How Do Large Vision-Language Models See Text in Image? Unveiling the Distinctive Role of OCR Heads

Ingeol Baek[1], Hwan Chang[1], Sunghyun Ryu[2], Hwanhee Lee[1†]
[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[2]Department of Computer Engineering, Sejong University, Seoul, Korea
`{ingeolbaek, hwanchang, hwanheelee}@cau.ac.kr`
`ryusunghyun1002@sju.ac.kr`

Abstract

Despite significant advancements in Large Vision Language Models (LVLMs), a gap remains, particularly regarding their interpretability and how they locate and interpret textual information within images. In this paper, we explore various LVLMs to identify the specific heads responsible for recognizing text from images, which we term the Optical Character Recognition Head (**OCR Head**). Our findings regarding these heads are as follows: (1) Less Sparse: Unlike previous retrieval heads, a large number of heads are activated to extract textual information from images. (2) Qualitatively Distinct: OCR heads possess properties that differ significantly from general retrieval heads, exhibiting low similarity in their characteristics. (3) Statically Activated: The frequency of activation for these heads closely aligns with their OCR scores. We validate our findings in downstream tasks by applying Chain-of-Thought (CoT) to both OCR and conventional retrieval heads and by masking these heads. We also demonstrate that redistributing sink-token values within the OCR heads improves performance. These insights provide a deeper understanding of the internal mechanisms LVLMs employ in processing embedded textual information in images.

1 Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2023; Li et al., 2023; Dong et al., 2024; Liu et al., 2024a) extend the capabilities of Large Language Models (LLMs) (Zhao et al., 2023; Achiam et al., 2023; Grattafiori et al., 2024) into multimodal domains. Typically, an LVLM integrates a pretrained vision encoder with pretrained LLM decoder via an adapter, enabling it to generate coherent text grounded in both image and text inputs. The power of LVLMs lies in their deep understanding of image-text relationships, which is evi-
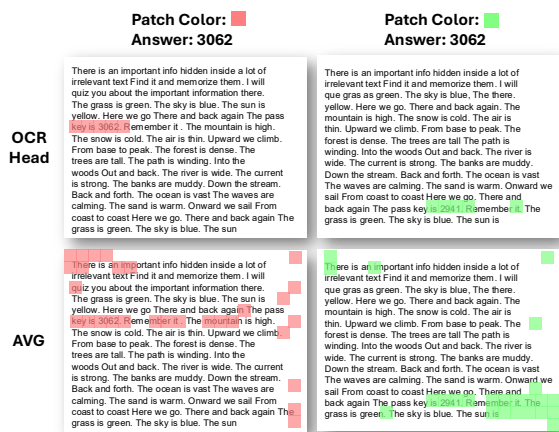

[†]Corresponding author.

Figure 1: Visualization of image attention maps from InternVL2-8B. L and H denote the layer and attention head of the LVLM, respectively.

dent in tasks like VQA (Antol et al., 2015; Goyal et al., 2017) and particularly those relying on text embedded in images, such as OCR-VQA. LVLMs have demonstrated impressive performance on these OCR-centric tasks (Mathew et al., 2021; Huang et al., 2024; Reddy et al., 2024). Despite these advancements, there is limited interpretability research investigating the internal mechanisms driving this performance, specifically how LVLMs handle textual information within images. Recent interpretability work has identified *retrieval heads* in LLMs—attention units that copy relevant tokens from long textual contexts to support long-context factuality (Wu et al., 2024b)—and visual grounding heads that localize image regions in response to textual queries (Xiao et al., 2024; Wu et al., 2024a; Kang et al., 2025b). However, these studies do not address how LVLMs internally locate and extract embedded text within images.

We hypothesize that LVLMs contain a distinct class of attention heads—*Optical Character Recognition (OCR) heads*—that operate independently of copy-paste retrieval heads. While retrieval heads attend to text tokens in the input sequence and replicate them in the output, OCR

heads should selectively attend to visual patches corresponding to characters or words and directly guide text extraction from images. To test this hypothesis, we convert the Passkey and Needle-in-a-Haystack (NIAH) (Kamradt, 2023) benchmarks into a multi-image QA setup requiring model to answer by copying text from rendered image patches. By analyzing attention distributions during answer generation (see Figure 1), we observe that certain heads consistently concentrate on the ground-truth text regions, supporting the existence of specialized OCR heads. Based on these observations, we formulate two core research questions:

RQ1. How can we identify the OCR heads?

RQ2. How do OCR heads differ from existing retrieval heads?

To answer these questions, we introduce an algorithm for detecting OCR heads in multi-image LVLMs. Our experiments reveal three key properties that distinguish OCR heads from retrieval heads: (1) Less Sparsity: OCR heads activate more densely across instances compared to the 3–5% active rate of conventional retrieval heads. (2) Qualitative Distinctiveness: OCR heads exhibit low overlap with retrieval heads, indicating they form an independent functional group. (3) Static Activation Patterns: OCR head activations remain consistent across diverse contexts, unlike context-sensitive retrieval heads.

To validate the specialized role of OCR heads, we evaluate their behavior in downstream tasks via CoT (Wei et al., 2022; Li et al., 2025) prompting, attention masking, and attention redistribution. These studies not only confirm the mechanistic role of OCR heads in reading embedded text but also demonstrate that manipulating their attention distributions can improve OCR-VQA performance. Our findings fill a critical gap in LVLM interpretability and provide actionable insights for enhancing multimodal reasoning and reducing hallucination in OCR-centric applications. We summarize our contributions as follows:

- We propose a scoring-based method for automatically identifying OCR heads in LVLMs.

- We demonstrate that OCR heads exhibit reduced sparsity, qualitative distinctiveness, and static activation patterns across diverse contexts.

- Through CoT prompting, targeted head masking, and strategic sink-token redistribution, we confirm OCR heads' specialized role in text extraction from images.

2 Preliminaries

2.1 Retrieval Heads

Retrieval heads are identified as a specific type of attention head primarily responsible for identifying and copying relevant information from the input context to the generated output. (Wu et al., 2024b) These heads are understood to play a crucial role in the model's ability to retrieve factual information, particularly from long input sequences. They are mechanistically linked to the process of extracting and repeating input tokens.

2.2 Retrieval Score

To quantitatively identify which attention heads are responsible for copying relevant tokens from the input, a retrieval score has been defined. (Wu et al., 2024b) This score serves to measure how consistently a specific attention head engages in this copy-paste behavior. Given a question q, an answer k, and a context x, the model addresses a NIAH task—locating k within x based on q. Let h denote attention head, w the generated token, and $a \in \mathcal{R}^{|x|}$ attention score; then two conditions must be satisfied: (1) the generated token w is part of the target sentence k, and (2) the input token x_j with the highest attention aligns exactly with w, formally $x_j = w, \quad j = \mathrm{argmax}(a)$. Under these conditions, the retrieval score for a head h is calculated as:

$$\text{Retrieval Score for head } h = \frac{|g_h \cap k|}{|k|} \quad (1)$$

where g_h represents the set of all correctly identified (copied-and-pasted) tokens by head h. Intuitively, this retrieval score measures the token-level recall capability of a specific head. Retrieval score is computed over many NIAH instances, and heads achieving a score above 0.1 are classified as retrieval heads.

3 Identifying Heads for Text Recognition

In this section, we describe our method for identifying attention heads within LVLMs that are primarily responsible for retrieving textual information from images. We term these specialized heads as Optical Character Recognition Heads (OCR heads). We adapt the concept of the retrieval score (Wu et al., 2024b), originally used to quantify copy-paste behavior from text contexts
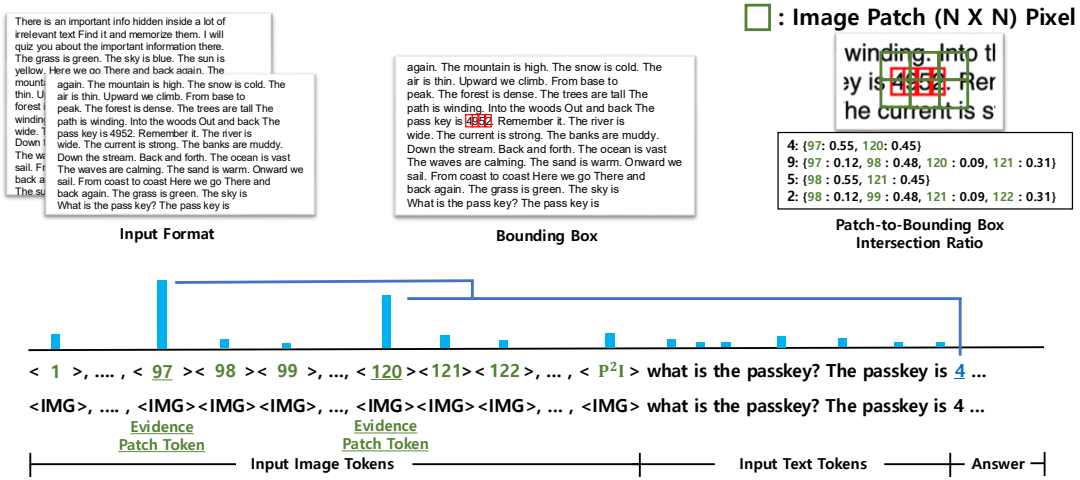
Figure 2: The figure illustrates the image input format designed to identify tokens responsible for copy-pasting text within the image. Tokens corresponding to the patch-to-bounding box intersection ratio serve as evidence patch tokens. The retrieval score computation utilizes attention between generated tokens and evidence patch tokens.

in LLMs, to the multimodal domain of LVLMs processing image inputs. Our adapted metric, the OCR score, measures how consistently a head retrieves text from image regions. Heads with higher OCR scores are more frequently involved in copying token information from image patches corresponding to text.

3.1 Constructing Image Passkey and Needle in a Haystack Dataset

To identify heads involved in image text retrieval, we generate an image-based passkey and NIAH dataset. The original task embeds a specific phrase—for example, "the passkey is [number]"—within a long text context and asks a question requiring the model to retrieve the number. We convert this text-based task into an image-based one by rendering the text context into multi-images.

To convert text data into image data, we render the text content into images, breaking lines and moving to the next image based on preset rules for length and line count. We then generate bounding boxes around the two parts of the ground-truth answer-the passkey numbers and the needle sentence-within each image using a rule-based approach. This involves detecting the digits of the answer text and calculating their coordinates based on text width and line position. This process yields separate bounding boxes x_min, y_min, x_max, and y_max for each character of the correct answer.

As shown in the bounding box of Figure 2, this

process generates images with associated bounding box information for the answer. Note that while bounding boxes are used for analysis like calculating intersection with patches, they are not provided as input to the model during inference.

3.2 Patch Token Preprocessing

As in Figure 2, we process each image into a sequence of patch tokens by the visual encoder. The encoder scans the image from top-left to bottom-right, generating one token representation for each non-overlapping patch of size $N \times N$. If an image has dimensions $w \times h$ and patch size $N \times N$, it yields $w/N \times h/N$ tokens. For example, a 294×196 image with 14×14 patches results in $294/14 \times 196/14 = 294$ tokens. When processing multiple images, each resized to a fixed dimension, the total number of patch tokens is $w/N \times h/N \times I$, where I is the number of images. These patch tokens form the visual part of the model's input sequence.

3.3 Evidence Patch Tokens

We adjust the bounding box positions of the ground truth answer based on any resizing applied to the original images. For example, if an image is resized from 588×392 to 294×196, the bounding box coordinates are scaled by 0.5. After scaling, we calculate the Intersection over Union IoU between each answer bounding box and each image patch token. Figure 2's "Patch-to-Bounding Box Intersection Ratio" illustrates this overlap. Patch tokens with an IoU greater than 0.1 with any
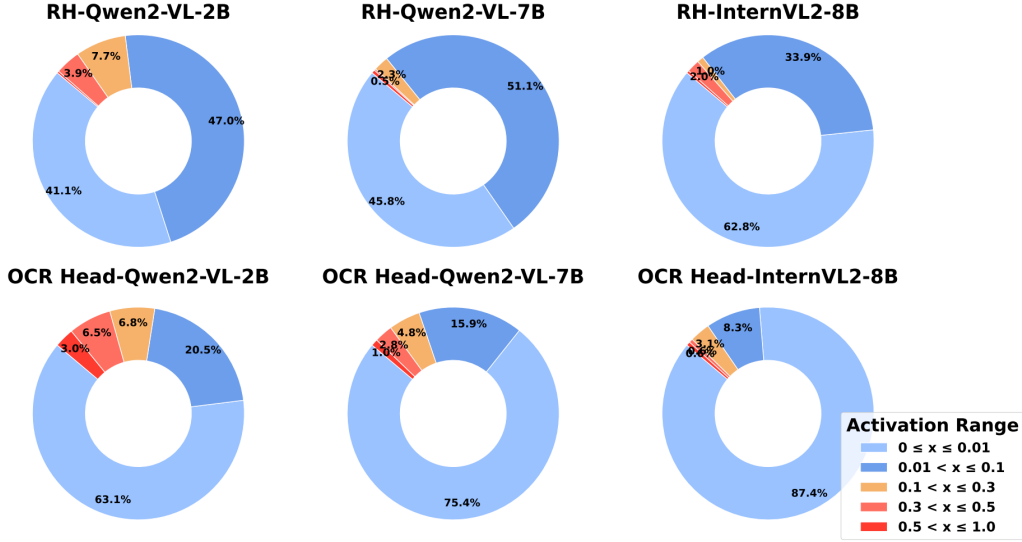
Figure 3: Proportion of OCR Heads and Retrieval Heads (RH) identified in the InternVL2 and Qwen2-VL models.

ground truth answer bounding box are designated as *Evidence Patch Tokens*. These tokens represent the image regions containing the text we expect the model to retrieve. Patch tokens with IoU below 0.1 are discarded from this set.

3.4 OCR Score

Following the concept of retrieval score introduced in (Wu et al., 2024b) to identify copy-paste behavior from text contexts, we define the OCR score to quantify this behavior when retrieving text from images. The original retrieval score measures when a head's highest attention aligns with an input token that is part of the correct generated output, specifically from a text context.

Our OCR score adapts this concept for image input by focusing on attention to Evidence Patch Tokens. Let q be the question, k be the correct answer text, and w be a generated token. Let E be the set of Evidence Patch Tokens identified through the Patch-to-Bounding Box Intersection Ratio. We assign the OCR score to a head when the following two conditions hold for a generated token w:

1. The generated token w must be part of the correct answer text k: $w \in k$.

2. The patch token x_j in the input sequence that receives the highest attention a for w must belong to the set of Evidence Patch Tokens E: $x_j \in E$, where $j = \arg\max(a)$.

The OCR score for a head h is then calculated using the same token-level recall formula as the orig-

inal retrieval score:

$$\text{OCR Score for head } h = \frac{|g_h \cap k|}{|k|} \quad (2)$$

where g_h is the set of generated tokens for which head h met the two conditions above, i.e., correctly "copied" a token by attending to an Evidence Patch Token. This score measures the token-level recall capability of a specific head for retrieving text from image regions identified as evidence.

3.5 Detecting OCR Head

We calculate the OCR score for each attention head across a diverse subset of the image passkey and NIAH dataset. This dataset includes instances with varying "context lengths" simulated using 2 to 12 images. All hyper-parameter settings are kept identical to those previously used for configuring retrieval heads. We analyze 1,200 sampled instances at a fixed ratio. To classify a head as an OCR head, we require it to achieve an OCR score greater than 0.1 in at least 10% of these instances, i.e., in 120 or more cases. For any head meeting this frequency criterion, we then compute its average OCR score across all 1,200 instances. If this average score also exceeds 0.1, that head is classified as an OCR head.

4 How OCR Heads Differ in Their Operational Characteristics from Retrieval Heads

In this section, we discuss three key properties that distinguish the identified OCR heads from conventional retrieval heads: 1) reduced sparsity, 2)
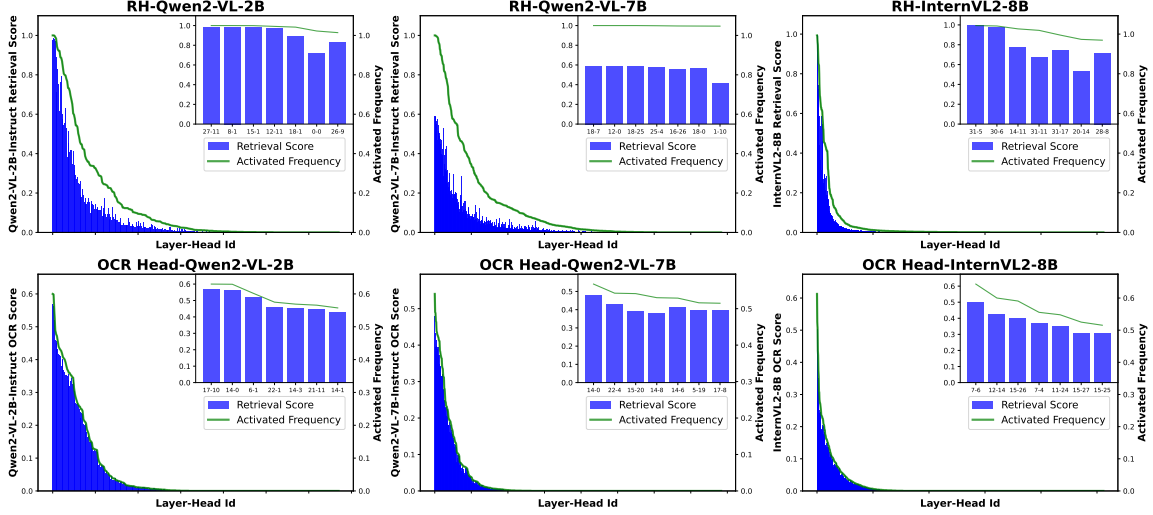
Figure 4: Visualization comparing the OCR Score for OCR Heads and Retrieval Score for Retrieval Heads. OCR Score and Retrieval Score measure token-level recall, how many correct tokens were copied, while Activation Frequency indicates how often a head activates on at least one token above a threshold.

qualitative distinctiveness, and 3) static activation. For the experiments, we use the Qwen2-VL (Wang et al., 2024) and InternVL2 (Cai et al., 2024) models to compare the characteristics of the OCR head and the retrieval head.

4.1 Reduced Sparsity and Enhanced Specialization

Figure 3 demonstrates that the OCR head is less sparse than the retrieval head. In the case of the retrieval head, when text-based passkey and NIAH datasets are used as input, only 3–6% of the heads that activate during passkey retrieval are involved, as stated in Wu et al. (2024b)'s paper (we consider heads with a retrieval score above 0.1 as actively participating in retrieval). We observe similar results in models with more than 7 billion parameters. Consistently, when image-based inputs containing text are used for the OCR head, we find that 91.3% more heads are involved on average compared to when text-based datasets are used. On the other hand, for heads that show low-frequency retrieval scores between 0.01 and 0.1, i.e., those affected by copy-paste, the retrieval head involves 192% more heads than the OCR head. Based on these experimental results, we conclude that the OCR head activates stronger heads more clearly, and its involvement in low-frequency retrieval is smaller, making it less sparse and more specialized. This indicates that OCR heads exhibit enhanced focus and are more robust in attending to relevant image regions containing text, whereas

Model	Avg Retrieval Score	Similarity
Qwen2-VL-7B Instruction	0	0.507
	0-0.1	0.196
	0.1-0.3	0.104
	0.3-0.5	0.083
	0.5-1.0	0.076
	$0.1 \leq$	**0.347**
Model	Avg Retrieval Score	Similarity
InternVL2-8B Instruction	0	0.593
	0-0.1	0.093
	0.1-0.3	0.000
	0.3-0.5	0.000
	0.5-1.0	0.000
	$0.1 \leq$	**0.133**

Table 1: Jaccard similarity of the average retrieval scores between OCR Heads and Retrieval Heads.

retrieval heads are more diffusely activated across a wider range of scenarios.

4.2 Do OCR Head and Retrieval Heads Exist as Distinct Entities?

Building on the previous experiment, one might ask whether the heads involved in retrieval heads or OCR heads are distributed similarly. To investigate this, we check how commonly the heads are distributed based on similarity. We measure the similarity between heads using the Jaccard similarity coefficient, as shown in the formula below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3)$$

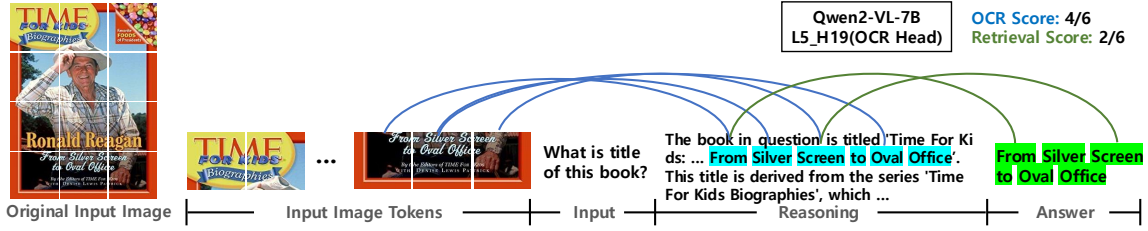In Table 1, for the Qwen2-VL-7b model, the

Figure 5: In the OCR-VQA task, we calculate the OCR score and Retrieval score based on the CoT prompting. L5_H19 refers to Layer 5, Head 19. The blue and green lines indicate the token positions with the highest attention.

Jaccard similarity for heads with a retrieval score above 0.1 is 0.347, which is relatively low. For the InternVL-8b model, we observe an even lower value of 0.133. This shows that there is a low similarity between the retrieval heads and the OCR heads, confirming that they are clearly distinct.

4.3 Activation Frequency Mirrors Retrieval Score

Heads serve multiple roles beyond a single task. We conduct experiments to assess how much attention the retrieval heads focus on tasks like copy-paste, comparing OCR heads and retrieval heads in terms of their sensitivity. Sensitivity refers to how a head activates in specific contexts; a highly sensitive head activates in only one context and not in others. For instance, if a passkey context is "the passkey is 1234," a head with high sensitivity would activate strongly for "1" but not for "234." To evaluate this, we compare activation frequency with the average retrieval score, as in previous studies (Wu et al., 2024b). Activation frequency measures how often activation occurs for any token, giving a score of 1 if any token activates. If activation occurs in just one token, the average retrieval score would be 1/4. Therefore, the difference between frequency and average retrieval score indicates whether a head performs roles beyond retrieval in the context. In Figure 4, we observe a significant difference in the frequency and average score for the retrieval head, suggesting that it plays additional roles outside of retrieval. This is what we refer to as "dynamically activated." On the other hand, the OCR head shows little difference, indicating that these heads, specialized for OCR tasks, are "statically activated."

5 Unveiling the Functional Roles of OCR Heads in Downstream Applications

To compare the impact of OCR heads and retrieval heads on downstream tasks, we conduct

Qwen2-VL-7B	Method	Score
Top10 OCR Head	Avg OCR Score	0.4004
	Avg Retrieval Score	0.0123
Top10 Retrieval Head	Avg OCR Score	0.2929
	Avg Retrieval Score	0.4705

Table 2: The average scores of the top 10 OCR and Retrieval heads when CoT prompting is applied in the OCR-VQA task.

three complementary analyses: CoT prompting, attention head masking, and sink token redistribution. We also use the OCR-VQA (Mishra et al., 2019), Multi-Page DocVQA (Tito et al., 2023), DocVQA (Mathew et al., 2021), NQ (Kwiatkowski et al., 2019), and HotpotQA (Yang et al., 2018) datasets for the downstream tasks. Furthermore, the evaluation metric for all the datasets in the downstream tasks is the F1 score.

5.1 Impact of OCR Head for Reasoning in OCR-VQA

We investigate the contribution of OCR and retrieval heads to the reasoning capabilities of LVLMs within the OCR-VQA task using CoT prompting. As illustrated in Figure 5, CoT prompting enables the generation of intermediate reasoning steps before producing the final answer based on the input of the image. The blue and green lines represent the positions where maximum attention occurs in the tokens. We calculate the OCR score between the image tokens and the reasoning part using the method in Section 3, and calculate the retrieval score by copying the answer from the reasoning and answer sections, as described in Section 2. In Table 2, we measure the scores based on the top 10 heads with the highest average scores for each method. The OCR head shows a high OCR score, while the retrieval head not only shows a high retrieval score but also a high OCR score. The fact that the OCR head only shows a

Masked Heads	DocVQA			MP-DocVQA			NQ			HotpotQA		
Qwen2-VL-7B	OCR	RH	Random	OCR	RH	Random	OCR	RH	Random	OCR	RH	Random
Baseline		76.3			69.1			58.0			49.6	
5	74.1	76.1	75.2 ± 0.3	68.1	69.4	67.9 ± 0.2	57.7	57.2	57.6 ± 0.6	51.2	47.8	49.5 ± 0.1
10	73.9	74.6	74.2 ± 0.2	66.7	67.6	66.6 ± 0.4	59.0	53.8	56.4 ± 0.2	49.7	45.3	47.9 ± 1.2
20	59.7	67.1	75.1 ± 0.4	47.7	62.4	67.7 ± 1.7	56.8	36.4	57.3 ± 0.5	47.5	37.3	48.9 ± 0.4
Intern VL2-8B	OCR	RH	Random	OCR	RH	Random	OCR	RH	Random	OCR	RH	Random
Baseline		39.1			35.0			58.3			35.8	
5	36.6	39.7	39.2 ± 0.2	32.8	35.5	35.1 ± 0.6	60.3	56.3	55.9 ± 0.3	33.5	32.8	34.6 ± 0.1
10	30.0	39.7	38.8 ± 0.2	25.1	35.7	34.7 ± 0.4	59.6	55.8	57.8 ± 0.3	33.5	28.2	29.2 ± 0.3
20	24.2	38.2	38.3 ± 0.6	20.7	36.1	34.6 ± 0.3	56.6	58.9	57.6 ± 0.2	34.8	25.7	33.5 ± 1.2

Table 3: Performance on VQA and QA datasets when masking the OCR head, the retrieval head (RH), or a randomly selected head (Random: mean ± variance across 5 selections).

high OCR score, as discussed in Section 4.1, suggests that it is more specialized. On the other hand, the retrieval head consistently shows high scores because it is dynamically activated as we discuss in Section 4.3. This means that the retrieval head is context-sensitive and performs multiple roles. Based on these observations, we highlight the significant influence of the OCR head when performing CoT reasoning on text information in LVLMs.

5.2 Dissecting the Impact of Attention Head Masking on Task Performance

To understand the functional contributions of OCR heads and retrieval heads across multiple downstream tasks, we perform a systematic head-masking analysis. Specifically, we evaluate how LVLM performance degrades when we mask the top-scoring OCR heads, retrieval heads (RH), or an equal number of randomly selected heads. We apply this intervention to two VQA tasks—DocVQA, which requires copying specific text from a single image, and Multi-Page DocVQA, which involves reasoning over multiple images—and two QA tasks, Natural Questions (NQ) and HotpotQA, where we include the evidence passage and randomly sample 0–5 negative passages with shuffled ordering. We mask the top 5, 10, and 20 heads of each type (OCR, RH, and random). As shown in Table 3, masking OCR heads causes a pronounced decline in VQA performance: with 20 heads masked, DocVQA and MP-DocVQA F1 scores fall to 59.7% and 47.7% (nearly a 20% drop), whereas masking RH yields much smaller declines of 9.2% and 7.3%, respectively. By contrast, on the QA tasks, masking RH leads to a substantial performance decrease. These results confirm that vision–text VQA tasks depend critically on OCR heads, while retrieval heads play

		DocVQA	MP-DocVQA
Qwen2-VL-7B	Baseline	76.3	69.1
	OCR	76.6 (+0.3)	69.7 (+0.6)
	RH	76.2 (-0.1)	69.4 (+0.3)
InternVL2-8B	Baseline	39.1	35.0
	OCR	40.0 (+0.9)	35.8 (+0.8)
	RH	39.3 (+0.2)	34.7 (-0.3)

Table 4: Performance results after redistributing the attention sink token value in the OCR head and retrieval head (RH) across MP-DocVQA and DocVQA tasks.

a more pivotal role in text-based QA.

5.3 Enhancing OCR Head Performance via Sink Token Value Redistribution

In this experiment, we investigate the potential for enhancing downstream task performance by strategically redistributing the attention values associated with the sink token. An attention sink token (Kang et al., 2025a) refers to a token in the input sequence that receives disproportionately high attention scores despite its lack of semantic relevance. Although it carries little meaningful information, its attention value diverts focus from more important tokens in the sequence. Recent research (Yu et al., 2024; Kang et al., 2025a; Gu et al., 2025) suggests that optimizing the attention score distribution and reducing the impact of such attention sinks can enhance performance, especially in tasks requiring more accurate and meaningful attention allocation. Based on this, we redistribute the sink token value proportionally to the attention scores of other tokens. This process can be expressed as:

$$\hat{A}_h^l[T,t] = A_h^l[T,t] + \beta \cdot S \cdot \frac{A_h^l[T,t]}{\sum_{i \in \{1,\cdots,T\}} A_h^l[T,i]} \quad (4)$$
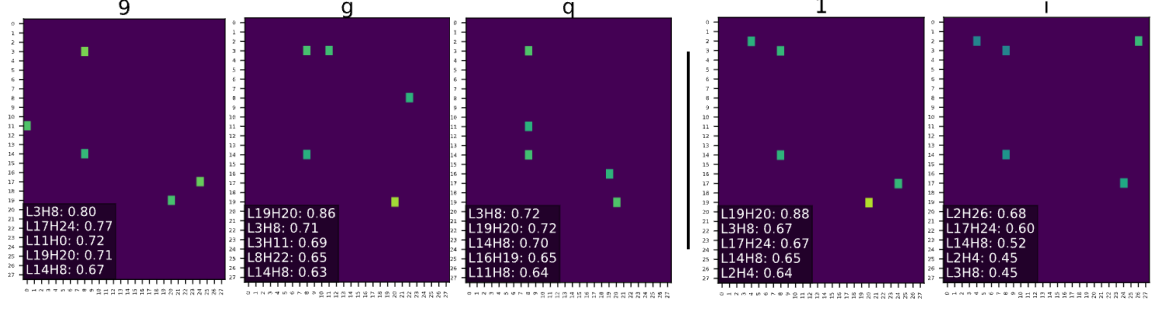
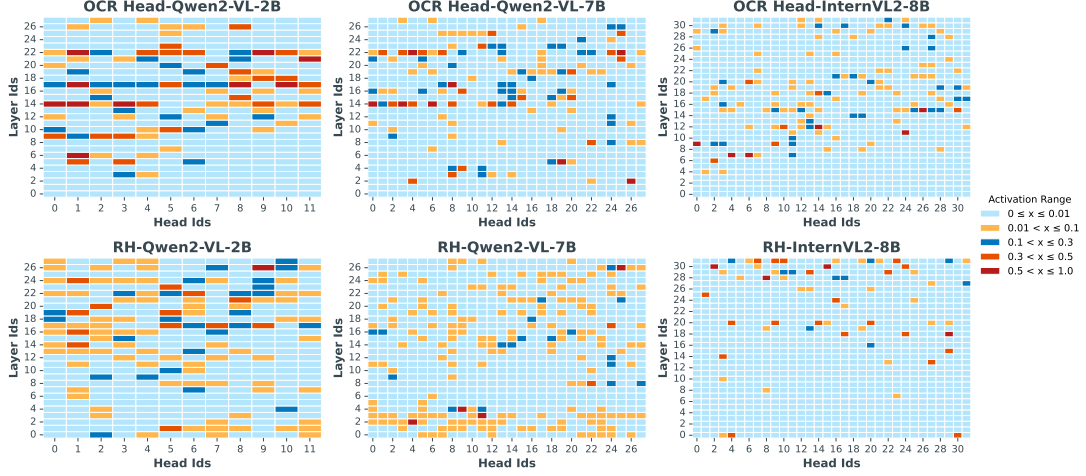Figure 6: Character-wise activation of top-5 OCR heads; visually similar shapes co-activate common heads.

Figure 7: Heatmap visualization of OCR and Retrieval heads.

Here, l and h refer to the layer and head, respectively, and A represents the attention score. T refers to the total number of tokens in the attention score, and S is the original value of the sink token. β is a hyperparameter that controls how much we want to remove excess attention scores from the attention sink. Following prior studies (Yu et al., 2024), we set $\beta = 0.4$. First, we compute the attention score ratios for all tokens, excluding the 0th sink token. The computed ratios are then multiplied by S to calculate the redistributed attention scores for each token, which are added to the original attention score to obtain $\hat{A}$. This new $\hat{A}$ is used for subsequent operations, and after inference, the F1 score is measured. The heads selected for redistribution are the top 4 heads with the highest average OCR or retrieval scores. Table 4 shows that redistributing the sink token among OCR heads boosts performance on DocVQA and MP-DocVQA for both QwenVL and InternVL models. In contrast, redistributing among retrieval heads lowers performance in some cases and yields slight gains in others.

6 Case Study

6.1 Explaining Character Recognition via OCR Heads

We investigate whether each character—from 0 to 9 and a to z—tends to activate a specific attention head. To do this, we replaced the original multi-digit pass keys in our dataset with single characters (0–9, a–z), rendered the prompts as images, and then applied the OCR scoring procedure described in Section 3. Figure 6 shows the activation heatmaps along with the top-5 heads per character. Characters with similar shapes consistently activate the same OCR heads—for instance, '1' and 'i' share four heads (L3H8, L2H4, L14H8, and L17H24), and '9', 'g', and 'q' share three. This consistent co-activation demonstrates that OCR heads mechanistically encode visual similarity to guide text extraction. We show that the behavior of OCR heads can directly elucidate the mechanism by which LVLMs perform optical character recognition on embedded text. The complete set of character-wise activation heatmaps for all tokens is presented in Appendix A.

6.2 Distribution of OCR and Retrieval Heads

In Figure 7, we visualize the distribution of OCR and Retrieval heads. The OCR heads are not uniformly distributed; rather, they are found to be concentrated in specific areas of the middle-to-late layers. Figure 7 displays the number of OCR heads per layer (average score > 0.1) for the Qwen2-VL series, revealing a clear concentration. Specifically, in the Qwen2-VL-2B model, there are 11 OCR heads in the 17th layer and 8 OCR heads in the 22nd layer. Similarly, for the Qwen2-VL-7B model, 10 OCR heads are located in the 14th layer and 11 OCR heads in the 22nd layer. This demonstrates that OCR heads tend to cluster within certain middle-to-late layers, rather than being evenly distributed across all layers.

7 Related Work

Attention Head Analysis in LVLM Interpretability research in LVLMs has investigated how attention mechanisms integrate visual and textual information. These studies have identified different types of heads involved in processing visual input: Image-centric heads (Kang et al., 2025a) focus on semantically meaningful visual features and manage attention distribution. The other work (Liu et al., 2024b) explores amplifying visual attention to reduce issues like hallucination. Localization heads (Kang et al., 2025b) are shown to be effective for visual grounding, pinpointing image regions based on text queries without explicit training. Furthermore, Visual heads actively processing visual information have been identified using entropy-based scores (Bi et al., 2024). While these studies shed light on LVLM's visual processing and visual-text alignment, they do not specifically investigate attention heads dedicated to the unique task of recognizing and retrieving text embedded within images. Understanding these multimodal mechanisms builds upon interpretability insights gained from analyzing LLMs.

In-Context Identification Heads Transformer-based models utilize attention heads that focus on different aspects of input data to extract meaningful information (Zheng et al., 2024). In-Context Identification heads are attention heads that selectively attend to and aggregate structural, syntactic, and semantic cues within the prompt, writing these distilled signals back into the residual stream to guide subsequent generation. For example, the Previous Head (Olsson et al., 2022)

captures positional relationships within token sequences, while the Duplicate Head (Wang et al., 2022) identifies repeating content, becoming more active with increased repetition of tokens. In contrast, the Rare Words Head (Voita et al., 2019) targets tokens with low frequency. Additionally, Retrieval Heads (Wu et al., 2024b) precisely pinpoint the positions of specific tokens. Identifying these specialized attention heads typically involves replacement-based methods such as ablation (Yu and Ananiadou, 2024; Jin et al., 2024) or patching (Merullo et al., 2023; Todd et al., 2023), which test their influence on model outputs, or scoring-based methods (Jin et al., 2024; Crosbie and Shutova, 2024) that evaluate their functionality. In this study, we propose the OCR Score, a scoring-based approach for efficiently detecting attention heads with unique capabilities.

8 Conclusion

This paper identifies and characterizes Optical Character Recognition (OCR) heads as the specialized mechanisms that Large Vision-Language Models (LVLMs) use to read text directly from images. We demonstrate that these heads are functionally distinct from conventional retrieval heads, defined by three key properties: lower sparsity, qualitative distinctiveness, and static activation patterns. Our experiments validate their critical role; targeted head-masking severely degrades performance on OCR-VQA tasks while leaving text-based QA unaffected, confirming their specialized function. Furthermore, we show that manipulating these heads via sink-token redistribution can enhance performance on downstream tasks. This work provides a direct, mechanistic insight into how LVLMs process embedded text, paving the way for more interpretable and reliable multimodal models.

Limitations

We validate the OCR head properties through five empirical experiments. However, many open-source LVLMs support only a single image input or rely on thumbnail-style training settings. Some models also produce image tokens that mismatch with patch counts due to encoder variations. These constraints limit the diversity of models we evaluate. Additionally, due to computational constraints, we conduct experiments on the passkey and NIAH dataset only up to 8k samples.

Ethics Statement

This study involves open-source LVLMs for QA tasks, and some models may generate inappropriate outputs. Also, the VQA and QA datasets used in our experiments may also include potentially harmful or biased content.

References

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. 2024. Unveiling visual perception in language models: An attention head analysis approach. *arXiv preprint arXiv:2412.18108*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Joy Crosbie and Ekaterina Shutova. 2024. Induction heads as an essential mechanism for pattern matching in in-context learning. *arXiv preprint arXiv:2407.07011*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *CoRR*, abs/2401.16420.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*.

Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025a. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025b. Your large vision-language model only needs a few attention heads for visual grounding. *arXiv preprint arXiv:2503.06287*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023. Circuit component reuse across tasks in transformer language models. *arXiv preprint arXiv:2310.08744*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. *arXiv preprint arXiv:2401.06915*.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024a. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024b. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2024. Towards visual grounding: A survey. *arXiv preprint arXiv:2412.20206*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024. How do large language models learn in-context? query and key matrices of in-context heads are two towers for metric learning. *arXiv preprint arXiv:2402.02872*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. 2024. Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.

A Character-wise Analysis

We investigate whether each character—from 0 to 9 and a to z—tends to activate a specific attention head. To do this, we replaced the original multi-digit pass keys in our dataset with single characters (0–9, a–z), rendered the prompts as images, and then applied the OCR scoring procedure described above. As shown in Figure 8, we found that visually similar characters share common high-activation heads among their top 5. For instance, for the characters "1" and "i," four heads (L14H8, L2H4, L17H24, and L3H8) appear jointly in their top 5. Similarly, for "9," "g," and "q," the heads L3H8, L14H8, and L19H20 co-occur among the top 5 activations. This suggests that when visually similar shapes enter the ViT encoder, they produce similar representations that in turn trigger the same decoder heads in the LLM. In Figure 6, we use the Qwen2-VL-7B model; the horizontal axis denotes the head index, and the vertical axis denotes the layer index.

B CoT Prompt

Let's think about it step by step. The response must be a valid JSON object with two fields: "thinking" and "answer".

The "answer" field should provide the final answer, starting with "Answer:".

Here is the question:
"Question": questions[n]
Respond in the following JSON format:
{
"**thinking**": "Thinking: ...",
"**answer**": "Answer: ..."
}

Table 5: CoT Prompt.

In Section 4.3, we use Chain-of-Thought (CoT) prompting to simultaneously measure the OCR score and the retrieval score. To obtain structured responses, we instruct the model to generate reasoning and answers in JSON format, as illustrated in Table 5. Specifically, we extract the OCR score from the input image tokens and reasoning component, and compute the retrieval score from the reasoning and answer components.
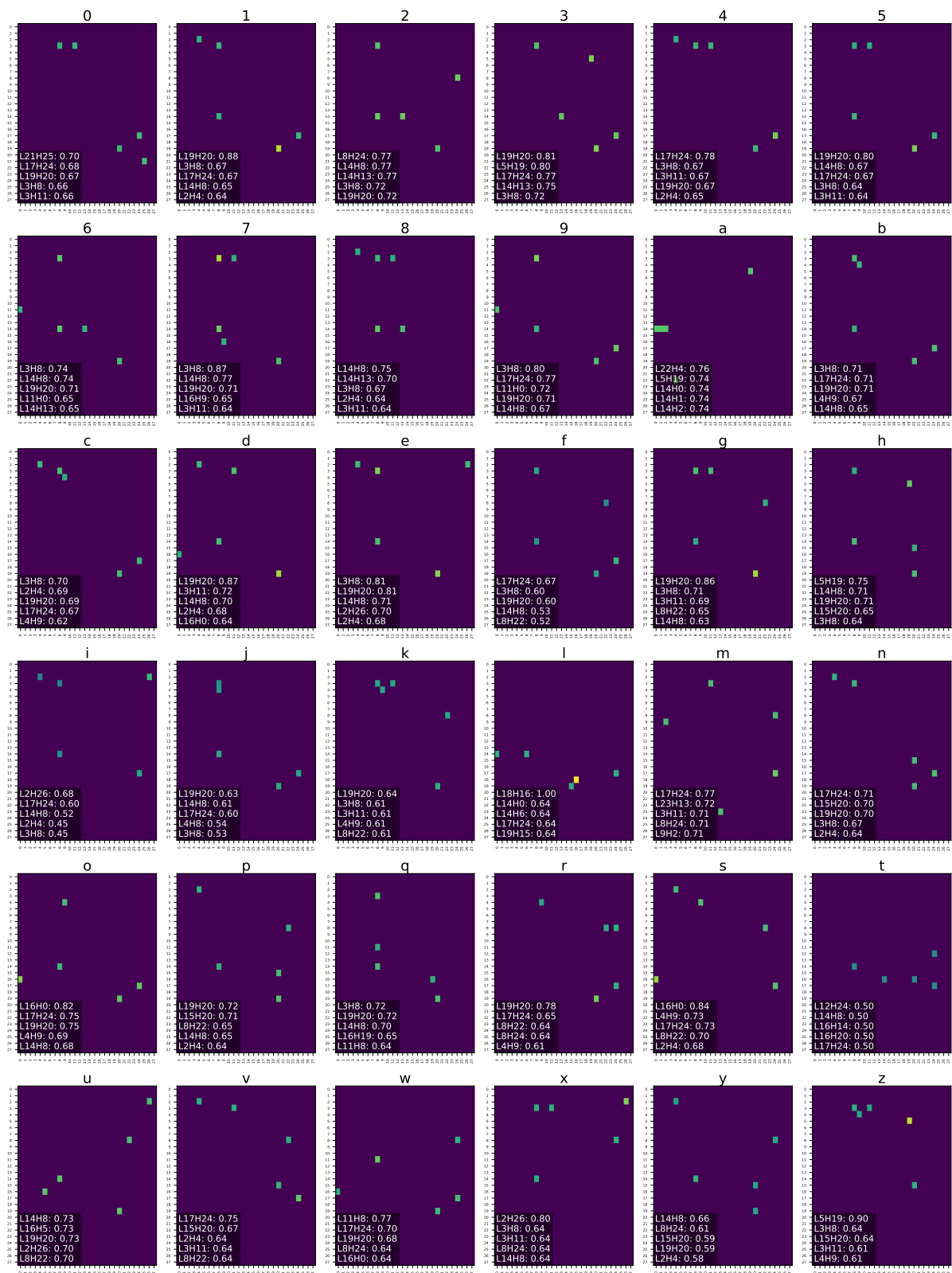
Figure 8: Heatmap visualization of top5 OCR scores for each character from 0 to 9 and a to z.