# AGILEPRUNER：大型视觉语言模型中自适应视觉词元剪枝的注意力与多样性实证研究

**Changwoo Baek**[1*] **JouwonSong**[2*] **SohyeonKim**[1*] **Kyeongbo Kong**[1†]釜山国立大学，{higok18, shkim0503, kbkong}@pusan.ac.kr[2]LG电子，juwon05.song@lge.com

arXiv:2603.01236v1 [cs.CV] 1 Mar 2026

摘要

大型视觉语言模型（LVLMs）采用视觉词元剪枝策略，以减轻由大量视觉词元序列带来的巨大计算开销。尽管以往研究主要关注基于注意力或基于多样性的剪枝方法，但对这些方法的特性和局限性进行深入分析，仍鲜有探索。在本研究中，我们使用有效秩（erank）作为特征多样性的度量，并结合注意力分数熵，对视觉词元处理机制进行全面的实证分析，进而探究各类方法的优势与不足。我们的分析揭示了两点洞见：（1）基于有效秩的定量分析表明，许多面向多样性的剪枝方法所保留的特征多样性远低于预期；此外，对 CHAIR 数据集的分析发现，与基于注意力的剪枝相比，这些方法保留的多样性与幻觉频率升高密切相关。（2）我们进一步观察到，基于注意力的方法在视觉证据集中的简单图像上效果更好，而基于多样性的方法更擅长处理特征分布分散的复杂图像。基于这些实证洞见，我们表明，在现有混合剪枝策略中加入图像感知调整，能够持续提升其性能。我们还通过一种简单的自适应剪枝机制，对实证发现进行了最简实现；该机制在标准基准和幻觉专项评估中均取得了强劲且可靠的表现。我们的项目页面：https://cvsp-lab.github.io/AgilePruner。

## 1 引言

大型视觉语言模型（LVLMs）（Liu等人，2023；Wang等人，2024；Liu等人，2024a）通过整合图像、文本和视频等多种模态，实现了人类水平的视觉语言推理能力，因此备受关注。具体而言，视觉信息被编码为词元嵌入，供语言模型处理；在这一过程中，会生成数百个视觉词元。词元数量的增加使基于注意力的计算复杂度呈二次增长，显著影响推理速度和效率。

为了解决这些问题，许多研究者尝试通过词元剪枝方法移除不必要或冗余的视觉词元，以降低计算成本（Xing等人，2024；Chen等人，2024；Zhang 等，2025b）。现有方法通常采用两种主要策略。第一种是基于注意力的方法（Zhang 等，2024；Yang 等，2025；Arif 等，2025），这类方法将注意力分数较高的词元视为重要信息，并移除其余词元。第二种是基于多样性的方法（Alvar 等，2025），这类方法根据视觉词元之间的特征相似度来减少冗余。两种方法各有不同倾向。基于注意力的方法优先保留权重较高的词元，这可能导致选择结果集中，但有时会出现重复。相比之下，基于多样性的方法力求覆盖更广，

*共同贡献
†通讯作者

但往往会因此忽略重要词元。也有研究尝试通过混合方案结合这两种策略（Zhang 等，2025a；Shang 等，2025）。

尽管基于注意力、基于多样性以及混合词元剪枝策略相继出现，但人们对它们的实际表现仍缺乏充分了解。具体而言，（i）这些方法究竟保留了多少特征空间多样性，以及（ii）保留词元的属性如何影响大型视觉语言模型中的幻觉倾向，尚未得到系统考察。此外，（iii）不同图像类型是否天然适合采用基于注意力或基于多样性的剪枝策略，仍不明确。

为阐明这些问题，我们开展了两部分实证分析。首先，我们刻画现有剪枝范式的内在表现——使用有效秩（erank）（Roy 和 Vetterli，2007）量化保留的多样性，并考察其与不同图像类型下幻觉模式的关系。其次，我们分析剪枝效果如何随图像级复杂度变化，从而揭示各范式分别在何种情况下更为适用。

通过使用有效秩和注意力熵，我们的分析表明：

- **（方法层面的表现：多样性与幻觉）**许多具备多样性感知能力的剪枝方法所保留的多样性远低于预期。更重要的是，保留的多样性越高，幻觉频率（CHAIR（Rohrbach 等，2018））就越高；而基于注意力的剪枝保留多样性较低的词元集合，能够生成更保守的输出，并减少幻觉。
- **（图像级行为：依赖复杂度的偏好）**基于注意力的剪枝在简单图像上更为有效，因为关键信息集中在少量词元中；而基于多样性的剪枝在复杂图像上表现更优，因为语义信息分布广泛。

基于这些实证发现，我们接下来考察所观察到的行为能否转化为实际改进。通过将已识别的倾向应用于现有剪枝策略——包括混合方法以及基于注意力与基于多样性的方法的直接组合——我们引入了一种源自分析的简单图像感知调整。该调整在各项基准测试中持续提升性能，表明本研究揭示的实证模式稳健且具有广泛适用性。

我们进一步提出一种简单的基于阈值的剪枝过程，用以落实上述已识别的实证行为。该方法按注意力顺序降序探索词元，并依据相似度移除冗余token，同时使用由图像级复杂度指导的自适应设定阈值。尽管设计上刻意保持极简，这一实例在九个标准数据集上取得了强劲性能——往往与现有剪枝方法持平或超越——并且与我们的实证分析一致，它进一步缓解了幻觉倾向，这一点已在 CHAIR 基准上得到验证。这些结果表明，本研究揭示的实证原则不仅具有解释力，而且切实有效。此外，我们在规模更大、架构不同的大型视觉语言模型上观察到相同的行为趋势与改进，包括 LLaVA-1.5-13B、LLaVA-NeXT-7B 和 Qwen2.5-VL-7B，表明所发现的原则稳健且与模型无关。

总而言之，我们的贡献有三点：

- 我们首次基于有效秩刻画了现有剪枝方法如何保留特征多样性，以及这种保留的多样性与幻觉行为之间的关系。
- 我们揭示了基于注意力的剪枝与基于多样性的剪枝之间存在一种与图像复杂度相关的稳定偏好，并解释了各自范式何时有效、何时失效。
- 我们证明，这些经验原则具有实践指导价值：它们既能改进现有剪枝方法，也能指导我们提出一种简洁的自适应实例化方案，从而实现出色且稳定的性能。

## 2 相关工作

**大型视觉语言模型**大型语言模型的进步(大语言模型）（Bai 等，2023；Touvron 等，2023；Yu 等，2024；Cai 等，2024）已拓展至大型视觉语言模型（Liu等人，

2023；Wang等人，2024；Liu等人，2024a），如今已广泛应用于各个领域，以推理多种模态的信息。其中，专为视觉语言融合而设计的大型视觉语言模型尤其受到关注。典型的大型视觉语言模型架构由视觉编码器（Radford 等，2021；Tschannen 等，2025）、模态投影器和大语言模型组成。视觉编码器将输入图像转换为视觉词元，投影器则将这些词元映射至大语言模型的词嵌入空间，使大语言模型能够有效地理解和处理视觉信息。然而，视觉词元不仅数量庞大，而且高度冗余；在高分辨率图像或视频输入的情况下，这一倾向更加明显。再加上大语言模型解码器的自回归特性，推理效率会大幅下降。

**视觉词元缩减** 减少冗余且不必要的视觉词元，是降低计算量与内存占用、从而提升大型语言模型推理效率的有效途径。特别是，许多研究采用了无需额外训练的令牌剪枝方法，这些方法大体可分为以下几类。**（i）基于注意力的方法：**这些方法在词元输入大语言模型之前，利用视觉编码器倒数第二层的注意力分布对视觉词元进行剪枝（Zhang 等，2024；Shang 等，2025；Yang 等，2025；Zhang 等，2025a；Arif 等，2025）。基于视觉编码器中的图像信息往往集中于少量关键词元这一观察，这些方法利用输出层的注意力分数，选出数量有限、能够汇聚全局信息的词元。然而，它们往往保留集中于特定区域的相似词元，从而导致多样性不足，无法充分代表整个令牌集。**（ii）基于多样性的方法：**这些方法利用词元间相似度来增强所选令牌集合的多样性（Alvar 等，2025），从而促使选择更加多样化的词元。然而，它们会引入额外的计算开销，并存在丢弃重要词元的风险。近期已有研究尝试通过混合剪枝方案将这两种策略结合起来（Zhang 等，2025a；Shang 等，2025）。这些方法代表了预剪枝的主流策略，与此同时，也已有

人们越来越关注如何动态确定要保留的词元数量，近期研究如 ATP-LLaVA（Ye 等，2025）对此进行了探索。

## 3 预备知识

**视觉词元剪枝。** 由于视觉编码器通常会生成数百个词元，将所有词元传入大语言模型会产生高昂的计算成本。视觉词元剪枝通过移除冗余词元，仅保留 $K \ll N$ 词元来解决这一问题。形式上，剪枝定义了一个函数~ ~ ~$g$，使得 $V' = g(V, K)$，其中 $V'$ 表示保留的子集。在现有方法中，我们重点关注两类：（*i*）基于注意力的方法，根据 [CLS] 词元的注意力分数选择词元；以及（*ii*）基于多样性的方法，旨在减少冗余并保留特征多样性。这些策略为后续的实证分析奠定了基础。图 1 概述了我们的分析框架，该框架从两个角度评估视觉编码器：注意力集中和 token 嵌入多样性。

**通过注意力熵衡量注意力集中。** 为评估视觉编码器内的注意力集中程度，我们计算类别token注意力分数的香农熵。给定从倒数第二层获得的头平均注意力分数 $\alpha \in \mathbb{R}^N$，我们排除类别token的自注意力分数，并将剩余分数重新归一化为有效的概率分布：

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

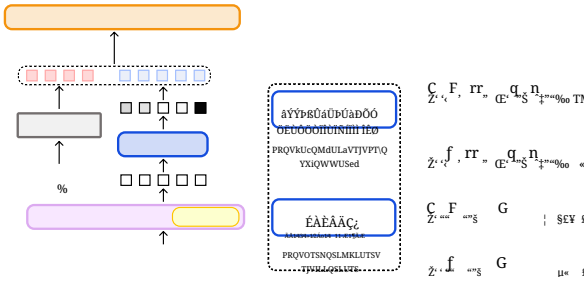


图1：注意力熵与有效秩概述。

随后将熵计算为

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

熵值 $H(p)$用于量化注意力在各词元之间的分布情况：值越低，表示类别token对少数区域的注意力越强；值越高，则表示注意力在多个视觉词元之间的分布越均匀

在本文后续部分，我们将这一指标称为**注意力熵**。

**通过有效秩衡量 token 嵌入多样性**。为了定量评估 token 嵌入的多样性，我们采用有效秩的概念（Roy & Vetterli,2007）。与传统的矩阵秩不同，有效秩是一种基于熵的度量，用于评估矩阵有效利用的维度数量。

给定一个 token 嵌入矩阵 $A \in \mathbb{R}^{N\times d_l}$，我们首先通过奇异值分解（SVD）得到其奇异值 $\{\sigma_i\}$。令

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

有效秩定义为

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

erank($A$) 的取值范围为 1 到 $L$。有效秩较低表示嵌入表示集中在少数几个主导维度上，而有效秩较高则表明嵌入空间在多个维度上的分布更加均匀。

## 4 实证研究

本节对基于注意力的和基于多样性的词元剪枝方法进行实证分析。我们重点关注两个方面：现有方法保留的多样性及其与幻觉的关系（第 4.1 节），以及图像复杂度对词元选择策略的影响（第 4.2 节）。基于这两项分析所得的见解，我们随后提出一种简单的自适应剪枝框架（第 4.3 节）。

### 4.1 基于注意力和基于多样性的剪枝实证分析

为了更好地理解现有剪枝范式的内在行为，我们使用有效秩开展实证研究，该指标能够捕捉词元级视觉特征的多样性。通过分析不同剪枝方法实际选出的词元集合中的这些信号，我们旨在刻画各种方法保留了多少语义多样性，以及多样性方面的差异最终如何影响模型行为，并通过幻觉分析揭示这一影响。

#### 4.1.1 基于有效秩分析现有剪枝范式中的多样性保持

**通过有效秩定量比较多样性保留**我们首先通过测量有效秩（见表 1），评估现有剪枝方法保留语义多样性的能力。这项比较揭示了各种方法之间存在明显的差异。DivPrune（Alvar 等，2025）的平均有效秩最高（21.84），体现了其明确旨在最大化词元间几何分散度的目标。与之相反，PruMerge+ （Shang 等，2025）的有效秩最低（10.91），表明其空间采样策略下的多样性有限。VisPruner（Zhang 等，2025a）（14.35）和 VisionZip（Yang 等，2025）（14.02）则处于中间水平，多样性相近——高于 PruMerge+，但仍低于 DivPrune。结果表明，即使令牌预算相同，不同的剪枝策略所保留的语义多样性程度也有显著差异。

| | PruMerge+ | VisionZip | VisPruner | DivPrune |
|---|---|---|---|---|
| 有效秩 | 10.91 | 14.02 | 14.35 | **21.84** |

表 1：POPE 上保留的 64 个词元的平均有效秩。

**不同方法中多样性机制的相对优势**进一步考察这些方法，可以更清楚地理解这一规律。尽管 VisPruner、VisionZip 和 PruMerge+ 均采用优先选择注意力分数较高词元的策略，但其辅助多样性机制的

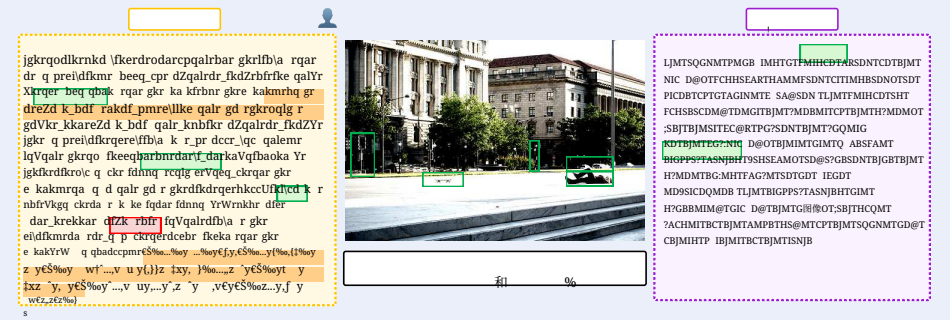


图 2：**DivPrune（基于多样性的）与 FasterVLM（基于注意力的）的响应模式**。DivPrune 的响应更为全面，但存在幻觉风险，而 FasterVLM 生成的描述更安全、更聚焦。在标注中，真实物体和幻觉物体标注物体词；标记 DivPrune 特有措辞；红色文本表示错误短语。

强度却大不相同。VisPruner 会先移除特征冗余词元，再进行注意力过滤，从而以更直接的方式保持多样性，因此在注意力驱动方法中取得了略高的有效秩。相比之下，VisionZip 的词元合并和 PruMerge+ 的空间采样只能带来有限的分散效果，导致整体多样性较低。尽管如此，这三种方法从根本上都受限于对注意力引导的词元选择的依赖，因此与 DivPrune 这类以最大化几何分散度为主要目标的方法相比，它们能够保留的语义多样性天然较低。

#### 4.1.2 剪枝方法与幻觉之间的关系

大型视觉语言模型中经常出现物体幻觉，这是损害其可靠性的关键问题。本节从幻觉的角度比较并分析基于注意力分数的方法和基于多样性的方法的特点，旨在确定这两种剪枝方法在诱发幻觉方面有何不同。为此，我们评估了 LLaVA-1.5-7B 模型中的物体幻觉，使用的数据集不仅包括以往令牌缩减研究中常用的数据集，还包括其他数据集，并展示相应的结果。

**物体幻觉**。为评估图像描述任务中的物体幻觉程度，我们采用 CHAIR 数据集。CHAIR 量化生成的描述中提及但未出现在真实标注中的物体所占比例，并提供两个子指标 $C_I$ 和 $C_S$，如公式 5 所定义。

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

这两个指标分别在实例级和句子级评估幻觉，数值越低表示性能越好。作为辅助指标，我们还报告召回率和长度，其中召回率表示生成的描述中提及的真实物体所占比例，长度表示生成的描述中的平均词数。

**在 CHAIR 数据集上的结果**。 表 2 展示了基于注意力的方法和基于多样性的方法在 CHAIR 数据集上的结果。尽管长度差异较小，基于多样性的方法的幻觉指标 $C_S$ 和$C_I$ 均高于基于注意力的方法，这表明选择特征相似度较低的多样化词元可能会增加产生幻觉的可能性。相比之下，基于多样性的方法取得了更高的召回率，说明它们能够捕捉到比基于注意力的方法更多的物体。

| 方法 | $C_S \downarrow$ | $C_I \downarrow$ | 召回率↑ | Len |
|---|---|---|---|---|
| **LLaVA-1.5-7B** | 51.0 | 13.9 | 78.7 | 101.4 |
| 基于注意力的方法 | | | | |
| 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 |
| 基于多样性的方法 | | | | |
| DivPrune（CVPR'25） | 57.4 | 18.0 | 76.4 | 101.1 |
| FPSPruner† | 58.6 | 18.6 | 76.0 | 100.5 |

表 2：**在 CHAIR 上的比较**。
†FPSPruner 基于最远点采样（FPS），通过迭代选择距离最远的词元来保证多样性。

**两种方法在响应模式方面的比较**。这两种方法在响应模式上也有所不同。图 2 展示了这一差异，表明基于多样性的方法

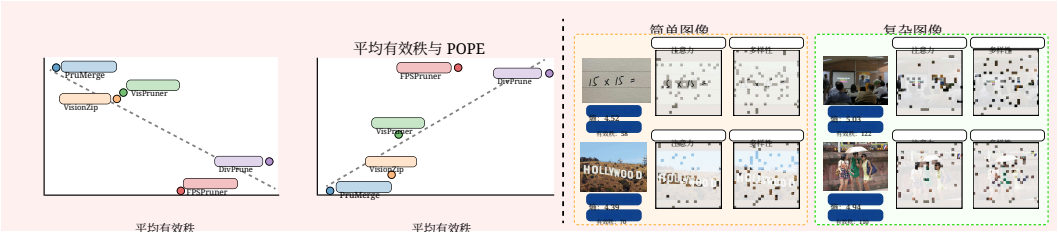



图 3：**不同数据集和图像复杂度下剪枝时多样性与注意力的对比**。(a) 有效秩较高的方法在复杂数据集（POPE）上表现更好，而有效秩较低的方法在简单数据集（ScienceQA）上表现出色。(b) 简单图像的熵和有效秩较低，因此集中注意力适合采用基于注意力的剪枝。复杂图像的熵和有效秩较高，此时基于多样性的剪枝更有效。

DivPrune 会生成内容更广泛、更加开放式的描述，并且常常包含推测性表达，如黄色标注所示。此外，基于多样性的响应提及了更多以绿色标出的真实物体，但同时也引入了以红色标出的幻觉物体，以及以红色文字强调的不准确表述。相比之下，基于注意力的方法 FasterVLM 聚焦于主要物体，并提供更为审慎、可靠的解释，从而抑制基于多样性的方法输出中经常出现的幻觉。

**基于注意力的选择对物体幻觉的影响**。基于注意力的词元选择往往会减少幻觉，我们据此开展了实验，通过调整基于多样性的选择与基于注意力的选择之间的平衡，定量评估基于注意力的选择对幻觉的影响。

| 方法 | $C_S \downarrow$ | $C_I \downarrow$ | 召回率↑ | Len | 平均有效秩 | 平均注意力 |
|---|---|---|---|---|---|---|
| 保留64个词元 | | | | | | |
| 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 |

表3：**基于注意力的选择比例的影响 $R$对 CHAIR 指标的影响**。更高的基于注意力的选择比例会减少幻觉（$C_S, C_I$），但会降低召回率。

如表3所示，我们将令牌预算固定为64，并逐步减少 DivPrune 选择的词元数量，同时用注意力分数更高的词元替换它们。实验结果表明，增加基于注意力的选择比例会使幻觉指标 $C_S$ 和 $C_I$逐渐降低。

这些发现表明，仅根据多样性选择词元可能会导致相对较高的幻觉，而注意力分数较高、能够集中关键信息的词元，对于生成可靠的描述和减轻幻觉起着关键作用。

### 4.2 分析剪枝策略的样本依赖效应

我们此前考察了不同剪枝方法在保留语义多样性方面的差异（以有效秩 erank 衡量），以及这些差异与幻觉等现象之间的关系。基于这些观察，我们现在探究这些特征如何影响模型在下游推理任务中的表现。

**多样性何时有益，何时基于注意力的选择更占优势**。我们的实验表明，基于多样性的剪枝和基于注意力的剪枝都并非在所有情况下占优。如图3 (a)所示，我们观察到明显的数据集依赖模式：保留更多样化词元（即有效秩 erank 更高）的方法，在 POPE 等数据集上往往表现更好；而基于注意力的方法在 ScienceQA 等数据集上准确率更高。受这些相反趋势的启发，我们进一步分析哪些类型的图像样本更适合基于多样性的方法，哪些更受益于基于注意力的词元选择。

**图像复杂度会影响注意力熵和多样性**。 我们首先分析了图像复杂度如何影响大型视觉语言模型。为此，我们使用注意力熵衡量注意力的集中程度，并使用有效秩评估令牌特征的多样性。我们在 LLaVA-v1.5-7B 上使用 MME基准（Yin等人，2024）开展分析，其中 光学字符识别、数值计算和文本翻译 等任务涉及背景单一且关键对象较少的图像，而 位置、场景和计数 等任务则涉及背景混杂且包含多

| 方法 | MME | | | ScienceQA |
|---|---|---|---|---|
| | OCR | 数值 Cal. | Text 翻译 | |
| **指标** | | | | |
| 注意力熵 | 4.61 | 4.47 | 4.39 | 4.45 |
| 有效秩 | 78 | 58 | 49 | 74 |
| **剪枝后 576 → 64 个词元的分数** | | | | |
| **基于注意力** | **140** | **55** | **100** | **69.51** |
| 基于多样性 | 130 | 40 | 80 | 67.53 |

(a) 在包含**简单图像**的数据集上的结果。

| 方法 | MME | | | POPE |
|---|---|---|---|---|
| | 位置 | 场景 | 计数 | |
| **指标** | | | | |
| 注意力熵 | 4.90 | 4.86 | 4.82 | 4.87 |
| 有效秩 | 109 | 103 | 102 | 106 |
| **剪枝后 576 → 64 个词元的分数** | | | | |
| 基于注意力 | 105 | 157 | 120 | 77.4 |
| **基于多样性** | **111** | **168** | **140** | **86.0** |

(b) 复杂图像数据集上的结果。

表4：**简单和复杂图像数据集上的注意力熵与有效秩**。 简单图像的熵和有效秩较低，而复杂图像的相应数值较高；两种剪枝方法在简单图像和复杂图像上的性能表现则截然不同。

| 方法 | GQA | $SQA^{IMG}$ | POPE | MME | Rel. |
|---|---|---|---|---|---|
| 全部*576*个令牌 | | | | | |
| LLaVA-1.5-7B | 61.9 | 69.5 | 85.9 | 1862 | 100.0% |
| 保留*128*个词元 | | | | | |
| BAT (CVPR'23) | 58.6 | 69.3 | 85.3 | 1737 | 96.75% |
| BAT + 逆向 | 58.4$_{-0.2}$ | 69.1$_{-0.2}$ | 85.0$_{-0.3}$ | 1734$_{-3}$ | 96.46%$_{-0.19}$ |
| **BAT + 我们的方法** | **58.8**$_{+0.2}$ | **69.4**$_{+0.1}$ | **86.8**$_{+0.9}$ | **1782**$_{+45}$ | **97.91%**$_{+1.11}$ |
| Vispruner (ICCV'25) | 58.2 | 69.0 | 84.6 | 1768 | 96.72% |
| VisPruner +逆向 | 57.9$_{-0.3}$ | 68.8$_{-0.2}$ | 85.3$_{+0.7}$ | 1744$_{-24}$ | 96.35%$_{-0.37}$ |
| **Vispruner +我们的方法** | **58.6**$_{+0.4}$ | **69.1**$_{+0.1}$ | **85.5**$_{+0.9}$ | **1787**$_{+19}$ | **97.32%**$_{+0.60}$ |
| 保留*64*个词元 | | | | | |
| BAT (CVPR'23) | 56.6 | 68.8 | 81.5 | 1683 | 93.91% |
| BAT + 逆向 | 55.8$_{-0.8}$ | 68.7$_{-0.1}$ | 79.7$_{-1.8}$ | 1602$_{-81}$ | 91.98%$_{-1.93}$ |
| **BAT + 我们的方法** | **56.9**$_{+0.3}$ | **69.1**$_{+0.3}$ | **83.5**$_{+2.0}$ | **1692**$_{+9}$ | **94.85%**$_{+0.94}$ |
| VisPruner (ICCV'25) | 55.4 | 69.1 | 80.4 | 1650 | 92.78% |
| VisPruner +逆向 | 55.2$_{-0.2}$ | 68.9$_{-0.2}$ | 78.0$_{-2.4}$ | 1643$_{-7}$ | 91.83%$_{-0.95}$ |
| **Vispruner +我们的方法** | **55.9**$_{+0.5}$ | **69.3**$_{+0.2}$ | **81.5**$_{+1.1}$ | **1671**$_{+21}$ | **93.76%**$_{+0.98}$ |

表5：结合基于多样性和基于注意力的剪枝方法的性能结果。

| 方法 | GQA | $SQA^{IMG}$ | POPE | MME | Rel. |
|---|---|---|---|---|---|
| 全部*576*个令牌 | | | | | |
| LLaVA-1.5-7B | 61.9 | 69.5 | 85.9 | 1862 | 100.0% |
| 保留*128*个词元 | | | | | |
| DivPrune（CVPR'25） | 59.4 | 68.6 | 87.0 | 1707 | 96.90% |
| FasterVLM（ArXiv'24） | 57.9 | 68.5 | 83.2 | 1757 | 95.80% |
| DivPrune + FasterVLM | 58.4 | 68.5 | 84.9 | 1768 | 96.66% |
| **Divprune + FasterVLM（自适应）** | 58.9 | 68.8 | 86.0 | 1787 | **97.55%** |
| 保留*64*个词元 | | | | | |
| Divprune（CVPR'25） | 57.5 | 68.0 | 85.5 | 1615 | 94.25% |
| FasterVLM（ArXiv'24） | 55.0 | 69.0 | 77.4 | 1665 | 91.91% |
| DivPrune + FasterVLM | 56.8 | 68.6 | 82.2 | 1681 | 94.08% |
| **DivPrune + FasterVLM（自适应）** | 57.2 | 68.9 | 83.5 | 1690 | **94.87%** |

表6：基于注意力与基于多样性的剪枝的自适应组合优于固定方案。

tiple物体，使其相对复杂。此外，我们纳入了来自外部基准测试的*ScienceQA*（简单）和*POPE*（复杂），以进行更全面的分析。事实上，如表4所示，简单图像表现出更低的注意力熵，例如*OCR*（4.61）的情况，其中视觉编码器能够轻易地将信息集中到少数主导区域。相比之下，复杂图像如*POPE*达到更高的值（4.87），表明注意力更分散地分布在多个区域。一致地，简单图像也表现出更低的有效秩，例如*OCR*（78），而复杂图像达到更高的值，例如*POPE*（106），反映出冗余与多样的令牌表示。

**按图像复杂度的性能分化**。 我们的分析表明，这两种方法的性能会因数据集特征而分化，如表4所示。因此，基于多样性的方法在高有效秩任务中表现更优，而基于注意力的方法在低秩任务中更为出色。这种性能逆转似乎由图像复杂度的差异所驱动。如图3(b)所示，简单图像使视觉编码器能够轻松地将注意力集中在特定区域，从而形成集中信息。在这种情况下，基于注意力的方法可以有效地选择这些集中的词元。相比之下，复杂图像包含多个物体和混合背景，导致信息分散在整幅图像中。在此情形下，能够捕获更广泛特征的基于多样性的方法变得更加有效。

### 4.3 从实证洞见到自适应词元剪枝

**基于分析改进剪枝方法**。 第 4.1 节和第 4.2 节揭示了两种一致的实证规律：(i) 不同的剪枝范式保留的特征多样性水平各异，并呈现出各自典型的幻觉行为；(ii) 图像复杂度系统性地决定了基于注意力还是多样性导向的选择更为合适。为考察这些发现是否具有实际应用价值，我们首先将其应用于现有的剪枝策略。VisPruner（Zhang 等，2025a）和 BAT（Long 等人，2023）等混合方法采用固定混合比例，而我们的方法将有效秩映射到线性加权函数：对于低有效秩（简单）图像，分配较大比例的基于注意力的词元；对于高有效秩（复杂）图像，则分配较大比例的基于多样性的词元。

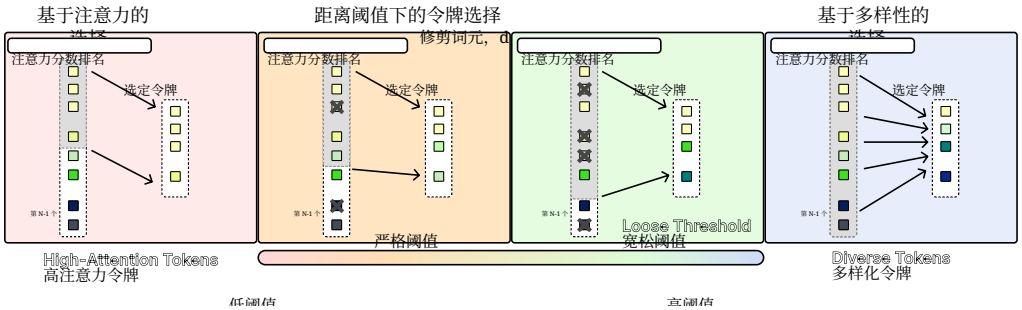



图 4：**相似度阈值对 $\tau$ 词元选择的影响**。较低（严格） $\tau$ 会优先选择高注意力词元，而较高（宽松） $\tau$ 会增加选定词元的多样性。

将这一自适应规则应用于 VisPruner 和 BAT 后，在 128 和 64 词元设置下均能持续提升准确率，如表5所示。相反，逆向自适应——即刻意为高有效秩图像赋予更多重要词元——会导致各基准测试上的性能明显下降。此外，如表6所示，我们还将同一自适应规则应用于一种纯基于注意力的方法与一种纯基于多样性的方法的成对组合：FasterVLM + DivPrune。即使在这种设置下，自适应规则的准确率仍高于固定的非自适应组合。

**实证指导的自适应相似度阈值处理**。我们进一步提出一种简单的基于阈值的剪枝流程，将上述实证发现付诸实践。我们的方法迭代地选择高注意力词元并剪除相似的邻近词元，从而根据所选阈值调节最终词元集的多样性。流程如下：

> 1. 按注意力分数降序排列所有词元。2. 从秩最高的词元开始，选中该词元，然后剪除所有与选定词元的余弦距离 $d$ 小于自适应阈值 $\tau_i$的其他候选词元。3. 接着处理下一个尚未被剪除且秩最高的词元，并重复剪枝步骤，直至选出所需数量的词元。

在这一框架中，阈值 $\tau$ 是直接决定最终令牌集多样性的参数。如图 4 所示，采用低（严格）阈值只会剪枝少量高度相似的词元。相反，高（宽松）阈值会移除范围更广的相似词元，从而构建出多样性更高的最终令牌集。附录B.4提供了详细的定量实验，验证了不同图像复杂度下 $\tau$ 与令牌多样性（通过有效秩衡量）及性能之间的这种相关性。

然而，最佳多样性水平取决于图像的内部特征，因此会产生截然不同的结果。对于有效秩和注意力熵较低（即信息集中）的图像，关键信息集中在高注意力令牌中，即使高度相似的词元也可能包含重要的细粒度细节。因此，采用高阈值（$\tau$）进行激进剪枝会降低性能，而保守地采用低阈值则效果更好。相比之下，对于有效秩和注意力熵较高（即信息分散）的图像，信息分布更广且存在冗余；较高的阈值能够有效消除冗余，并有助于选出多样性更高的令牌集，从而提升性能。

基于这些实证洞见，我们提出了一种统计驱动的自适应策略，可根据图像复杂度进行内在调整。为了将研究发现直接转化为稳健的机制，我们根据归一化后的图像复杂度定义动态阈值 $\tau_i$：

$$\tau_i = \text{order}_i \times \left( \frac{\text{erank}_{\text{input}}}{\text{erank}_{\text{avg}}} \times 0.01 \right) \tag{6}$$

其中，$\text{order}_i$ 表示词元的秩（从 1 开始），$\text{erank}_{\text{avg}}$ 表示 LLaVA 训练集的平均有效秩。为确保稳定性，最终阈值受统计上界$\tau_{max}$限制。该公式通过以下机制自然地体现了我们的发现：

| 方法 | 视觉问答 v2 | GQA | VizWiz | SQA$^{IMG}$ | TextVQA | POPE | MME | MMB | MMB$^{CN}$ | 平均值 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Vanila 576 个令牌* | | | | | | | | | | |
| LLaVA-1.5-7B | 78.5 | 61.9 | 50.1 | 69.5 | 58.2 | 85.9 | 1862 | 64.7 | 58.1 | 100.00% |
| 保留*128*个词元 | | | | | | | | | | |
| 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 | 69.0 | 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 | 59.4 | 53.0 | 68.6 | 57.0 | 87.4 | 1748 | 61.8 | 55.5 | **98.04%** |
| 保留*64*个词元 | | | | | | | | | | |
| 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 | 1640 | 59.6 | 52.1 | 92.76% |
| VisionZip (CVPR'25) | 72.4 | 55.1 | 52.9 | 68.7 | 55.5 | 77.0 | 1690 | 60.1 | 55.4 | 94.46% |
| VisPruner（ICCV'25） | 72.7 | 55.4 | 53.3 | 69.1 | 55.8 | 80.4 | 1650 | 61.3 | 55.1 | 95.07% |
| DivPrune（CVPR'25） | 74.1 | 57.5 | 53.6 | 68.0 | 54.5 | 85.5 | 1615 | 60.1 | 52.3 | 95.02% |
| **Ours** | 75.5 | 57.4 | 54.0 | 68.6 | 56.0 | 84.1 | 1703 | 60.7 | 55.8 | **96.76%** |
| 保留*32*个令牌 | | | | | | | | | | |
| 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 | 53.0 | 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** | 74.0 | 54.1 | 53.4 | 69.0 | 54.5 | 80.1 | 1603 | 60.4 | 53.6 | **94.02%** |

表 7：**9 个多模态基准测试上不同词元剪枝方法的结果**。平均值以全词元 **LLaVA-1.5-7B** 为基准进行归一化（设为 100%）。MME 使用原始分数单位报告。

- **复杂图像**($\text{有效秩}_{\text{输入}} > \text{有效秩}_{\text{平均值}}$)：较大的缩放因子会提高 $\tau_i$，从而增强剪枝效果并促进令牌多样性。
- **简单图像** ($\text{有效秩}_{\text{输入}} < \text{有效秩}_{\text{平均值}}$)：较小的因子会使 $\tau_i$ 保持较低，从而保留细粒度的高注意力令牌。

该公式通过将阈值建立在数据集统计之上，确保在不同图像复杂度下实现一致的自适应。

## 5 实验

**基线与模型**我们将该方法应用于广泛采用的开源模型LLaVA-1.5-7B（Liu等人，2024b），它是LLaMA系列的微调变体。基于该架构，我们将我们的方法与多种视觉词元剪枝技术进行比较。基线包括利用大语言模型内部注意力分数的方法（FastV（Chen等人，2024）、SparseVLM（Zhang 等人，2025b）、PyramidDrop（Xing等人，2024））、利用视觉编码器注意力分数的方法（VisionZip（Yang 等，2025）、VisPruner（Zhang 等，2025a）），以及基于多样性的策略，如DivPrune（Alvar 等，2025）。为确保一致、公平且可复现的评估，我们固定了LLaVA-1.5-7B的预训练权重，并在所有实验中将温度设为0，以产生确定性输出。
**数据集** 我们在总共九个多模态基准测试上进行评估。VQAv2（Goyal等人，2017年）和GQA（Hudson与Manning，2019年）是大规模视觉问答数据集，用于评估一般的视觉-语言理解能力。VizWiz（Gurari等人，2018年）和TextVQA（Singh等人，2019年）引入了更具挑战性的场景，涉及无障碍相关问题以及图像中的文本识别。ScienceQA（Lu等人，2022）要求模型具备科学知识，以完成复杂的推理任务。MME（Yin等人，2024）提供了一个用于细粒度多模态理解的综合指标。最后，MMBench和MMBench-CN（Liu等人，2024c）作为多语言基准测试，用于评估模型在不同任务和语言上的整体性能。此外，如第4.2节所述，我们进一步使用CHAIR 数据集（Rohrbach 等，2018）分析幻觉问题，以定量评估模型中出现物体幻觉的情况。

**主要结果** 我们在 LLaVA-1.5 上评估了自适应阈值方法，重点考察动态调整阈值对令牌多样性和下游任务性能的影响。如

表5所示，在激进剪枝的情况下，我们的方法始终能够保持准确率。在保留128个词元时，我们的方法取得了具有竞争力的性能，分别比 VisionZip 和 DivPrune 高出 0.85% 和 0.79%。当词元数减至64时，基于注意力的剪枝方法性能下降超过25%，而我们的方法仅下降3.24%，且性能略优于 VisionZip 和 DivPrune，分别高出2.2%和1.74%。尽管 VisPruner（Zhang 等，2025a）等近期混合方法主要采用基于注意力的方法，并适度补充远距离词元以提高多样性，但它们仍不具备自适应能力，因此缺乏鲁棒性，尤其是在基于注意力的剪枝效果较弱的数据集上，例如 POPE 和 MME 数据集。效率分析见附录A；在更广泛的大型视觉语言模型上的补充结果——包括 Qwen2.5-VL-7B、LLaVA-1.5-13B 和 LLaVA-NeXT-7B——见附录B.1。

总之，我们的实证分析揭示了现有剪枝方法的不同倾向，并表明自适应方法对于平衡信息保留与多样性至关重要。基于这些发现，我们确立了自适应阈值是固定或非自适应方法的一种有原则且有效的替代方案。

**幻觉分析**。如表8所示，且与第4.2节中的观察一致，我们对 CHAIR 数据集进行的实证分析揭示了不同剪枝策略之间的明显差异。基于多样性的方法通常能取得更高的幻觉分数（$C_S$, $C_I$）和更高的召回率，而基于注意力的方法则呈现相反趋势。这种权衡可能是因为注意力分数较高的词元往往包含更可靠的视觉信息，而这对于缓解幻觉至关重要。

| 方法 | 保留64个词元 | | | | 保留128个词元 | | | |
|---|---|---|---|---|---|---|---|---|
| | $C_s\downarrow$ | $C_i\downarrow$ | 召回率↑ | Len | $C_s\downarrow$ | $C_i\downarrow$ | 召回率↑ | Len |
| LLaVA-1.5-7B | 51.0 | 13.9 | 78.7 | 101.4 | 51.0 | 13.9 | 78.7 | 101.4 |
| 基于注意力的方法 | | | | | | | | |
| 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 |
| 基于多样性的方法 | | | | | | | | |
| 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** | | **75.7** | **99.1** | **54.4** | | **78.1** | **101.1** |

表8：CHAIR 评估（64 /128 个词元）。

进一步分析表明，主要依赖词元多样性的方法在整体基准测试中的性能往往较弱，因为它们没有纳入注意力分数，也不太善于捕捉集中且可靠的信息。相反，基于注意力的方法会保留这类集中的词元，但缺乏多样性，因此难以处理涉及多个物体的问题。

基于这些实证发现，我们提出了一种平衡注意力与多样性的自适应方法。该方法在 $C_S$ 上取得 52.2，在 $C_I$ 上取得 15.9，召回率为 75.7，接近使用完整视觉词元集时得到的数值。这些结果凸显了实证分析在阐明不同剪枝策略倾向方面的作用，并表明需要采用自适应方法，以更好地适应不同的图像特征。

## 6 结论

本研究对大型视觉语言模型中的视觉词元剪枝进行了系统性的实证研究。通过有效秩和注意力熵进行分析，我们发现了此前尚未被充分刻画的两种稳定行为模式：（i）不同剪枝方法保留的多样性存在显著差异，而这种保留的多样性与幻觉倾向密切相关；（ii）基于注意力的剪枝与基于多样性的剪枝，其有效性会随着图像复杂度呈现可预测的变化：简单图像更适合基于注意力的选择，而复杂图像则受益于基于多样性的保留。这些发现提供了统一的视角，解释了不同剪枝策略在何时以及为何会成功或失败。在这一实证认识的基础上，我们表明，将图像感知调整融入现有的混合剪枝方法，能够在各项基准测试中带来稳定提升，证明我们确定的实证原则具有广泛适用性，且与模型无关。我们还通过一种自适应阈值处理程序对这些原则进行了最简实现，该程序在显著降低计算成本的同时，实现了优异的性能并减少了幻觉。总而言之，本研究强调了理解大型视觉语言模型中剪枝范式底层行为的重要性，并提出了可指导未来自适应剪枝策略设计的实证原则。

致谢

本研究获得韩国国家研究基金会（NRF）资助，资助来源为教育部（编号：RS-2025-25433078）和韩国科学技术信息通信部（MSIT）（编号：RS-2024-00456152）。本研究还获得了LG电子以及国家IT产业振兴院（NIPA）高性能计算支持NPU计划的支持（编号：N2025-0158）。计算资源由釜山国立大学计算科学集群服务器提供。

参考文献

Saeed Ranjbar Alvar、Gursimran Singh、Mohammad Akbari 和 Yong Zhang。DivPrune：面向大型多模态模型的基于多样性的视觉词元剪枝。载于计算机视觉与模式识别会议论文集，第 9392–9401 页，2025 年。

Kazi Hasan Ibn Arif、JinYi Yoon、Dimitrios S Nikolopoulos、Hans Vandierendonck、DeepuJohn 和 Bo Ji。Hired：面向高分辨率视觉语言模型高效推理的注意力引导的词元丢弃。载于*AAAI*人工智能会议论文集，第 39 卷，第 1773–1781 页，2025 年。

Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang 等。通义千问技术报告。*arXiv* 预印本 *arXiv:2309.16609*，2023 年。

Zheng Cai、Maosong Cao、Haojiong Chen、Kai Chen、Keyu Chen、Xin Chen、Xun Chen、Zehui Chen、Zhi Chen、Pei Chu 等。InternLM2 技术报告。*arXiv* 预印本 *arXiv:2403.17297*，2024。

Liang Chen、Haozhe Zhao、Tianyu Liu、Shuai Bai、Junyang Lin、Chang Zhou 和 Baobao Chang。第 2 层之后，一张图像相当于 1/2 个词元：面向大型视觉语言模型的即插即用推理加速。载于欧洲计算机视觉会议，第 19–35 页。施普林格，2024。

Tri Dao、Dan Fu、Stefano Ermon、Atri Rudra 和 Christopher R´e。FlashAttention：具备 IO 感知能力的快速、节省内存的精确注意力。神经信息处理系统进展，35：16344–16359，2022。

Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra 和 Devi Parikh。让 VQA 中的 V 发挥作用：提升图像理解在视觉问答中的作用。载于计算机视觉与模式识别会议（*CVPR*），2017。

Danna Gurari、Qing Li、Abigale J. Stangl、Anhong Guo、Chi Lin、Kristen Grauman、Jiebo Luo 和 Jeffrey P. Bigham。Vizwiz 大挑战：为盲人回答视觉问题。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），2018年6月。

Dan Hendrycks 和 Thomas Dietterich。基准测试神经网络对常见损坏和扰动的鲁棒性。发表于国际学习表征会议论文集，2019年。

Drew A Hudson 和 Christopher D Manning。GQA：一个用于真实世界视觉推理和组合式问答的新数据集。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 6700–6709 页，2019年。

Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*，2024a。

Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。神经信息处理系统进展，36:34892–34916，2023。

Haotian Liu、Chunyuan Li、Yuheng Li和Yong Jae Lee。通过视觉指令微调改进的基线。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第26296–26306页，2024b。

Yuan Liu、HaodongDuan、Yuanhan Zhang、Bo Li、Songyang Zhang、WangboZhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu等。MMBench：你的多模态模型是全能选手吗？载于欧洲计算机视觉会议，第216–233页。施普林格，2024c。

Sifan Long、Zhen Zhao、Jimin Pi、Shengsheng Wang和Jingdong Wang。超越注意力令牌：融入令牌重要性与多样性，实现高效视觉变换器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10334–10343页，2023。

Pan Lu、Swaroop Mishra、Tanglin Xia、Liang Qiu、Kai-Wei Chang、Song-Chun Zhu、Oyvind Tafjord、Peter Clark和Ashwin Kalyan。学会解释：借助思维链进行多模态推理以完成科学问答。神经信息处理系统进展，35:2507–2521，2022。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark等。从自然语言监督中学习可迁移的视觉模型。载于国际机器学习会议，第8748–8763页。PmLR，2021。

Anna Rohrbach、Lisa Anne Hendricks、Kaylee Burns、Trevor Darrell 和 Kate Saenko。图像描述中的物体幻觉。第 4035–4045 页，2018年10月至11月。doi：10.18653/v1/ D18-1437。URL：https://aclanthology.org/D18-1437/。

Olivier Roy 和 Martin Vetterli。有效秩：一种有效维数的度量。载于第*20*届欧洲信号处理会议，*07.15*
第 606–610 页。IEEE，2007年。

Yuzhang Shang、Mu Cai、Bingxin Xu、Yong Jae Lee 和 Yan Yan。Llava-prumerge：面向高效大型多模态模型的自适应词元缩减。载于*ICCV*，2025。

Amanpreet Singh、Vivek Natarajan、Meet Shah、Yu Jiang、Xinlei Chen、Dhruv Batra、Devi Parikh 和 Marcus Rohrbach。迈向能够阅读的视觉问答模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8317–8326页，2019年。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar 等。《LLa MA：开放且高效的基础语言模型》。*arXiv*预印本 *arXiv:2302.13971*，2023。

Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muhammad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy、Talfan Evans、Lucas Beyer、Ye Xia、Basil Mustafa 等。《Siglip 2：具备改进的语义理解、定位和密集特征的多语言视觉语言编码器》。*arXiv*预印本 *arXiv:2502.14786*，2025。

Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。《Qwen2-vl：增强视觉语言模型在任意分辨率下对世界的感知》。*arXiv*预印本 *arXiv:2409.12191*，2024。

LongXing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu 等。Pyramiddrop：通过金字塔视觉冗余削减加速大型视觉语言模型。*arXiv* 预印本 *arXiv:2410.17247*，2024。

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu 和 Jiaya Jia。Visionzip：在视觉语言模型中，越长越好，但并非必需。载于计算机视觉与模式识别会议论文集，第 19792–19802 页，2025。

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang 和 Yansong Tang。ATP-LLaVA：面向大型视觉语言模型的自适应令牌剪枝。载于计算机视觉与模式识别会议论文集，第24972–24982页，2025。

Shukang Yin、Chaoyou Fu、Sirui Zhao、Ke Li、Xing Sun、Tong Xu 和 Enhong Chen。多模态大语言模型综述。《国家科学评论》，11(12):nwae403，2024。

Weihao Yu、Zhengyuan Yang、Linjie Li、Jianfeng Wang、Kevin Lin、Zicheng Liu、Xinchao Wang 和 Lijuan Wang。《Mm-vet：评估大型多模态模型的综合能力》。发表于第 *41* 届国际机器学习会议论文集，第 57730–57754 页，2024 年。

Qizhe Zhang、Aosong Cheng、Ming Lu、Zhiyong Zhuo、Minqi Wang、Jiajun Cao、Shaobo Guo、Qi She 和 Shanghang Zhang。 [cls] 注意力就是训练免视觉词元剪枝所需的一切：加快视觉语言模型推理速度。*arXiv* 电子预印本，第 arXiv–2412 页，2024 年。

Qizhe Zhang、Aosong Cheng、Ming Lu、Renrui Zhang、Zhiyong Zhuo、Jiajun Cao、Shaobo Guo、Qi She 和 Shanghang Zhang。超越文本-视觉注意力：利用视觉线索实现视觉语言模型的有效词元剪枝。载于*IEEE/CVF*国际计算机视觉会议论文集，第 20857–20867 页，2025a。

Yuan Zhang、Chun-Kai Fan、Junpeng Ma、Wenzhao Zheng、Tao Huang、Kuan Cheng、Denis Gu- dovskiy、Tomoyuki Okuno、Yohei Nakata、Kurt Keutzer 等。Sparsevlm：通过视觉词元稀疏化提升视觉语言模型推理效率。发表于国际机器学习会议，2025b。

附录目录


- **A 效率分析**
- **B 补充结果** – B.1 在其他模型上的评估– B.2 基于熵的自适应阈值设置– B.3 依赖图像复杂度的剪枝差异补充示例– B.4 相似度阈值对令牌多样性的影响– B.5 将分析扩展到自适应令牌数量
- **C 最远点采样**
- **D 注意力熵与有效秩尺度**
- **E 输入损坏下有效秩的鲁棒性分析**
- **F 细粒度推理任务中的令牌选择定性分析** – F.1 令牌选择行为及其对细粒度推理的影响– F.2 自适应方法在不同图像复杂度下进行细粒度推理时的失效模式
- **G CHAIR的更多示例**

| 方法 | 保留词元 | 浮点运算次数 (T) | 延迟（毫秒/样本） | GPU 内存（GB） | 准确率 |
|---|---|---|---|---|---|
| 原始（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** |

表9：效率与准确率对比**在单张RTX 4090上的TextVQA数据集**。 所有模型均在相同设置下进行评估。

| 批次大小 | 1 | 3 | 5 | 10 |
|---|---|---|---|---|
| 有效秩开销（毫秒） | 3.65 | 10.21 | 16.84 | 33.40 |

表10：不同批次大小下的有效秩计算时间，**在单张 RTX 4090 上**。

A 效率分析

**注意力熵和有效秩的计算开销**。有效秩根据奇异值谱的分布来量化特征矩阵 $X \in \mathbb{R}^{N\times D}$ 的表示复杂度。为提高计算效率，我们使用快速$N \times N$ 公式计算词元之间的协方差矩阵。定义如下：

$$C = XX^\top, \quad S = \sqrt{\lambda(C)}, \quad p_i = \frac{S_i}{\sum_j S_j}, \quad \mathrm{erank}(X) = \exp\left(-\sum_i p_i \log p_i\right). \tag{7}$$

其中， $C$ 表示 $N \times N$ 协方差矩阵， $\lambda(C)$表示其特征值谱， $S$ 表示这些特征值的平方根（即奇异值）， $p_i$ 表示归一化谱比率。因此，有效秩对应于归一化谱的香农熵的指数。由于 $X$ 的奇异值等于 $XX^\top$的特征值平方根，公式（7）与基于 SVD 的定义等价。

在典型的大型视觉语言模型设置中，词元数量远小于特征维度（$N \ll D$）。例如，LLaVA-7B-1.5 使用 $N$=576 个视觉词元和 $D$=4096 嵌入。这使得能够利用较小的协方差矩阵 $C \in \mathbb{R}^{N\times N}$高效地计算有效秩。

为评估这些指标的实际开销，我们将其理论复杂度与实测运行时间进行比较。直接对 $X \in \mathbb{R}^{N\times D}$ 应用 SVD 的复杂度为 $\mathcal{O}(ND^2)$，因为分解是在一个高矩阵 $N \times D$ 上执行的。相比之下，快速计算方法对较小的协方差矩阵 $C = XX^\top$求谱，从而在 $N \ll D$时将总体开销降至 $\mathcal{O}(N^2D + N^3)$。

我们还测量了 LLaVA-1.5-7B 配置下的单张图像运行时间。计算有效秩平均需要 3.4 毫秒，而完整的模型推理需要 115 毫秒。因此，**有效秩仅占总推理时间的~3.2%，在实际应用中开销仍然很小**。在批量推理过程中，如表10所示，开销大致随图像数量线性增长，因为协方差构造和特征值计算都是针对各个样本独立进行的。

**效率比较**。 如表9所示，在保留 64 个词元的设置下，所提出的方法将浮点运算次数减少了 **89%**，同时仍保留了原始 LLaVA-1.5-7B 模型 **96.2%** 的性能。值得注意的是，在准确率方面，我们的方法优于 SparseVLM（Zhang 等，2025b）和 PyramidDrop（Xing 等，2024）等大语言模型内部剪枝方法，实现了更优的效率与性能权衡。与此同时，与 VisPruner（Zhang 等，2025a）和 DivPrune（Alvar 等，2025）等近期预剪枝方法相比，我们的方法在计算指标（浮点运算次数、延迟、GPU 内存）几乎相同的情况下，仍取得了其中最高的准确率。

| 方法 | GQA | SQA$^{IMG}$ | POPE | MME | TextVQA | Rel. |
|---|---|---|---|---|---|---|
| 全部 *576*个令牌 | | | | | | |
| LLaVA-1.5-7B | 61.9 | 69.5 | 85.9 | 1862 | 58.2 | 100.0% |
| 保留*128*个词元 | | | | | | |
| 基于 Erank | 59.4 | 68.6 | 87.4 | 1748 | 57.0 | 98.13% |
| 基于注意力熵 | 59.4 | 68.6 | 87.0 | 1721 | 56.9 | 98.06 % |
| 保留*64*个词元 | | | | | | |
| 基于 Erank | 57.4 | 68.6 | 84.1 | 1703 | 56.0 | 95.97% |
| 基于注意力熵 | 57.7 | 68.7 | 83.2 | 1690 | 56.0 | 95.84% |

表 11：基于有效秩与基于熵的阈值设定。

这三种方法——VisPruner、DivPrune 和我们的方法——都具有在输入大语言模型之前进行预剪枝这一特点，从而大幅减少大语言模型的内部计算量。与在大语言模型中间层进行剪枝相比，预剪枝的效率提升显著更大，因为词元削减会作用于所有后续层，从而以极低的额外开销大幅节省浮点运算次数、内存和延迟。相比之下，以 SparseVLM 和 PyramodDrop 等方法为例，在大语言模型内部的中间层进行剪枝，可以在移除词元之前进行更充分的上下文建模，因此丢弃重要信息的风险较低；但由于较早的层仍需处理完整的词元集，其效率收益有限。因此，从效率角度来看，预剪枝更可取。
此外，我们的方法与 FlashAttention（Dao 等，2022）完全兼容，与先进的加速技术结合时还能进一步提升效率。总体而言，这些结果表明，我们的方法在计算效率与准确性之间实现了有效平衡。

## B 补充结果

### B.1 在其他模型上的评估

除 LLaVA-1.5-7B 外，我们还在规模更大、架构更多样的大型视觉语言模型上进行了实验，包括 LLaVA-1.5-13B（576个令牌）、LLaVA-NeXT-7B（2880个令牌）和 Qwen2.5-VL-7B。在所有这些设置下，我们的方法始终表现出稳定且优异的性能，进一步验证了我们方法的有效性和通用性。详细结果见表14、表15和表16。

### B.2 基于熵的自适应阈值设定

我们首先考察了注意力熵与有效秩（有效秩）之间的关系。如图 5 所示，在整个 MME 数据集上，这两个指标的皮尔逊相关系数为 0.63，呈中等相关性，表明二者都能捕捉视觉信息分散度的相关特征。熵反映了注意力权重的分布范围，而有效秩还通过其奇异值谱纳入了特征空间几何结构。
受这一相关性的启发，我们研究了熵能否在自适应阈值规则中直接替代有效秩。具体而言，我们在公式 6 中以注意力熵替代有效秩，并使用 LLaVA 训练集计算一次参考平均熵。

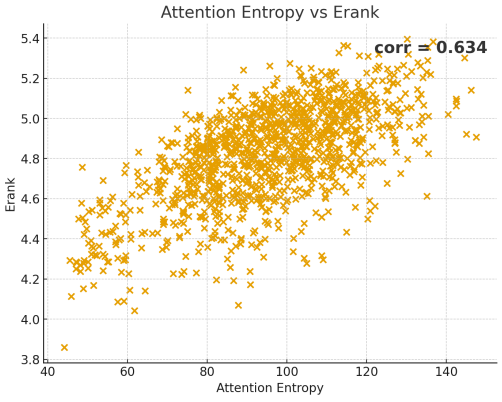



图 5：MME 数据集上的注意力熵与有效秩。

表 11 展示了多个基准测试在词元预算为 64 和 128 时的结果。基于熵的




图 6：**简单图像与复杂图像的扩展示例。** 观察到与图 3 相同的趋势：基于注意力的方法在简单图像上表现良好，而基于多样性的方法能更广泛地覆盖复杂图像。

自适应所呈现的性能趋势与基于 erank 的自适应非常接近，仅有约 0.13 的微小差异。

### B.3 图像复杂度影响剪枝差异的补充示例

如图 6 所示，在其他样本中也始终观察到相同的定性模式。对于简单图像（熵和有效秩较低），基于注意力的剪枝能够有效捕捉集中的区域，而基于多样性的剪枝带来的额外收益有限。相反，对于复杂图像（熵和有效秩较高），基于多样性的剪枝能够确保更广泛的覆盖，凸显其在分散场景中的优势。这些补充示例进一步表明，图像复杂度是决定剪枝效果的关键因素，并说明了采用整合两种方法的自适应策略的必要性。

### B.4 相似度阈值对令牌多样性的影响

我们将 $\tau$ 从 0 调整到 0.25，并观察其对令牌多样性的影响。如表 12 所示，结果表明，相似度阈值 $\tau$ 与

| 指标 | 相似度阈值 ($\tau$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | **0** | **0.01** | **0.05** | **0.1** | **0.15** | **0.20** | **0.25** |
| | **有效秩 / 性能** | | | | | | |
| | 高有效秩（复杂图像）数据集 | | | | | | |
| **MME（高）** | 13.85 / 1362 | 16.72 / 1358 | 17.44 / 1368 | 18.58 / 1379 | 19.55 / 1363 | 20.13 / **1381** | 19.60 / 1380 |
| **POPE** | 14.40 / 77.5 | 17.13 / 77.7 | 17.81 / 79.1 | 18.94 / 81.5 | 19.83 / 83.4 | 20.43 / 85.3 | 20.49 / **86.1** |
| | 低秩（简单图像）数据集 | | | | | | |
| **MME（低）** | 11.32 / 309 | 14.46 / **311** | 15.38 / 307 | 16.39 / 292 | 17.21 / 304 | 17.00 / 284 | 15.50 / 290 |
| **ScienceQA** | 11.91 / **69.1** | 15.29 / 68.9 | 16.22 / 68.9 | 17.33 / 68.2 | 18.09 / 68.3 | 18.45 / 68.7 | 16.50 / 68.5 |

表 12：不同相似度阈值下的性能比较。对于每项指标，粗体表示最佳性能。

| 方法 | 平均词元数 | GQA | $SQA^{IMG}$ | POPE | MME | MMBench | Rel. |
|---|---|---|---|---|---|---|---|
| LLaVA-1.5-7B | 576 | 61.9 | 69.5 | 85.9 | 1510 | 64.3 | 100.0% |
| ATP-LLaVA (CVPR'25) | 88 | 56.8 | 67.2 | 82.8 | 1401 | 64.7 | 95.1% |
| 我们的方法（固定计数） | 88 | 58.1 | 68.0 | 84.2 | 1405 | 62.3 | 95.4 % |
| **我们的方法（自适应计数）** | 85.5 | 58.4 | 68.2 | 85.2 | 1408 | 63.1 | **96.0%** |

表 13：自适应数量剪枝与固定数量剪枝的结果

所选令牌集合的多样性呈正相关。随着 $\tau$ 增大，所有数据集上选定词元的有效秩均持续上升。这表明，更高的阈值会使更多词元被视为冗余并被剪枝，进而提升最终集合的多样性。值得注意的是，当阈值接近零时，令牌多样性最低，因为在这种情况下几乎不会进行基于相似度的剪枝。

尽管增加 $\tau$ 能够持续提升令牌多样性，但其对性能的影响会因图像复杂度而异。对于有效秩更高的复杂图像，更大的 $\tau$ 有助于消除冗余token，并通过促进更多样化的令牌表示来提升性能。相比之下，对于复杂度较低的图像，信息内容集中在少量高注意力令牌中，使用较大的 $\tau$ 进行过于激进的剪枝可能会移除细粒度但重要的细节，从而导致性能下降。

### B.5 将我们的分析扩展至自适应词元数量

近期方法——包括 ATP-LLaVA（Ye 等人，2025）——积极探索了自适应词元数量策略，根据每个输入调整保留的视觉词元数量。我们的实证分析自然延伸至这一方向。高有效秩（复杂）图像包含更分散的视觉信息，因此适合保留更多样化的词元；而低有效秩（简单）图像的信息更为集中，选择较少且重点突出的词元即可更好地表示。基于这一分析，我们设计了一种自适应数量策略，为复杂图像保留更多词元，并对简单图像进行更激进的剪枝。为便于与 ATP-LLaVA 比较，我们采用相同的 88 个词元参考预算：低有效秩图像的词元数量最多减少 20%，高有效秩图像则按比例增加。由于不同输入保留的词元数量各异，我们报告数据集级平均值。如表 13 所示，自适应数量变体平均保留约 85.5 个词元，在 GQA、SQA、POPE、MME 和 MMBench 上的表现均优于 ATP-LLaVA 和我们的固定数量基线，同时也略微提升了效率。这些结果证实，调整词元数量与我们基于有效秩的分析自然契合，并能根据图像复杂度更有效地分配计算预算。

| 方法 | 视觉问答 v2 | GQA | VizWiz SQA$^{IMG}$ | TextVQA | POPE | MME | MMB | MMB$^{CN}$ | 平均值 |
|---|---|---|---|---|---|---|---|---|---|
| 原始 *576* 个词元 | | | | | | | | | |
| LLaVA-1.5-13B | 80.0 | 63.3 | 53.6 | 72.8 | 61.2 | 86.0 | 1531 | 68.5 | 63.5 | 100% |
| 保留*128*个词元 | | | | | | | | | |
| 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 | 59.1 | 52.5 | 72.8 | 58.9 | 86.9 | 1481 | 67.6 | 61.9 | **97.6%** |
| 保留*64*个词元 | | | | | | | | | |
| 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 | 75.7 | 57.5 | 54.2 | 72.0 | 58.6 | 82.0 | 1437 | 66.2 | 61.6 | **96.0%** |

表 14：**9 个多模态基准测试上不同词元剪枝方法的结果**。平均值以全词元 **LLaVA-1.5-13B** 为基准进行归一化（设为 100%）。MME 按原始分数单位报告，且仅纳入感知部分，以便与现有方法进行更广泛的比较。

| 方法 | 视觉问答 v2 | GQA | VizWiz SQA$^{IMG}$ | TextVQA | POPE | MME | MMB | MMB$^{CN}$ | 平均值 |
|---|---|---|---|---|---|---|---|---|---|
| 原始 *2880* 词元（上界） | | | | | | | | | |
| LLaVA-NeXT-7B | 81.3 | 62.5 | 55.2 | 67.5 | 60.3 | 86.8 | 1512 | 65.8 | 57.3 | 100% |
| 保留 *640* 词元 | | | | | | | | | |
| 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 | 62.0 | 56.0 | 67.8 | 59.0 | 86.1 | 1502 | 65.9 | 58.3 | **99.64%** |
| 保留 *320* 个词元 | | | | | | | | | |
| 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 | 60.1 | 55.8 | 67.3 | 58.4 | 84 | 1475 | 64.5 | 57.1 | **97.94%** |

表 15：**9 个多模态基准测试上不同词元剪枝方法的结果**。平均值以全词元 **LLaVA-NeXT-7B** 为基准进行归一化（设为 100%）。MME 以其原始分数单位报告，且仅列于感知部分，以便与现有方法进行更广泛的比较。

## C 最远点采样

最远点采样（FPS）是能够保证多样性的最简单方法之一。它从一个初始点开始，通过测量每个点到已选点集中最近点的距离，迭代地选出距离该集合最远的点。对于每个点 $p_j$，其到已选点集 $S$ 的最小距离定义为

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

下一个点选择为

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

| 方法 | TextVQA | 图表问答 | AI2D | OCRBench | MME（感知） | MME（认知） | MMB-EN | MMB-CN | Rel. |
|---|---|---|---|---|---|---|---|---|---|
| 原始 *1225* 词元（100%） | | | | | | | | | |
| Qwen2.5-VL-7B | 82.1 | 77.5 | 83.0 | 84.1 | 1691 | 642 | 83.2 | 83.2 | 100.0% |
| 保留 *512* 个词元（削减 60.5%） | | | | | | | | | |
| DivPrune(CVPR25) | 79.7 | 72.0 | 81.9 | 78.5 | 1704 | 620 | 81.8 | 81.0 | 96.8% |
| **Ours** | 79.8 | 71.5 | 82.1 | 78.5 | 1700 | 630 | 81.3 | 80.8 | **96.9%** |
| 保留 *256* 个词元（削减 79.1%） | | | | | | | | | |
| DivPrune(CVPR25) | 75.2 | 62.9 | 79.7 | 66.4 | 1703 | 542 | 80.3 | 79.8 | 90.7% |
| **Ours** | 74.2 | 62.4 | 80.5 | 65.8 | 1670 | 630 | 80.9 | 80.3 | **92.1%** |

表 16：**在不同词元保留比例下，Qwen2.5-VL-7B 的性能比较**。相关性是相对于全词元 Qwen2.5-VL-7B 的归一化相对性能（设为 100%）。MME 分数分别报告为感知（Perc.）和认知（Cogn.）。

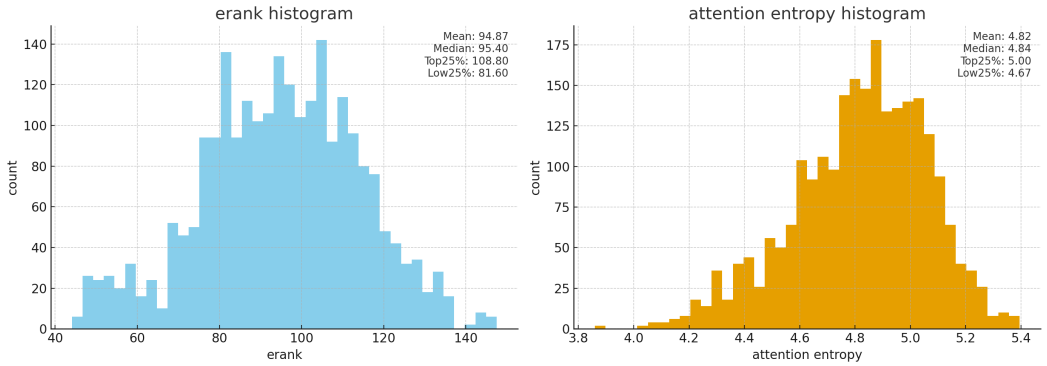



图 7：在 MME 数据集上测量的有效秩和注意力熵直方图。

重复此过程，直到达到所需数量 $k$，即可确保选定点均匀分布在数据空间中，相较于简单的随机采样，能够提供更加均衡的表示。

## D 注意力熵与有效秩的尺度

熵和有效秩表现出截然不同的缩放行为。如图 7 所示，从 CLIP-L（使用 576 个视觉词元）计算得到的熵集中在一个非常狭窄的范围内。具体而言，所有样本的平均熵为 4.80，中位数为 4.78，上四分位数（Q3）为 4.96，下四分位数（Q1）为 4.63。换句话说，熵值主要在约 4.0–5.4 的区间内变化。因此，即使看似很小的差异（例如 0.3–0.5），在这一受限尺度内也代表着具有实际意义的相对变化。

相比之下，有效秩的取值范围宽得多。平均有效秩为 94.87，中位数为 95.40，上四分位数为 108.80，下四分位数为 81.59。这种较大的离散程度表明，有效秩能够捕捉词元相关结构和谱离散度中更粗粒度的变化，与熵的细粒度敏感性形成互补。

### 输入损坏下有效秩的鲁棒性分析

**有效秩对噪声或退化输入的鲁棒性** 为评估有效秩指标在视觉条件退化下的稳定性，我们遵循 COCO-C（Hendrycks 和 Dietterich，2019年）的损坏协议进行了敏感性分析。我们将十五种损坏类型应用于 MME 数据集中的全部 2,374 张图像，并将所得有效秩值与在干净图像上计算的值进行比较。我们使用了两种损坏严重程度（1 和 3）（见图 8），分别对应鲁棒性研究中常考察的轻度与中度扰动强度。

| 损坏类型 | 严重程度 = 1 | | 严重程度 = 3 | |
|---|---|---|---|---|
| | 平均有效秩差 | 相对变化 | 平均有效秩差 | 相对变化 |
| 缩放模糊 | 5.85 | 6.14% | 7.62 | 8.12% |
| 亮度 | 1.48 | 1.56% | 2.52 | 2.70% |
| 对比度 | 2.38 | 2.54% | 4.38 | 4.62% |
| 散焦模糊 | 2.06 | 2.28% | 3.74 | 4.16% |
| 弹性变换 | 2.18 | 2.54% | 3.38 | 4.05% |
| Fog | 3.24 | 3.73% | 4.05 | 4.71% |
| 霜冻 | 3.96 | 4.56% | 6.10 | 7.24% |
| 高斯噪声 | 2.83 | 3.23% | 4.17 | 4.73% |
| 脉冲噪声 | 3.34 | 4.04% | 4.37 | 5.10% |
| JPEG 压缩 | 2.31 | 2.48% | 2.77 | 2.98% |
| 运动模糊 | 2.11 | 2.33% | 3.38 | 3.91% |
| 像素化 | 1.22 | 1.32% | 1.61 | 1.76% |
| 散粒噪声 | 2.74 | 3.14% | 4.06 | 4.62% |
| Snow | 3.84 | 4.17% | 5.37 | 5.76% |
| 眼镜模糊 | 2.15 | 2.46% | 4.07 | 4.79% |
| **平均值** | **2.78** | **3.10%** | **4.11** | **4.62%** |

表 17：在 MME 数据集上评估的有效秩对 COCO-C 中 15 种损坏类型的敏感性。




(a) 严重程度 = 1　　(b) 严重程度 = 3

图 8：将 COCO-C 的 15 种损坏类型以严重程度 1 和 3 应用于 MME 图像的可视化。

表 17 总结了有效秩的平均绝对偏差以及相对于干净图像均值的相应相对变化。总体而言，有效秩在所有损坏类型上均表现出高度稳定性。严重程度 1 时的平均偏差为 2.78，严重程度 3 时为 4.11，仅占有效秩自然变化的一小部分。作为参考，干净 MME 数据集的有效秩取值范围很广（均值 94.86，最小值 44.08，最大值 147.49，标准差 19.46）。

改变图像全局空间结构的损坏（如缩放模糊、霜冻、雪和弹性变换）会产生略大的偏差（通常为 4–7 个百分点），这反映了它们对奇异值谱的更广泛影响。相比之下，仅改变局部像素级外观的失真，包括亮度变化、像素化和 JPEG 压缩，影响很小（1–2.5 个百分点）。将损坏严重程度从 1 提高到 3，偏差仅略有增加，这表明即使在更严重的退化情况下，有效秩仍保持稳定。

# F 细粒度推理任务中的令牌选择定性分析

## F.1 词元选择行为及其对细粒度推理的影响

基于注意力和基于多样性的剪枝所选词元的特征，不仅会影响幻觉生成，也会影响模型整体的推理行为。如图 9 的示例所示，第 4.1 节中发现的趋势，在计数和空间关系推断等更细粒度的推理任务中也同样存在。因此，我们通过定性分析比较基于多样性的、基于注意力的和自适应剪枝方法如何选择词元，以及这些选择如何影响它们生成的开放式回答。

示例 1 和 4 表明，基于注意力的剪枝由于过于聚焦于特定区域，可能会忽略某些物体并给出错误答案。相比之下，基于多样性的剪枝保留了更广泛的空间信息，能够捕捉更多物体，并且往往能提供更详细的描述，因此在这类情况下预测更加准确。然而，在示例 2 和 5 中，情况正好相反：基于注意力的剪枝通过聚焦于主要实例，正确预测了物体数量；而基于多样性的剪枝由于词元选择较为分散，错误地推断出还有其他物体，从而产生幻觉。自适应方法的表现介于两者之间。尽管它在示例 1 和 4 中因聚焦区域比基于多样性的方法更窄而未能成功，但这三种方法在这些特别具有挑战性的案例中都表现不佳。相反，在示例 2 和 5 中，自适应方法与基于注意力的方法一样，正确预测了数量，而基于多样性的方法则幻觉出其他物体。

空间推理中也出现了类似的模式。如示例 3 和 6 所示，基于多样性的方法利用了更广泛的空间线索，并生成了更丰富的关系描述，但这也增加了推断实际并不存在的关系的可能性。相比之下，基于注意力的方法能够提供稳定的、以物体为中心的关系描述，但与基于多样性的方法相比，覆盖范围更有限。自适应方法则再次介于两者之间，既能捕捉关键对象，又能减少基于多样性的方法引入的幻觉，最终实现更稳定的空间推理。

为进一步考察自适应方法是否会忽略小物体或罕见物体，我们在存在性任务上评估了三种剪枝策略。该任务要求判断图像中是否存在某个特定物体（图 10）。在示例 1 中，自适应方法成功保留了检测小物体所需的词元，而基于注意力的方法未能检测到该物体。同样，在示例 2 中，基于注意力的方法和基于多样性的方法都未能识别一个非常小的物体，而自适应方法保留了相关线索并给出了正确答案。这些结果表明，自适应策略能够可靠地保留重要词元，即使小物体是推理的关键证据，也依然有效。

总体而言，在各种推理任务中，自适应方法表现出稳健且可靠的行为，这一点得到了定量结果和定性示例的共同支持。




图 9：**三种剪枝策略——基于注意力的剪枝、基于多样性的剪枝和我们的自适应方法——在计数和空间推理任务上的定性比较。基于多样性的剪枝保留了更广泛的空间线索，通常能生成细致的描述，但偶尔也会引入幻觉物体或关系。基于注意力的剪枝聚焦于显著区域，预测结果较为稳定，但有时范围过窄。自适应方法平衡了这两种机制，在保留关键视觉线索的同时减少幻觉，从而支持可靠推理。**

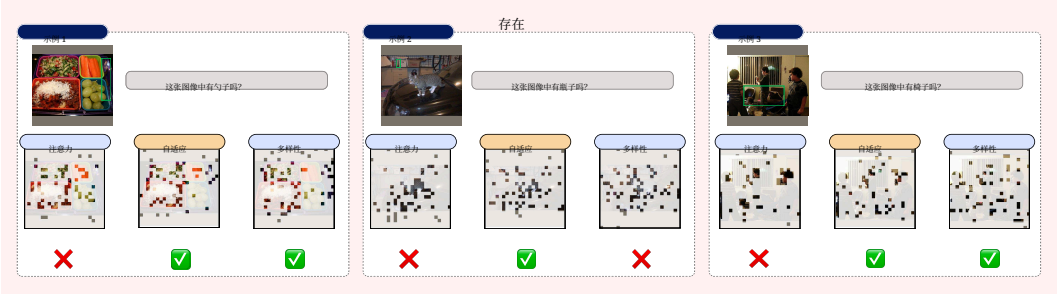



图10：**存在性任务上的定性比较。**自适应剪枝能够正确保留识别小物体所需的词元，在具有挑战性的情形中优于基于注意力和基于多样性的方法。



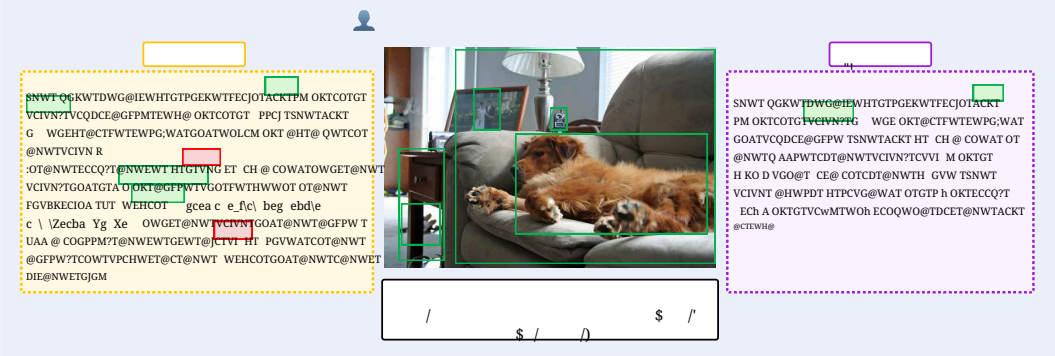

图11：**示例说明图像复杂度如何决定自适应剪枝方法在细粒度推理中的成败**。对于简单的低 Erank 图像，当关键对象在空间上集中时，该方法表现良好（a），但当许多相似物体分散在整个场景中时则会遇到困难（b）。对于复杂的高 Erank 图像，当语义线索广泛分布时它能够成功（c），但当细粒度推理所需的关键证据高度集中于杂乱场景中的局部区域时则会失败（d）。

### F.2 自适应方法在不同图像复杂度下细粒度推理中的失败模式

我们分析自适应方法——其令牌选择由图像复杂度引导——在细粒度推理任务上失败的情形。为此，我们使用同时需要对象计数和空间关系推理、且覆盖不同复杂度图像的问题，并在下文总结观察到的失败模式。代表性示例如图 11 所示。

**（1）包含许多物体的低 Erank 图像**。在低秩的简单图像中，自适应方法会为高注意力词元分配更大权重，并以有限的多样性加以补偿。这会产生与基于注意力的剪枝相似的选择模式。当物体较少且位置集中时（例如图 11a），这种聚焦式选择能够实现准确的计数和空间推理。然而，当许多物体散布于整幅图像中时（图 11b），集中式选择无法捕捉更广泛的空间布局，从而导致计数错误和推理不完整。

**（2）关键证据局部集中的高有效秩图像**。对于有效秩较高的复杂图像，自适应方法的行为更接近基于多样性的剪枝，倾向于选择分布广泛的词元。当物体或语义线索广泛分布时（图 11c），这种分散的选择具有优势，有助于准确计数和进行关系推理。然而，当关键证据高度集中在场景中的局部区域时（图 11d），广泛分布的词元选择会分散重要区域周围的注意力，导致方法忽略关键线索，在计数和空间推断方面均告失败。



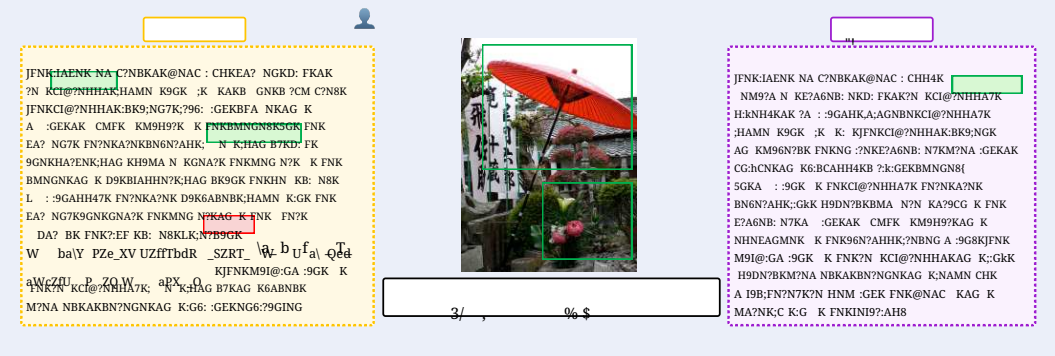


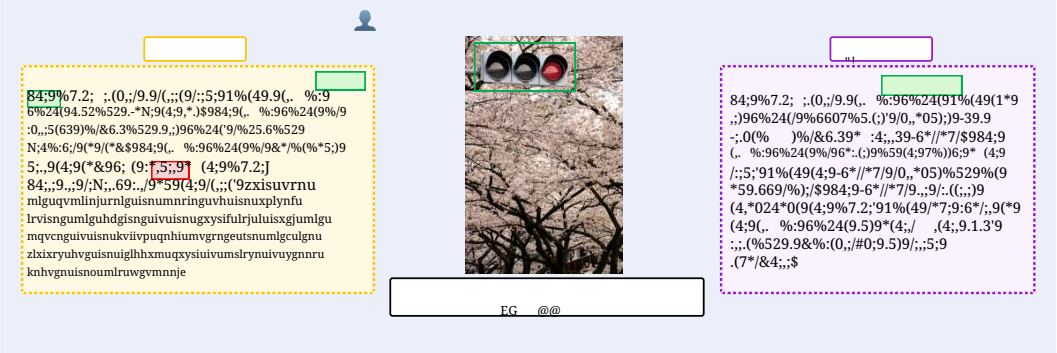


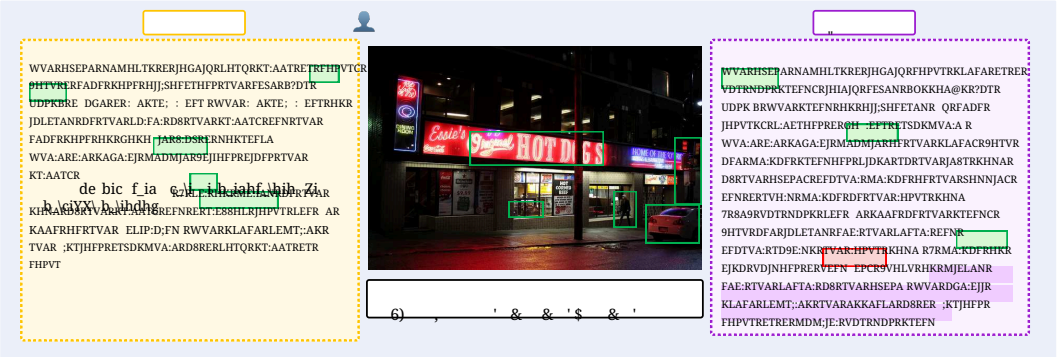

图 12： **CHAIR 定性比较：FasterVLM 与 DivPrune（集合 1）**。在标注中，真实物体和幻觉物体用于标记物体词，红色文本表示错误短语。

## G CHAIR 上的更多示例

为了进一步展示图像描述任务中基于注意力的剪枝与基于多样性的剪枝之间的差异，我们从 CHAIR 数据集中提供了更多定性样例，对比作为基于注意力的方法的 FasterVLM 和作为基于多样性的方法的 DivPrune（图 12 和图 13）。这些案例展示了这两种方法在响应风格和幻觉倾向上的差异：DivPrune 往往生成内容更丰富、描述性更强的描述，但也会引入幻觉物体；而 FasterVLM 生成的描述则更加保守、聚焦。

此外，图 14 展示了一项受控实验：令牌预算固定为 64，并以 25% 为步长调整 DivPrune 选定令牌与注意力选定的令牌之间的比例（DivPrune 与注意力的比例， $R$）。我们观察到，纯 DivPrune（$R = 0$）经常生成的幻觉物体会随着注意力选定的令牌占比增加而逐渐减少，并在 $R \geq 50\%$ 时完全消失。与此同时，随着 $R$ 增加，响应风格也从推测性、探索性转变为更加客观、可靠，体现了基于注意力的令牌选择所带来的稳定作用。





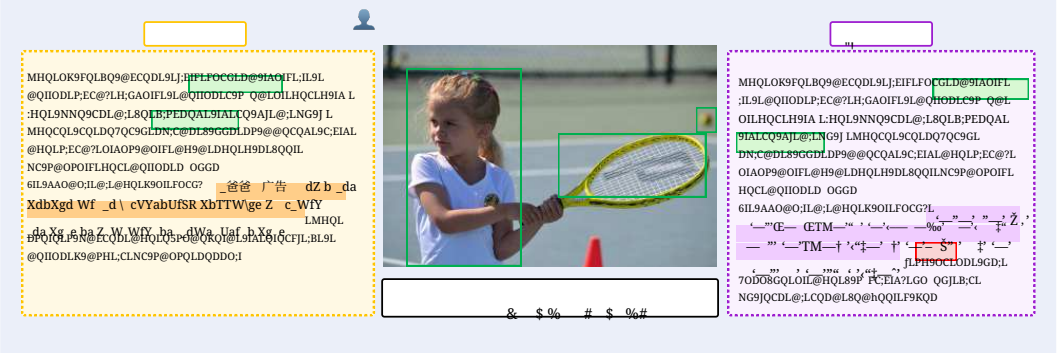


图 13： **CHAIR 定性比较：FasterVLM 与 DivPrune（第 2 组）**。在标注中，真实物体和幻觉物体用于标记物体词；表示 DivPrune 的措辞表示 FasterVLM 的措辞；红色文本表示错误短语。

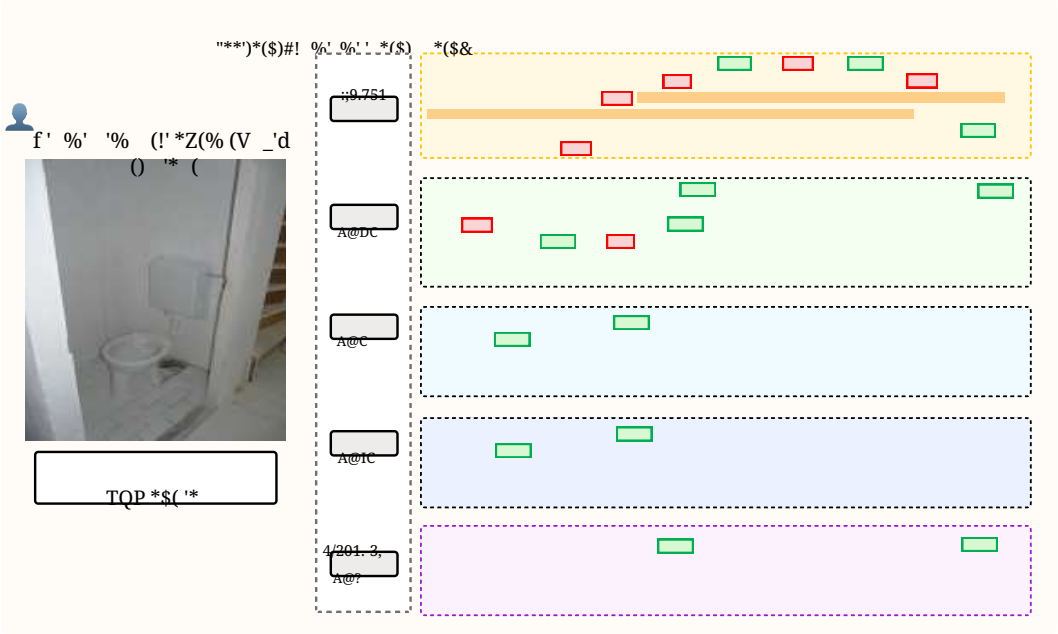



图 14：**改变基于注意力的选择比例的影响 $R$ 在 64 个 token 预算下**。随着 $R$ 增大，DivPrune 产生的幻觉物体被逐步抑制，回答从探索性描述转向面向事实的描述。在标注中，真实物体和幻觉物体用于标注物体词；表示 DivPrune 特有的措辞。

关于大型语言模型使用的声明

为确保透明度并遵守 ICLR 2026 指南，我们报告称，大语言模型（大语言模型）被用于协助润色本文的文本。

**使用范围**。 该模型的作用严格限于写作助手。其贡献包括：

- 纠正语法错误、拼写和标点符号。
- 改进句子结构与行文流畅度，以提升清晰度。
- 优化措辞，以提高精准度和简洁性。