VisionThink: Smart and Efficient Vision Language Model via Reinforcement Learning

李莹璐 2025.8.27

目录

- 作者介绍
- 研究背景
- 研究方法
- 实验效果

VisionThink: Smart and Efficient Vision Language Model via Reinforcement Learning

Senqiao Yang*[1] Junyi Li*[2] Xin Lai*[1] Bei Yu[1] Hengshuang Zhao[2] Jiaya Jia[1,3]

[1]CUHK [2]HKU [3]HKUST

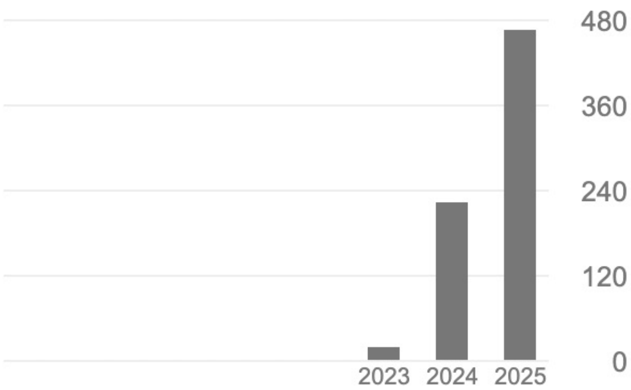
Codes and models: `https://github.com/dvlab-research/VisionThink`

Senqiao Yang

The Chinese University of Hong Kong
Verified email at link.cuhk.edu.hk - Homepage

Cited by

	All	Since 2020
Citations	715	715
h-index	12	12
i10-index	13	13



TITLE	CITED BY
Step-dpo: Step-wise preference optimization for long-chain reasoning of llms X Lai, Z Tian, Y Chen, S Yang, X Peng, J Jia arXiv preprint arXiv:2406.18629	148
Lidar-llm: Exploring the potential of large language models for 3d lidar understanding S Yang, J Liu, R Zhang, M Pan, Z Guo, X Li, Z Chen, P Gao, Y Guo, ... AAAI 2025	95
Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation J Liu, M Liu, Z Wang, P An, X Li, K Zhou, S Yang, R Zhang, Y Guo, ... Advances in Neural Information Processing Systems 37, 40085-40110	79
Visionzip: Longer is better but not necessary in vision language models S Yang, Y Chen, Z Tian, C Wang, J Li, B Yu, J Jia Proceedings of the Computer Vision and Pattern Recognition Conference, 19792 ...	66

Multi-Modality Language Models、Representation Learning

研究背景

- 视觉语言模型（VLMs）中Visual Token 消耗过高
  - 一张2048*1024的图片在LLaVA 1.5中需要576个visual token， 在Qwen2.5-VL中需要2678个，避免过度使用视觉 Token 势在必行
- 现实统一压缩Visual Token 的问题
  - 现有方法通常采用固定剪枝率或阈值压缩 Token

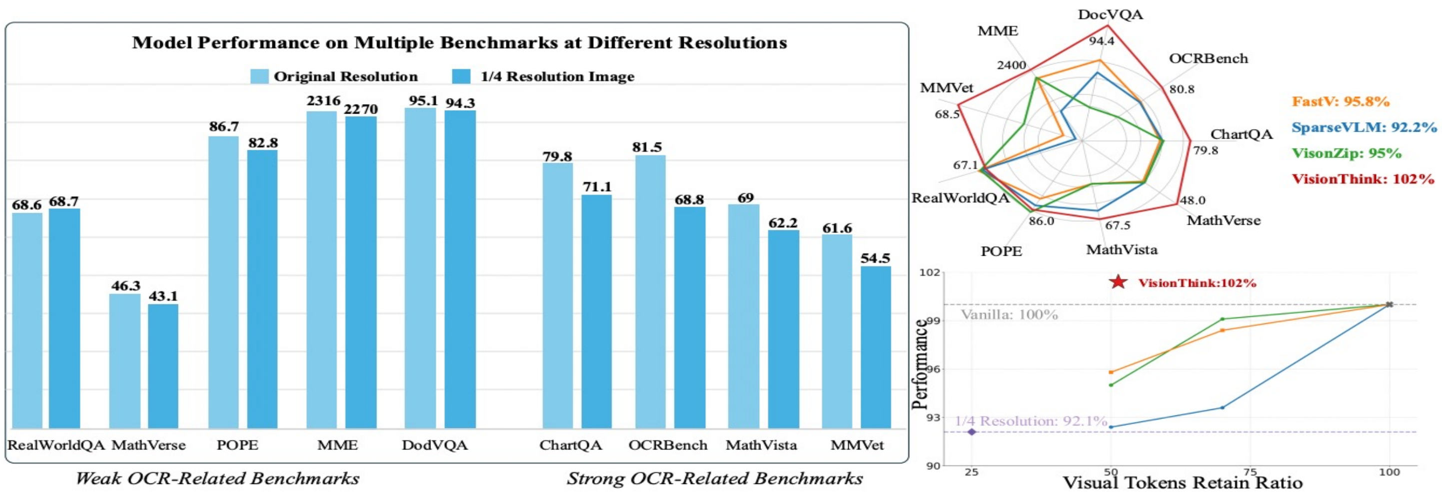


Figure 1: **Our key observations and VisionThink performance and efficiency**. **Left**: We find that in most general scenarios, even reducing visual tokens by a factor of four results in only minimal performance drop. However, token compression leads to a significant performance drop on strong OCR-related benchmarks. **Right**: Our VisionThink significantly outperforms previous work in both performance and efficiency.

不同任务对分辨率的需求差异较大，导致性能下降，特别是在OCR等需要细粒度视觉理解的任务中。

若能够动态区分需要高分辨率处理和不需要处理的样本，将存在显著的效率优化潜力。

研究背景

- 传统基于规则的强化学习难以处理通用VQA任务的多样性和复杂性。

GRPO

- 在训练过程中，GRPO根据给定的问题q从旧策略中采样一组输出，然后通过最大化以下目标函数来优化策略模型：

$$\mathcal{I}_{GRPO}(\theta) = \mathbb{E}_{[q\sim\mathcal{D},\{\sigma_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|q)]}$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(\sigma_i|q)}{\pi_{\theta_{old}}(\sigma_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(\sigma_i|q)}{\pi_{\theta_{old}}(\sigma_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)$$

$$D_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1$$

计算复杂度

To evaluate the computational complexity of VLMs, we analyze key components, including the self-attention mechanism and the feedforward network (FFN). The total floating-point operations (FLOPs) are given by:

$$\text{Total FLOPs} = T \times (4nd^2 + 2n^2d + 2ndm)$$

- T表示Transformer层数，n是序列长度，d是hidden层维度，m是FFN的中间维度
- 显然，计算复杂度主要受序列长度n的影响，在一般VLM任务中，n由 $n_{\text{sys}} + n_{\text{img}} + n_{\text{question}}$ 组成，image的token数量通常远大于其他两个分量。因此，控制图像token数量是提升VLM效率的关键。

研究方法

1. LLM-as-Judge、多轮GRPO
2. 设计动态Reward函数与惩罚机制

- VisionThink: 动态判断是否需要高分辨率图像
  - 首先处理低分辨率图像以降低计算成本。当下采样图像中的信息不足以回答问题时，它会智能地请求原始高分辨率输入

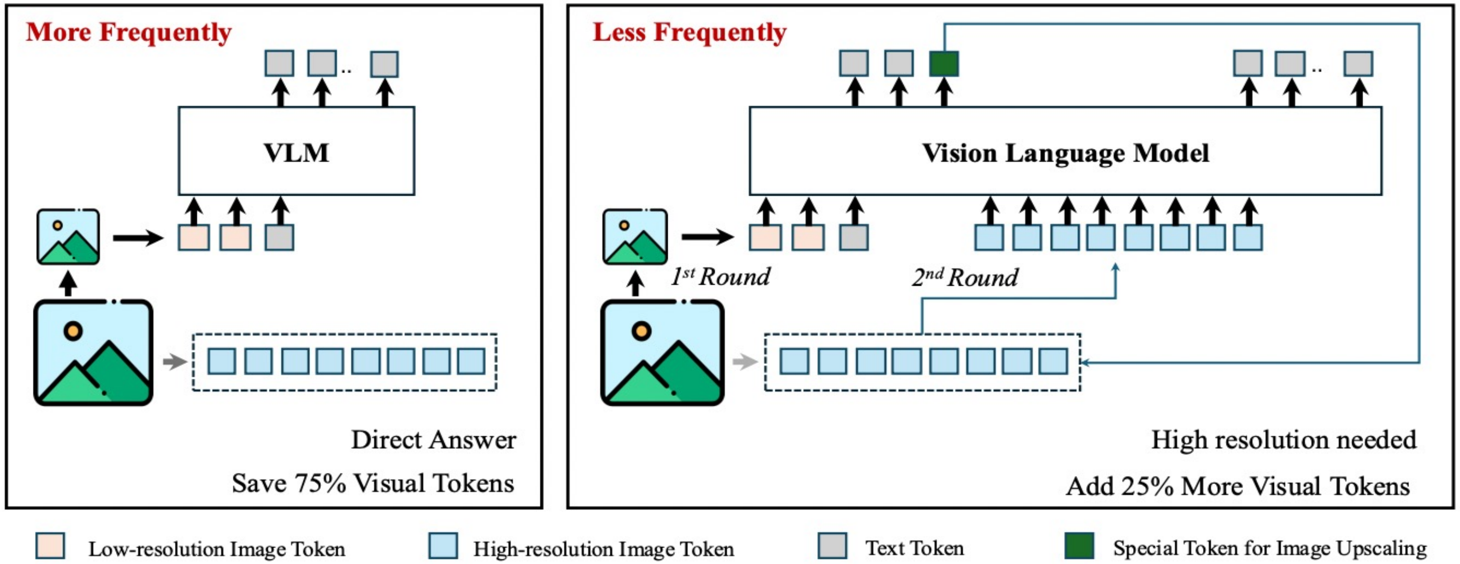


Figure 2: **Framework of VisionThink.** (a) The left image illustrates VisionThink processing an image with resolution reduced by a factor of four, where the VLM directly provides an answer. (b) The right image shows a case where the model detects insufficient information and requests a high-resolution image to answer the question.

实验室
Computing Lab

研究方法

1. LLM-as-Judge、多轮GRPO
2. 设计动态Reward函数与惩罚机制

- LLM-as-Judge
  - 利用LLM广泛的知识及语言理解能力评估模型输出的正确性。评估完全通过文本进行，将模型的答案与真实值进行比较，避免了视觉内容的偏差以及VLM性能的限制。Reward为0或1。

Table 3: **Judgment Prompt Template.** Question, Ground Truth and Prediction are dynamically replaced with the specific question, ground truth and model prediction during evaluation.

SYSTEM PROMPT:
You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs.
Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:
INSTRUCTIONS:
- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

USER PROMPT:
I will give you a question related to an image and the following text as inputs:
1. **Question Related to the Image**: Question
2. **Ground Truth Answer**: Ground Truth
3. **Model Predicted Answer**: Prediction
Your task is to evaluate the model's predicted answer against the ground truth answer, based on the context provided by the question related to the image. Consider the following criteria for evaluation:
- **Relevance**: Does the predicted answer directly address the question posed, considering the information provided by the given question?
- **Accuracy**: Compare the predicted answer to the ground truth answer. You need to evaluate from the following two perspectives:
(1) If the ground truth answer is open-ended, consider whether the prediction accurately reflects the information given in the ground truth without introducing factual inaccuracies. If it does, the prediction should be considered correct.
(2) If the ground truth answer is a definitive answer, strictly compare the model's prediction to the actual answer. Pay attention to unit conversions such as length and angle, etc. As long as the results are consistent, the model's prediction should be deemed correct.
Output Format:
Your response should include an integer score indicating the correctness of the prediction: 1 for correct and 0 for incorrect. Note that 1 means the model's prediction strictly aligns with the ground truth, while 0 means it does not.
The format should be Score: 0 or 1

多轮GRPO

- VisionThink框架中，首先将问题和下采样图像输入到VLM，如果信息不足以回答当前问题，模型将自主请求更高分辨率的图像并生成新的响应，本质上是一个多轮交互的问题

$$\mathcal{I}_{GRPO}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|q;\mathcal{T})}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{\sum_{t=1}^{|o_i|}\mathbb{I}\left(o_{i,t}\right)}\sum_{t=1}^{|o_i|}\mathbb{I}\left(o_{i,t}\right)$$

$$\cdot\min\left(p_{i,t}\hat{A}_{i,t},\text{clip}\left(p_{i,t},1-\epsilon,1+\epsilon\right)\hat{A}_{i,t}\right)-\beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{\text{ref}}\right]\Bigg],$$

Reward函数

$$\mathcal{R}_{\text{overall}} = \mathcal{R}_{\text{accuracy}} + \mathcal{R}_{\text{format}} - \mathcal{P}_{\text{control}},$$

$\mathcal{R}_{accuracy}$：LLM-as-Judge，正确的答案打1分，错误的为0分

$\mathcal{R}_{format}$：要求推理过程被包含在“<think></think>”，答案包含于“<answer></answer>”，只有所有格式都正确才能获得0.5分，否则为0

$\mathcal{P}_{control}$：

$$\mathcal{P}_{control} = 0.1 \cdot \left[\mathbf{1}_{\text{direct}}\mathbb{I}(r < \theta) + \mathbf{1}_{\text{high}}\mathbb{I}(r \geq \theta)\right], \qquad r = \frac{C_{\text{direct}}}{C_{\text{direct}} + C_{\text{high}}},$$

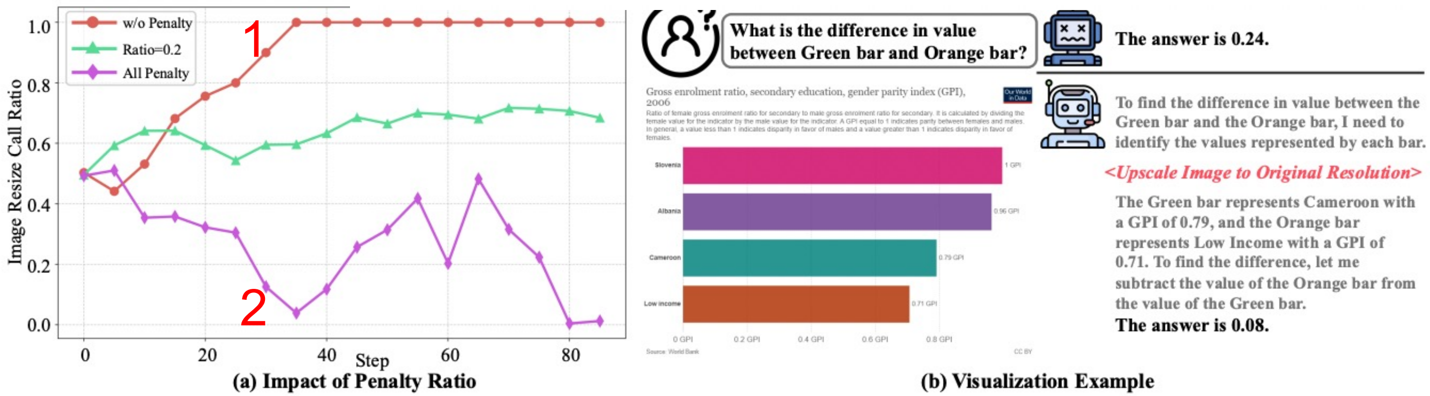


Figure 3: (a) Impact of the Penalty Ratio. Applying a penalty to all resize image requests or removing the penalty entirely will both lead to model collapse. (b) VisionThink correctly solves OCR-related problems by autonomously requesting high-resolution images.

1. 若无惩罚，模型会一直倾向于请求高分辨率图像
2. 遵循Search-R1对依赖高分辨率图像的正确答案采用0.1惩罚，导致模型倾向于直接回答，会崩溃

实验效果

VisionThink与SOTA的对比实验：
RL使VLM更有效

Table 1: **Effective Performance Compared to the Sota Model.** Our model is based on Qwen2.5-VL-7B-Instruct. VisionThink‡ represents a model trained on general VQA tasks using full image resolution with the LLM-as-Judge strategy, which does not contain efficiency capabilities. Qwen2.5-VL-7B* reports the results evaluate by lmms-eval[86].

Method	MMMU	MMMU-Pro	MMBench	RealWorldQA	POPE	MME	MathVista	MathVerse	MMVet
	val	test	en_test	test	test	test	testmini	testmini	test
Closed-Source Model									
GPT-4o [48]	69.1	54.0	83.4	58.6	85.6	2329	63.8	50.2	69.1
Claude-3.5 Sonnet [2]	68.3	55.0	82.6	59.9	-	1920	67.7	41.2	70.1
Gemini-1.5-Pro [57]	62.2	49.4	73.9	70.4	88.2	-	63.9	-	64.0
Open-Source General Model									
Cambrain-1-8B [60]	42.7	-	75.9	60.0	86.4	1803	49.0	-	-
InternVL2-8B [12]	49.3	32.5	81.7	64.4	84.2	2210	58.3	-	60.0
LLaVA-OneVision-7B [28]	48.8	-	-	66.3	88.4	1998	63.2	-	57.5
MiniCPM-Llama-V-2.5-8B [81]	45.8	19.6	77.2	63.0	86.7	2025	54.3	-	-
MiniCPM-V-2.6-8B [81]	49.8	27.2	78.0	65.0	83.2	2348	60.6	-	-
IXC-2.5 [87]	42.9	-	82.2	67.8	-	2229	63.8	-	51.7
InternVL2.5-8B [11]	56.0	38.2	84.6	70.1	90.6	2344	64.4	39.5	62.8
Reasoning Model									
LLaVA-CoT-11B [71]	-	-	75.0	-	-	-	54.8	-	60.3
LLaVA-Reasoner-8B [89]	-	-	-	-	-	-	50.6	-	-
Insight-V-8B [14]	50.2	24.9	82.3	-	-	2312	59.9	-	-
Mulberry-7B [78]	55.0	-	-	-	-	2396	63.1	-	-
Vision-R1-LlamaV-CI-11B [19]	-	-	-	-	-	2190	62.7	27.1	-
VisionThink									
Qwen2.5-VL-7B* [5]	50.3	37.7	82.6	68.6	86.7	2316	68.2	46.3	61.6
VisionThink ‡	51.0	40.1	82.9	68.6	87.9	2307	71.2	48.8	67.5
VisionThink	51.2	38.9	80.0	68.5	86.0	2400	67.5	48.0	67.1

实验效果

- RL使VLM效率更高
  - 在大多数基准测试中，VisionThink的推理时间接近使用1/4图像Token的QwenRL 1/4，并且显著优于处理所有图像Token的QwenRL模型
  - 在像ChartQA这样高度依赖OCR的基准测试中，VisionThink消耗的时间比Baseline QwenRL更多。这是因为VisionThink识别出大多数问题在低分辨率下无法正确回答，因此自主请求高分辨率图像。

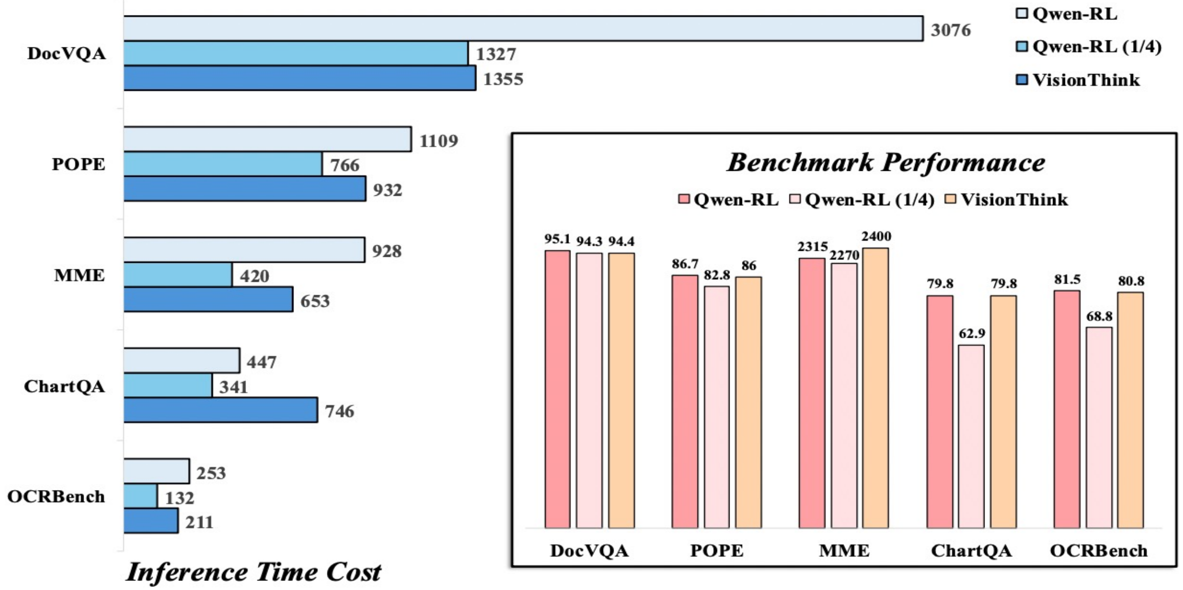


Figure 4: **Inference Time Cost and Benchmark Performance Comparison for Reasoning Model.** Qwen-RL and Qwen-RL (1/4) represent leveraging the LLM-as-Judge on the Qwen2.5-VL-Instruct Model and inference on full resolution image and 1/4 resolution image, respectively.

容计算实验室
ontent Computing Lab

实验效果

□ RL使VLM效率更高

Table 2: **Comparison with Traditional Efficient VLM Methods.** Vanilla represents the Qwen2.5-VL-7B-Instrcut. The retained ratio of the baseline methods is a predefined hyperparameter, while for VisionThink, the ratio is determined autonomously by the model and reported as a statistical value. Note that *Down-Sample* refers to the model's performance when directly fed images with their resolution reduced by half. Additional baseline comparison results are shown in Table. 7

Method	ChartQA†	OCRBench	DocVQA	MME	MMVet	RealWorldQA	POPE	MathVista	MathVerse	Avg.
	test	test	val	test	test	test	test	testmini	testmini	
Retain 100% Visual Tokens Across All Benchmarks										
Vanilla	79.8 100%	81.5 100%	95.1 100%	2316 100%	61.6 100%	68.6 100%	86.7 100%	68.2 100%	46.3 100%	100%
Retain 25% Visual Tokens Across All Benchmarks										
Down-Sample	62.9 78.8%	68.8 84.4%	94.3 99.1%	2270 98.0%	54.5 88.5%	68.8 100.3%	82.8 95.5%	62.2 91.2%	43.1 93.1%	92.1%
Retain 50% Visual Tokens Across All Benchmarks										
SparseVLM (ICML 2025)	73.2 91.7%	75.6 92.7%	66.8 70.2%	2282 98.5%	51.5 83.6%	68.4 99.7%	85.5 98.6%	66.6 97.6%	45.1 97.4%	**92.2%**
FastV (ECCV 2024)	72.6 91.0%	75.8 93.0%	93.6 98.4%	2308 99.6%	52.8 85.7%	68.8 100.3%	84.7 97.7%	63.7 93.4%	45.0 97.2%	**95.8%**
Retain 70% Visual Tokens Across All Benchmarks										
SparseVLM (ICML 2025)	75.8 94.9%	79.3 97.3%	68.7 72.2%	2276 98.3%	53.7 87.2%	68.5 99.8%	85.4 98.5%	66.3 97.2%	45.1 97.4%	**93.6%**
FastV (ECCV 2024)	77.2 96.7%	82.2 100.8%	94.4 99.3%	2342 101.1%	56.0 90.9%	68.6 100%	85.9 99.1%	65.9 96.6%	46.9 101.3%	**98.4%**
Retain Approximately 51.3% Visual Tokens Across All Benchmarks										
VisionThink	79.8 100%	80.8 99.1%	94.4 99.3%	2400 103.6%	68.5 111.2%	67.1 97.8%	86.0 99.2%	67.5 99.0%	48.0 103.7%	101.4%

- FastV和SparseVLM都需要计算注意力分数来剪枝visual token，因此无法使用FlashAttention2进行优化，并可能导致内存使用增加，且与一些VLLM不兼容。
- VisionThink在九个基准测试中平均优于先前方法，且先前方法需要预定义剪枝率阈值，而VisionThink可以根据问题和图像内容自主决定是否减少token。

实验效果

- RL使VLM更智能
  - 实验结果：在ChartQA和OCRBench等需要详细视觉理解的基准测试上，VisionThink请求高分辨率图像的比例更高。相比之下，对于MME和DocVQA等基准测试，至少有70%的样本可以直接使用原始分辨率1/4的低分辨率图像进行回答。
  - 直觉：大多数日常问题并不需要高分辨率图像，而只有OCR相关的任务真正依赖于它们。

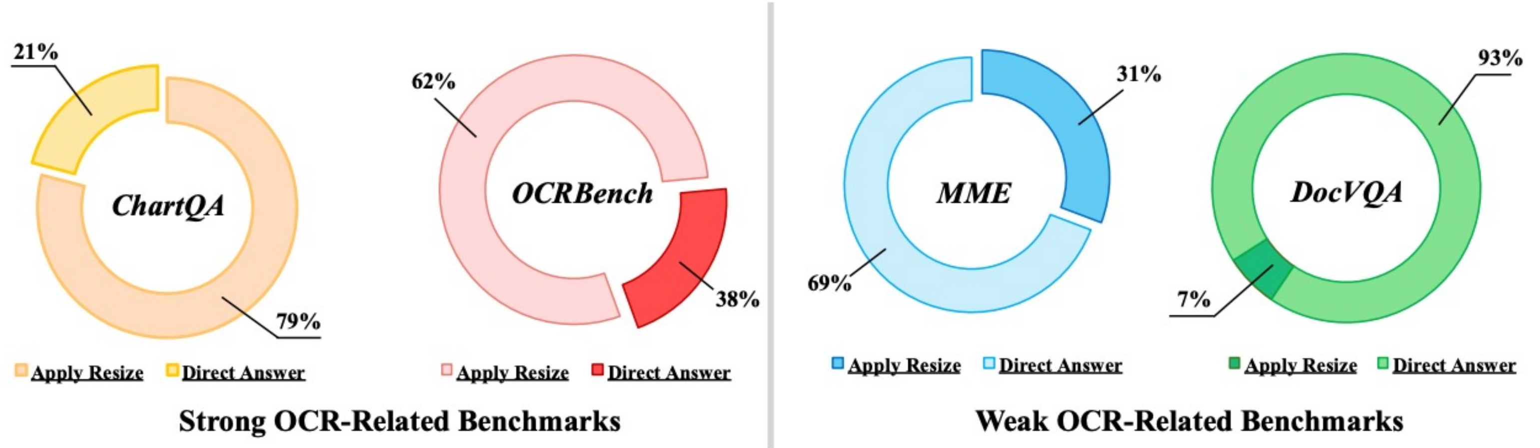


Figure 5: **VisionThink smartly determine the high-resolution image ratio.** Apply Resize indicates that the model autonomously requests to view the original high-resolution image, while Direct Answer indicates that the model is able to answer the question using only the 1/4-sized image.

总结

- VisionThink能够根据图像内容智能判断是否需要更高分辨率，例如在OCR任务中请求高分辨率的比例显著高于其他任务
- 目前仅支持 2 倍分辨率提升和最多 2 轮对话，未来可探索更灵活的分辨率调整机制
- 虽然 LLM-as-Judge 提升了强化学习效果，但其依赖外部模型，未来可探索更轻量或自适应的评判机制

Thanks！