RATION: Entropy-Driven Task-Adaptive Visual Attention Allocation Framework for Multimodal Reasoning

Xingle Xu, Fanheng Kong, Dexian Cai, Shi Feng*,
Xiaocui Yang, Daling Wang, Yifei Zhang
School of Computer Science and Engineering, Northeastern University,
Shenyang 110819, China
{xuxingle, kongfanheng}@stumail.neu.edu.cn, 2301840@stu.neu.edu.cn
{fengshi, yangxiaocui, wangdaling, zhangyifei}@cse.neu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) integrate visual encoders with Large Language Models (LLMs) and enable multimodal reasoning. However, for tasks that heavily rely on visual information, the model's utilization of visual information remains unstable, which leads to reasoning failures. Prior works mainly strengthen multimodal reasoning by improving representation alignment or increasing computation. These methods do not explicitly characterize the differences in visual demands across tasks, making it difficult for the model to decide where and how strongly to attend to visual information. Consequently, visual attention allocation for different task becomes a key factor that affects multimodal reasoning. To address these, we propose RATION, an entropy-driven task-adaptive visual attention allocation framework. First, we use a task routing strategy to infer the task type of each sample and identify the key layers. We use visual attention entropy as a control signal to dynamically allocate attention according to task demands. Experiments show that RATION achieves consistent performance gains across diverse reasoning tasks, datasets, and models, providing a clear direction toward more reliable multimodal reasoning. Our code is available at https://github.com/betterfly123/RATION.

1 Introduction

Recently, Multimodal Large Language Models (MLLMs) that follow the alignment paradigm between vision encoders and Large Language Models (LLMs) demonstrate strong capabilities in unifying visual understanding and language reasoning (Bai et al., 2024; Chen et al., 2024b). Under this paradigm, MLLMs make rapid progress in application scenarios such as GUI understanding (Wang et al., 2025b), video understanding (Chen et al., 2024a), and embodied interaction (Li et al., 2024b).

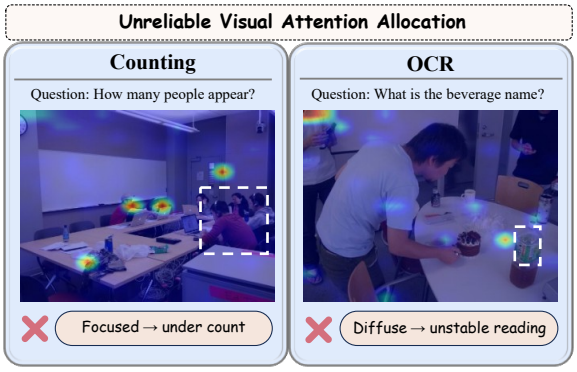

Figure 1: Visual attention allocation mismatch across tasks. We show the attention heatmaps on Qwen3-VL. White boxes indicate misallocated attentions.

Despite substantial progress in multimodal alignment, multimodal reasoning remains fundamentally constrained by the unreliability of visual information utilization, such as deficiencies in visual perception (Tong et al., 2024) and difficulty in utilizing representations (Liu et al., 2025a). In tasks that highly depend on visual evidence, such as counting, spatial reasoning, and OCR, models often fail (Yang et al., 2025a; Schulze Buschoff et al., 2025). These failures arise from the internal visual attention allocation mechanisms within models. As shown in Figure 1, in counting tasks, focused attention pattern leads to missed instances by limiting global coverage, whereas in OCR tasks, diffuse attention prevents the model from focusing on dense textual regions. These observations highlight the importance of matching task-specific visual demands with the model's implicit attention allocation strategy.

To mitigate reasoning failures, existing works mainly encourage models to attend to visual inputs more frequently (Lin et al., 2024; Liu et al., 2024) or allocate additional inference computation (Zhang et al., 2024b; Jia et al., 2024; Yang et al., 2025b). However, these methods often operate at

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 24726–24744
July 2-7, 2026

the token or prompt level, treating the model's internal attention mechanism as a "black box". While these enhancements improve performance in specific scenarios, they do not explicitly model the distributional differences in visual demands across tasks. As a result, models remain unable to decide **where** and with **what** intensity visual information should be attended during inference. This mismatch between task demands and attention allocation leads to both under-utilization and misutilization of visual information, limiting the reliability of multimodal reasoning.

To fill the gap in visual attention allocation, we propose **RATION**, an entropy-d**R**iven t**A**sk-adap**T**ive v**I**sual attenti**ON** allocation framework. Specifically, we design a task routing strategy that automatically infers the task type of each sample from layer activation change patterns within a single forward pass, and identifies key layers that directly affect the task, thereby determining **where** visual attention allocation should be applied. Unlike prior works that require expensive fine-tuning or remains task-agnostic, RATION performs training-free sample-specific modulation of internal attention states. To further determine **what** intensity visual attention should be allocated, we leverage visual attention entropy as a sample specific signal. Entropy serves as an intrinsic measure of the attention distribution (Bao et al., 2024; Zhang et al., 2025), characterizing how focused or diffuse the attention is. Building on this signal, RATION dynamically adjusts visual attention allocation in a task-aware manner. Overall, RATION performs training-free, task-adaptive visual attention allocation during inference, explicitly aligning internal attention behavior with task-specific visual demands. This leads to more stable and reliable utilization of visual information and consistently improves multimodal reasoning capacity. Our contributions are summarized as follows:

- We propose RATION, a training-free framework that leverages attention entropy to perform dynamical visual attention allocation based on the demands of different tasks, improving the reliability of visual information utilization in multimodal reasoning.
- We introduce a task routing strategy based on layer activation changes, enabling task type inference and decisions of task-specific key layers.
- We conduct extensive evaluations on 4 MLLMs and 7 benchmarks, and the results demonstrate that RATION achieve consistently performance gains across diverse MLLMs and tasks, validating its effectiveness and generalization.

2 Related Work

Reasoning in language-only models. Text-only reasoning research mainly focuses on how to elicit and stabilize multi-step computation over discrete linguistic tokens. Classic inference-time strategies, such as Chain-of-Thought and self-consistency, improve reliability by explicitly unfolding intermediate steps and aggregating diverse reasoning traces (Wei et al., 2022; Wang et al., 2023). Tool-based paradigms connect reasoning with external operations and their feedback (Yao et al., 2022). Recently, reasoning models enabled by large-scale math-oriented pretraining and RL-based post-training demonstrate stronger reasoning ability, but also depend more heavily on inference-time compute (Shao et al., 2024; Guo et al., 2025). Unlike text-only reasoning, which is often constrained primarily by the construction and consistency of reasoning chains, multimodal reasoning failures more often arise when the model cannot reliably use key visual information. Consequently, even with sufficient reasoning ability, the model can still produce various errors due to unreliable utilization of visual evidence.

Reasoning in multimodal models. Recent progress in multimodal reasoning mainly manifests as enhancements on both the training side and the inference side. Training-side improvements primarily scale up training (Chen et al., 2024b), refine data selection (Xu et al., 2024; Kong et al., 2025b), and introduce stronger training objectives (Li et al., 2024a; Wang et al., 2026; Kong et al., 2025a) to improve general multimodal capability and reasoning performance. Although these methods yield measurable gains, they do not explicitly constrain how visual information is used. Consequently, models can still produce errors due to unreliable utilization of visual information. On the inference side, existing methods mainly improve performance through different reasoning strategies. MM-CoT (Zhang et al., 2024b) constructs explicit chains of thought to encourage joint use of text and images, and VCTP (Chen et al., 2024c) iteratively performs multi-step operations at inference to strengthen vision interaction. ControlMLLM (Wu et al., 2024) uses learnable latent variables to control the contributions of visual and textual tokens. Another line of work introduces additional information to reinforce

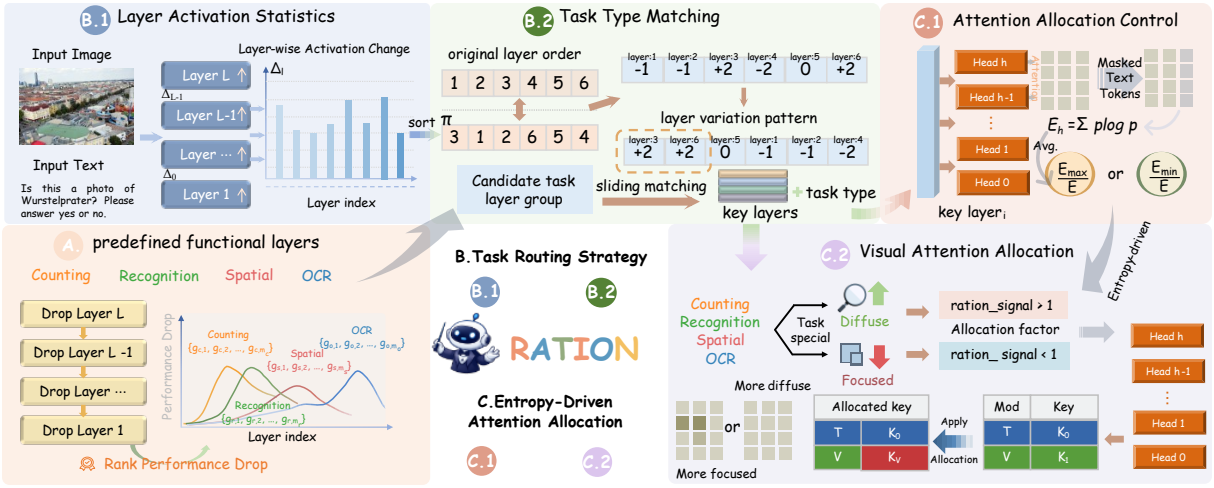

Figure 2: Framework overview of RATION. RATION consists of three components: predefined functional layers, task routing strategy and entropy-driven attention allocation.

attention to visual evidence. DCoT (Jia et al., 2024) decomposes questions to guide the model to focus on key cues, and Socratic-MCTS (Acuna et al., 2025) explicitly formulates reasoning as search and uses MCTS to generate high-value question pairs. MAGIC-VQA (Yang et al., 2025b) retrieves external commonsense knowledge to support question understanding. Despite these advances, existing works often overlook task-specific differences in visual information demand and cannot allocate internal computation accordingly. We address this issue by allocating visual attention in a task-adaptive manner.

3 Method

In this section, we describe how RATION determines where to intervene, and selects what strength to allocate visual attention. As shown in Figure 2, RATION consists of three components: predefined functional layers, task routing strategy, and entropy-driven attention allocation. Predefined functional layers identify functional layers for a specific task. Task routing strategy infers the task type at the sample level and selects the corresponding key layers for subsequent intervention. Entropy-driven attention allocation then uses entropy as a control signal to perform attention allocation at the selected key layers based on the task demands.

3.1 Predefined Functional Layers

MLLMs exhibit functional stratification across decoder depths in their visual capability (Shi et al., 2025). To support subsequent task routing, we draw on prior layer analysis methods (Lin et al., 2025; Wu et al., 2025) by zeroing out the attention states at specific layers and defining task-specific functional layer groups based on the performance changes. The dropping operation is defined:

$$(K_{\mathcal{V}}^{(l)}, V_{\mathcal{V}}^{(l)}) = \mathbf{0},\ \forall l \geq k, \tag{1}$$

where l denotes the layer index, k is the layer where dropping starts, and $\mathcal{V}$ denotes the set of visual token positions. The key and value states at layer l are denoted as $(K^{(l)}, V^{(l)})$. We follow the four tasks taxonomy in (Shi et al., 2025), covering recognition, counting, spatial, and OCR. This taxonomy provides a relatively comprehensive coverage of the common atomic perceptual capabilities that determine the visual-side computational bottlenecks. We define a mapping for the functional layer groups:

$$\mathcal{G}[t] = (g_{t,1}, \ldots, g_{t,m_t}), \tag{2}$$

where $t \in \{recognition, counting, spatial, ocr\} = \mathcal{T}$, and m_t is the layer number for task t. Detailed results are provided in Appendix A.

3.2 Task Routing Strategy

We propose a task routing strategy to infer the task type of each sample and accordingly identify the key layers for subsequent intervention. Since different tasks exhibit distinct visual demands, identifying the task type is necessary to enable targeted attention allocation in later stages. Meanwhile, layers in the model serve different functions, so we must determine where to intervene to avoid applying adjustments to non-key layers and perturbing their normal representation update process.

3.2.1 Layer Activation Statistics

To adaptively infer the task type of each sample, we collect layer activation statistics from a single forward pass. For a given sample, the input sequence has length N, and the set of visual token positions is denoted as $\mathcal{V} \subseteq \{1, \ldots, N\}$. Let $H^{(l)} \in \mathbb{R}^{B\times N\times D}$ denote the output hidden states at layer l, where B is the batch size, and D is the hidden dimension. For an input sample, at layer l, the input and output are denoted as:

$$H^{(l-1)},\ H^{(l)} \in \mathbb{R}^{B\times N\times D}. \tag{3}$$

Since decoder layers typically adopt a residual structure, the difference between the layer input and output directly characterizes the update magnitude. The activation change at layer l as:

$$\Delta^{(l)} = H^{(l)} - H^{(l-1)}. \tag{4}$$

This statistic relies only on the hidden state caches available in a single forward pass and does not introduce extra decoding steps.

3.2.2 Task Type Matching

In layer decoding, the absolute change magnitudes vary across layers, and directly comparing absolute scores does not reliably characterize which layer is relatively more active. Therefore, we introduce the layer relative offset to measure transformation intensity, highlighting the most dramatic updates. Specifically, given the layer change scores $\Delta^{(l)}$, we sort layers in descending order by change scores:

$$\pi = \mathrm{argsort}_{l\in\{1,\ldots,L\}}\big(\Delta^{(l)}\big), \tag{5}$$

where π denotes the change ranking, and L denotes the total number of layers. We then compare each layer position in π with its original layer index, and define their difference as the relative offset:

$$\delta^{(l)} = l - \pi_{(l)}, \tag{6}$$

where $\pi_{(l)}$ denotes the position of layer l in π. We use $\delta^{(l)}$ as a relative activity score and rank layers accordingly to form a layer variation pattern:

$$\mathbf{c} = (c_0, \ldots, c_l) = \mathrm{argsort}_{l\in\{1,\ldots,L\}}\big(\delta^{(l)}\big). \tag{7}$$

We then perform sliding matching between $\mathbf{c}$ and the predefined functional layer group of each task t. For each window size j, we define:

$$\begin{aligned} \mathbf{w}_j &= (c_0, \ldots, c_{j-1}), \\ d(t,j) &= \sum_{u=1}^{j} \mathbb{I}[w_{j,u} = g_{t,u}], \end{aligned} \tag{8}$$

where $d(t,j)$ denotes the hit count within the window, $w_{j,u}$ is the u th element of window w_j, and $g_{t,u}$ is the u th element of the functional layer group for task t. $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the condition holds and 0 otherwise.

$$\begin{aligned} &\hat{t} = t \\ \text{when}\quad &j^* = \min\{\, j \mid d(t,j) = m_t \,\}. \end{aligned} \tag{9}$$

where $\hat{t}$ denotes the sample task type, and j^* denotes the first window size that matches successfully. Through sliding matching, we adaptively select the most suitable task type for each sample and accordingly determine the corresponding key layers $\mathcal{G}[\hat{t}]$. Task routing strategy process runs within a single forward pass, enabling low overhead and adaptive sample level task routing.

3.3 Entropy-Driven Attention Allocation

After the task routing strategy components predicts the task type and the key layers, we further perform adaptive visual attention allocation. As an intrinsic measure of the attention distribution, entropy stably indicates how focused or diffuse attention is, and naturally aligns with allocation mechanism. Therefore, we introduce entropy into attention allocation control, so that the adjustment is driven by each sample's own attention distribution, enabling customized visual attention allocation.

3.3.1 Attention Allocation Control

After identifying the key layers, we compute an attention entropy control signal within these layers and map it to an attention allocation factors. For key layer l, we first compute attention distribution:

$$P^{(l)} = \mathrm{softmax}\left(\frac{Q^{(l)}\big(K^{(l)}\big)^{\top}}{\sqrt{d}}\right) \in \mathbb{R}^{B\times A\times N\times N}, \tag{10}$$

where $Q^{(l)}$ and $K^{(l)}$ denote the queries and keys, respectively, A is the number of attention heads, N is the sequence length, and $\frac{1}{\sqrt{d}}$ is the scaling factor. To avoid interference from text positions, we first restrict the attention weights to $\mathcal{V}$ and renormalize them within $\mathcal{V}$ to obtain a conditional distribution:

$$\tilde{P}^{(l)}_{b,a,n,i} = \frac{P^{(l)}_{b,a,n,i}}{\sum_{r\in\mathcal{V}} P^{(l)}_{b,a,n,r}}, \quad i \in \mathcal{V}, \tag{11}$$

where b denotes the sample index, a denotes the attention head index, n denotes the query position,

and i denotes the key position. Thus, the normalized visual attention entropy for each sample and each head is given by:

$$E_{b,a}^{(l)} = \frac{1}{N \log |\mathcal{V}|} \sum_{n=1}^{N} \left(- \sum_{i \in \mathcal{V}} \tilde{P}_n^{(l)} \log \tilde{P}_n^{(l)} \right). \tag{12}$$

To characterize the overall distribution of entropy on the key layer, we further compute its minimum, maximum, and mean values:

$$\bar{E}_{mean} = \frac{1}{A} \sum_{a=1}^{A} E_{b,a}^{(l)},$$
$$E_{\min}^{(l)} = \min_{a \in A}{}^{0.1} \hat{E}_{b,a}^{(l)}, \quad E_{\max}^{(l)} = \max_{a \in A}{}^{0.9} \hat{E}_{b,a}^{(l)}, \tag{13}$$

where 0.1 and 0.9 denote the quantiles, which mitigate the impact of outlier heads. Based on the entropy statistics, we obtain two control signals.

$$rs_{focused}^{(l)} = \frac{E_{\min}^{(l)}}{\bar{E}_{mean}}, \quad rs_{Diffuse}^{(l)} = \frac{E_{\max}^{(l)}}{\bar{E}_{mean}}, \tag{14}$$

where $rs_{focused}^{(l)}$ is focused allocation factor, $rs_{Diffuse}^{(l)}$ is diffuse allocation factor.

3.3.2 Visual Attention Allocation

This section applies the control signal directly to the attention computation. We perform position allocation only on the key vectors at visual token positions, thereby directly modulating the subsequent attention allocation over the visual region.

We adaptively select the allocation factor based on the sample's task type $\hat{t}$. For the counting task, we increase attention coverage to obtain a more dispersed distribution. For recognition or spatial tasks, we strengthen attention focus to form a more concentrated distribution.

$$\alpha(\hat{t}) = \begin{cases} rs_{Diffuse}^{(l)}, & \hat{t} = counting \\ rs_{focused}^{(l)}, & \hat{t} = rec/spa/ocr \end{cases} \tag{15}$$

where, $\alpha(\hat{t})$ denotes the allocation factor, rec is recognition, spa is spatial. Based on $\alpha(\hat{t})$, we adjust the key vectors at visual token positions:

$$\widetilde{K}^{(l)} = K^{(l)} \odot (1 - z^{kv}) + \alpha(\hat{t})\, K^{(l)} \odot z^{kv}, \tag{16}$$

where, z^{kv} is the binary mask over visual positions. $\odot$ denotes element-wise multiplication. By allocating attention appropriately, the model adopts a better matched attention resource computation pattern for different tasks, thereby effectively improving the reliability of visual information utilization.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate on seven benchmarks. MMMU (Yue et al., 2024) tests broad multimodal knowledge and reasoning over diverse subjects. MME (Fu et al., 2025) evaluates both perceptual understanding and cognitive reasoning. POPE (Li et al., 2023b) focuses on object level recognition and hallucination sensitive perception. SEED (Li et al., 2023a) measures general visual recognition and understanding. OmniSpatial (Jia et al., 2025) targets spatial reasoning and geometric relations. VizWiz (Gurari et al., 2018) evaluates real world OCR and VQA robustness on images captured by blind users. ChartQA (Masry et al., 2022) assesses chart-based question answering that requires reading and reasoning over visualized data. Performance is measured using the standard metrics pertinent to each benchmark. More details are in Appendix B.

4.2 Baselines and Implementation Details

We evaluate our method on four MLLMs, including LLaVA-v1.6-7B (Liu et al., 2023), InternVL3.5-8B (Wang et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a), to assess the stability and transferability of the method. As a training-free approach, we mainly compare against three representative inference-time enhanced training-free baselines. MM-CoT (Zhang et al., 2024b) adopts a two-stage pipeline for explicit multimodal reasoning chain generation. DCoT (Jia et al., 2024) performs joint reasoning via question decomposition. MAGIC-VQA (Yang et al., 2025b) retrieves commonsense knowledge to assist reasoning and is among the most recent methods for multimodal reasoning.

We conduct all evaluations using the lmms-eval framework (Zhang et al., 2024a). For all datasets, we follow the official evaluation protocols and the default prompts and input formats provided by the evaluation scripts, and keep the decoding configuration consistent across methods. For MM-CoT, DCoT, and MAGIC-VQA, we reproduce their original inference pipelines without any parameter updates, and evaluate them in the zero-shot setting to ensure a fair comparison. More baselines and implementation details are provided in Appendix B.

4.3 Overall Analysis

As shown in Table 1, RATION demonstrates consistent overall gains across four MLLMs and seven

Model, Method		General		Recognition		Spatial	OCR		Avg.
		MMMU	MME	POPE	SEED	OmniSpatial	VizWiz	ChartQA	
LLaVA-v1.6	Orign	36.22	1787.59	84.79	62.76	36.07	71.51	54.20	00.00
	MM-CoT	36.05	1659.60	83.84	61.34	30.93	69.25	40.85	07.58 ↓
	DCoT	32.48	1578.30	72.55	58.24	26.84	63.40	43.11	14.44 ↓
	MAGIC-VQA	33.14	1565.70	83.07	62.19	28.06	70.34	45.02	09.23 ↓
	RATION	**37.63**	**1798.20**	**86.12**	**63.33**	**37.18**	**72.72**	**55.41**	02.00 ↑
InternVL3.5	Orign	55.78	2341.45	87.63	77.09	47.16	61.68	87.20	00.00
	MM-CoT	56.03	2311.40	86.09	75.80	40.33	53.09	62.00	08.80 ↓
	DCoT	49.70	1869.32	75.18	72.39	35.33	55.35	48.14	18.79 ↓
	MAGIC-VQA	52.27	2159.45	86.58	76.54	36.58	59.04	53.34	11.65 ↓
	RATION	**56.11**	**2350.14**	**87.98**	**77.17**	**48.23**	**62.59**	**87.67**	00.82 ↑
Qwen2.5-VL	Orign	50.22	2315.82	85.56	77.76	36.86	71.23	83.20	00.00
	MM-CoT	50.11	2290.37	84.52	76.33	31.77	57.28	59.16	09.52 ↓
	DCoT	45.06	1841.36	73.23	73.20	27.66	63.60	36.53	20.40 ↓
	MAGIC-VQA	47.11	2148.30	83.87	77.17	28.51	62.26	49.66	13.10 ↓
	RATION	**50.67**	**2320.92**	**86.90**	**78.26**	**37.96**	**72.01**	**84.08**	01.21 ↑
Qwen3-VL	Orign	69.33	2353.61	86.62	78.15	38.51	67.65	74.08	00.00
	MM-CoT	69.22	2324.14	86.26	76.70	36.20	54.18	50.87	08.70 ↓
	DCoT	62.47	1968.29	73.92	73.88	35.80	60.73	31.52	17.30 ↓
	MAGIC-VQA	64.86	2172.84	85.41	78.10	37.40	59.63	44.10	12.97 ↓
	RATION	**70.11**	**2356.67**	**87.52**	**79.20**	**39.65**	**68.92**	**75.12**	01.41 ↑

Table 1: The zero-shot performance of RATION and baselines. Each dataset evaluated using its official metrics (higher values are better). Significance testing on RATION and the original model results in following p-values: 2.3×10^{-4}, 0.01, 0.001, $2.7 \times 10^{-5} < 0.05$ indicates significant differences. The Avg. column represents the average performance improvement percentage relative to the original model. Green indicates improvement, and red indicates decline. Grey shading indicates the results of our framework, and the best results are highlighted in bold.

benchmarks. In all backbone–dataset combinations, RATION improves upon the original MLLMs, indicating that explicitly allocating visual attention enables models to utilize visual information more reliably during inference. The seven benchmarks cover a wide range of task types, and some of them, such as MMMU and MME, involve hybrid tasks that require multiple abilities simultaneously. Even under such complex evaluation settings, RATION still delivers consistent performance gains. This suggests that our task taxonomy can provide relatively comprehensive coverage of the core demands in multimodal reasoning. At the same time, these improvements remain consistent across diverse tasks further validates the effectiveness of the task-adaptive attention allocation and the robustness of the proposed routing strategy.

In contrast, existing inference enhancement baselines exhibit a performance degradation under this unified evaluation setting. While these methods can yield improvements on the specific tasks they are originally designed for, their gains do not generalize when evaluated across various backbones and tasks without parameter updates. Consistent with recent findings (Liu et al., 2025b; Zeng et al., 2025), test-time scaling strategies based on explicit reasoning chains or prompt heuristics often exhibit insufficient robustness and poor generalization across tasks and models, particularly under zero-shot settings. Our results demonstrate that merely increasing inference computation or introducing external reasoning structures does not reliably strengthen multimodal reasoning. In contrast, our approaches that operate on the model's internal mechanisms, provide a more stable and generalizable path for inference enhancement.

4.4 Ablation Study

As shown in Table 2, we conduct ablation studies of each component of RATION on Qwen3-VL across seven datasets, further validating the effectiveness. **Removing the task routing strategy (w/o Task Routing)** It causes the most pronounced overall degradation. In this setting, the task type is randomly generated, which confuses the decision of whether to encourage focused or diffuse attention. Meanwhile, key layer matching is no longer constrained, so allocation is triggered at inappropriate layers and in inappropriate directions, leading to severe computation mismatches. This results in substantial performance drops across task categories, confirming that task routing is essential for attention allocation. Therefore, visual attention allocation only becomes meaningful after the task demand is correctly identified.

Removing key-layer selection (w/o Key Layer) We allocate over all layers, which also consistently harms performance. Intuitively, a sample typically relies primarily on a certain capability, while still requiring other basic capabilities to support reason-

Method	General		Recognition		Spatial	OCR	
	MMMU	MME	POPE	SEED	OmniSpatial	VizWiz	ChartQA
Qwen3-VL	69.33	2353.61	86.62	78.15	38.51	67.65	74.08
+RATION	**70.11**	**2356.67**	**87.52**	**79.20**	**39.65**	**68.92**	**75.12**
w/o Task Routing	62.34	2327.55	80.03	67.26	32.12	63.10	67.20
w/o Key Layer	68.01	2350.14	84.67	74.29	37.00	64.28	70.17
w Random Allocation	68.92	2350.00	85.95	76.95	37.44	66.32	73.10
w Fixed Allocation	69.10	2345.98	86.20	77.03	35.37	66.40	71.45

Table 2: Ablation study of different components. We conduct experiments with Qwen3-VL as the backbone and report ablation results on seven benchmarks. The evaluation metrics are consistent with the main experiments.

Model	Recognition	Counting	Spatial	OCR
LLaVA_{T}	85.90	87.43	82.69	84.10
$\text{InternVL3.5}_{\text{T}}$	87.13	88.51	90.22	86.29
$\text{Qwen2.5-VL}_{\text{T}}$	87.20	86.35	87.23	81.59
$\text{Qwen3-VL}_{\text{T}}$	87.96	90.18	85.67	89.11

Table 3: Task type prediction result. Each column corresponds to the classification accuracy. We perform task classification for four MLLMs by integrating a task routing strategy.

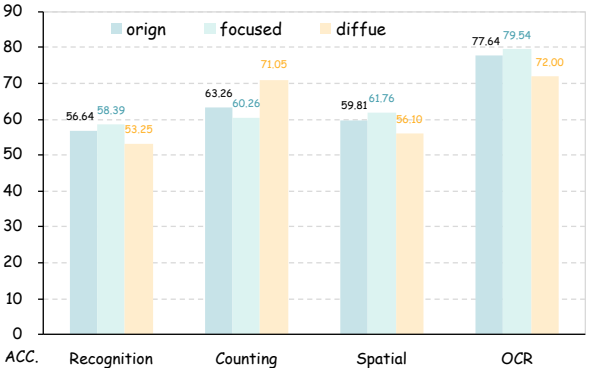

Figure 3: Allocation performance across tasks. We conduct experiments on Qwen3-VL. Focused indicates results under focused allocation for the task, while diffuse indicates results under diffuse allocation.

ing. Applying allocation to all layers can perturb general representation updates that are not specific to the current task yet remain necessary, thereby introducing side effects and degrading overall performance. This result indicates that it is not enough to perform allocation. It must be applied at the correct key layers, otherwise it can disrupt the model's original capability structure.

For the ablation that removes entropy-driven attention allocation, we conduct experiments by replacing it with random and fixed allocation.

Random allocation (w Random Allocation) We randomly sample the allocation factor for attention allocation. Since the ratio is entirely unstable, the allocation strength varies drastically across samples. Moreover, attention adjustment is sensitive: both over-focusing and over-diffusing can lead to an performance drop.

Fixed allocation (w Fixed Allocation) We use constant factors for allocation. The results show that this strategy cannot adapt to the demand differences across tasks, leading to more pronounced fluctuations and degradation across datasets. This indicates that a fixed strategy cannot cover the diverse requirements of visual tasks.

The four ablations show that the components of RATION work jointly. The Task Routing strategy ensures that both the allocation direction and the key layers are correctly identified, while the entropy-driven attention allocation ensures that the allocation is adaptive and stable. As a result, RATION is more robust than the baselines.

4.5 Routing and Allocation Analysis

To further analyze the RATION, we construct a task labeled diagnostic dataset. Specifically, we randomly sample 60 examples from each of MME, SEED, OmniSpatial, and VizWiz, corresponding to four task type. Based on this dataset, we conduct more analyses. The first examines whether the task routing strategy correctly infers the task type. The second examines whether different tasks are better served by focused or diffuse attention allocation.

Task routing strategy analysis As shown in Table 3, we compare the task type predicted by the routing strategy, denoted as $\hat{t}$, with the ground truth labels in our diagnostic dataset, and use classification accuracy to quantify routing quality. Without introducing any additional supervision, the task routing strategy achieves over 80% task type prediction accuracy across all task categories. This result indicates that layer activation variation patterns provide a stable and reliable signal for coarse grained task identification, enabling accurate sample level routing decisions in a single forward pass.

Task allocation demand analysis As shown in Figure 3, we verify task specific needs for attention al-

Figure 4: Attention allocation comparison on Qwen3-VL. We compare the orign with RATION. Green boxes denote the key layers. Red boxes denote the differences in attention to the target visual information.

location. For each task category, we construct two allocation variants that force either focused allocation or diffuse allocation. The results exhibit a clear task dependent trend. Counting benefits substantially from diffuse allocation, while recognition, spatial, and OCR consistently prefer focused allocation. These findings support our design choice in RATION. We encourage more diffuse attention for counting to increase coverage, and strengthen attention focus for the other tasks to improve the utilization of critical visual information.

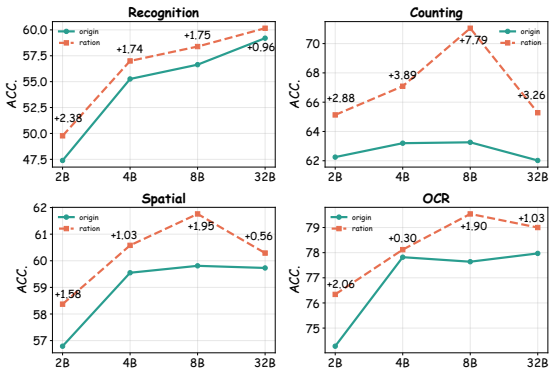

Figure 5: Performance across model scales. We conduct experiments on Qwen3-VL and report accuracy.

4.6 Case Study

As shown in Figure 4, we conduct a case study and use attention heatmaps to analyze how RATION changes visual allocation. Based on the routing results, layers 16/18 are the key layers for counting, and layers 23/24 are the key layers for OCR. We visualize layers 0/16/18/25/35 for counting and 0/23/24/30/35 for OCR. For counting, the original model exhibits strong attention peaks only at a few locations after layers 16/18, which leads to missed instances. With RATION, the allocation at layers 16/18 strengthens attention while keeping it relatively diffuse and covering multiple target locations, thereby reducing object omission errors. For OCR, the original model still distributes attention to irrelevant regions after layers 23/24, making text easy to overlook. RATION at layers 23/24 yields a more concentrated attention pattern that precisely locks onto text regions, enabling more reliable extraction of OCR relevant information.

4.7 Different Scale Analysis

As shown in Figure 5, we evaluate RATION on the Qwen3-VL family at different parameter scales, including 2B, 4B, 8B, and 32B. The results show that RATION achieves performance gains across all model scales and all task categories, indicating that the proposed mechanism is not only effective for models of a specific scale, but can also adapt well to models with different capacities, demonstrating strong cross scale robustness. In addition, although the gains become smaller on some tasks for larger models, the overall performance continues to improve. This suggests that even as the base

model becomes stronger and reasoning ability is enhanced, RATION can still unlock its potential and bring additional gains. Overall, RATION shows strong generalization across different model scales, and its consistent improvements across all task categories further demonstrate its broad applicability.

5 Conclusion

This work targets a key bottleneck of unstable utilization of visual information at inference time in multimodal reasoning, and proposes RATION, an entropy-driven task-adaptive visual attention allocation framework. Unlike prior inference strategies, RATION explicitly matches task demands with visual attention allocation during inference. By coupling a task routing strategy with entropy-driven attention modulation, RATION improves the reliability and stability of visual information usage in multimodal reasoning. Overall, RATION achieves task-adaptive visual attention allocation in a training-free manner during inference, and demonstrates effectiveness and generalization across multiple MLLMs and datasets, offering a new direction toward more reliable multimodal reasoning.

Limitations

Our evaluation mainly focuses on public benchmarks and mainstream open-source MLLMs. In more open-domain real world interactive scenarios, the gains may be affected by input noise and task distribution shift. As a training-free method, our approach is currently better suited as a complementary component. Future work can further explore incorporating this mechanism into training to enhance models' utilization of visual information.

Ethical Considerations

Our research focuses on developing multimodal reasoning methods using public datasets. We do not collect any information from human subjects, private information, or sensitive data. We carefully select datasets that are widely used in the community and comply with their respective license agreements. Our method aims to improve reasoning performance and does not introduce any foreseeable misuse risks or social harms beyond those inherent to MLLMs themselves.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (No. 62272092, No. 62172086), and the Fundamental Research Funds for the Central Universities under Grants (N25XQD004).

References

David Acuna, Ximing Lu, Jaehun Jung, Hyunwoo Kim, Amlan Kar, Sanja Fidler, and Yejin Choi. 2025. Socratic-MCTS: Test-time visual reasoning by asking the right questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24158–24171, Suzhou, China. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Han Bao, Ryuichiro Hataya, and Ryo Karakida. 2024. Self-attention networks localize when qk-eigenspectrum concentrates. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2903–2922.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024a. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024c. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng,

Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. 2024. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Fuzheng Zhang, Guorui Zhou, and 1 others. 2025a. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*.

Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Fuzheng Zhang, and 1 others. 2025b. Tuna: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1810–1839.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Shengzhi Li, Rongyu Lin, and Shichao Pei. 2024a. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14188–14200.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. 2024b. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*.

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

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. 2025a. On the perception bottleneck of VLMs for chart understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10829–10841, Suzhou, China. Association for Computational Linguistics.

Yexiang Liu, Zekun Li, Zhi Fang, Nan Xu, Ran He, and Tieniu Tan. 2025b. Rethinking the role of prompting strategies in LLM test-time scaling: A perspective of probability theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27962–27994, Vienna, Austria. Association for Computational Linguistics.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. 2025. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Cheng Shi, Yizhou Yu, and Sibei Yang. 2025. Vision function layer in multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Peidong Wang, Zhiming Ma, Xin Dai, Yongkang Liu, Shi Feng, Xiaocui Yang, Wenxing Hu, Zhihao Wang, Mingjun Pan, Li Yuan, and 1 others. 2026. Safeqaq: End-to-end slow-thinking audio-text fraud detection via reinforcement learning. *arXiv preprint arXiv:2601.01392*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Ziwei Wang, Weizhi Chen, Leyang Yang, Sheng Zhou, Shengchu Zhao, Hanbei Zhan, Jiongchao Jin, Liangcheng Li, Zirui Shao, and Jiajun Bu. 2025b. Mp-gui: Modality perception with mllms for gui understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29711–29721.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2024. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234.

Qiong Wu, Wenhao Lin, Yiyi Zhou, Weihao Ye, Zhanpeng Zeng, Xiaoshuai Sun, and Rongrong Ji. 2025. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15271–15342.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Shuo Yang, Caren Han, Siwen Luo, and Eduard Hovy. 2025b. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16967–16986.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. 2025. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4651–4665, Vienna, Austria. Association for Computational Linguistics.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Tao Zhang, Jinyong Wen, Zhen Chen, Kun Ding, Shiming Xiang, and Chunhong Pan. 2025. UNIP: Rethinking pre-trained attention patterns for infrared semantic segmentation. In *The Thirteenth International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

A Dropping Experiment

A.1 Experiment Setting

We evaluate the distribution of task functional layers across decoder layers via a dropping experiment. We conduct experiments on POPE, MME, OmniSpatial, and ChartQA following their official evaluation protocols and default metrics. We apply interventions to the visual keys and values in the decoder attention modules. Let the input sequence length be N, and let $\mathcal{V} \subseteq \{1, \ldots, N\}$ denote the set of visual token positions. For the decoder layer l, we denote the key and value tensors as $K^{(l)}$ and $V^{(l)}$. Given a start layer $s \in \{0, \ldots, L-1\}$, we progressively remove visual information by masking the visual keys and values from layer s onward, i.e., for all layers $l \geq s$, we disable $K^{(l)}_{\mathcal{V}}$ and $V^{(l)}_{\mathcal{V}}$. By sweeping s over all layers, we obtain a performance curve as a function of the removal start layer, which reveals where each task most relies on visual information during decoding. We repeat each configuration five times. All other experimental settings remain consistent with the main experiments across all reported results.

A.2 Dropping Results

We present the dropping experiment results across different MLLMs. Table 4, Table 5, Table 6, Table 7 correspond to LLaVA, InternVL3.5, Qwen2.5-VL, and Qwen3-VL, respectively. We observe that even on the same task, applying dropping at the same decoder layer leads to inconsistent performance changes across models. This phenomenon indicates that the injection timing of visual information and the cross-modal fusion pathway during decoding are strongly model specific.

Meanwhile, within the same MLLM, different tasks exhibit markedly different degradation trends when dropping starts from the same layer. This finding suggests that tasks differ in both the strength and the stage of their dependence on visual information, resulting in distinct functional-layer distributions and visual intervention windows. Overall, the results validate the importance of performing task-specific key layer localization.

A.3 Task Layer Group

Based on the experimental results, we summarize the functional layers for each task on each MLLM. Figure 6, Figure 7, Figure 8, Figure 9 show the task functional layers for LLaVA, InternVL3.5, Qwen2.5-VL, and Qwen3-VL, respectively.

B More Details

B.1 Experiment Setting

We conduct evaluation based on the lmms-eval framework. All datasets follow the official evaluation protocols and the default prompts formats provided by the evaluation scripts, and we keep identical inference settings across methods to ensure fair comparisons. We use a zero-shot setting. For decoding, we adopt deterministic generation with do_sample=false and temperature=0. All experiments run in bfloat16 precision. For visual preprocessing, we resize each image with the longest side set to 2048. Experiments are conducted on 8× RTX PRO 6000 GPUs. To reduce evaluation variance and improve robustness, we independently run each configuration 5 times and perform statistical significance analysis on the repeated results.

B.2 Dataset

We introduce the evaluation metrics used by each dataset. MMMU uses accuracy as the primary metric. SEED-Bench uses multiple-choice accuracy. OmniSpatial uses accuracy. VizWiz uses VQA-style accuracy. ChartQA uses relaxed accuracy. POPE uses F1 as the primary metric. MME uses its official overall score as the primary metric, which consists of the scores from the perception and cognition parts. The prompt templates used for each dataset are shown in Figure 10 and Figure 11.

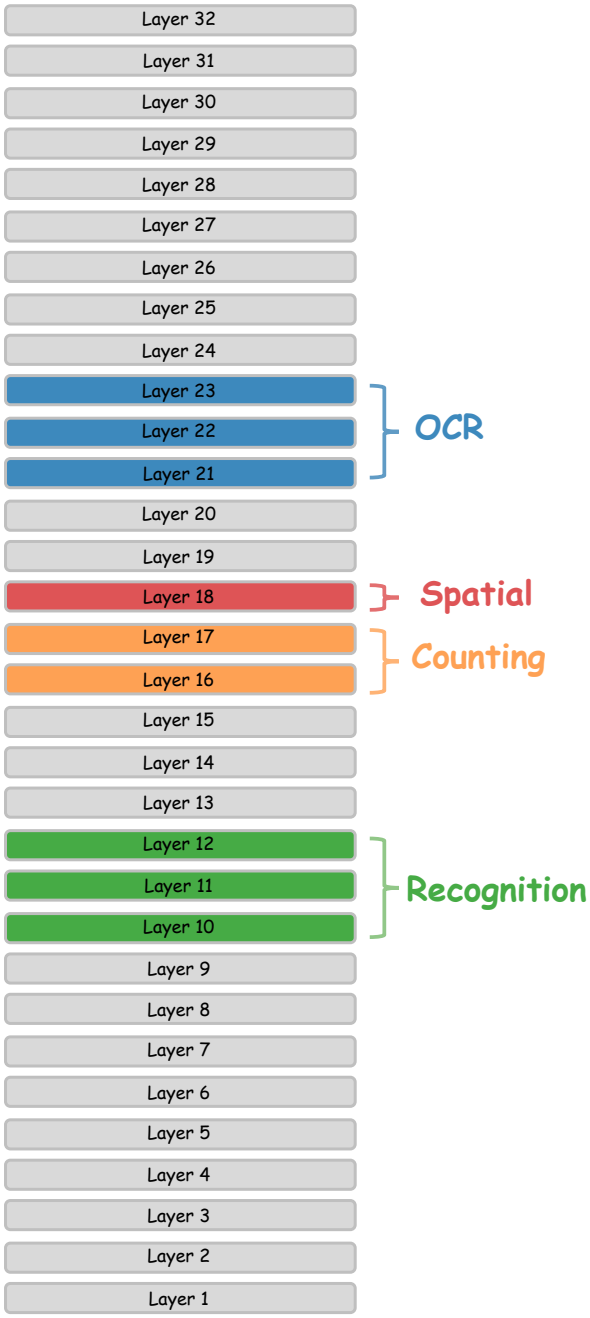

Figure 6: LLaVA task layer group.

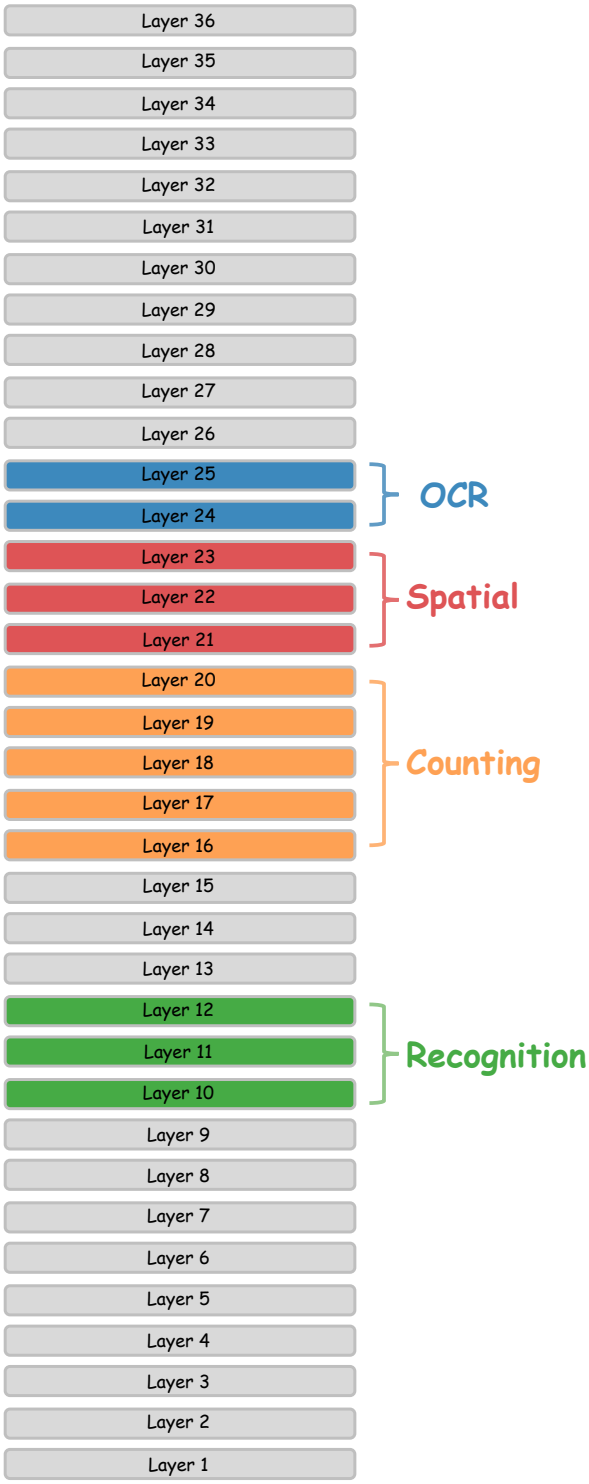

Figure 7: InternVL3.5 task layer group.

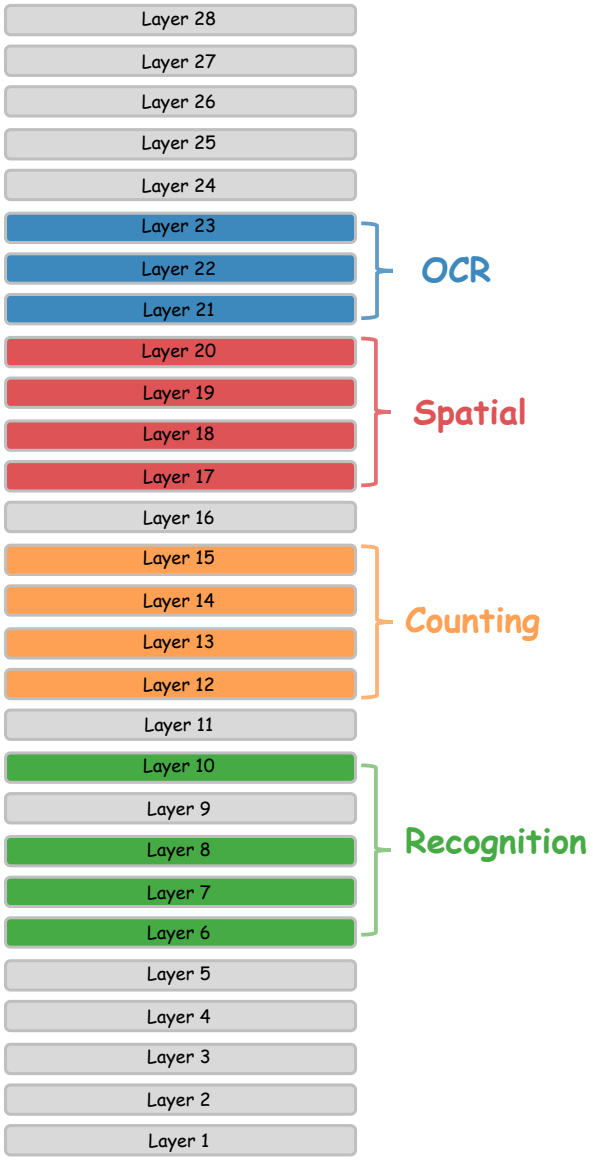

Figure 8: Qwen2.5-VL task layer group.

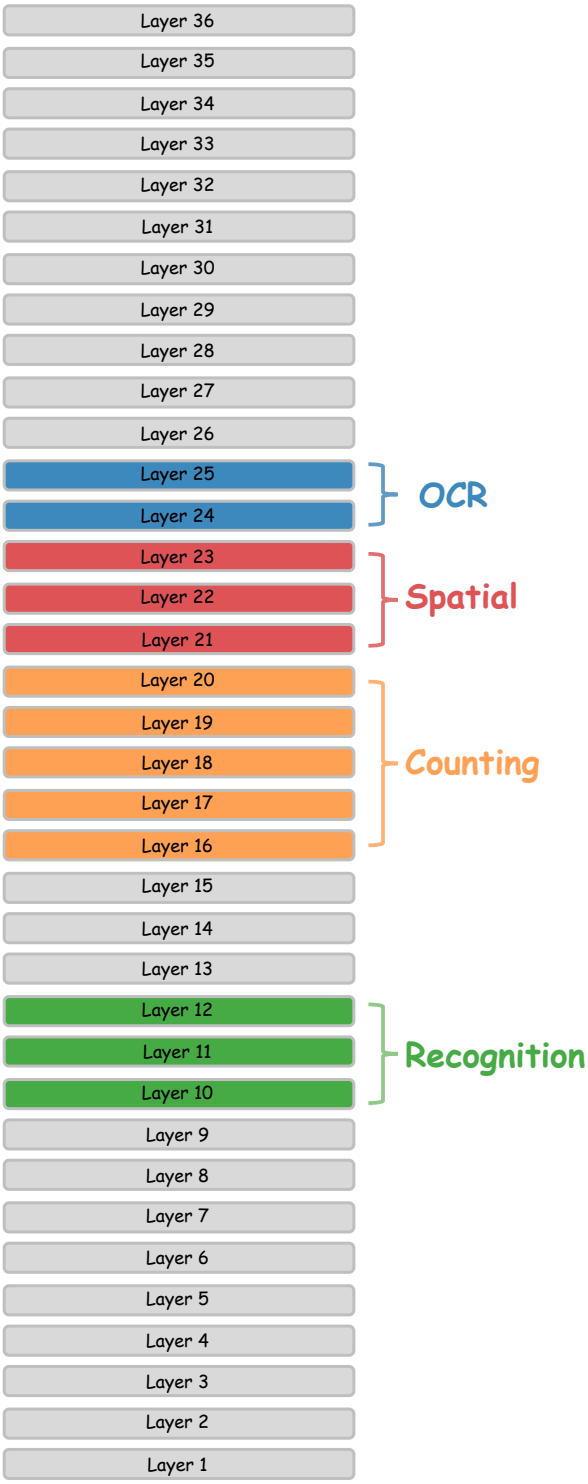

Figure 9: Qwen3-VL task layer group.

MMMU Prompt

{Question}
Answer with the option's letter from the given choices directly.

MME Prompt

{Question}
Answer the question using a single word or phrase.

POPE Prompt

{Question}
Answer the question with a single word.

SEED Prompt

{Question}
Answer with the option's letter from the given choices directly.

VizWiz Prompt

{Question}
When the provided information is insufficient, respond with 'Unanswerable'.
Answer the question using a single word or phrase.

ChartQA Prompt

{Question}
Answer the question with a single word.

Figure 10: Task prompt.

OmniSpatial Prompt

{Question}
You are a spatial-reasoning assistant.

Task
—
You will receive
1. **Image** - a single RGB frame depicting a scene.
2. **Question** - a natural-language query about spatial relationships between objects in the image.
3. **Options** - $\geq$2 answer candidates, each tagged by a capital letter (A, B, C, D...).

Guidelines
———-
Please follow these steps to analyze the image and answer the question:
1. First, carefully observe the image and identify all relevant objects and their spatial relationships.
2. Next, break down the question into key components that need to be addressed.
3. Think through the spatial reasoning step-by-step to arrive at your answer. It may be necessary to transfer perspective to better understand the scene.
4. Finally, select the most appropriate option (A, B, C, or D) based on your analysis.

Always ground your answer in the visual evidence; do not hallucinate unseen objects.
If uncertain, pick the most plausible option—never refuse or reply "insufficient information."

End your answer with a separate line formatted exactly as:

Answer: X
where X $\in$ A, B, C, D.

Figure 11: OmniSpatial task prompt.

Layer	POPE	MME_{count}	OmniSpatial	ChartQA
32	82.00	135.00	37.38	44.00
31	82.00	125.00	35.49	40.00
30	0.00	0.00	28.20	0.00
29	84.00	140.00	33.80	44.00
28	84.00	150.00	35.00	50.00
27	84.00	140.00	35.20	44.00
26	82.00	135.00	36.60	46.00
25	84.00	140.00	35.00	44.00
24	82.00	150.00	35.40	40.00
23	82.00	140.00	36.40	42.00
22	84.00	140.00	24.20	38.00
21	74.00	135.00	32.80	40.00
20	82.00	135.00	35.75	40.00
19	6.00	145.00	28.50	4.00
18	84.00	135.00	36.70	40.00
17	78.00	55.00	29.10	20.00
16	0.00	36.60	0.00	0.00
15	82.00	46.60	0.00	6.00
14	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00
11	10.00	96.60	0.00	26.67
10	50.00	43.30	0.00	8.00
9	0.00	1.60	0.00	2.00
8	6.00	1.60	0.00	8.00
7	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00
4	0.00	3.30	0.00	0.00
3	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00
1	20.00	1.60	0.00	0.00

Table 4: The results of LLaVA dropping.

Layer	POPE	MME_{count}	OmniSpatial	ChartQA
36	87.63	165.00	47.16	71.00
35	87.00	160.00	46.60	70.00
34	86.50	165.00	46.80	70.50
33	62.00	85.00	38.00	47.00
32	85.00	165.00	46.00	67.00
31	80.00	155.00	45.00	67.00
30	81.00	155.00	47.60	67.00
29	81.00	165.00	46.20	66.00
28	80.00	165.00	47.60	65.00
27	83.00	165.00	47.00	66.00
26	86.00	165.00	47.60	61.00
25	84.00	165.00	46.20	57.00
24	82.00	160.00	46.80	24.00
23	82.00	160.00	45.40	24.00
22	91.00	170.00	45.00	25.00
21	23.00	53.30	32.20	4.00
20	59.00	165.00	41.80	17.00
19	83.00	143.30	41.00	20.00
18	86.00	123.30	41.40	17.00
17	88.00	90.00	39.40	19.00
16	86.00	81.60	42.60	17.00
15	81.00	83.30	41.40	23.00
14	64.00	65.00	36.00	15.00
13	70.00	76.60	30.80	15.00
12	60.00	61.60	32.40	14.00
11	72.00	51.60	35.50	18.00
10	66.00	71.60	36.40	17.00
9	66.00	73.30	31.20	16.00
8	65.00	68.30	31.20	18.00
7	61.00	76.60	39.00	14.00
6	38.00	60.00	31.20	17.00
5	0.00	6.60	30.80	0.00
4	0.00	23.30	32.40	3.00
3	54.00	80.00	31.00	11.00
2	56.00	65.00	31.00	13.00
1	0.00	8.30	31.80	0.00

Table 5: The results of InternVL3.5 dropping.

Layer	POPE	MME_{count}	OmniSpatial	ChartQA
28	86.00	170.00	35.12	80.25
27	86.00	170.00	34.18	80.00
26	86.00	170.00	34.69	78.00
25	86.00	170.00	33.40	75.50
24	85.50	170.00	33.76	70.00
23	85.00	170.00	31.20	23.50
22	86.00	169.00	34.60	28.00
21	85.00	165.00	33.10	26.50
20	87.00	165.19	30.00	23.50
19	85.00	165.19	29.00	24.00
18	85.00	159.38	28.20	18.50
17	83.00	158.69	26.85	17.00
16	69.00	100.15	25.34	18.50
15	71.50	96.38	28.00	19.00
14	69.25	66.67	28.00	17.50
13	69.00	62.36	28.00	19.00
12	69.10	61.67	29.18	16.50
11	55.69	59.32	29.36	17.00
10	47.20	88.33	29.20	18.00
9	47.00	68.33	30.00	14.00
8	43.00	91.67	28.60	13.00
7	40.06	53.33	29.40	12.00
6	0.00	0.00	27.80	12.50
5	33.00	50.00	28.20	11.50
4	33.00	67.58	30.96	14.50
3	35.00	100.00	29.80	16.00
2	36.12	132.33	26.80	13.00
1	30.00	135.33	26.80	16.00

Table 6: The results of Qwen2.5-VL dropping.

Layer	POPE	MME_{count}	OmniSpatial	ChartQA
36	86.52	173.33	38.11	73.68
35	86.50	173.33	38.35	73.50
34	86.50	173.33	36.20	73.50
33	87.60	175.00	31.80	72.00
32	87.00	173.33	32.60	73.00
31	87.00	173.33	31.60	71.00
30	86.50	173.33	53.40	69.00
29	86.50	173.33	53.80	68.50
28	86.50	173.33	32.00	69.50
27	86.00	173.33	30.40	70.50
26	86.00	173.33	31.80	69.50
25	86.00	173.33	33.40	58.00
24	85.50	173.33	32.40	26.00
23	85.00	173.33	31.20	27.50
22	86.00	173.33	34.60	26.00
21	85.00	173.33	29.60	26.50
20	87.00	148.33	37.40	20.50
19	80.50	146.67	30.20	21.50
18	81.50	108.33	37.40	19.50
17	83.00	103.33	29.40	19.50
16	67.00	98.33	31.40	19.50
15	71.50	76.67	32.00	18.50
14	65.00	66.67	32.60	18.50
13	69.00	83.33	34.40	19.00
12	51.00	71.67	31.00	17.00
11	50.50	61.67	32.20	17.00
10	48.00	88.33	29.20	18.00
9	43.00	68.33	30.00	14.00
8	43.50	91.67	28.60	14.00
7	45.50	53.33	29.40	12.00
6	0.00	0.00	27.80	0.00
5	47.50	50.00	28.20	16.50
4	50.00	58.33	31.40	16.00
3	38.50	100.00	29.80	14.50
2	45.50	66.67	26.80	12.00
1	75.00	123.33	33.00	15.50

Table 7: The results of Qwen3-VL dropping.