An Empirical Study of Attention and Diversity for Adaptive Visual Token Pruning in Multimodal Reasoning Models

Changwoo Baek[1*] **Jouwon Song**[2*] **Sohyun Kim**[1*] **Kyungbo Kong**[1†]
[1]Department of Electrical and Electronic Engineering, Pusan National University
[2]LG Electronics
{higok18, shkim0503, kbkong}@pusan.ac.kr, juwon05.song@lge.com

Abstract

With the rapid progress of Large Reasoning Models (LRMs), interest in multimodal reasoning has grown substantially. However, multimodal reasoning often requires processing a large number of visual tokens, leading to significant computational overhead. To alleviate this issue, recent studies have explored visual token pruning strategies. Most prior works primarily focus on either attention-based or diversity-based pruning methods. However, in-depth analysis of their characteristics and limitations remains largely unexplored. In this work, we conduct thorough empirical analysis using effective rank (erank) as a measure of feature diversity and attention score entropy to investigate visual token processing mechanisms and analyze the strengths and weaknesses of each approach. Our analysis reveals two insights: (1) Attention-based methods demonstrate superior performance on simple images where information is easily concentrated, whereas diversity-based methods excel in handling complex images with distributed features. (2) Analysis using the hallucination dataset (CHAIR) shows that attention-based methods generate more conservative answers with lower hallucination rates compared to diversity-based methods which produce more exploratory responses with higher hallucination tendencies. Motivated by these observations, we propose a novel token pruning framework that adaptively combines the strengths of both methods. Extensive experiments show that our method delivers consistent high performance with efficient reasoning across both standard benchmarks and hallucination evaluation datasets.

1 Introduction

Recent advances in Large Reasoning Models (LRMs) [1, 2] highlighted the potential of scaling reasoning capabilities through long chains of thought, enabling strong performance on complex tasks such as mathematics and scientific problem solving. As these models expand beyond purely textual domains, there is a growing interest in multimodal reasoning, which integrate visual and linguistic reasoning to handle richer and more realistic inputs.

Therefore, Reinforcement Learning (RL) techniques such as s Proximal Policy Optimizatio (PPO) [3, 4, 5, 6] and Group Relative Policy Optimization (GRPO) [7] have been applied to Multimodal Large Language Models (MLLMs) to enhance their reasoning capabilities. These approaches strengthen step-by-step reasoning beyond purely supervised training. However, in multimodal reasoning, visual information is converted into token embeddings that can be processed by language models, producing

* Equal contribution. † Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

hundreds of visual tokens in the process. The large number of these tokens increases the complexity of attention-based computations quadratically, imposing a significant burden on inference speed and efficiency. Therefore, in the context of multimodal reasoning, pruning redundant visual tokens becomes essential for enabling efficient reasoning, as it reduces unnecessary computational overhead while preserving the core information required for complex multimodal understanding.

To address these issues, numerous researchers have attempted to reduce computational costs by removing unnecessary or redundant visual tokens through token pruning methods [8, 9, 10]. These existing methods typically employ two main methods. The first is attention-based methods [11, 12, 13, 14, 15], which consider tokens with high attention scores as important information and remove the rest. The second is diversity-based methods [16], which reduce redundancy based on feature similarity between visual tokens. Each approach exhibits distinct tendencies. Attention-based methods prioritize the preservation of highly weighted tokens, which can result in concentrated but sometimes repetitive selections. In contrast, diversity-based methods encourage broader coverage, often at the cost of overlooking important tokens.

To provide a clearer understanding, this work empirically analyzes the tendencies of token pruning methods. We employ effective rank (erank) [17] and attention score entropy as metrics to quantify image complexity and token concentration, and analyze the effect of these factors on different pruning approach. Our analysis demonstrates characteristic tendencies of different methods: (i) Attention-based pruning is more effective on simple images, where essential information is concentrated in a few tokens, whereas diversity-based pruning performs better on complex images, where information is more widely distributed. This is confirmed through experiments on MME [18], POPE [19], and ScienceQA [20]. (ii) We further examine hallucination tendencies using the CHAIR dataset [21]. Attention-based methods yield more conservative answers with lower hallucination rates, while diversity-based methods produce more exploratory responses with higher hallucination rates. We also observe that progressively increasing the proportion of high-attention tokens reduces hallucinations.

Based on our empirical analysis, we propose a token pruning method that builds on the observed tendencies of different approaches. To determine the adaptive setting of thresholds, we first analyze the effect of varying similarity thresholds for redundancy pruning in relation to erank. The analysis shows that simple images, where essential information is concentrated in high-attention tokens, benefit from stricter thresholds. In contrast, complex images, with more distributed and redundant information, require looser thresholds to enhance diversity. Guided by these observations, our method explores tokens in order of attention scores while removing redundant tokens based on similarity, with thresholds adaptively adjusted to balance information concentration and diversity. Experimental results demonstrate that the proposed method achieves performance comparable to state-of-the-art methods across nine standard datasets. Moreover, unlike existing approaches that exhibit dataset-specific behavior, it effectively mitigates hallucinations as validated on the CHAIR benchmark.

In summary, our contributions are three-fold:

- We provide a systematic empirical analysis to demonstrate that image complexity, as measured by erank and attention entropy, is the key factor dictating the choice between attention-based and diversity-based pruning strategies.
- In addition to standard benchmarks, we conduct a targeted analysis on a hallucination benchmark to reveal the distinct trade-offs of each strategy, particularly between factual conservatism and descriptive recall.
- Based on these insights, we propose a novel adaptive pruning framework that adjusts to image complexity, and validate its robustness and superior performance across both general and hallucination-specific benchmarks.

2 Related Works

Multimodal reasoning Recent studies have applied reinforcement learning (RL) not only for aligning models with human preferences but also as a direct mechanism to strengthen their reasoning ability. Early approaches predominantly adopted policy-gradient algorithms such as PPO [3, 4, 5, 6], while more recent work has advanced toward GRPO [7], which provides more stable and efficient optimization for reasoning tasks by comparing groups of reasoning trajectories and computing relative advantages without relying on a value function. Moreover, several approaches have specifically explored RL as a means to enhance the visual reasoning performance of MLLMs. However, a

Table 1: **Attention entropy and erank on simple and complex image datasets.** Simple images exhibit lower entropy and erank, while complex images show higher values, and the two pruning methods show contrasting performance between simple and complex images.

Method	MME			ScienceQA
	OCR	Numerical Cal.	Text Translation	
Metric				
Att. entropy	4.61	4.47	4.39	4.45
Erank	78	58	49	74
Scores after pruning 576 → 64 tokens				
Att. based	**140**	**55**	**100**	**69.51**
Div. based	130	40	80	67.53

(a) Results on datasets with **simple images.**

Method	MME			POPE
	Position	Scene	Count	
Metric				
Att. entropy	4.90	4.86	4.82	4.87
Erank	109	103	102	106
Scores after pruning 576 → 64 tokens				
Att. based	105	157	120	77.4
Div. based	**111**	**168**	**140**	**86.0**

(b) Results on datasets with **complex images.**

persistent challenge in multimodal reasoning is that raw images must be transformed into high-dimensional token embeddings, typically producing hundreds of visual tokens. While this detailed representation improves semantic comprehension, the sheer volume of tokens significantly amplifies the cost of attention computations, resulting in substantial computational overhead and slower inference speeds. [9, 11]

Visual token reduction Reducing redundant and unnecessary visual tokens is an effective way to decrease computation and memory usage, thereby improving the inference efficiency of LLMs. In particular, many studies have adopted token pruning methods that require no additional training, and these methods can largely be categorized as follows. **(i) Attention-based method:** These methods prune visual tokens by leveraging the attention distribution of [CLS] token in the penultimate layer of the vision encoder before the tokens are fed into the LLM [11, 12, 13, 14, 15]. Based on the observation that image information in the vision encoder tends to concentrate on a small set of key tokens, these methods utilize attention scores from the output layer to select a limited number of tokens that aggregate global information. However, they tend to retain similar tokens concentrated in specific regions, which results in insufficient diversity to fully represent the entire token set. **(ii) Diversity-based method:** These methods leverage inter-token similarity to enhance the diversity of the selected token set [16], thereby encouraging the selection of more diverse tokens. However, they introduce additional computational overhead and risk discarding important tokens.

3 Preliminaries

Attention concentration via attention entropy. To assess the concentration of attention within the vision encoder, we compute the Shannon entropy of the class token's attention score. Given the head-averaged attention score $\alpha \in \mathbb{R}^N$ obtained from the penultimate layer, we exclude the self-attention score of the class token and renormalize the remaining score into a valid probability distribution:

$$p_i = \frac{\alpha_i}{\sum_{j \neq \text{CLS}} \alpha_j}, \quad \sum_i p_i = 1. \tag{1}$$

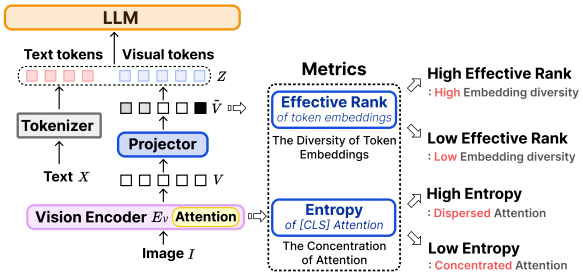


Figure 1: Overview of attention entropy and erank.

The entropy is then computed as

$$H(p) = -\sum_i p_i \log p_i. \tag{2}$$

The entropy value $H(p)$ quantifies how attention is distributed across tokens: a lower value indicates that the class token attends strongly to a few regions, whereas a higher value suggests a more uniform distribution over multiple visual tokens. We refer to this measure as **attention entropy** in the rest of this paper.

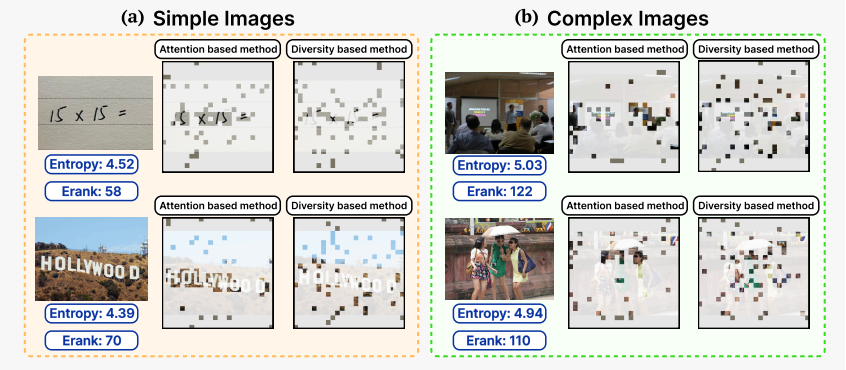


Figure 2: **Selection methods by image complexity.** (a) On simple images, attention-based methods capture concentrated information. (b) On complex images, diversity-based methods achieve broader coverage as attention disperses.

Token embedding diversity via erank. To quantitatively assess the diversity of token embeddings, we adopt the notion of erank [17]. Unlike the conventional matrix rank, the erank is an entropy-based measure that evaluates the number of dimensions effectively utilized by a matrix.

Given a token embedding matrix $A \in \mathbb{R}^{N \times d_l}$, we first obtain its singular values $\{\sigma_i\}$ via singular value decomposition (SVD). Let

$$L = \min(N, d_l), \quad q_i = \frac{\sigma_i}{\sum_{j=1}^{L} \sigma_j}, \quad q_i \in \mathbb{R}^L. \tag{3}$$

The erank is then defined as

$$\text{erank}(A) = \exp\Big(-\sum_{i=1}^{L} q_i \log q_i \Big). \tag{4}$$

The value of erank(A) ranges between 1 and L. A low erank indicates that the embedding representation is concentrated in a few dominant dimensions, whereas a high erank suggests that the embedding space is more evenly distributed across multiple dimensions.

4 Emprical Studies

This section presents empirical analyses of attention-based and diversity-based token pruning methods. We focus on two aspects: the impact of image complexity on token selection strategies (Sec. 4.1) and the relationship between pruning strategies and hallucination (Sec. 4.2). Building on the insights from these two analyses, we then propose an adaptive pruning framework (Sec. 4.3).

4.1 Impact of Image Complexity on Token Selection Strategies

Attention-based methods select tokens where information is concentrated, whereas diversity-based methods aim to reduce redundancy among the selected features and secure a broader range of representations. we set out to analyze whether image complexity causes the two methods to yield contrasting results.

Image complexity affects attention entropy and diversity. We first analyzed how image complexity affects MLLMs in their reasoning process. To this end, we measured the concentration of attention using attention entropy and assessed the diversity of token features using erank. The analysis was conducted on LLaVA-v1.5-7B using the MME Benchmark [18], where tasks such as *OCR, Numerical calculation, and Text translation* involve images with plain backgrounds and few key objects, whereas tasks such as *Position, Scene, and Count* involve images with mixed backgrounds and multiple objects, making them relatively complex. In addition, we included *ScienceQA* (simple) and *POPE* (complex) from external benchmarks for a more comprehensive analysis. Indeed, as shown in Table 1, simple images exhibited lower attention entropy, as in the case of *OCR* (4.39),

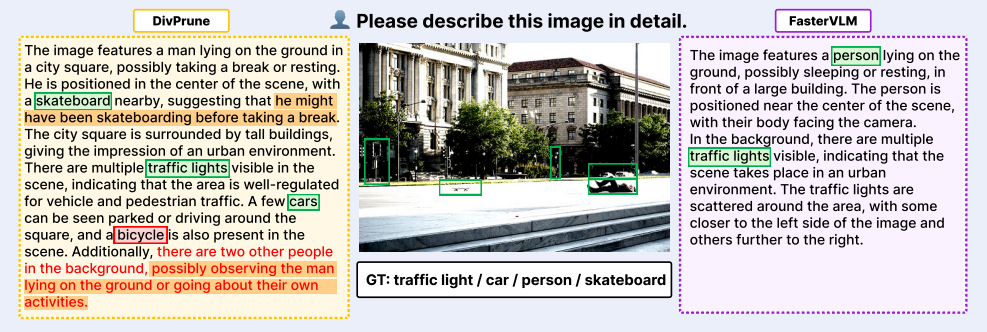


Figure 3: **Response patterns of DivPrune (diversity-based) vs. FasterVLM (attention-based).** DivPrune's responses are more comprehensive but risk hallucination, whereas FasterVLM produces safer, more focused descriptions. In the annotations, GT Obj. and Hallucinated Obj. label object words; marks DivPrune-specipic phrasing; **red text** indicates incorrect phrases.

where the vision encoder could readily concentrate information into a few dominant regions. In contrast, complex images such as *POPE* reached higher values (4.90), indicating more dispersed attention across multiple regions. Consistently, simple images also showed lower erank, such as *OCR* (49), while complex images reached higher values, such as *POPE* (109), reflecting redundant versus diverse token representations.

Performance divergence by image complexity. Our analysis reveals that the performance of these two approaches diverges depending on dataset characteristics, as shown in Table 1. As a result, diversity-based methods outperformed in high erank tasks, while attention-based methods were superior in low erank tasks. This performance reversal appears to be driven by differences in image complexity. As illustrated in Figure 2, simple images allow the vision encoder to easily concentrate attention on specific regions, leading to concentrated information. In such cases, attention-based methods can effectively select these concentrated tokens. In contrast, complex images contain multiple objects and mixed backgrounds, causing information to be dispersed across the entire image. In this scenario, diversity-based methods that capture a broader range of features become more effective.

In conclusion, our analysis shows that image complexity guides the token selection strategy. Attention-based methods are more effective for simple images with concentrated information, while diversity-based methods are superior for complex images with dispersed features.

4.2 The Relationship between Pruning Methods and Hallucination

Object hallucination occurs frequently in MLLMs and is a critical issue that undermines their reliability. In this section, we compare and analyze the characteristics of attention score–based and diversity-based methods from the perspective of hallucination, aiming to identify how the two pruning methods differ in inducing hallucinations. To this end, we evaluate object hallucination in the LLaVA-1.5-7B model using not only the datasets commonly employed in prior token reduction studies but also additional datasets, and we present the corresponding results.

Object hallucination. To assess the degree of object hallucination in the image captioning task, we employ the CHAIR dataset. CHAIR quantifies the proportion of objects mentioned in generated captions that are absent in the ground-truth annotations, providing two sub-metrics, C_I and C_S, as defined in Eq. 5.

$$C_I = \frac{\{\text{hallucinated objects}\}}{\{\text{all mentioned objects}\}}, \qquad C_S = \frac{\{\text{captions with hallucinated objects}\}}{\{\text{all captions}\}}. \tag{5}$$

Each metric evaluates hallucination at the instance level and the sentence level, respectively, and lower values indicate better performance. As auxiliary metrics, we also report recall and len, where recall denotes the proportion of ground-truth objects mentioned in the generated captions, and len represents the average number of words in the generated captions.

Results on the CHAIR dataset. Table 2 presents the results of the attention-based and diversity-based methods on the CHAIR dataset. Despite small differences in Len, the diversity-based methods

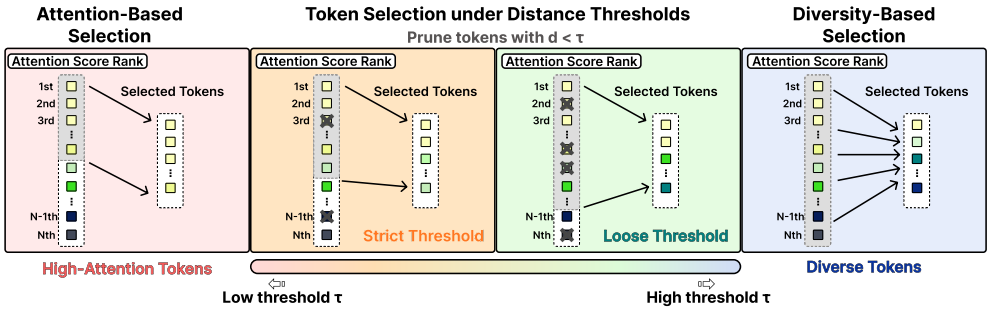


Figure 4: **Effect of similarity threshold τ on token selection.** A low (*strict*) τ prioritizes high-attention tokens, while a high (*loose*) τ increases the diversity of the selected tokens.

exhibit higher values of the hallucination metrics C_S and C_I compared to the attention-based methods, suggesting that selecting diverse tokens with low feature similarity may increase the likelihood of hallucination. In contrast, the diversity-based methods achieve higher recall, indicating that they are able to capture a larger number of objects than the attention-based ones.

Comparison of the two methods in terms of response patterns. The two methods also differ in their response patterns. Fig. 3 illustrates this distinction, showing that the diversity-based method DivPrune generates broader and more open-ended descriptions and often includes speculative expressions, as highlighted in yellow. In addition, the diversity-based response refers to a larger set of ground-truth objects marked in green, but at the same time it also introduces hallucinated objects highlighted in red and incorrect phrases emphasized in red text. In contrast, the attention-based method FasterVLM focuses on the main objects and provides more conservative and reliable explanations, thereby suppressing hallucinations that frequently appear in diversity-based outputs.

Effect of attention-based selection on object hallucination. Based on the observation that attention-based token selection tends to reduce hallucination, we conducted experiments to quantitatively assess the effect of attention-based selection on hallucination by varying the balance between diversity-based and attention-based selection. As shown in Table 3a, we fixed the token budget at 64 and gradually reduced the number of tokens selected by the diversity-based method, DivPrune, while replacing them with tokens having higher attention scores. The experimental results demonstrate that increasing the proportion of attention-based selection leads to a gradual decrease in the hallucination metrics C_S and C_I. *These findings suggest that selecting tokens solely based on diversity can induce relatively higher hallucination, whereas tokens with high attention scores, which concentrate critical information, play a pivotal role in generating reliable captions and mitigating hallucination.*

Table 2: **Comparison on CHAIR.** Diversity-based methods yield higher recall but also higher hallucination (C_S, C_I), whereas attention-based methods reduce hallucination at the cost of recall. †FPSPruner is based on farthest point sampling (FPS), which iteratively selects the farthest token to guarantee diversity.

Method	$C_s \downarrow$	$C_i \downarrow$	Recall↑	Len
LLaVA-1.5-7B	51.0	13.9	78.7	101.4
Attention-based methods				
FasterVLM (arXiv'24)	45.4	13.5	69.3	94.0
PruMerge+ (ICCV'25)	45.2	15.6	66.7	91.4
Vispruner (ICCV'25)	49.8	15.0	72.6	96.7
Diversity-based methods				
DivPrune (CVPR'25)	57.4	18.0	76.4	101.1
FPSPruner†	58.6	18.6	76.0	100.5

Effect of image complexity on object hallucination. To analyze object hallucination performance with respect to image complexity, we compared FasterVLM and DivPrune on two subsets of the CHAIR dataset, which consists of 500 samples in total. We categorized the samples according to image complexity using erank, based on the analysis in Sec. 4.1. The Low-erank subset consists of 75 samples that fall into the bottom 15% in terms of erank , whereas the High-erank subset consists of 75 samples that belong to the top 15%. As shown in Table 3b, DivPrune generally achieves higher recall values; however, within the Low-erank subset, FasterVLM outperforms DivPrune. This finding is consistent with the analysis in Sec. 4.1, which suggests that when erank is low, and information is concentrated, FasterVLM performs relatively better than DivPrune. Meanwhile, for the hallucination metrics C_S and C_I, FasterVLM also demonstrates superior performance over DivPrune in the High-

Table 3: Side-by-side results: (a) Mixtures of DivPrune/attention-selected tokens; (b) Comparison of FasterVLM and DivPrune on low erank vs. high erank subsets.

Method	$C_s \downarrow$	$C_i \downarrow$	Recall↑	Len	Mean erank	Mean attn.
Retain 64 Tokens						
R=0	57.4	18.0	76.4	101.1	21.14	0.0035
R=0.25	50.8	16.8	74.5	97.6	14.98	0.0065
R=0.50	46.2	14.5	73.7	95.5	14.38	0.0072
R=0.75	45.2	14.1	70.5	94.0	13.58	0.0076

(a) Hallucination performance across different attention-based selection ratio R.

		Low erank (15%)	High erank (15%)
Recall↑	FasterVLM	89.1	65.3
	DivPrune	88.8	73.3
$C_S \downarrow$	FasterVLM	26.7	60.0
	DivPrune	29.3	81.3
$C_I \downarrow$	FasterVLM	7.46	15.8
	DivPrune	8.54	20.6

(b) CHAIR metrics across low erank vs. high erank (15%) subsets.

Table 4: Comparison of performance across different similarity thresholds. For each metric, the boldface indicates the best performance.

Metric	Similarity Threshold (τ)							
	0	0.01	0.05	0.1	0.15	0.25	0.35	0.5
	Erank / Performance							
MME (High)	15.1 / 1351	18.2 / 1354	22.5 / 1358	25.8 / **1384**	28.1 / 1374	30.4 / 1352	32.7 / 1348	35.0 / 1333
POPE	14.9 / 83.1	17.8 / 83.4	21.9 / 84.0	24.5 / 85.2	27.8 / 85.5	30.1 / **86.1**	31.9 / 85.2	34.2 / 83.9
MME (Low)	16.2 / **316**	19.1 / 304	23.3 / 307	26.2 / 294	28.9 / 290	31.5 / 280	33.1 / 275	35.8 / 277
ScienceQA	15.5 / **69.5**	18.8 / 68.9	22.8 / 68.5	25.9 / 68.0	28.5 / 67.9	31.0 / 67.5	32.9 / 67.2	35.2 / 66.9

erank subset. *This indicates that even in complex and dispersed image settings, selecting tokens with diverse features does not provide a clear advantage in mitigating hallucination.*

4.3 Towards Adaptive Token Similarity Thresholding

In this section, we use the datasets with varying image complexities established in Section 4.1 to analyze the relationship between erank and the similarity threshold used for pruning redundant tokens. Our analysis method iteratively selects high-attention tokens and prunes similar neighbors, thereby modulating the diversity of the final token set based on the chosen threshold. The process is as follows:

1. All tokens are sorted in descending order based on their attention score.
2. Starting with the highest ranked token, we select it and then prune all other candidate tokens whose cosine distance d to the selected token is smaller than a predefined threshold τ.
3. The process moves to the next highest-ranked token that has not been pruned and repeats the pruning step until the desired number of tokens is selected.

In this framework, the threshold τ is the parameter that directly governs the diversity of the final token set. As illustrated in Figure 4, applying a low (strict) threshold results in the pruning of only a few, highly similar tokens. Conversely, a high (loose) threshold removes a wider range of similar tokens, constructing a final token set with greater diversity.

To quantitatively measure this effect, we varied τ from 0 to 0.5 and observed its impact on token diversity. As shown in Table 4, the results demonstrate a direct positive correlation between the similarity threshold τ and the diversity of the selected token set. As τ increases, the erank of selected tokens consistently rises across all datasets. This indicates that a higher threshold causes more tokens to be treated as redundant and pruned, which in turn enhances the diversity of the final set. Notably, token diversity was at its lowest when the threshold was close to zero, as very little similarity-based pruning occurs under this condition.

However, the optimal level of diversity is dictated by the image's internal characteristics, leading to contrasting outcomes. For images with low attention entropy (i.e., concentrated information), where critical information is focused in high-attention tokens, even highly similar tokens can contain vital fine-grained details. Consequently, aggressive pruning with a high threshold (τ) degraded performance, whereas a conservative low threshold proved more effective. In stark contrast, for images with high attention entropy (i.e., dispersed information), where information is more distributed and redundant, a higher threshold improved performance by effectively eliminating this redundancy and facilitating the selection of a more diverse token set.

Table 6: **Results of different token pruning methods on 9 multimodal benchmarks.** *Average* is normalized to the full-token **LLaVA-1.5-7B** (set to 100%). MME is reported in its original score units.

Method	VQA^{v2}	GQA	VizWiz	SQA^{IMG}	TextVQA	POPE	MME	MMB	MMB^{CN}	Average
Vanila 576 Tokens										
LLaVA-1.5-7B	78.5	61.9	50.1	69.5	58.2	85.9	1862	64.7	58.1	100.00%
Retain 128 Tokens										
FastV (ECCV'24)	71.0	54.0	51.9	69.2	56.4	68.2	1490	63.0	55.9	92.31%
PDrop (CVPR'25)	74.3	57.1	49.4	70.1	56.7	77.5	1696	62.3	55.3	95.17%
SparseVLM (ICML'25)	75.1	57.3	49.7	69.0	56.3	83.1	1761	62.6	56.9	96.61%
PruMerge+ (ICCV'25)	75.0	58.2	53.7	69.1	54.0	83.1	1554	61.8	55.8	95.64%
VisionZip (CVPR'25)	75.6	57.6	51.6	68.7	56.9	83.3	1763	62.1	57.0	97.19%
VisPruner (ICCV'25)	75.8	58.2	52.7	69.1	57.0	84.6	1768	62.7	57.3	98.01%
DivPrune (CVPR'25)	76.0	59.4	52.8	68.6	54.5	85.5	1707	60.1	52.3	97.25%
Ours	76.4	59.3	52.6	68.5	57.0	86.5	1755	62.3	56.6	**98.29%**
Retain 64 Tokens										
FastV (ECCV'24)	55.9	46.0	49.1	70.1	51.6	35.5	1256	50.1	42.1	76.86%
PDrop (CVPR'25)	56.3	46.1	46.3	68.8	49.2	40.8	1505	48.0	36.6	76.41%
SparseVLM (ICML'25)	66.9	52.0	49.4	69.2	52.1	69.7	1561	58.3	49.6	88.60%
PruMerge+ (ICCV'25)	71.3	55.4	53.7	69.5	52.0	75.7	1549	59.6	52.1	92.22%
VisionZip (CVPR'25)	72.4	55.1	52.9	68.7	56.9	77.0	1690	62.1	57.0	94.46%
VisPruner (ICCV'25)	72.7	55.4	53.3	69.1	57.0	80.4	1650	62.7	57.3	95.07%
DivPrune (CVPR'25)	74.1	57.5	53.6	68.0	55.9	85.5	1615	61.5	56.6	95.02%
Ours	74.7	57.4	53.9	68.8	56.0	84.8	1715	61.4	55.8	**96.93%**
Retain 32 Tokens										
PruMerge+ (ICCV'25)	65.6	52.9	53.5	67.9	49.2	66.7	1550	55.1	45.9	87.01%
VisionZip (CVPR'25)	67.1	51.8	52.4	69.1	53.1	69.4	1579	57.0	50.3	89.41%
Vispruner (ICCV'25)	67.7	52.2	53	69.2	53.9	72.7	1538	58.4	52.7	90.75%
DivPrune (CVPR'25)	71.2	54.9	53.3	68.6	52.9	81.5	1594	57.6	49.1	92.16%
Ours	72.1	55.2	53.5	69.8	54.3	80.9	1656	61.2	53.2	**94.53%**

Notably, this performance trend is consistent with our findings in Section 4.1. This observation suggests that the effectiveness of a pruning strategy is closely linked to an image's characteristics, namely its feature diversity (measured by erank) and attention dispersion (measured by attention entropy). This applies whether the choice is between different methods (attention vs. diversity) or different thresholds (low vs. high τ).

These contrasting outcomes reveal the limitations of a fixed-threshold approach and indicate the need for a token pruning strategy that adaptively adjusts the threshold to the characteristics of each image. To this end, We introduce a simple heuristic where the similarity threshold τ is adapted based on the erank. In particular, τ is determined through a logarithmic mapping function:

$$\tau_{\text{adaptive}} = \alpha \ln(R) + \beta \tag{6}$$

where R is the erank, and α and β are scaling coefficients. The detailed coefficient values and implementation specifics are provided in Appendix A. In the following section, we analyze the performance of this adaptive thresholding strategy.

5 Experiments

Baselines and Models In the context of multimodal reasoning, MLLMs are often trained or enhanced with reinforcement learning. As a representative base model for such studies, we adopt LLaVA-1.5-7B, which serves as a standard foundation for multimodal reasoning research. Based on this architecture, we compare our approach with several vision token pruning techniques. The baselines include methods leveraging attention scores within the LLM (FastV [9], SparseVLM [10], PyramidDrop [8]), approaches utilizing attention scores from the vision encoder (VisionZip [13], VisPruner [14]), and Diversity-based strategies such as DivPrune [16]. To ensure consistent, fair, and reproducible evaluation, we fixed the

Method	Retain 64 Tokens				Retain 128 Tokens			
	$C_s\downarrow$	$C_i\downarrow$	Recall↑	Len	$C_s\downarrow$	$C_i\downarrow$	Recall↑	Len
LLaVA-1.5-7B	51.0	13.9	78.7	101.4	51.0	13.9	78.7	101.4
Attention-based methods								
FasterVLM (arXiv'24)	45.4	13.5	69.3	94.0	45.8	13.3	75.4	97.0
PruMerge+ (ICCV'25)	45.2	15.6	66.7	91.4	46.8	14.4	71.5	95.2
Vispruner (ICCV'25)	49.8	15.0	72.6	96.7	52.8	15.4	77.1	98.7
Diversity-based methods								
DivPrune (CVPR'25)	57.4	18.0	76.4	101.1	58.6	18.1	78.4	103.1
FPSPruner	58.6	18.6	76.0	100.5	59.4	18.8	81.1	104.1
Ours	**52.2**	**15.9**	**75.7**	**99.1**	**54.4**	**16.5**	**78.1**	**101.1**

Table 5: CHAIR evaluation (64 /128 tokens).

pretrained weights of LLaVA-1.5-7B and set the temperature to 0 across all experiments to produce deterministic outputs.

Datasets We conduct evaluations on a total of nine multimodal benchmarks. VQAv2 [22] and GQA [23] are large-scale visual question answering datasets that assess general vision-language understanding. VizWiz [24] and TextVQA [25] introduce more challenging scenarios involving accessibility-related queries and text recognition in images. ScienceQA [20] requires scientific knowledge for complex reasoning tasks. MME [18] provides a comprehensive metric for fine-grained multimodal understanding. Finally, MMBench and MMBench-CN [26] serve as multilingual benchmarks that evaluate overall performance across diverse tasks and languages. In addition, as introduced in Section 4.2, we further analyze the hallucination problem by using the CHAIR dataset [21] to quantitatively evaluate the occurrence of object hallucination in the model.

Main results We evaluate our adaptive thresholding approach on LLaVA-1.5, focusing on the effect of dynamic threshold adjustment on token diversity and downstream task performance. As shown in Table 5, our method consistently preserves accuracy under aggressive pruning. With 128 tokens, our method achieves competitive performance, showing modest gains of 1.6% over VisionZip and 1.3% over DivPrune. When reduced to 64 tokens, attention-based pruning methods suffer more than 25% degradation, whereas our method incurs only a 3.1% drop and achieves a slight performance edge over VisionZip and DivPrune by 2.2% and 1.9%, respectively. While recent hybrids such as VisPruner [14], are primarily attention-based approaches that modestly complement their selection with distant tokens for diversity, they remain non-adaptive and thus lack robustness, particularly on datasets where attention-based pruning is weak, such as POPE and MME datasets. The efficiency analysis is provided in Appendix B, and additional results on larger models are reported in Appendix C.

In summary, our empirical analysis reveals the distinct tendencies of existing pruning approaches and demonstrates that an adaptive approach is essential to balance information preservation and diversity. Building on these findings, we establish that adaptive thresholding provides a principled and effective alternative to fixed or non-adaptive methods.

Hallucination analysis. As shown in Table 5 and consistent with the observations in Section 4.2, our empirical analysis on the CHAIR dataset reveals clear contrasts between pruning strategies. Diversity-based methods generally achieve higher hallucination scores (C_S, C_I) with higher recall, whereas atteneion-based methods show the opposite trend. This trade-off is likely because tokens with high attention scores tend to contain more reliable visual information, which is essential for mitigating hallucination.

Further analysis indicates that methods primarily relying on token diversity tend to show weaker performance on overall benchmarks, as they do not incorporate attention scores and are less effective in capturing concentrated and reliable information. Conversely, attention-based methods preserve such concentrated tokens but lack diversity, which restricts their ability to handle questions involving multiple objects.

Building on these empirical findings, we propose an adaptive method that balances attention and diversity. This approach achieves 52.2 on C_S, 15.9 on C_I, and a recall of 75.7, which are close to the values obtained with the full set of visual tokens. These results emphasize findings from empirical analysis in clarifying the tendencies of different pruning strategies and provide evidence that adaptive approaches are needed to better align with varying image characteristics.

6 Conclusion

In this paper, we conducted an empirical study of visual token pruning in MLLMs, studying the influence of image complexity on the effectiveness of different pruning strategies. Our analysis with attention entropy and erank led to three main insights: (1) Attention-based pruning is more effective on simple images with concentrated information, whereas diversity-based pruning works better on complex images with distributed information. (2) On the CHAIR dataset, attention-based methods reduce hallucinations with more conservative answers, while diversity-based methods increase them with exploratory responses. We also find that increasing the proportion of high-attention tokens further reduces hallucinations.

Based on these findings, we propose an adaptive thresholding approach that adjusts pruning according to image complexity using erank. Experiments across nine benchmarks and CHAIR confirm that our

method maintains accuracy under aggressive pruning, mitigates hallucination, and approaches the performance of full-token models while cutting computational cost.

The adaptive token pruning principles explored in this work—balancing attention concentration with diversity under varying input complexity—were validated through experiments on MLLMs. These principles can also be extended to more advanced multimodal reasoning scenarios, including models trained with CoT supervision, enhanced through reinforcement learning techniques such as PPO or GRPO, or designed for complex reasoning tasks, where efficient token utilization remains essential for scaling to increasingly challenging reasoning problems.

Acknowledgments and Disclosure of Funding

This research was supported by the High-Performance Computing Support NPU Program of the National IT Industry Promotion Agency(NIPA) (N2025-0158), by LG Electronics, and by the Regional Innovation System & Education(RISE) program through the Institute for Regional Innovation System & Education in Busan Metropolitan City, funded by the Ministry of Education(MOE) and the Busan Metropolitan City, Republic of Korea (2025-RISE-02-004-12880001-02).

References

[1] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[2] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

[5] Pranav Agarwal, Aamer Abdul Rahman, Pierre-Luc St-Charles, Simon JD Prince, and Samira Ebrahimi Kahou. Transformers in reinforcement learning: a survey. *arXiv preprint arXiv:2307.05979*, 2023.

[6] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[7] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

[9] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

[10] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, 2025.

[11] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pages arXiv–2412, 2024.

[12] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *ICCV*, 2025.

[13] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802, 2025.

[14] Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *arXiv preprint arXiv:2412.01818*, 2024.

[15] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1773–1781, 2025.

[16] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025.

[17] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[18] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023.

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

[21] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019.

[24] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[27] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Algorithm 1 Pseudo Code for Adaptive Token Prunings

Image features F, attention scores A, erank r, attention entropy H, maximum tokens T Selected token set S
Normalize F and compute distance matrix D
$d_{len} \leftarrow f(H), \quad \tau \leftarrow g(r)$
Sort tokens by descending $A \rightarrow \pi$
$S \leftarrow \emptyset$, $M \leftarrow$ all True
for each token i in π **do**
  if $|S| \geq T$ **then**
    break
  end if
  if $M[i] =$ False **then**
    continue
  end if
  Add i to S
  if position$(i) < d_{len}$ **then**
    $\theta \leftarrow \tau$ {adaptive threshold}
  else
    $\theta \leftarrow$ base_threshold
  end if
  Find neighbors C with $D[i, j] \leq \theta$
  Suppress redundant tokens in C by updating M
end for
return S

A Proposed method detail

Algorithm 1 outlines the proposed adaptive token pruning strategy, which integrates attention-based importance with diversity-aware redundancy control. Given image features $F \in \mathbb{R}^{N \times D}$ and [CLS] token attention score $\alpha \in \mathbb{R}^N$ in the penultimate layer of the vision encoder, we first normalize features and compute the pairwise cosine distance matrix $D \in \mathbb{R}^{N \times N}$, defined as

$$D_{ij} = 1 - \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|}. \tag{7}$$

Our strategy is governed by two parameters that adapt to image characteristics: a threshold length d_{len} derived from attention entropy H, and a similarity threshold τ derived from erank r.

First, the similarity threshold τ controls the aggressiveness of redundancy pruning. A high effective rank (r) indicates that the token representations are already diverse. This implies we can safely employ a looser similarity criterion (a larger τ) to prune potential redundancies more aggressively without risking the loss of unique information. This behavior is modeled by a monotonic increasing function of r:

$$\tau(r) = \begin{cases} \tau_{\min}, & r \leq r_{\min}, \\ \alpha_\tau \ln(r) + \beta_\tau, & r_{\min} < r < r_{\max}, \\ \tau_{\max}, & r \geq r_{\max}. \end{cases} \tag{8}$$

Here, $\tau_{\min}$ and $\tau_{\max}$ are set to 80 and 100, respectively, while α is 0.32 and β is -1.36.

Second, d_{len} defines the number of top-ranked tokens to which the adaptive threshold τ is applied, while the remaining tokens are handled by a fixed base threshold. The rationale is that for high-attention entropy (H) images, information is broadly distributed, so the special significance of the highest-attention tokens is diminished. We therefore apply a more consistent pruning strategy across all tokens by reducing d_{len} as H increases. This is captured by a monotonic decreasing function:

$$d_{len}(H) = \begin{cases} d_{\max}, & H \leq H_{\min}, \\ \alpha_d \ln(C - H) + \beta_d, & H_{\min} < H < H_{\max}, \\ d_{\min}, & H \geq H_{\max}. \end{cases} \tag{9}$$

Function	POPE	SQA	MME
Linear	84.30	68.34	1391
Exponential	84.26	68.77	1388
Logarithmic	**84.76**	**68.80**	**1400**

Table 7: Ablation study on the functional form of $\tau(R)$. Logarithmic mapping consistently outperforms linear and exponential alternatives.

Here,C $H_{\min}$ and $H_{\max}$ are set to 5.1, 4.5 and 5.0, respectively, while α is 0.3 and β is 3.6.

Tokens are then sorted by attention scores in descending order, yielding a permutation $\pi \in \{1, \ldots, N\}^N$. For each selected token, the algorithm applies either the adaptive threshold τ (for the top d_{len} tokens) or a fixed base threshold to suppress neighboring tokens that are highly similar according to D. A binary mask $\mathcal{M} \in \{0, 1\}^N$ is updated at each step to ensure redundancy removal. The process terminates when the maximum number of tokens $T \in \mathbb{Z}$ is reached, and the algorithm returns the final selected set $\mathcal{S} \subseteq \{1, \ldots, N\}$ with $|\mathcal{S}| \leq T$.

A.1 Ablation study

In this section, we analyze the scoring function, a key component of our adaptive pruning framework, and present the empirical basis for our choice of a logarithmic function. An appropriate scoring function plays a crucial role in transforming the image complexity metric (erank) into a meaningful value that the model can effectively leverage.

To identify the optimal scoring technique, we transformed the image complexity metric using three representative functions—linear, exponential, and logarithmic—and compared their impact on model performance. As summarized in Table 8, our experimental results indicated that the logarithmic function achieved the best performance, followed by the linear and exponential approaches. This performance difference appears to stem from how each function handles the full spectrum of image complexity.

- A **linear function** tends to treat the difference between complexity scores of 10 and 20 the same as the difference between 100 and 110. However, for high-complexity images, it is often more ideal for the impact of score increases to diminish gradually. Therefore, a purely linear approach may not be optimal.
- An **exponential function**, on the other hand, can potentially amplify this issue. It risks exaggerating high complexity scores, which can generate outliers and lead to instability in the thresholding system, where a few complex images might dominate the outcome.
- In this context, a **logarithmic function** appears to be a more suitable approach for modeling this "diminishing returns" phenomenon. The logarithmic transformation tends to dampen the effect of high values, allowing for more stable thresholding, especially among images with very high complexity.

In summary, our experiments suggest that logarithmic scoring offers a more balanced approach. It provides sufficient sensitivity to distinguish between less complex images while mitigating the impact of extreme values from highly complex images, thereby ensuring greater system stability.

B Efficiency Analysis

Computation Overhead of Attention entropy and erank. The erank quantifies the representational complexity of a feature matrix $X \in \mathbb{R}^{N \times D}$ based on the distribution of its singular spectrum. To improve computational efficiency, we compute the covariance matrix across tokens using a fast $N \times N$ formulation. The definition is given as follows:

$$C = XX^\top, \quad S = \sqrt{\lambda(C)}, \quad p_i = \frac{S_i}{\sum_j S_j}, \quad \mathrm{erank}(X) = \exp\left(-\sum_i p_i \log p_i\right). \tag{10}$$

Method	Retain Tokens	FLOPs (T)	Latency (ms/sample)	GPU Memory (GB)	Accuracy
Vanilla (LLaVA-1.5-7B)	576	3.14	172	13.60	58.2
PDrop (CVPR'25)	64	0.51	128	13.30	55.0
SparseVLM (ICML'25)	64	0.52	129	16.26	55.2
DivPrune (CVPR'25)	64	0.48	110	13.30	55.8
VisPruner (ICCV'25)	64	0.48	115	13.30	55.4
Ours	64	0.48	115	13.30	**56.0**

Table 8: Efficiency and accuracy comparison **on single RTX 4090 at TextVQA dataset.** All models are evaluated under identical settings.

Here, C denotes the $N \times N$ covariance matrix, $\lambda(C)$ represents its eigenvalue spectrum, S corresponds to the square roots of eigenvalues (singular values), and p_i is the normalized spectral ratio. Thus, the effective rank approximates the intrinsic dimensionality by exponentiating the Shannon entropy of the normalized spectrum. Eq. (10) is mathematically equivalent to the SVD-based definition of the effective rank (Eq. 4), since the singular values of X correspond to the square roots of the eigenvalues of $XX^\top$.

To quantify the computational cost of adaptive pruning metrics, we estimate the FLOPs required to compute attention entropy and erank under the LLaVA-1.5-7B configuration with 576 visual tokens and a 4096-dimensional embedding. Following the fast formulation in Eq. (10), constructing the covariance matrix $C = XX^\top \in \mathbb{R}^{N \times N}$ requires approximately 1.36 GFLOPs, and the subsequent eigenvalue decomposition adds 0.19 GFLOPs, yielding a total of about 1.5 GFLOPs. This corresponds to roughly 0.05% of the full 3.14 TFLOPs inference cost of LLaVA-1.5-7B. In comparison, attention-entropy computation involves only simple logarithmic and summation operations, requiring about 0.02 GFLOPs, which accounts for approximately 0.0007% of the total inference cost.

Efficiency Comparison. As shown in Table 8, the proposed method reduces FLOPs by **89%** under the 64-token setting, while still preserving **96.2%** of the original performance compared to the vanilla LLaVA-1.5-7B model. Notably, our method outperforms in-LLM pruning approaches such as SparseVLM [10]and PyramodDrop [8]n terms of accuracy, achieving a better efficiency–performance trade-off. Meanwhile, when compared with recent pre-pruning approaches such as VisPruner [14] and DivPrune [16], the computational indicators (FLOPs, latency, GPU memory) remain nearly identical, while our method still attains the highest accuracy among them.

These three methods—VisPruner, DivPrune, and ours— share the property of performing pre-pruning before the LLM input, which substantially reduces the internal computation of the LLM. Compared to pruning inside intermediate layers of the LLM, pre-pruning provides a much stronger efficiency gain since the token reduction applies to all subsequent layers, yielding significant savings in FLOPs, memory, and latency with minimal overhead. In contrast, pruning at intermediate layers inside the LLM, as exemplified by methods such as SparseVLM and PyramodDROP, allows richer contextualization before tokens are removed and thus carries a lower risk of discarding important information, but its efficiency benefit is limited because the early layers still process the full set of tokens. Therefore, pre-pruning is preferable in terms of efficiency.

In addition, our method is fully compatible with FlashAttention [27], enabling further efficiency gains when combined with state-of-the-art acceleration techniques. Overall, these results demonstrate that our method strikes an effective balance between computational efficiency and accuracy.

C Additional Results

C.1 Evaluation on others model

In addition to LLaVA-1.5-7B, we also conducted experiments on larger models, namely LLaVA-1.5-13B (576 tokens) and LLaVA-NeXT-7B (2880 tokens). Across these settings, our method consistently demonstrated stable and strong performance, further validating the effectiveness of our approach (see Table 9 and Table 10).

Method	VQA^{v2}	GQA	VizWiz	SQA^{IMG}	TextVQA	POPE	MME	MMB	MMB^{CN}	Average
Vanilla 576 Tokens										
LLaVA-1.5-13B	80.0	63.3	53.6	72.8	61.2	86.0	1531	68.5	63.5	100%
Retain 128 Tokens										
FastV (ECCV'24)	75.3	58.3	54.6	74.2	58.6	75.5	1460	66.1	62.3	96.0%
PDrop (CVPR'25)	78.2	61.0	53.8	73.3	60.2	83.6	1489	67.5	62.8	98.4%
SparseVLM (ICML'25)	77.6	59.6	51.4	74.3	59.3	85.0	1488	68.4	62.6	97.8%
PruMerge+ (ICCV'25)	76.2	58.3	52.8	73.3	56.1	82.7	1446	66.3	61.2	95.8%
VisionZip (CVPR'25)	76.8	57.9	52.3	73.8	58.9	82.7	1450	67.4	62.5	96.7%
DivPrune (CVPR'25)	77.1	59.2	53.5	72.8	58.0	86.8	1458	66.3	60.7	97.0%
Ours	77.5	58.7	53.0	73.9	58.9	86.3	1480	67.6	62.1	**97.8**%
Retain 64 Tokens										
FastV (ECCV'24)	65.3	51.9	53.8	73.1	53.4	56.9	1246	59.2	55.1	85.8%
PDrop (CVPR'25)	70.8	54.1	50.5	73.1	55.3	66.1	1247	63.1	56.6	88.7%
SparseVLM (ICML'25)	73.2	55.9	52.1	73.0	57.1	77.9	1374	65.2	60.3	93.5%
PruMerge+ (ICCV'25)	72.6	56.3	52.4	73.5	54.4	75.7	1338	65.0	59.3	92.3%
VisionZip (CVPR'25)	73.7	56.2	53.2	74.2	57.4	75.7	1380	64.9	61.3	93.9%
DivPrune (CVPR'25)	75.2	57.9	54.4	71.7	57.4	84.5	1454	64.1	59.8	95.6%
Ours	75.9	57.8	54.4	72.2	58.5	81.8	1433	65.7	61.7	**96.0**%

Table 9: **Results of different token pruning methods on 9 multimodal benchmarks.** *Average* is normalized to the full-token **LLaVA-1.5-13B** (set to 100%). MME is reported in its original score units, and it is included only in the *Perception* section to enable broader comparison with existing methods.

C.2 Supplementary Examples of Image Complexity–Dependent Pruning Differences

As shown in Fig. 5, the qualitative patterns observed consistently reproduced across additional samples. For simple images (with low entropy and erank), attention-based pruning effectively captures the concentrated regions, while the additional benefits of diversity-based pruning are limited. Conversely, for complex images (with higher entropy and erank), diversity-based pruning ensures broader coverage, highlighting its strength in dispersed scenarios. These supplementary examples reinforce that image complexity is a key determinant of pruning effectiveness and motivate the need for an adaptive strategy that integrates both approaches.

C.3 More Examples on CHAIR

To further illustrate the differences between attention-based and diversity-based pruning in the image captioning task, we provide additional qualitative samples from the CHAIR dataset comparing FasterVLM as an attention-based method and DivPrune as a diversity-based method(Fig. 6 and Fig. 7). These cases illustrate how the two approaches differ in response style and hallucination tendency: DivPrune often yields broader and more descriptive captions but introduces hallucinated objects, whereas FasterVLM produces more conservative and focused descriptions.

Moreover, Fig.8 presents a controlled experiment where the token budget is fixed at 64 and the ratio between DivPrune- and attention-selected tokens (DivPrune-to-Attention ratio, R) is varied in steps of 25%. We observe that the hallucinated objects frequently generated under pure DivPrune ($R = 0$) gradually diminish as the share of attention-selected tokens increases, disappearing entirely when $R \geq 50\%$. In parallel, the response style evolves from speculative and exploratory to more factual and reliable as R increases, showing the stabilizing effect of attention-based token selection.

D Farthest Point Sampling

Farthest Point Sampling (FPS) is one of the simplest methods that guarantees diversity. Starting from an initial point, it iteratively selects the point that is farthest from the already chosen set by measuring the distance to the nearest selected point. For each point p_j, the minimum distance to the selected set S is defined as

$$d(p_j) = \min_{s \in S} \|p_j - s\|_2,$$

Method	VQA^{v2}	GQA	VizWiz	SQA^{IMG}	TextVQA	POPE	MME	MMB	MMB^{CN}	Average
Vanilla 2880 Tokens (Upper Bound)										
LLaVA-NeXT-7B	81.3	62.5	55.2	67.5	60.3	86.8	1512	65.8	57.3	100%
Retain 640 Tokens										
FastV (ECCV'24)	77.0	58.9	53.9	67.4	58.1	79.5	1412	63.1	53.5	95.2%
PDrop (CVPR'25)	79.1	60.0	53.8	66.7	57.8	83.8	1475	64.1	55.2	97.0%
SparseVLM (ICML'25)	79.2	61.2	53.6	67.6	59.7	85.3	1456	65.9	58.6	98.8%
PruMerge+ (ICCV'25)	78.2	60.8	57.9	67.8	54.9	85.3	1480	64.6	57.3	98.3%
VisionZip (CVPR'25)	79.1	61.2	57.1	68.1	59.9	86.0	1493	65.8	58.1	99.8%
DivPrune (CVPR'25)	79.3	61.9	55.7	67.8	57.0	86.9	1469	65.8	57.3	98.9%
Ours	79.3	61.9	56.3	69.0	59.7	86.3	1489	66.6	58.0	**100.0**%
Retain 320 Tokens										
FastV (ECCV'24)	61.5	49.8	51.3	66.6	52.2	49.5	1099	53.4	42.5	79.9%
PDrop (CVPR'25)	66.8	50.4	49.7	66.7	49.0	60.8	1171	55.5	44.7	82.5%
SparseVLM (ICML'25)	74.6	57.9	54.2	67.2	56.5	76.9	1386	63.1	56.7	94.6%
PruMerge+ (ICCV'25)	75.3	58.8	57.7	68.1	54.0	79.5	1444	63.0	55.6	95.7%
VisionZip (CVPR'25)	76.2	58.9	56.2	67.5	58.8	82.3	1397	63.3	55.6	96.4%
DivPrune (CVPR'25)	77.2	61.1	55.6	67.7	56.2	84.7	1423	63.9	55.7	97.0%
Ours	77.8	60.3	55.9	67.8	59.1	84.2	1453	65.5	57.5	**98.3**%

Table 10: **Results of different token pruning methods on 9 multimodal benchmarks.** *Average* is normalized to the full-token **LLaVA-Next-7B** (set to 100%). MME is reported in its original score units, and it is included only in the *Perception* section to enable broader comparison with existing methods.

and the next point is chosen as

$$p_{i_t} = \arg \max_{p_j \in P \setminus S} d(p_j).$$

Repeating this process until the desired number k is reached ensures that the selected points are evenly distributed across the data space, providing a more balanced representation than simple random sampling.

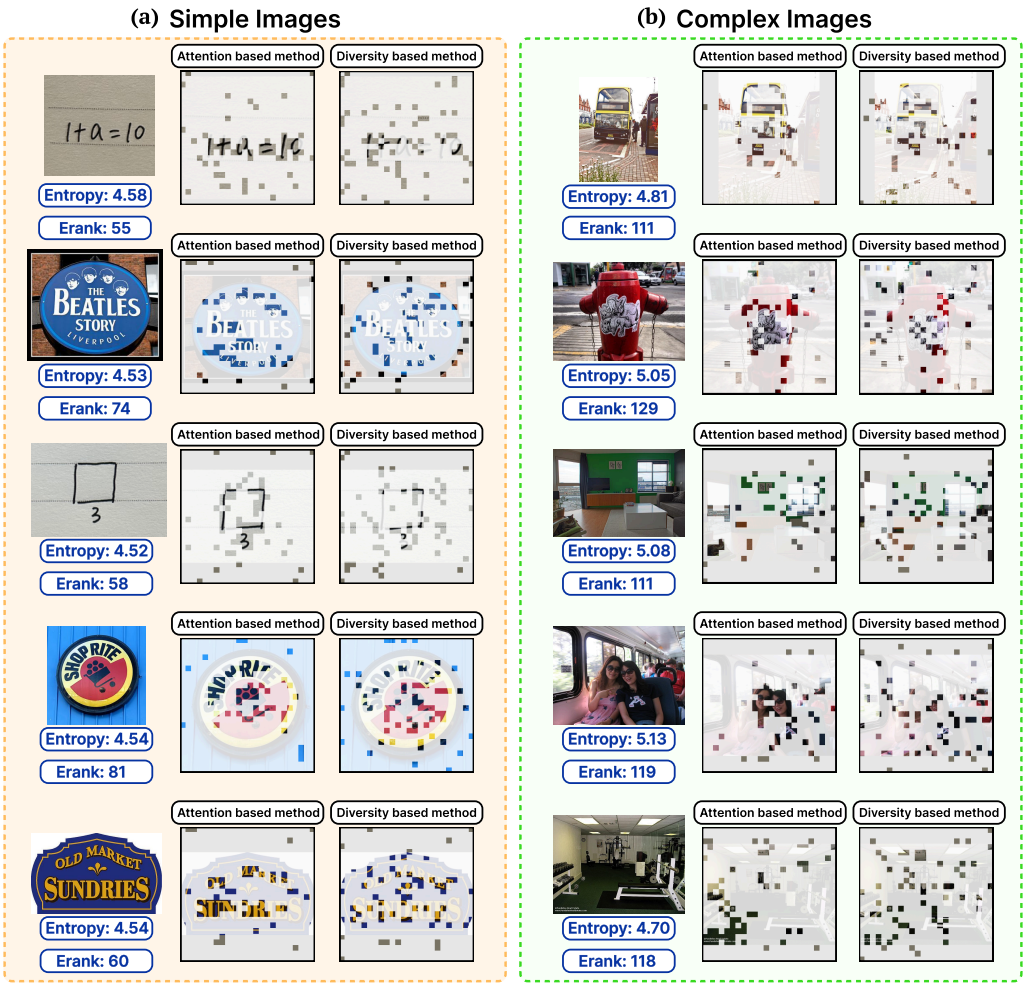


Figure 5: **Extended examples for simple vs. complex images.** The same trend as in Fig. 2 is observed: attention-based methods work well on simple images, while diversity-based methods cover complex images more broadly.

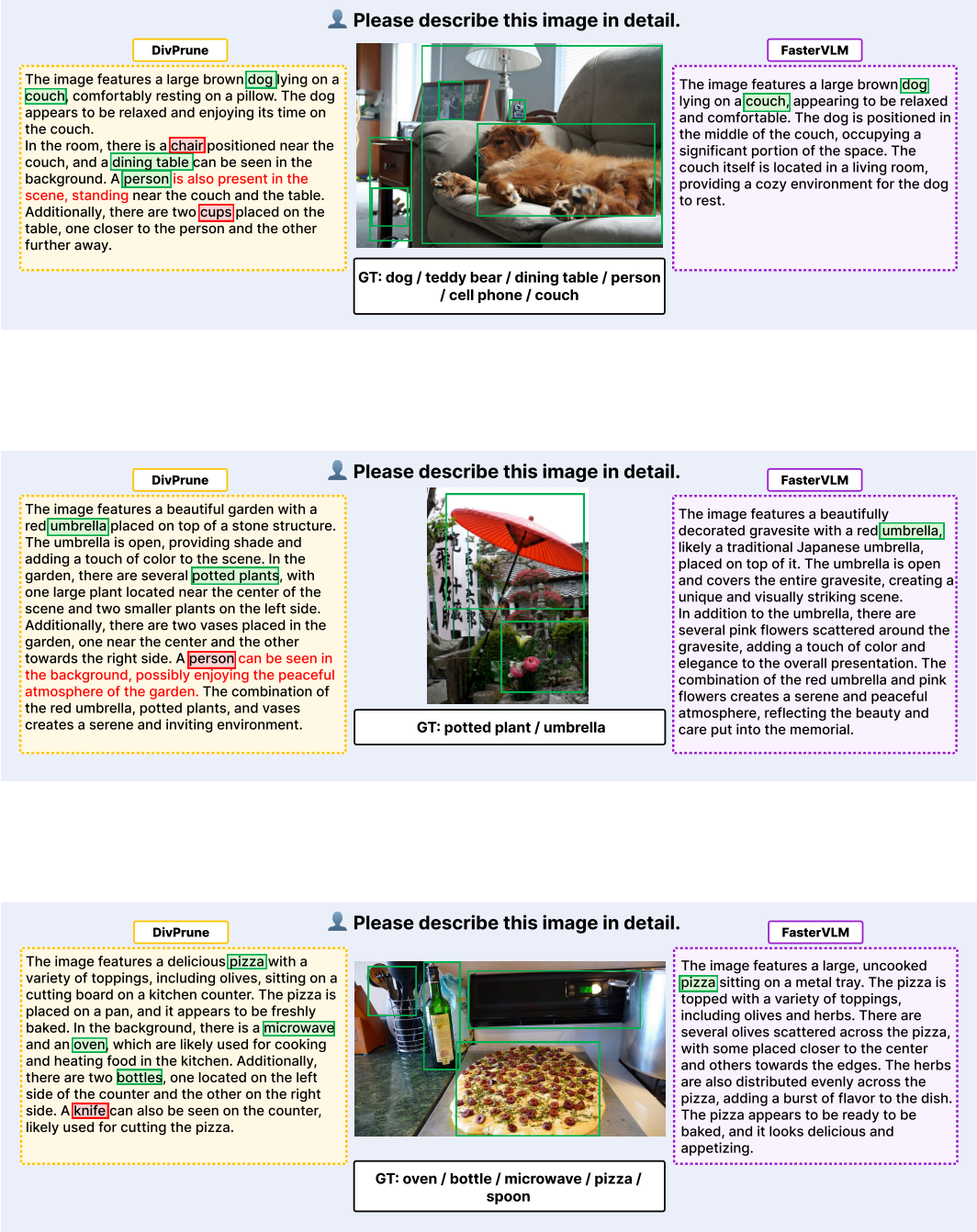
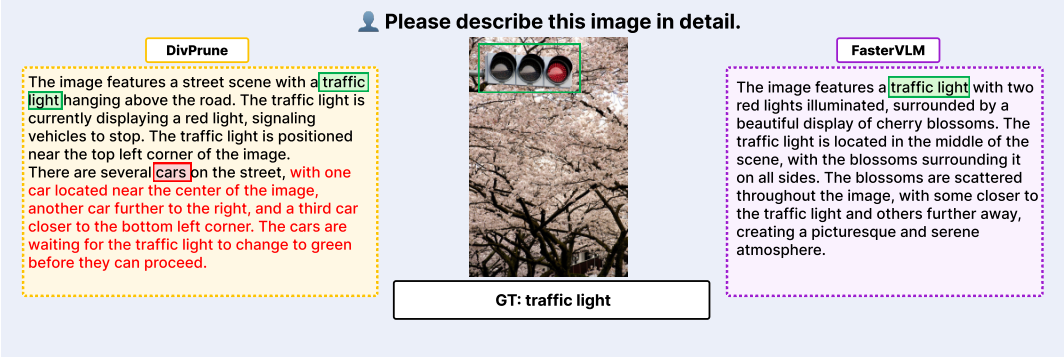


Figure 6: **CHAIR qualitative comparisons: FasterVLM vs. DivPrune (Set 1).** In the annotations, GT Obj. and Hallucinated Obj. label object words; **red text** indicates incorrect phrases.



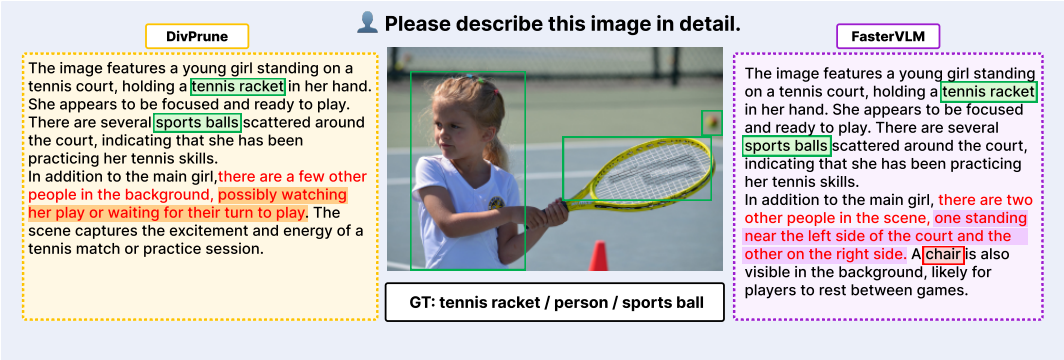


Figure 7: **CHAIR qualitative comparisons: FasterVLM vs. DivPrune (Set 2).** In the annotations, GT Obj. and Hallucinated Obj. label object words; marks DivPrune's phrasing; marks FasterVLM's phrasing; **red text** indicates incorrect phrases.

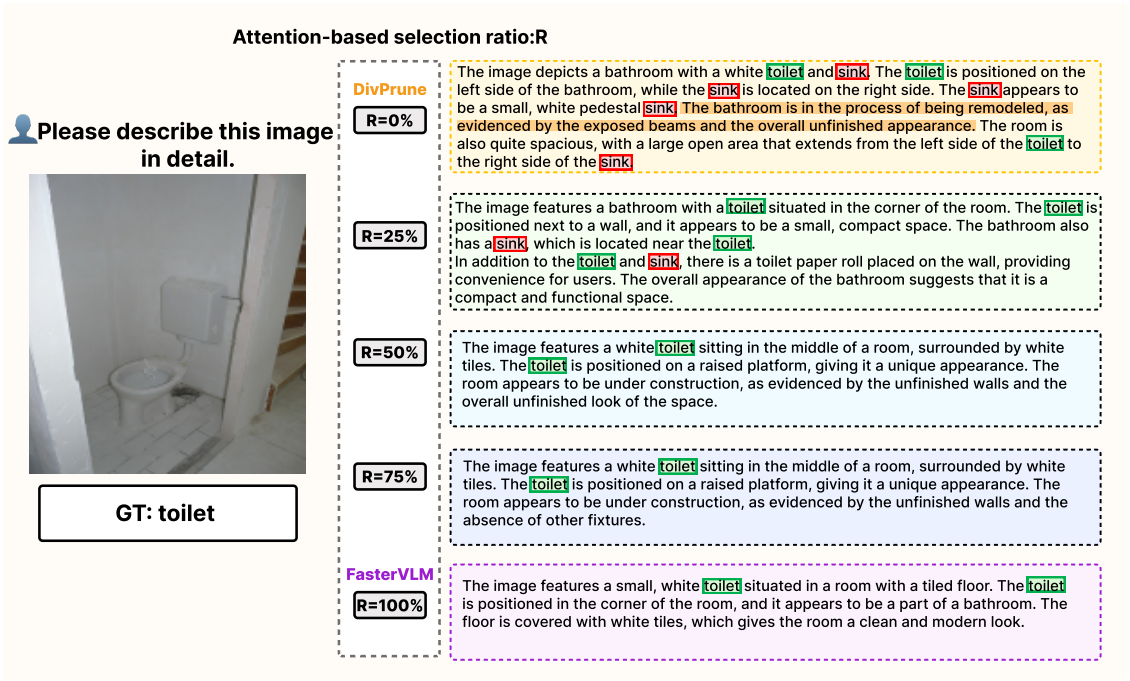


Figure 8: **Effect of varying the Attention-based selection ratio R under a 64-token budget.** As R increases, hallucinated objects produced by DivPrune are progressively suppressed, and the responses shift from exploratory to fact-oriented descriptions. In the annotations, GT obj. and Hallucinated obj. label object words; denotes DivPrune-specific phrasing.