RAR: Reversing Visual Attention Re-Sinking for Unlocking Potential in Multimodal Large Language Models

Zhehan Kan[1, 2*] **Xin Li**[2*] **Yanli Liu**[1*] **Xiaochen Yang**[3] **Xinghua Jiang**[2]
Yinsong Liu[2] **Deqiang Jiang**[2] **Xing Sun**[2] **Qingmin Liao**[1] **Wenming Yang**[1†]
[1]Tsinghua University [2]Tencent Youtu Lab [3]The University of Glasgow

一、Task

VQA（visual question answer）：
1. Input：images + Open question
2. Output：describe
3. Metric：ACC（match）

VQA_OCR：
1. Input：image include text
2. Output：text
3. Metric：ACC（match）

Visual Grounding：
1. Input：images + question
2. Output：bbox
3. Metric：IOU

Visual Hallucination：
1. Input：image + question（induce hallucinatior
2. Output：text
3. Metric：Faithfulness

二、related work & insight

1. Best layer in MLLMs Decoder

mid–to–late vision encoder layers often outperform the final one across tasks

2. Visual attention sink in MLLMs

In LLMs, attention sink involves low–semantic tokens drawing excessive weights

3、insight

Consequently, gradients for vision tokens rely solely on backpropagation of textual losses through the attention mechanism, constraining its learning capacity on vision tokens and rendering the overall gradient distribution increasingly sparse.

三、Method

Sink Attention Dynamic Sparsification framework

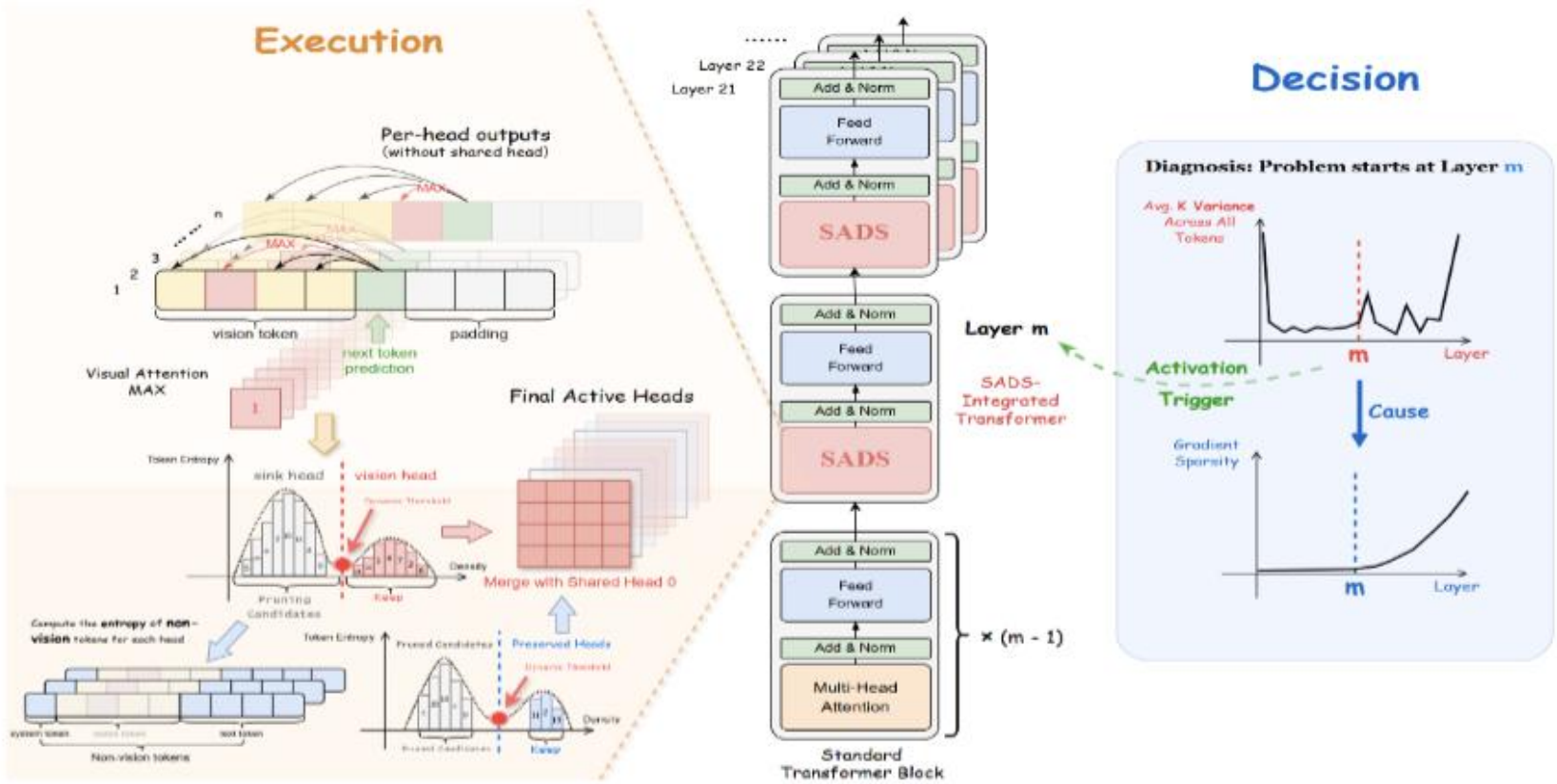
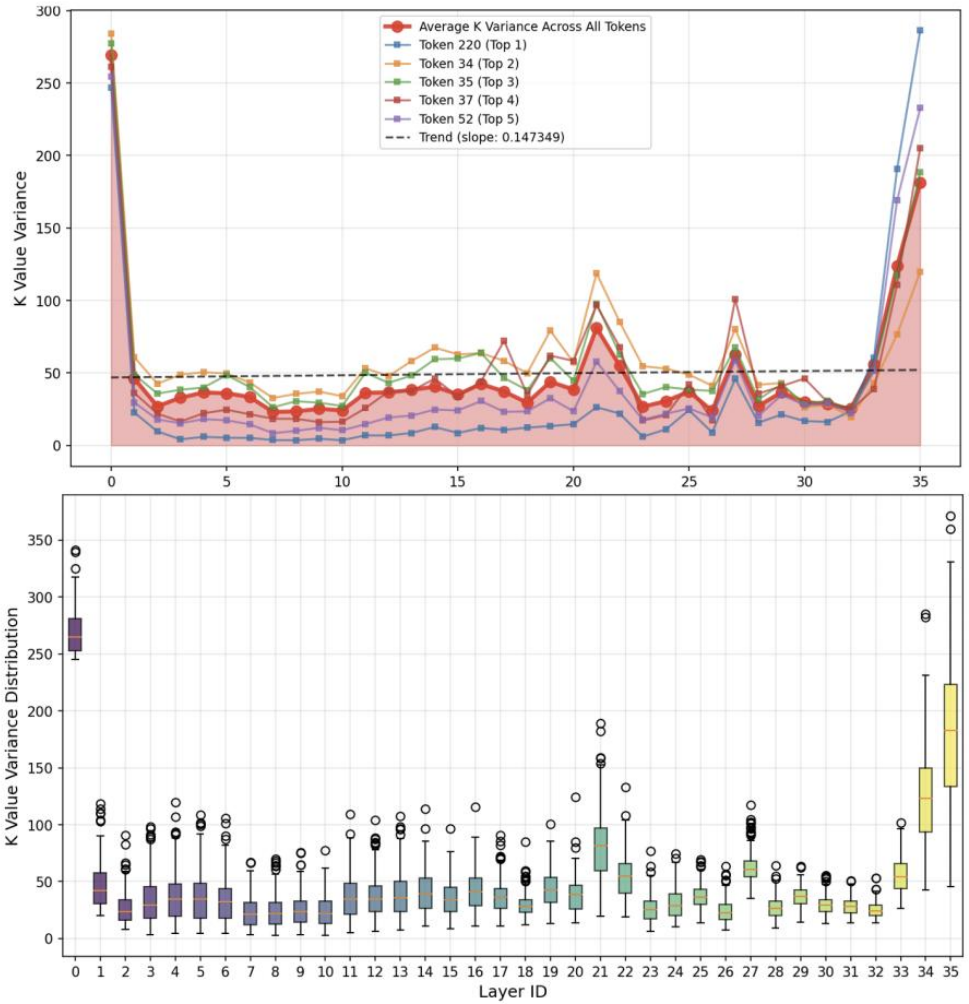
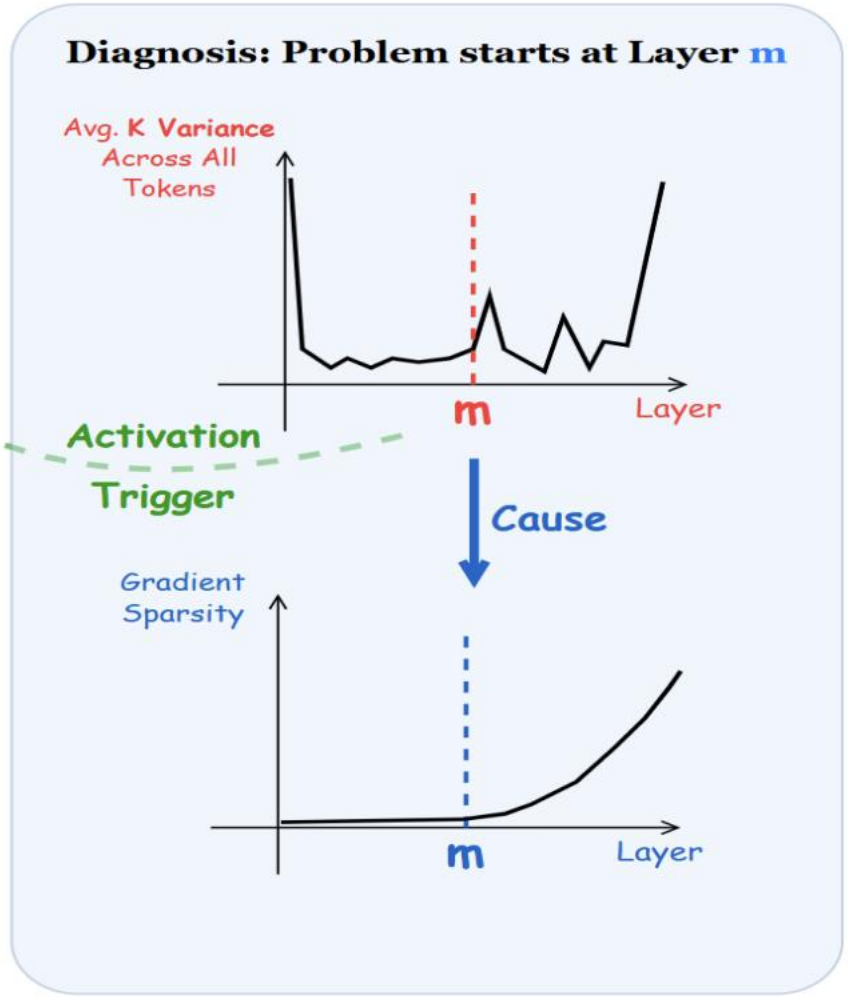


Figure 6: **Overview of the SADS framework. Decision (Right):** Triggers activation at Layer m upon detecting gradient sparsity and key variance anomalies. **Execution (Left):** Filters heads using bimodal thresholds on maximum visual attention (separating vision and sink heads) and non-vision token entropy (distinguishing sink_G from sink_S within sinks). Merges all vision heads, retained sink_G heads, and a shared head for computation, preserving global context.

三、Method

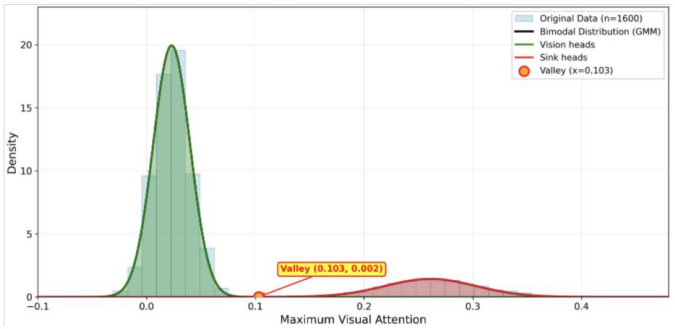
Observation → proxy



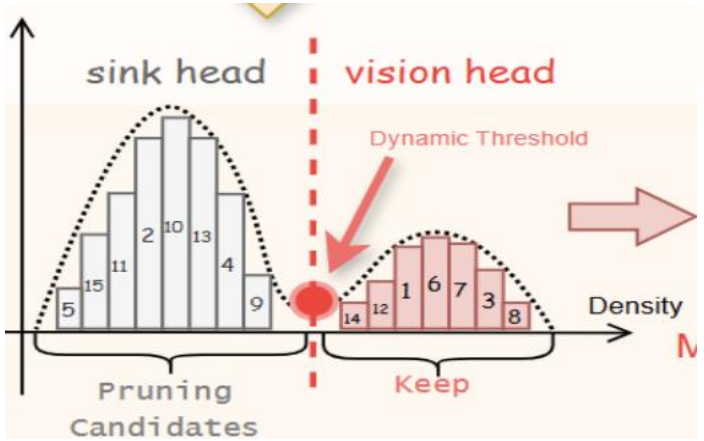


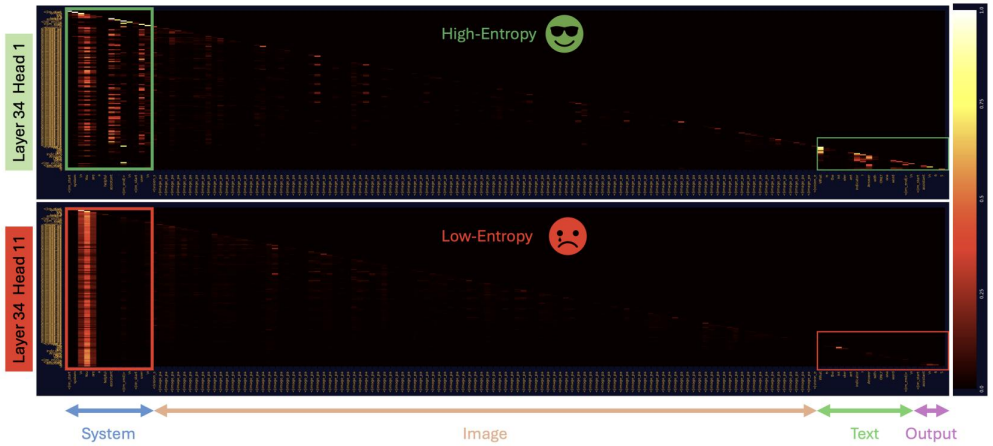
三、Method

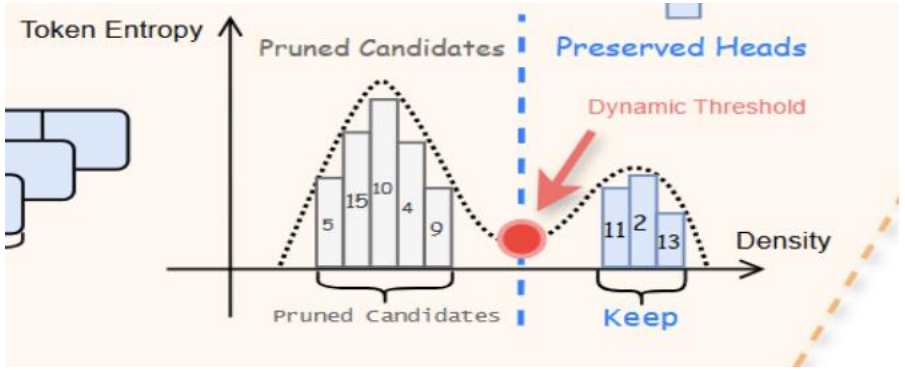
Observation → proxy



(b) Distribution of maximum visual attention for vision heads and sink heads, showing a bimodal distribution.







四、experiments

Table 2: Benchmark performance comparison on general VQA and OCR VQA tasks.

Model	General VQA Task							OCR VQA Task		
	VQA^{v2}	GQA	VQA^{vg}	MME	MMB	MMStar	AI2D	InfoVQA	TextVQA	DocVQA
LLaVA-1.5-7B	78.3	61.1	54.6	1808.4	61.1	33.2	55.7	41.2	64.7	69.4
+SFT	79.1	63.1	55.7	1899.6	61.9	34.5	56.2	43.7	65.5	71.2
+ Ours	**80.8**	**65.2**	**58.5**	**2018.8**	**63.2**	**36.3**	**57.1**	**46.3**	**67.5**	**74.4**
Qwen2.5VL-3B	76.7	60.4	54.3	2184.1	75.4	53.0	77.9	75.1	78.7	93.0
+SFT	77.9	62.0	55.2	2199.9	75.9	53.7	78.4	75.9	79.0	92.9
+ Ours	**79.7**	**64.2**	**58.1**	**2276.3**	**76.9**	**55.4**	**79.5**	**77.3**	**80.4**	**93.5**
InternVL2-2B	72.9	55.6	50.1	1864.3	69.1	48.9	73.1	58.8	73.4	86.4
+SFT	74.2	56.9	52.3	1899.1	70.0	49.6	73.9	59.1	73.8	86.6
+ Ours	**75.9**	**59.0**	**55.4**	**2006.5**	**71.6**	**50.8**	**75.7**	**60.5**	**75.9**	**88.2**
Qwen2.5VL-7B	81.6	65.8	60.5	2276.3	82.2	64.2	84.1	81.7	80.2	94.8
+SFT	81.9	66.1	61.0	2230.2	82.0	64.5	84.4	82.0	80.7	94.2
+ Ours	**82.6**	**67.9**	**62.1**	**2289.8**	**83.3**	**66.0**	**84.8**	**82.9**	**81.3**	**95.0**
Qwen2.5VL-32B	82.9	68.4	63.6	2297.4	83.8	70.3	85.2	83.4	82.8	94.8
+SFT	83.0	68.6	63.9	2255.4	83.8	69.6	85.1	83.0	82.9	94.4
+ Ours	**83.8**	**69.9**	**64.5**	**2326.6**	**84.5**	**71.3**	**85.7**	**83.8**	**83.9**	**95.1**

Table 3: Benchmark performance comparison on visual perception tasks.

Model	Visual Grounding Task					Vision Centric Task			Visual Hallucination Task	
	RefCOCO/+/g	LISA	RefGTA	OD^{VG}	OVDEval	MMVP	CVBench	CLEVER	CHAIR↓	POPE↑
LLaVA-1.5-7B	76.2	44.2	64.1	19.4	22.7	3.1	57.4	43.6	44.7	85.6
+SFT	77.1	44.7	64.6	20.2	23.0	9.7	57.8	44.1	45.2	85.7
+ Ours	**78.9**	**50.1**	**66.2**	**24.8**	**27.1**	**15.1**	**60.4**	**46.6**	**41.7**	**86.4**
Qwen2.5-VL-3B	84.2	55.3	70.8	32.1	39.5	50.4	67.3	68.7	35.6	86.1
+SFT	84.6	55.3	71.0	32.5	39.9	52.1	68.1	70.0	35.4	86.4
+ Ours	**86.8**	**58.1**	**72.9**	**36.7**	**43.8**	**54.9**	**70.1**	**72.5**	**32.6**	**87.4**
InternVL2-2B	77.8	46.1	66.4	21.7	24.9	39.6	56.5	57.1	37.8	86.2
+SFT	78.1	45.6	66.9	23.2	25.3	40.4	57.2	57.9	37.9	86.0
+ Ours	**80.1**	**48.2**	**68.9**	**26.6**	**29.9**	**42.7**	**59.2**	**59.6**	**34.3**	**87.1**
Qwen2.5VL-7B	87.1	60.3	74.4	39.3	44.8	55.1	73.6	74.4	32.6	88.9
+SFT	87.3	60.1	74.8	39.7	44.4	55.8	73.8	74.8	33.1	89.2
+ Ours	**88.2**	**63.6**	**76.0**	**42.1**	**47.2**	**57.0**	**75.2**	**75.9**	**29.7**	**89.6**
Qwen2.5VL-32B	89.8	65.9	77.5	43.1	49.3	60.4	77.2	78.5	28.2	90.3
+SFT	89.7	63.3	77.8	43.3	49.3	60.8	77.1	78.7	30.1	90.1
+ Ours	**90.6**	**67.4**	**79.3**	**44.5**	**51.8**	**62.5**	**79.0**	**80.0**	**21.2**	**90.6**

ZEROTUNING: UNLOCKING THE INITIAL TOKEN'S POWER TO ENHANCE LARGE LANGUAGE MODELS WITHOUT TRAINING

Feijiang Han[1*] **Xiaodong Yu**[1,2] **Jianheng Tang**[3]
Delip Rao[1] **Weihua Du**[4] **Lyle Ungar**[1]

[1]University of Pennsylvania [2]AMD [3]Peking University [4]Carnegie Mellon University

一、Task

SST-2/MR: Sentiment Classification, instruction: Classify the sentiment into 'positive' or 'negative', sample fields: Sentence/Review;

SST-5: Fine-grained Sentiment Classification, categories: terrible/negative/neutral/positive/great;

SUBJ: Subjective/Objective Classification, instruction: Classify into 'subjective' or 'objective';

TREC: Question Type Classification, categories: Description/Entity/Expression/Person/Number/Location;

CB: Commitment Detection, instruction: Determine if the hypothesis is true (based on premise);

BoolQ: Boolean Question Answering, instruction: Answer by True or False (based on text)

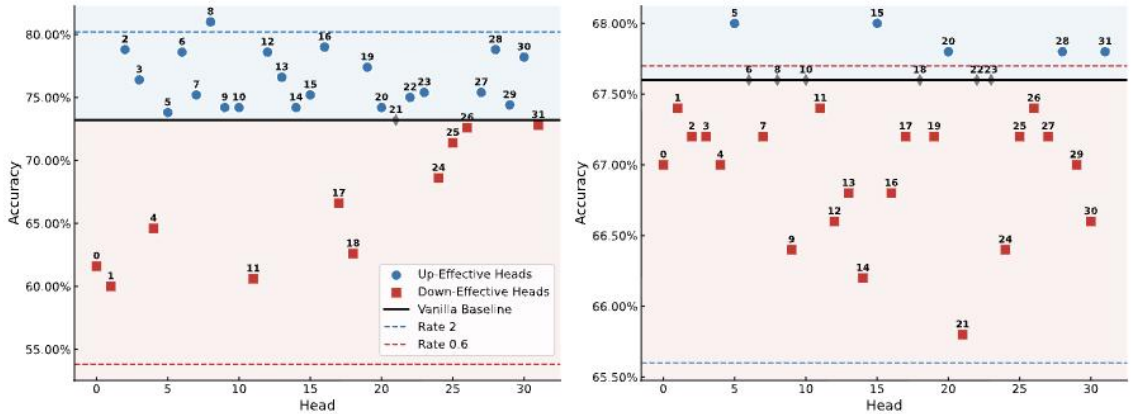
二、related work

1. Post-hoc Attention Steering (PASTA)：identify and up-weight key tokens
2. Attention Calibration (ACT)：down-weights non-initial sink tokens

fundamentally rely on heuristics to identify task-specific tokens

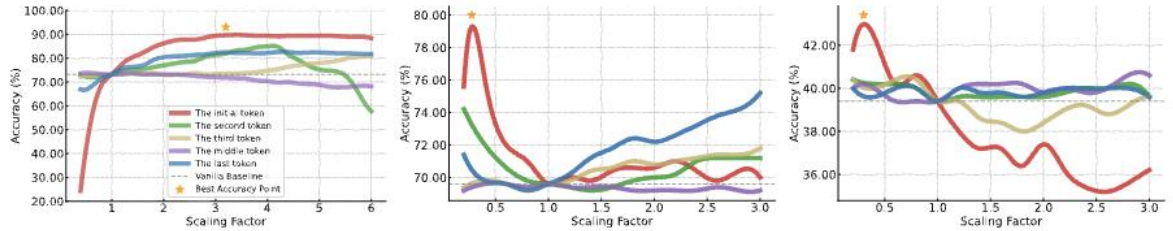
1. introduces the risk of bias
2. limits applicability when token importance is ambiguous or when optimized attention kernels make attention maps inaccessible.

三、Observation



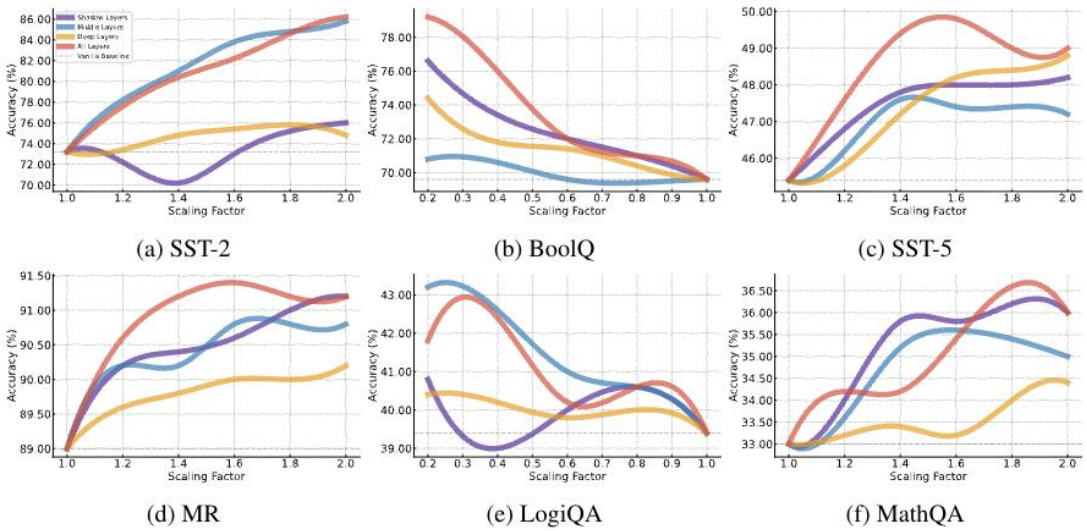
(a) SST-2 (b) MMLU

Figure 5: Accuracy of scaling the initial token's attention in individual heads using $\gamma = 1.5$ across (a) SST-2, (b) BoolQ, (c) MMLU, and (d) MathQA. Results reveal heterogeneous behavior among heads, motivating head-specific tuning strategies.



(a) SST-2 (b) BoolQ (c) LogiQA

Figure 2: Impact of attention scaling factor γ on different token positions across three tasks: (a) SST-2, (b) BoolQ, and (c) LogiQA. Modifying the initial token's attention consistently yields significant accuracy improvements, often surpassing adjustments to other tokens.



(a) SST-2 (b) BoolQ (c) SST-5

(d) MR (e) LogiQA (f) MathQA

Figure 4: Accuracy trends when scaling the initial token's attention across different layer groups: shallow (Layers 1–10), middle (Layers 11–21), and deep (Layers 22–31). Different depths exhibit a consistent accuracy trend with varying magnitudes.

三、Method

1. Supervised Calibration
    1. Head Behavior Profiling: Assessing each attention head's sensitivity to the initial token's attention scaling. A head is classified as up-effective if increased attention improves accuracy, and down-effective otherwise.
    2. Selective Rescaling: Applying a scaling factor γ, identified by searching for the value that maximizes accuracy on the calibration set, exclusively to the dominant head type.

2. Unsupervised Calibration via Entropy Minimization
    1. a model's output entropy strongly correlates with its accuracy
    2. performing entropy-based search over the current batch of test-time queries.

三、Method

$$\boldsymbol{a} = [a_0, a_1, \ldots, a_{T-1}], \quad \text{where} \quad a_i \geq 0 \quad \text{and} \quad \sum_{i=0}^{T-1} a_i = 1. \tag{1}$$

Here, a_0 is the attention score assigned to the initial token, while $a_1, \ldots, a_{T-1}$ correspond to subsequent tokens. To control the influence of x_0, we introduce a tuning factor $\gamma > 0$ to scale its attention and re-normalize:

$$a'_0 = \frac{\gamma a_0}{D}, \quad a'_i = \frac{a_i}{D} \quad \text{for } i = 1, \ldots, T-1, \tag{2}$$

where the normalization constant $D = \gamma a_0 + \sum_{i=1}^{T-1} a_i = (\gamma - 1)a_0 + 1$.

This rescaling preserves the relative proportions among all non-initial tokens:

$$\frac{a'_i}{\sum_{j=1}^{T-1} a'_j} = \frac{\frac{a_i}{D}}{\sum_{j=1}^{T-1} \frac{a_j}{D}} = \frac{a_i}{\sum_{j=1}^{T-1} a_j}, \quad \text{for } i \geq 1, \tag{3}$$

but compresses or expands their differences as

$$a'_i - a'_j = \frac{a_i - a_j}{D} = \frac{a_i - a_j}{(\gamma - 1)a_0 + 1}, \quad \text{for } i, j \geq 1. \tag{4}$$

四、experiment

Table 1: Performance Comparison of Classification Tasks Across Models. The best performance in each dataset is **bolded** and the ZeroTuning method is highlighted in gray.

Model	Method	Datasets							
		SST2	SST5	MR	BoolQ	CB	TREC	SUBJ	Avg.
Llama-3.1-8B-Instruct	Vanilla	73.20	45.40	89.20	69.60	82.14	14.00	44.60	59.59
	ACT	85.00	43.80	90.80	58.60	82.14	15.80	44.60	60.11
	Auto-PASTA	89.60	47.20	91.40	72.60	83.93	16.00	45.40	63.73
	ZeroTuning	**91.60**	**52.00**	**92.00**	**82.40**	**89.29**	**26.20**	**66.60**	**71.44**
Qwen-2-7B (SDPA)	Vanilla	78.80	45.40	72.40	85.00	78.50	12.60	13.00	55.10
	ACT	–	–	–	–	–	–	–	–
	Auto-PASTA	89.00	47.00	77.70	85.00	**89.29**	14.00	**57.00**	65.57
	ZeroTuning	**89.60**	**47.20**	**87.40**	**86.40**	85.71	**26.60**	54.40	**68.19**
Deepseek-R1-14B (Flash)	Vanilla	91.20	49.40	89.20	83.40	89.29	20.80	50.40	67.67
	ACT	–	–	–	–	–	–	–	–
	Auto-PASTA	92.00	**52.20**	89.80	83.40	**92.86**	22.60	50.40	69.04
	ZeroTuning	**93.00**	51.20	**90.20**	**88.00**	**92.86**	**32.00**	**55.80**	**71.87**

Table 2: Performance Comparison of Multiple-Choice Tasks Across Models.

Model	Method	Datasets							
		MMLU	AQUA	MathQA	LogiQA	CQA	PIQA	ARCC	Avg.
Llama-3.1-8B-Instruct	Vanilla	67.40	25.69	33.60	39.40	77.60	83.60	84.62	58.84
	ACT	67.60	29.64	33.60	38.00	77.60	83.00	84.62	59.15
	Auto-PASTA	67.00	**31.23**	35.20	40.40	78.20	84.60	84.62	60.18
	ZeroTuning	**68.80**	30.43	**36.60**	**42.80**	**80.40**	**85.40**	**85.95**	**61.48**
Qwen-2-7B (SDPA)	Vanilla	69.80	36.76	39.20	45.00	78.80	85.20	86.96	63.10
	ACT	—	—	—	—	—	—	—	—
	Auto-PASTA	69.80	39.13	39.20	45.00	**82.60**	85.40	86.96	64.01
	ZeroTuning	**70.40**	**39.92**	**40.20**	**47.40**	81.80	**86.20**	**87.96**	**64.84**
Deepseek-R1-14B (Flash)	Vanilla	66.60	38.74	38.20	27.80	78.20	84.20	86.62	60.05
	ACT	—	—	—	—	—	—	—	—
	Auto-PASTA	66.60	38.74	39.40	28.20	78.20	84.40	86.62	60.31
	ZeroTuning	**70.00**	**39.13**	**39.80**	**35.60**	**78.60**	**85.00**	**87.29**	**62.20**