrieval algorithms. In retrieval tasks, selecting an appropriate distance function is critical for effective similarity measurement. Different distance functions can have a significant impact on similarity computation, thereby affecting the final retrieval results and the model's performance. We evaluated five different distance functions: L1 distance, L2 distance, Lp distance, Linf distance, and Inner product. Among these, Linf distance—highlighting the maximum individual difference—yielded the best results, demonstrating its suitability for tasks that require sensitivity to distinct feature variations.

Query Condition. Table 3b investigates the impact of different query structures on retrieval performance. By comparing the original prompt, keyword extraction, and template-based rewriting, we explore how query optimization can enhance retrieval. The original prompt serves as a baseline, while keyword extraction improves semantic clarity by eliminating redundancy. Template-based rewriting further aligns text and visual embeddings, boosting retrieval accuracy. Experimental results show that the predefined query rewriting framework significantly improves performance."

Retrieval Granularity. In table 3c we investigated the impact of four different retrieval approaches on the results.We first employ direct retrieval under different settings, and later explore a semantic retrieval strategy. For direct retrieval, we used token blocks of sizes 1×1, 2×2, and 3×3 as the basic retrieval units. In contrast, for semantic retrieval, we employed semantic clusters based on semantic clustering as the retrieval units. Experimental results show that our gragh-based semantic retrieval methods effectively incorporate surrounding semantic information, thereby achieving the highest benchmark accuracy.

Distance Metric	POPE	MMB
L1	84.18	64.24
L2	82.01	63.31
Lp	68.89	56.94
Linf	87.18	64.33
Inner Product	83.20	62.07

(a) **Distance Metric.**

Query Condition	POPE	MMB
Original Prompt	82.38	61.29
Key word Prompt	84.25	63.57
Prompt Rewriting	87.18	64.33

(b) **Query Prompt.**

Token number	POPE	MMB
1*1 Token	83.91	63.86
2*2 Tokens	84.34	62.88
3*3 Tokens	82.91	62.76
Semantic Tokens	87.18	64.33

(c) **Retrieval Granularity.**

Table 3: **Ablation study of PAR on POPE and MMBench.** We examine three key components: (a) distance metrics for clustering, (b) prompt conditions for retrieval, (c) retrieval granularity. The best-performing configurations are highlighted in gray, confirming the effectiveness of our design choices.

Method	FLOPs (TB)	Total Memory (GB)	Prefill Time (ms)	Activation (GB)	KV Cache (MB)
LLaVA	8.2	21.8	59.1	3.9	323
PruMerge+	2	14.8	19.3	0.68	81.8
PAR	**1.4**	**14.2**	**18.5**	**0.43**	**54.5**
	(↓ 6.8)	(↓ 7.6)	(↓ 40.6)	(↓ 3.47)	(↓ 268.5)

Table 4: **Efficiency Analysis**. We use LLaVA-v1.5-7B as a baseline, the precision is fp16 and batchsize=1. All the data are estimated using a theoretical model.

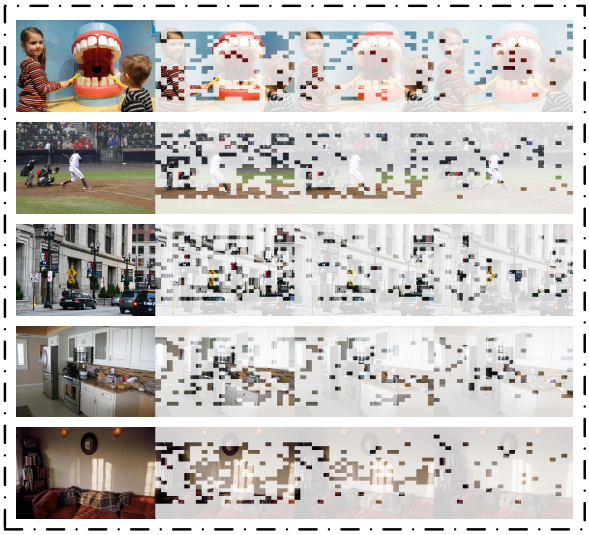

Figure 4: **Visualization of PAR**.From left to right, we change the ratio of retrieval and the visual tokens become increasingly sparse. In the utmost right is the final result of PAR.

4.4 Efficiency Analysis

To efficiently evaluate the computational performance of our method, we conducted a theoretical analysis of factors such as latency and memory usage using the Roofline tool based on LLMviewer(Yuan et al., 2024). Using the LLaVA-1.5 7B model as an example, we analyzed the multimodal large model inference process in typical scenarios. This model processes images with a resolution of 336×336 pixels, which are converted into 576 visual tokens through the CLIP model, combined with a prompt input of approximately 40 prompt tokens.

LLaVA-PruMerge+(Shang et al., 2024) achieved a compression ratio of about 20 %, reducing the visual tokens to 116. In contrast, our method, while maintaining similar accuracy on VQA tasks, achieved an 11% compression ratio, reducing the visual tokens to around 64.

As shown in Table 4, our method significantly improved model inference speed and reduced memory consumption. Specifically, for generating the first token, prefill time was reduced to 31.3% of the original, and the activation during inference was reduced to 11%. This makes our method more suitable for deploying large models in resource-constrained environments.

5 Conclusions

In this paper, we introduced PAR (Prompt-Aware Token Reduction), a lightweight and effective framework for reducing computational and memory overhead in multimodal large language models (MLLMs) by pruning redundant visual tokens. Motivated by an information-theoretic perspective on external and internal redundancy, PAR first performs semantic retrieval to identify task-relevant visual clusters, and is then followed by a token routing mechanism to eliminate semantic duplication while preserving representational diversity. Experimental results demonstrate that PAR achieves a superior trade-off between efficiency and performance, reducing FLOPs by 83%, shortening prefill latency to 31.3% of the original, and compressing 89% of the cache, while maintaining 97% of the baseline accuracy across diverse visual question answering and reasoning benchmarks.

Limitations

Currently, our work only supports the widely adopted LLaVA architecture. In the future, we plan to extend our approach to diverse model architectures, covering a variety of visual encoders. While PAR is training-free and plug-and-play, this constraint may limit its ability to further optimize token selection through learned parameters.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arxiv:abs2308.12966*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *arxiv:2005.14165*.

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. 2025. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. 2023. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2061–2070.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. 2023. Mobilevlm : A fast, strong and open vision language assistant for mobile devices. *arxiv:2312.16886*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arxiv:2305.06500*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arxiv:2306.13394*.

Joakim Bruslund Haurum, Sergio Escalera, Graham W. Taylor, and Thomas B. Moeslund. 2023. Which tokens to use? investigating token reduction in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 773–783.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient multimodal large language models: A survey. *arxiv:2405.10739*.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

OpenAI. 2024. Gpt-4 technical report. *arxiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 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:2403.15388*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. 2023. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arxiv:2311.07594*.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *arxiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arxiv:2302.13971*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arxiv:2308.02490*.

Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, and Kurt Keutzer. 2024. Llm inference unveiled: Survey and roofline model insights. *arxiv:2402.16363*.

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