Beyond Answer Symbols: Discovering Key Layers of VLMs for Visual Processing in Multiple Choice Question Answering

1st Zhiling Jin*†
College of Biomedical Engineering and Instrument Sciences
Zhejiang University
Hangzhou, PR China
22415038@zju.edu.cn

2nd Rongge Mao*
School of Biomedical Engineering
University of Science and Technology of China
Suzhou, PR China
ronggemao@mail.ustc.edu.cn

***Abstract*—While multiple-choice question answering (MCQA) has emerged as a standard paradigm for evaluating Vision-Language Models (VLMs), the internal mechanisms through which these models process visual information during decision-making remain largely opaque. Unlike text-only LLMs where critical layers for answer prediction have been established, it is unknown whether VLMs harbor specialized neural circuits dedicated to visual information encoding, or how such visual processing is hierarchically organized across transformer layers. In this study, we employ activation patching and vocabulary projection techniques to systematically investigate visual token processing in state-of-the-art VLMs. Our experiments reveal the existence of *visual processing key layers*—distinct, localized early-to-middle transformer layers where visual information undergoes intensive encoding and integration prior to downstream reasoning. Furthermore, these critical layers exhibit task-specific functional specialization: recognition tasks engage earlier layers (layers 14–16), while grounding and counting tasks recruit progressively deeper layers (layers 15–20). These findings delineate a hierarchical visual-to-linguistic processing architecture in VLMs, wherein distinct visual competencies activate discrete neural circuits at specific architectural depths.**

***Index Terms*—MCQA, VLM, Interpretability**

I. Introduction

Driven by unprecedented investment and soaring interest, multimodal large language models (MLLMs) have advanced rapidly in recent years, demonstrating increasingly sophisticated capabilities in cross-modal understanding and reasoning. Representative systems such as GPT-4V [1], Qwen-VL [2], and DeepSeek-VL [3] have achieved remarkable performance across diverse domains ranging from medical imaging and autonomous driving to business analytics and scientific document processing. Consequently, the rigorous evaluation of these models' capabilities has become a central research priority, with dedicated benchmarks emerging to assess their multimodal reasoning capacities. Among these, multiple-choice question answering (MCQA) has established itself as a particularly valuable evaluation paradigm, offering a controlled setting to assess models' world knowledge, visual understanding, and causal reasoning through structured discrimination tasks. Prominent examples include MMMU [4] and ScienceQA [5] for scientific reasoning, MMBench [6] for comprehensive vision-language evaluation, and GQA [7] for visual question answering.

The MCQA format presents a unique methodological advantage for assessing machine reasoning: by requiring models to select a single correct answer from multiple plausible candidates, it isolates the discriminative reasoning process from open-ended generation, enabling more precise measurement of specific capabilities such as visual grounding, attribute recognition, and relational reasoning. However, the evaluation of MCQA performance extends beyond mere accuracy metrics. Recent advances in mechanistic interpretability have revealed that large language models (LLMs) exhibit distinct internal mechanisms when processing multiple-choice prompts, particularly regarding *answer prediction key layers*—specific transformer layers where the model transitions from processing input information to committing to a specific output selection [8]. These critical layers, typically identified through activation patching and causal mediation analysis, represent neural circuits where answer probabilities are significantly concentrated. Understanding these mechanisms is crucial not only for fundamental interpretability research but also for practical applications such as model compression, targeted fine-tuning, and steering model behavior.

While the existence of answer prediction key layers has been empirically established in text-only LLMs [8], [9], the extension of these findings to vision-language models (VLMs) remains largely unexplored. This gap is particularly significant given the architectural and functional differences between unimodal and multimodal systems. Specifically, we must determine whether answer prediction in VLMs relies on the same transformer layers identified in text-only models, or whether the introduction of visual modalities creates distinct critical pathways that integrate cross-modal representations prior to answer selection.

Moreover, the heterogeneity of visual tasks introduces additional complexity to this investigation. Visual MCQA encompasses diverse cognitive operations: *recognition* tasks

Co-first author*. Corresponding author†

require identifying object categories and attributes; *grounding* tasks demand locating specific entities within spatial configurations; and *counting* tasks involve enumerating visual instances. These distinct competencies likely recruit different neural mechanisms and, consequently, may exhibit differentiated distributions of critical layers across the model's depth. For instance, low-level visual processing might predominantly occur in earlier layers, while complex cross-modal integration necessary for grounding may engage deeper transformer blocks. Understanding these task-specific activation patterns would provide unprecedented insight into the hierarchical organization of visual reasoning in MLLMs, moving beyond aggregate performance metrics to reveal the internal functional architecture of these systems.

In this work, we investigate the mechanistic underpinnings of visual MCQA in large vision-language models through systematic activation patching experiments. By employing causal intervention techniques across different layers of several representative VLMs—including LLaVA [10], Qwen-VL [2], and DeepSeek-VL [3]—we identify not only the answer prediction key layers analogous to those found in text-only LLMs, but also reveal the existence of visual processing key layers uniquely associated with multimodal integration. Furthermore, through disaggregated analysis across distinct visual task categories, we demonstrate that these critical layers exhibit functional specialization: recognition tasks primarily engage early-to-middle layers responsible for visual feature extraction, whereas grounding and counting tasks recruit deeper layers involved in cross-modal alignment and spatial reasoning.

In summary, this paper makes the following contributions:

1) We empirically confirm the existence of answer prediction key layers in VLMs through large-scale activation patching experiments across multiple model architectures. Crucially, we extend these findings beyond the linguistic domain to demonstrate that VLMs additionally possess specialized visual processing key layers dedicated to the integration and encoding of visual information prior to answer selection.
2) We conduct a comprehensive analysis of key layer distributions across fundamental visual competencies, revealing that recognition, grounding, and counting tasks exhibit distinct activation signatures and recruit differentially distributed neural circuits. These findings suggest a hierarchical organization of visual reasoning in MLLMs, wherein different cognitive operations are localized to specific architectural depths.
3) We provide the first systematic characterization of how visual information propagates through transformer-based VLMs during decision-making, offering both theoretical insights into multimodal neural mechanisms and practical implications for efficient model editing and capability-specific model optimization.

https://github.com/HawkFaust/Beyond-Answer-Symbols.git

II. Related Work

MCQA Testing of LLMs. Thilini et al. [11] evaluated the proportional reasoning capabilities of LLMs by formulating analogy problems as multiple-choice questions (MCQs). They tested nine large models on a dataset containing 15,000 samples under zero-shot, few-shot, structured knowledge prompting (SKP), and targeted knowledge prompting (TKP) paradigms, revealing persistent deficiencies in proportional analogy performance. Patrick Sutanto et al. [12] proposed a knowledge distillation framework that leverages LLM-generated MCQ data. By using option selection probabilities from large models to guide the training of smaller models, their approach achieved a 10% improvement in accuracy on the MMLU benchmark. Shi et al. [13] investigated the key layers for visual tasks in question answering by employing a vision token swapping method. Joshua Robinson et al. [9] demonstrated that Multiple-Choice Prompting, in which answer options are bound to symbolic tokens and co-presented with the question, outperforms Cloze-Prompting. Collectively, these studies indicate that MCQA serves as a stable and effective interface for probing the information processing and decision-making behaviors of large language models.

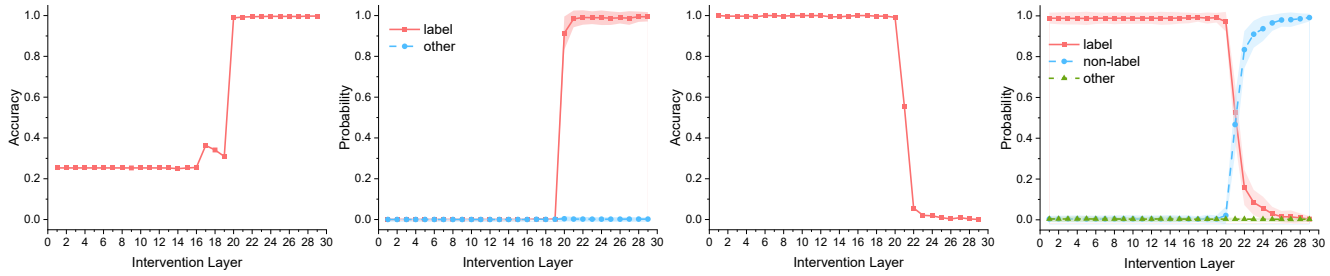
Key Layers in Transformer Models. Walmer et al. [14] analyzed CLIP [15], DINO [16] and MoCo-v3 [17] by examining the attention patterns of the CLS token and the clustering purity of CLS representations across layers. Their results suggest that image feature extraction in different encoder models occurs at slightly different stages, but generally emerges in relatively early layers. However, their analysis was limited to simple clustering-based metrics, leading to conclusions that are relatively coarse and less intuitive. Moreover, they did not systematically examine how different visual task characteristics correspond to distinct key layers, nor did they precisely localize task-specific critical layers. Sarah Wiegreffe et al. [8] investigated the decision mechanisms of LLMs in MCQA settings using three datasets and three open-source language models. Through an analysis of transformer layer activations in response to structured inputs, they identified critical hidden states located before the final token position that determine answer selection, with subsequent layers progressively amplifying the probability mass assigned to the correct answer symbols. Nevertheless, their study is confined to purely language models and does not explore the properties of key layers involved in processing visual information.

III. Method

A. Formatted Visual MCQA

Task Notation. We adopt formatted Visual MCQA (VMCQA), which is analogous to formatted MCQA. Each instance consists of a question, an image relevant to the question, and several answer options presented as an enumerated list, with each option associated with a corresponding answer symbol. The format is as follows:

Instance $= \{\text{Image}, \text{Question}, \{(\text{Symbol}_i, \text{Option}_i)\}_{i=1}^{n}\}$



(a) (b) (c) (d)

Fig. 1. Averaged Accuracy and Probability at each layer of Vocabulary Projection and Activation Patching LastToken for DeepSeek-VL-7B on the ScienceQA Dataset. (a)(b) for Vocabulary Projection and (c)(d) for Activation Patching. "label" represents the ground truth symbol token for Q1, "non-label" represents the ground truth symbol token for Q2, and the "other" denotes averaged result of all remaining symbols in the choose symbols set.

Datasets. We utilize two VMCQA benchmarks: MMMU and ScienceQA. For the MMMU dataset, we exclusively use questions with the `question_type` labeled as multiple-choice. For ScienceQA, we restrict to questions where the `task` is defined as closed choice. Furthermore, we filter for questions that have exactly four answer options and exactly one associated image.

To disentangle the effect of visual understanding and answer prediction from that of complex inference, we additionally construct an *EASY* dataset containing 600 samples from the Oxford-IIIT-Pet dataset [18], selecting images of cats and dogs. Each instance contains the same question with two answer options, where 300 questions have the answer "cat" and 300 have the answer "dog." Testing all models on this dataset yields nearly 100% accuracy. Using this dataset for analysis allows us to effectively control for interference from question and option complexity in our experiments.

Models. We use Qwen2.5-VL-7B-Instruct which contains 28 layers, DeepSeek-VL-7B-chat which contains 30 layers and LLaVA-V1.5-7B which contains 32 layers.

Answer Choice Modification. We apply an Option-Order-Rearrangement modification to answer choices. Following the approach in [19], the options for each question are rearranged to create four different versions per question. For the EASY dataset, only two rearrangements are possible. In total, there are 2,304 questions in MMMU, 2,752 in ScienceQA, and 1,200 in EASY after rearrangement. For all modified questions, we first perform model inference and retain only those questions that are answered correctly.

B. Vocabulary Projection

As discussed in the study by Wiegreffe et al. [8], the dot product between the hidden state output by the language model (LM), denoted as $\mathbf{h}$, and the language model head, denoted as $\mathbf{W}$, defines the vocabulary space. Vocabulary projection is a useful technique to reveal how predictions are formed and provides insight into the causal effect of hidden states on predictions from a different perspective. The steps of vocabulary projection are as follows:

1) For each instance, obtain the final token's output hidden state at layer L, denoted as $\mathbf{h}_L$.
2) Let $\mathrm{LN}(\cdot)$ denote the layer normalization function of the LLM and $\mathbf{W}$ denote the language model head. For each layer l, the layer logits are computed as $\mathbf{z}^l = \mathbf{h}^l\mathbf{W}^\top$, where $\mathbf{h}^l$ is the normalized hidden state at layer l. These logits can be converted into probabilities by applying the softmax function $\mathbf{p}^l = \mathrm{softmax}(\mathbf{z}^l)$. From $\mathbf{p}^l$, we extract the scores corresponding to the choice symbols A/B/C/D, denoted as $p^l_A, p^l_B, p^l_C, p^l_D$, and select the symbol with the highest score as the predicted answer.
3) Repeat step 2 for each layer l.

For overall analysis, let y denote the ground-truth choice symbol for an instance, and $\bar{y}$ denote the set of all other choice symbols. We compute the average of p^l_y, the average of $p^l_{\bar{y}}$, and the accuracy for each layer l.

C. Activation Patching

Activation patching is an effective causal analysis technique used to determine which components of a neural network critically influence model predictions. It has been widely applied by Meng et al. [20] and Vig et al. [21] to analyze the role of last token in language models. In this work, we extend activation patching to operate on visual tokens, called **Visual Tokens Activation Patching**, thereby enriching the original framework with fine-grained causal analysis of visual representations. The procedure is as follows:

1) Run inference on a dataset instance x^{Q_1} with image I_1 that the LLM predicts correctly. Assume the ground truth is symbol $y^{Q_1} = A$ among A/B/C/D. This produces a set of scores where the score for A is higher, and those for B, C, D are lower.
2) Run inference on the same instance x^{Q_2} with another image I_2 and the ground truth $y^{Q_2} = B$ among A/B/C/D, producing a second set of scores similarly with B having the highest score. Store hidden state activations of interest.
3) Run inference on both x^{Q_1} and x^{Q_2} again, but replace the output hidden state at the visual tokens positions of layer l, denoted $\mathbf{h}^l_{x^{Q_1}}$, with the corresponding hidden

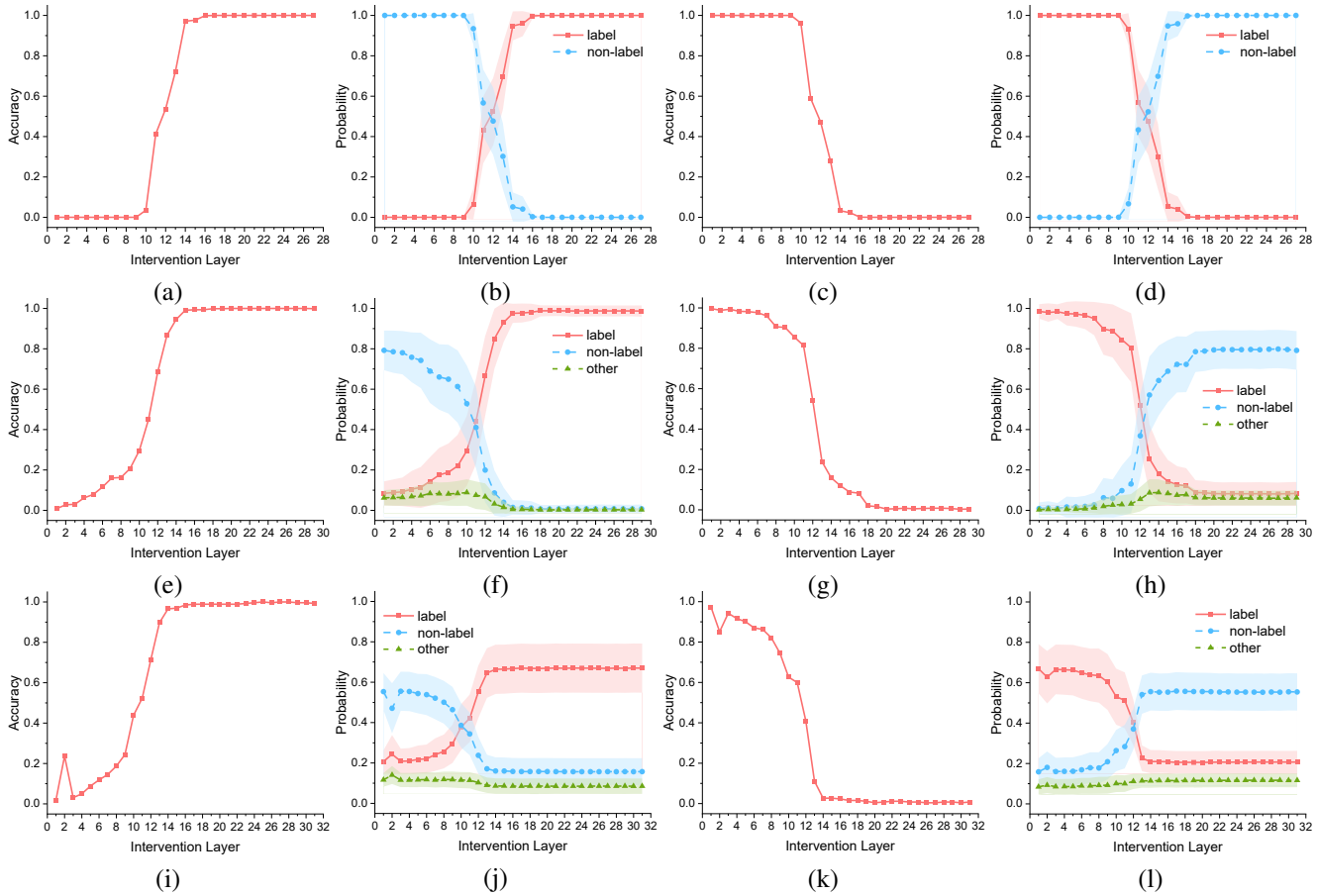


Fig. 2. Averaged Accuracy and Probability at each layer of Activation Patching ImageTokens and ImageTokens-Dual for Qwen2.5-VL-7B on the EASY Dataset (row 1), LLaVA-V1.5-7B on the MMMU Dataset (row 2) and DeepSeek-VL-7B on the ScienceQA Dataset (row 3). (a)(b)(e)(f)(i)(j) for Image Tokens Activation Patching and (c)(d)(g)(h)(k)(l) for Image Tokens Dual Activation Patching . The meanings of "label", "non-label", and "other" are as explained in Fig. 1.

state from x^{Q_2}, denoted $\mathbf{h}^l_{x^{Q_2}}$. Measure and record the scores $p^l_A, p^l_B, p^l_C, p^l_D$ after patching.

4) Repeat step 3 for each layer l.

For overall analysis, denote y^{Q_1} and y^{Q_2} as the ground-truth symbols for x^{Q_1} and x^{Q_2}, respectively, and let $\bar{y}$ denote the other symbols. We average the scores corresponding to y^{Q_1}, y^{Q_2}, and $\bar{y}$ across all trials and layers, obtaining $p_{Q_1\text{-label}}$, $p_{Q_2\text{-label}}$, and $p_{\text{other-mean}}$.

Because VLMs involve complex visual information processing, solely conducting image tokens activation patching is insufficient. Therefore, we introduce **Visual Tokens Dual Activation Patching**, where we use x^{Q_2} and x^{Q_1} to denote question-image pairs with mismatched and correctly matched images, respectively. Specifically, we first perform inference using the mismatched image (x^{Q_2}) and then replace the visual token activations layer-by-layer with those from the correctly matched image (x^{Q_1}). This allows us to observe the processing of visual information in VLMs from an alternative perspective during inference.

IV. Experiment and Results

Our comprehensive experiments covers all three models across all three datasets. However, for a streamlined presentation in the following sections, we illustrate the result of each model on a distinct dataset. This approach allows us to efficiently demonstrate the generalizability of our findings across both model architectures and task domains without redundant duplication of results.

Key Layers of Answer Symbol. From the results of vocabulary projection shown in Fig.1 a and b, DeepSeek-VL-7B begins to assign increasingly higher probabilities to the choice symbols starting from layers 18 to 20, with the probability assigned to the ground-truth answer symbol being significantly higher than those assigned to other choice symbols. Prior to these layers, the softmax probabilities corresponding to choice symbols remain very small and similar.

Qwen2.5-VL and LLaVA-V1.5 exhibit similar phenomena; however, the positions of the key layers differ, occurring at layers 20 to 22 for Qwen2.5-VL and layers 14 to 16 for LLaVA-V1.5. This indicates that the location of the critical answer prediction layers varies across different models.

The results from activation patching corroborate the findings from vocabulary projection (Fig.1 c and d). When the last token activation is patched before layer 20, DeepSeek-VL's accuracy remains nearly unchanged, with the probabilities

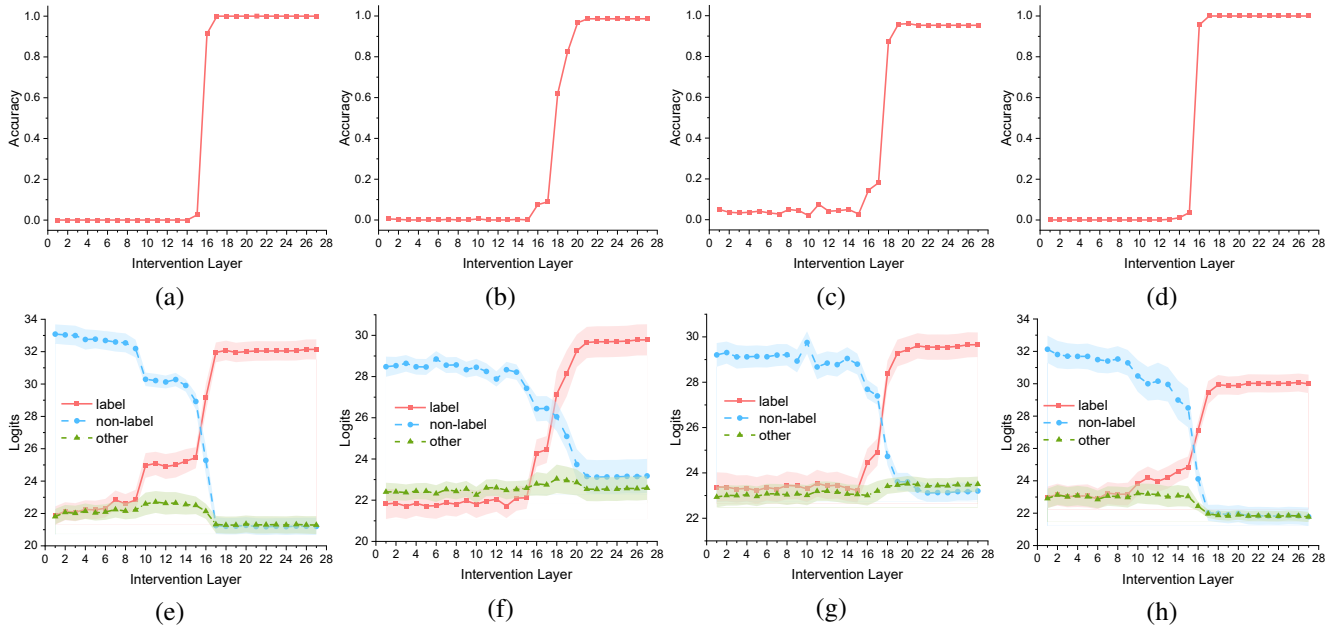


Fig. 3. Averaged Accuracy (row1) and Logits (row2) at each layer of Activation Patching ImageTokens for Qwen2.5-VL-7B on four visual tasks. (a)(e) for Color recognition, (b)(f) for counting, (c)(g) for grounding and (d)(h) for shape recognition. The meanings of "label", "non-label", and "other" are as explained in Fig. 1.

for the correct answer symbol consistently higher than for other options. However, patching after layer 20 causes the probabilities for the original answer option to decrease, while those for the replaced option increase. At layer 22, the order of probabilities reverses, resulting in a sharp drop in accuracy to zero. Qwen2.5-VL and LLaVA-V1.5 show similar behaviors, with model-specific differences consistent with those observed in vocabulary projection.

This observation aligns with the experimental findings of Wiegreffe et al., further indicating that VLMs share the same property. The EASY and MMMU datasets show analogous patterns, with key layers aligning with those found in the ScienceQA dataset. This suggests that the layers for predicting answer symbol are dataset-independent, providing valuable insight that could inform applications such as accelerating LLM inference.

Key Layers of Visual Information Processing. We perform activation patching and dual activation patching on the vectors at the image token positions as shown in Fig.2. The results indicate that for Qwen2.5-VL on the EASY dataset, replacing image token vectors after approximately layer 14 does not affect model accuracy or the probabilities of choice options, suggesting that the model ceases processing image information beyond this point. Conversely, between layers 9 and 13, accuracy and probabilities begin to show gradual increases, identifying these layers as critical for image token information processing. LLaVA-V1.5 and DeepSeek-VL exhibit similar trends, with the key difference being the precise location of the critical layers, indicating model dependence. Additionally, the same model ceases visual information processing at an identical layer across different datasets, suggesting that this property is dataset-agnostic.

V. Key Layers for Multiple Visual Tasks

To further investigate the distribution of key layers across different classic visual tasks, we selected four specific visual tasks: grounding, counting, color recognition, and shape recognition. We designed controlled VMCQA tasks with single variables for each to localize the key layers for these tasks.

Test Dataset Synthesis Pipeline. We designed a unified test dataset synthesis pipeline for each specific visual task. For each test sample, the background is a blank image, uniformly divided into four quadrants by light gray crosshairs, resembling a coordinate system. Shapes are placed in the image according to specific attributes. The set of shapes includes square, triangle, circle, and star. The color set includes red, blue, green, and yellow. The position set includes top-left, top-right, bottom-left, and bottom-right. The quantity set includes 1 to 4 objects. For each task, we iterate over combinations of the other attribute sets. For each combination, we generate four images differing only in the attribute corresponding to the task, while all other attributes are identical. These four images are then paired to form image pairs for vision activation patching. For example, for the counting task, we iterate over the shape, color, and position attributes, resulting in 64 unique combinations. For each combination, we generate images with 1, 2, 3, and 4 objects, creating 6 image pairs through pairwise combination. This unified data synthesis paradigm ensures that the evaluation across different visual tasks is consistent.

Experimental Results. From the changes in accuracy and the logits plots in Fig.3, we observed that at certain layers, the logits corresponding to the answer choice symbols for the

original and patched images undergo significant shifts, while at other layers they remain relatively stable. This eventually causes a reversal in the dominance of the token logits, manifesting as an abrupt change in accuracy. The layers where these significant logits shifts occur are identified as the key layers for the visual task. For the color and shape recognition tasks, despite targeting completely different object attributes, they share almost identical key layers. These are located in a narrow band around layers 14-16, with a minor shift also observed around layers 8-9. The layer with the steepest slope of change is layer 15. For the grounding task, its key layers follow closely after those of the recognition tasks, located in a slightly wider band from layers 15-18, with the steepest slope at layer 17. For the counting task, key layers span layers 15-20, with the steepest change also at layer 17. This task exhibits the longest band of key layers, causing the accuracy mutation to occur at a later stage.

Analysis of Results. For these fundamental visual tasks—recognition, grounding, and counting—there exist relatively narrow band of key layers. Within these layers, the model performs intensive visual information processing related to the image content. Outside these layers, virtually no visual information processing occurs for these specific task attributes. The processed information is then passed to the final several layers for reasoning and answering the MCQA question. This reveals the hierarchical information processing characteristics of Transformer-based MLLMs in MCQA tasks, both across different visual subtasks and from vision to language.

VI. Conclusion

Through experiments on multiple open-source VLMs across several datasets, we find that, analogous to text-based LLMs, VLMs exhibit answer prediction key layers during VMCQA tasks. Moreover, VLMs also possess critical layers dedicated to processing visual information. Specifically, our investigation into distinct visual tasks—including grounding, counting, color recognition, and shape recognition—reveals that each task is associated with a narrow band of key layers where visual information is intensively processed. These layers vary in location and width across tasks, reflecting a hierarchical organization of visual-to-linguistic processing in Transformer-based VLMs. Our experimental results contribute to a deeper understanding of VLM behavior on VMCQA tasks, which may facilitate applications such as inference acceleration, task-specific model editing, and answer calibration.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[3] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al., "Deepseek-vl: towards real-world vision-language understanding," *arXiv preprint arXiv:2403.05525*, 2024.

[4] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al., "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*, 2024, pp. 9556–9567.

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

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

[7] Drew A Hudson and Christopher D Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[8] Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal, "Answer, assemble, ace: Understanding how LMs answer multiple choice questions," in *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Joshua Robinson and David Wingate, "Leveraging large language models for multiple choice question answering," in *The Eleventh International Conference on Learning Representations*, 2023.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26296–26306.

[11] Thilini Wijesiriwardene, Ruwan Wickramarachchi, Sreeram Reddy Vennam, Vinija Jain, Aman Chadha, Amitava Das, Ponnurangam Kumaraguru, and Amit Sheth, "Knowledgeprompts: Exploring the abilities of large language models to solve proportional analogies via knowledge-enhanced prompting," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 3979–3996.

[12] Patrick Sutanto, Joan Santoso, Esther Irawati Setiawan, and Aji Prasetya Wibawa, "LLM distillation for efficient few-shot multiple choice question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, Nov. 2025, pp. 8502–8530.

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

[14] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava, "Teaching matters: Investigating the role of supervision in vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7486–7496.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[17] Xinlei Chen, Saining Xie, and Kaiming He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9640–9649.

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[19] Haochun Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu, "Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models," *arXiv preprint arXiv:2402.01349*, 2024.

[20] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov, "Locating and editing factual associations in gpt," *Advances in neural information processing systems*, vol. 35, pp. 17359–17372, 2022.

[21] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber, "Investigating gender bias in language models using causal mediation analysis," *Advances in neural information processing systems*, vol. 33, pp. 12388–12401, 2020.