FROM RETRIEVAL TO RECOGNITION: HOW VISION–LANGUAGE MODELS BECOME OCR SPECIALISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does OCR specialization create new reading heads or reuse existing ones? Prior OCR-head studies have primarily examined local-answer retrieval from images; here we study full-sequence transcription. Using an evidence-grounded protocol with held-out causal interventions, we identify sparse and stable OCR-head sets in GLM-OCR, MinerU2.5, and PaddleOCR-VL-1.6. We then investigate the mechanistic origin of these OCR heads by comparing them with independently identified textual retrieval/copy heads in general VLMs. Across two general VLMs, visual OCR heads strongly overlap independently identified textual retrieval/copy heads, yielding top-20 intersections of 73.3% and all-head Spearman correlations of 0.677–0.886. The overlap and causal interventions suggest that full-sequence OCR operates as dense sequential multimodal copy-and-paste, repeatedly retrieving visual evidence and routing it to the current output position. Finally, we examine how this shared circuit changes as a general VLM becomes an OCR specialist. Matched base-to-specialized comparisons show that OCR specialization largely preserves head identity, retaining 17–20 of the top 20 heads per task with all-head rank correlations of 0.874–0.942, while redistributing their functional and causal strengths.

1 INTRODUCTION

End-to-end OCR aims to transcribe diverse visible content in an image: characters in text passages, cell contents and structure in tables, and glyphs and syntax in formulas (Kim et al., 2022; Blecher et al., 2023; Poznanski et al., 2025). Intuitively, this involves repeatedly locating visual evidence relevant to the current output and passing that evidence into generation. In language models, *retrieval heads* (Wu et al., 2024) perform a closely related operation: they locate answer-bearing positions in context and support copying their contents into the output. The similarity suggests that visual transcription may build on a more general retrieval-and-copy mechanism. This leads to our central question: *does OCR specialization construct a new attention-head circuit, or reconfigure one that is already present in the model?*

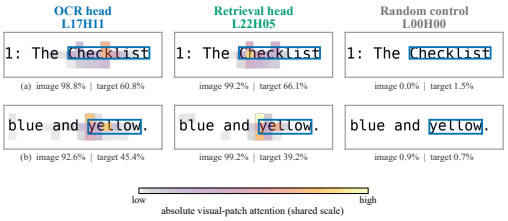

Figure 1: Token-level visual attention during Qwen2-VL-2B (Wang et al., 2024) OCR for generated tokens (a) "Checklist" and (b) "yellow." Columns compare the top text-OCR head, the retrieval head L22H05, and a head outside both rankings. Blue boxes mark target evidence, and all heatmaps use one absolute scale.

Prior work has identified attention heads involved in localized visual text retrieval in general VLMs, using tasks such as passkey retrieval and OCR-VQA (Baek et al., 2025). Those experiments reveal question-conditioned visual retrieval, but do not address full-sequence transcription. In our setting, OCR requires repeatedly retrieving the visual evidence needed at each generation step, across text passages, visible table content, and mathematical objects.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To study attention-head mechanisms under full transcription, we develop an evidence-grounded protocol that aligns correctly generated tokens with their corresponding visual evidence, with disjoint head discovery and held-out causal validation. Across GLM-OCR (Duan et al., 2026), MinerU2.5 (Niu et al., 2026), and PaddleOCR-VL-1.6 (Zhang et al., 2026), we identify sparse and reproducible OCR-head sets whose targeted ablation causes substantially greater degradation than matched random ablations across text, tables, and formulas.

Having established sparse and causally relevant OCR heads, we next ask whether they reflect a dedicated reading circuit or reuse a mechanism already present in the model. We therefore compare OCR heads identified from visual transcription with retrieval/copy heads independently identified from text-only retrieval tasks (Wu et al., 2024). As Figure 1 illustrates, both types of heads attend to the visual evidence corresponding to the current output token, whereas a random control head does not. This correspondence extends beyond individual examples, with strong overlap between the two head populations and supporting cross-task causal effects. These results suggest that full-sequence OCR recruits a shared copy-and-route mechanism, operating as dense sequential multimodal retrieval and copy.

If full-sequence OCR largely reuses a pre-existing retrieval-and-copy mechanism, what changes when a general VLM is specialized for OCR? Matched comparisons between the generalist and OCR-specialized versions of Qwen2-VL and Qwen3-VL show a consistent pattern across OCR tasks: specialization largely preserves the dominant OCR heads and their layer organization, while redistributing their relative functional and causal importance. This suggests that OCR specialization primarily reweights an existing retrieval-and-copy circuit rather than constructing a new attention-head circuit.

Our contributions are threefold:

1. We identify and causally validate sparse attention-head populations that support full-sequence OCR across text, tables, and formulas in three specialized OCR models.
2. We uncover substantial reuse between OCR heads and independently identified textual retrieval/copy heads in general VLMs, providing mechanistic evidence that full-sequence transcription operates through dense sequential multimodal retrieval and copy.
3. We show that OCR specialization largely preserves the underlying head population and its layer organization, while redistributing the functional and causal importance of the conserved heads.

2 Related Work

Document OCR. Generative document OCR has evolved from early image-to-sequence models such as Donut (Kim et al., 2022) and Nougat (Blecher et al., 2023) toward unified page-level transcription of heterogeneous document content. More recent systems such as GOT-OCR2.0 (Wei et al., 2024), olmOCR (Poznanski et al., 2025), and DeepSeek-OCR (Wei et al., 2025) directly generate page-level text or markup, broadening OCR from local recognition to end-to-end document transcription. In parallel, models such as MinerU2.5 (Niu et al., 2026) introduce hierarchical or coarse-to-fine processing to handle high-resolution pages efficiently, while GLM-OCR (Duan et al., 2026) and PaddleOCR-VL-1.6 (Zhang et al., 2026) further specialize vision–language architectures for unified document recognition. Building on these advances in recognition quality, structured decoding, resolution handling, and efficiency, we examine the internal mechanisms that sustain transcription and how they change under OCR specialization.

Functional attention heads. Functional heads have been linked to induction, copying, and long-context retrieval (Voita et al., 2019; Olsson et al., 2022; Wu et al., 2024). Multimodal studies find heads that route textual and visual evidence (Golovanevsky et al., 2025; Li et al., 2026; Sun et al., 2026). Most directly, Baek et al. (2025) identify OCR-related heads in general VLMs using OCR-VQA and visual passkey retrieval. We study full-sequence transcription, which requires repeated routing across text, tables, and formulas. We identify OCR and textual retrieval/copy heads independently, test their causal relationship, and track their organization under OCR specialization.

Causal validity. An attention head may point to the correct visual region without being necessary for transcription, because other heads may carry the same information. We therefore use attention alignment only to nominate candidates. We then zero their outputs at answer positions during held-out generation and compare the performance drop with that from zeroing layer-matched random heads in the same way. Using the same ablation for both sets controls for effects of the intervention itself (Li & Janson, 2024); a larger drop for the discovered heads supports their causal role.

3 IDENTIFYING OCR HEADS IN FULL-SEQUENCE TRANSCRIPTION

Let a VLM encode an image as a visual-token prefix and autoregressively generate a token sequence $y_{1:T}$. For decoder layer ℓ, attention head h, and generated token y_t, let $A_t^{\ell,h}(j)$ denote the attention weight assigned by head (ℓ, h) from the query token used to predict y_t to the key at position j. The key position j ranges over all positions available to that query token, including the visual-token prefix and previously generated textual positions. For each visible reference span z, we define its visual evidence set $E(z)$ as the visual-token positions whose patch rectangles have IoU greater than 0.1 with any annotated box belonging to z, following the evidence-patch criterion of Baek et al. (2025). These boxes cover characters in text passages and table cells, and rendered glyphs or mathematical objects in formulas. We exclude structural markup and layout elements, such as HTML tags and table borders, because they cannot be assigned a comparably precise visual patch target for an individual generated token.

Given these visual evidence sets, we first obtain the model prediction through greedy autoregressive decoding. We then replay the generated sequence to record the attention distributions along the same autoregressive trajectory. During this replay, each position is conditioned on the exact tokens produced by the original decoding process, thereby reconstructing the model's own autoregressive prefixes without substituting reference tokens. Further evidence-construction and replay-audit details are provided in Appendix B.

To score only content the model transcribed correctly, we align its freely generated text with the reference and identify contiguous exact-match character blocks. An annotated reference region is transferred to the prediction only if its full character span lies within one block. A generated BPE token is eligible only when its entire decoded span lies within one exact-match block and overlaps a transferred visible region; tokens crossing a mismatch are excluded. For an eligible token y_t, let z_t be its aligned reference span and $E(z_t)$ its visual evidence, and let $\mathcal{K}_t$ contain all key positions accessible to its query token. The OCR-hit indicator for head (ℓ, h) is

$$H_t^{\ell,h} = \mathbb{1}\left[\arg\max_{j \in \mathcal{K}_t} A_t^{\ell,h}(j) \in E(z_t)\right]. \tag{1}$$

For sample i, let $C_i^{\ell,h}$ denote the number of OCR hits recorded by head (ℓ, h), and let G_i denote the number of tokens in the official visible-content reference. We define the sample-level OCR score and its average over N samples as

$$s_i^{\ell,h} = \frac{C_i^{\ell,h}}{\max(G_i, 1)}, \qquad \bar{s}^{\ell,h} = \frac{1}{N}\sum_{i=1}^{N} s_i^{\ell,h}. \tag{2}$$

Following prior work (Baek et al., 2025), head (ℓ, h) is considered active on sample i when $s_i^{\ell,h} > 0.1$. We define the activation frequency of head (ℓ, h) as

$$f^{\ell,h} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[s_i^{\ell,h} > 0.1\right]. \tag{3}$$

We use the OCR-head selection thresholds of Baek et al. (2025): $\bar{s}^{\ell,h} > 0.1$ requires an average OCR score above 0.1, and $f^{\ell,h} \geq 0.1$ requires a sample-level score above 0.1 on at least 10% of instances. Both conditions must hold for a candidate OCR head. For rank-based analyses, we additionally rank all attention heads according to $\bar{s}^{\ell,h}$, irrespective of whether they satisfy these thresholds. Figure 2 summarizes the discovery procedure, from visual-evidence construction and autoregressive replay to token-level OCR hits and head-level scoring.

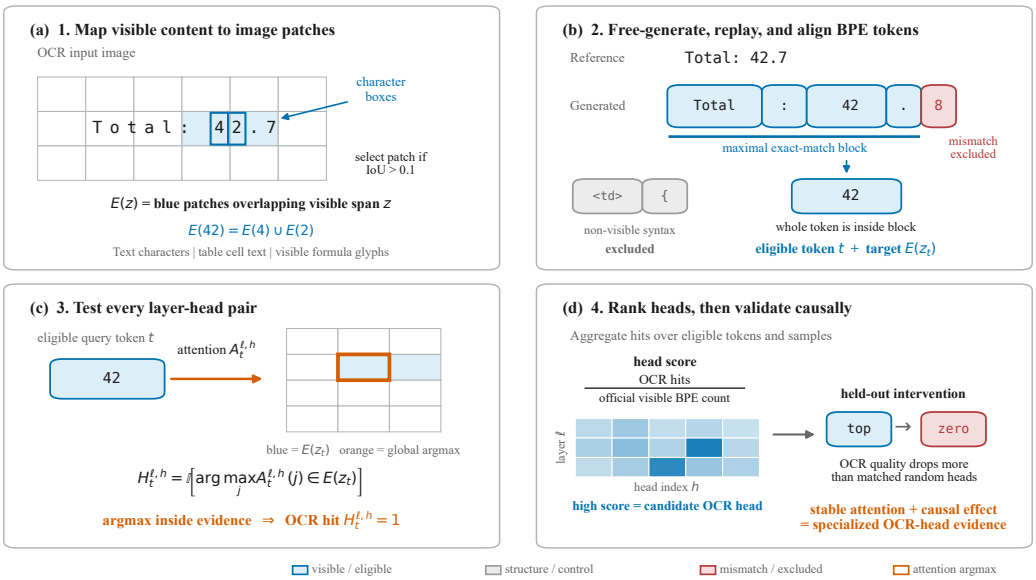

Figure 2: Evidence-grounded OCR-head discovery and causal validation. Character, table-cell, and formula-glyph boxes map reference spans to visual patches $E(z)$. We score only generated BPE tokens whose full decoded span matches the reference and aligns with visual evidence. A head records a hit when its global attention argmax falls in that token's evidence patches. Heads discovered on disjoint folds are ablated on held-out fold C and compared with layer-matched random head sets.

4 Validating OCR Heads in Specialized OCR Models

4.1 Experimental Setup and Validation Protocol

We validate our OCR-head identification procedure on GLM-OCR, MinerU2.5, and PaddleOCR-VL-1.6 using a shared, language-stratified OmniDocBench subset of 1,200 selected instances (400 each for text, table, and formula recognition). Text uses plain transcription, while tables and formulas follow each model's native output format. Within each task, a fixed-seed, page-grouped assignment divides the 400 instances into disjoint discovery folds A and B (160 each) and a held-out causal-evaluation fold C (80). Heads are discovered independently on A and B, with agreement measured by all-head rank correlation and top-k overlap; consensus heads are evaluated exclusively on C.

Causal damage is measured relative to the intact model using NED increase for text, TEDS (Zhong et al., 2020) decrease for tables, and CDM (Wang et al., 2025) decrease for formulas. For all three tasks, five random head sets exactly match the target set's head count in every layer and exclude the task's consensus top-20 heads. Full evaluation details and controls are provided in Appendix C.

4.2 Cross-Fold Stability and Held-Out Causal Validation

We first assess whether OCR-head discovery yields consistent high-ranking heads across independent data subsets. Across all nine model–task combinations, the independently discovered top-20 sets overlap by at least 19/20 heads across the two disjoint discovery folds A and B. This strong cross-fold agreement indicates that the highest-ranked OCR heads are stable across data subsets rather than driven by a particular fold. Full head-score maps, per-model overlap statistics, and cross-task intersections are provided in Appendix D.

We then ask whether this stable alignment corresponds to a functional contribution. Following head-ablation analyses (Voita et al., 2019), we intervene on the consensus OCR heads and evaluate their effects on the held-out fold C, which is excluded from head selection, against matched random head sets.

Targeted ablation of the discovered OCR heads causes consistently and substantially larger performance degradation than layer-matched random interventions across all three specialized OCR models and all three content types. For example, in GLM-OCR, ablating the discovered table heads produces a TEDS drop of 0.903, compared with 0.032 for matched random sets, while formula CDM drops by 0.846, compared with only 0.016 under random ablation. Similar selectivity is observed in MinerU2.5 and PaddleOCR-VL-1.6, indicating that the effect is not specific to a single OCR architecture. Detailed causal results and specificity controls are provided in Appendix E.

The held-out interventions establish the causal relevance of the discovered OCR-head populations. Since full-sequence transcription repeatedly requires retrieving visual evidence and routing its content to the current output position, we next ask whether these OCR heads reuse the retrieval/copy mechanism already present in general VLMs.

5 Do OCR Heads Reuse Retrieval/Copy Heads?

5.1 Independent OCR and retrieval/copy head discovery

We examine Qwen2-VL-2B (Wang et al., 2024) and Qwen3-VL-2B (Bai et al., 2025) to test whether the correspondence between OCR and retrieval/copy heads holds across two general VLMs. In each model, we independently identify OCR heads from visual transcription and retrieval/copy heads from text-only needle-in-a-haystack (NIAH) tasks (Wu et al., 2024).

The retrieval/copy criterion follows the standard NIAH setup, where a head is considered active when its maximum attention falls within the needle span and the attended source token matches the generated token, thereby operationalizing exact token-level copying from textual context. The OCR criterion follows the same source-selection principle, using spatial alignment to $E(z_t)$ as the visual analogue of exact source-token matching.

For each OCR task, we rank every decoder attention head by its OCR score and, independently, by its text retrieval/copy score. Both rankings are fixed before cross-task comparison. We quantify their correspondence using all-head rank correlation, top-k overlap, Jaccard similarity, and asymmetric containment $|O \cap R|/|O|$, where O and R denote the OCR and retrieval/copy head sets, respectively. A separate held-out needle is reserved for causal evaluation.

This comparison reveals substantial structural correspondence between the two head populations. In Qwen2-VL, the retrieval/copy-score ranking correlates with the OCR-score rankings for text, tables, and formulas, with all-head Spearman correlations of 0.846, 0.886, and 0.677, respectively.

Figure 3 provides a head-level view of this correspondence in Qwen2-VL. Each point represents the same attention head positioned by its retrieval/copy rank and task-specific OCR rank, with the shared top-ranked heads concentrated in the upper-right region. The independently obtained top-20 rankings overlap by 13, 17, and 15 heads for text, tables, and formulas, respectively. This pattern replicates structurally in Qwen3-VL: its 30-head retrieval/copy set yields top-20 overlaps of 15, 15, and 13 heads across the three OCR tasks, with all-head Spearman correlations ranging from 0.806 to 0.859. In addition, 74.1–84.6% of the selected OCR heads fall within the retrieval/copy set.

All six top-20 overlaps are well above random expectation, confirming that the structural correspondence is consistent across both models and all three OCR tasks.

5.2 Cross-task causal correspondence

Structural correspondence provides strong evidence that OCR and retrieval/copy heads may share a common mechanism, but structural overlap alone cannot determine whether they play causally related roles. We therefore examine cross-task causal transfer by intervening on heads identified in one task and evaluating their effects on the other.

In Qwen2-VL, ablating retrieval/copy heads substantially degrades OCR performance. Removing the retrieval top-20 increases text NED from 0.1106 to 0.6234 and reduces formula CDM from 0.9614 to 0.3502, while removing the retrieval top-20 lowers table TEDS from 0.6720 to 0.1597. The symmetric cross-task intervention further supports this functional relationship: ablating the

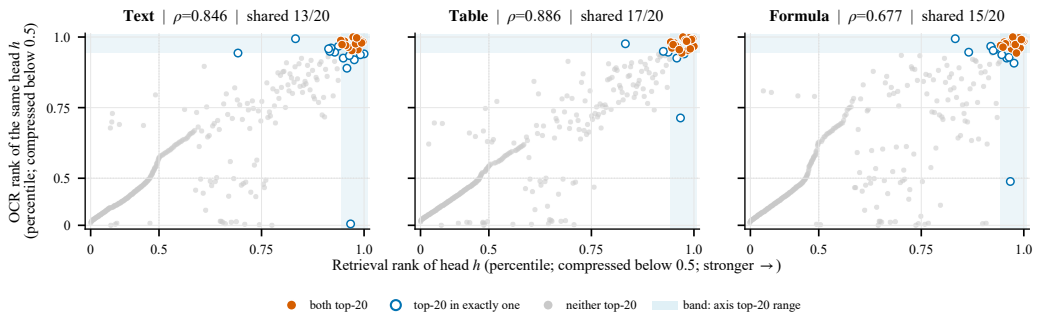

Figure 3: Structural correspondence between independently identified OCR and retrieval/copy heads in Qwen2-VL-2B. Each point represents the same head's retrieval and task-specific OCR rank percentiles. Orange points belong to both top-20 sets, blue rings to exactly one, and gray points to neither. Shaded bands mark the top-20 ranges. ρ denotes all-head Spearman correlation; ranks below the 50th percentile are compressed for readability.

top-20 OCR-head sets identified for text, table, or formula recognition reduces held-out retrieval ROUGE-1 recall from 0.9056 to 0.3556, 0.4000, and 0.2500, respectively.

To approximately match Qwen2-VL's top-20 budget ($20/336 \approx 6\%$), we ablate 27 of Qwen3-VL's 448 decoder query heads. At this budget, ablating retrieval heads increases text NED from 0.012 to 0.166 and reduces formula CDM from 0.970 to 0.528. Table TEDS also falls from 0.622 to 0.126. Conversely, ablating OCR heads ranked for text, table, and formula recognition reduces retrieval recall from 1.000 to 0.967, 0.406, and 0.339, respectively.

At comparable head budgets, these cross-task interventions support functional reuse of a retrieval/copy operation in both model generations. OCR heads are therefore better viewed as task-active components within a broader retrieval/copy mechanism, with substantial but incomplete correspondence to retrieval/copy heads rather than a separate OCR-specific circuit.

5.3 Specificity beyond generic visual importance

The structural and causal correspondence above could still admit a simpler explanation: the overlapping heads may be generally important for visual grounding rather than specifically involved in retrieval and copying. We therefore evaluate them on POPE (Li et al., 2023), which requires visual grounding but neither transcription nor source copying. Across all three task-specific overlap sets, targeted ablation does not significantly exceed layer-matched random controls, with all confidence intervals for the excess effects including zero. Thus, generic visual importance alone is unlikely to explain the OCR–retrieval correspondence (full results in Appendix E).

5.4 Counterfactual source-identity transmission

Structural overlap and cross-task ablation establish that OCR and textual retrieval rely on a shared, causally relevant head population. They do not reveal whether these heads convey which glyph appears at the visual location being read. We therefore use counterfactual value patching to test whether changing the value vectors at that location shifts the output toward the donor glyph while attention remains fixed, and whether the effect disappears when we patch unrelated visual locations.

For each OCR task, we construct minimal image pairs that differ only in one target glyph, keeping its location and surrounding content fixed. We retain pairs where the model correctly transcribes both variants, each target glyph maps to a single tokenizer token, and both images have visual evidence at the target position. This yields 184 text, 105 table, and 128 formula pairs. Additional construction details are provided in Appendix H.

Let A denote the recipient image and B the donor image, with target tokens y_A and y_B, respectively. For an overlap head h at target generation step t, α_{tj}^h is the attention weight on source position j, V_j^h is the value vector at that position, and W_O^h projects the weighted value into the model's residual stream. Summing over positions accessible to the query gives the head output

$$o_t^h = \sum_j \alpha_{tj}^h V_j^h W_O^h, \tag{4}$$

Source-identity transmission under fixed attention. Our first intervention asks whether these heads carry information about *what* is present at the visual source selected for the current output token. We keep the recipient attention pattern fixed and replace only the value contribution at the target visual evidence positions $E(z_t)$:

$$\tilde{o}_t^h = \sum_{j \notin E(z_t)} \alpha_{tj}^{A,h} V_j^{A,h} W_O^h + \sum_{j \in E(z_t)} \alpha_{tj}^{A,h} V_j^{B,h} W_O^h. \tag{5}$$

Thus, the recipient determines *where* the head attends, while the donor changes only the value content contributed by the selected source region. Queries, keys, attention weights, and all non-target value contributions remain those of the recipient. We jointly patch the task-specific OCR–retrieval overlap heads.

For $A \leftarrow B$, we measure the change in $\mathrm{logit}(y_B) - \mathrm{logit}(y_A)$ and average it with the reverse direction within each pair. Positive values indicate a donor-directed shift under fixed recipient attention. We compare against 20 random head sets matched in size and per-layer allocation.

As shown in Table 1, target-source value replacement through the OCR–retrieval overlap heads produces a large donor-directed shift in all three OCR tasks. The mean shifts are 4.476 for text, 2.425 for tables, and 2.498 for formulas, compared with 0.658, 0.812, and 0.168 for exactly layer-matched random head sets. Pair-bootstrap 95% confidence intervals for the target-source intervention are $[4.251, 4.701]$, $[2.076, 2.785]$, and $[2.347, 2.648]$, respectively. For every task, a one-sided paired sign-flip test comparing the overlap-head intervention with the within-pair mean of the matched-random sets gives $p < 10^{-5}$.

The effect is also consistent at the pair level: every retained pair shifts toward the donor identity in both patch directions. With recipient attention fixed, this supports source-identity transmission through the value pathway rather than a change in attention routing.

Spatial selectivity of value transmission. Having established source-identity transmission, we next ask whether the same heads transmit visual value content indiscriminately or whether their causal effect is tied to the visual source associated with the current output token. We therefore perform a second intervention using the *same OCR–retrieval overlap heads and the same recipient attention*, but replace values at an equal number of non-target visual positions while leaving the target-region values untouched. The non-target positions are chosen to maximize their minimum sequence-index distance from $E(z_t)$. Head identity and the number of patched positions are therefore matched across the two interventions, although attention mass and value-difference magnitude are not explicitly matched.

In sharp contrast to target-source replacement, patching unrelated visual locations produces essentially no donor-directed effect: 0.004, -0.003, and -0.003 for text, tables, and formulas, respectively, with all corresponding confidence intervals containing zero. The same heads therefore do not simply propagate arbitrary visual value perturbations toward the output. Their causal contribution depends on value content originating from the visual source associated with the current output token.

The target-source intervention does not flip the pairwise preference between y_A and y_B, nor does it change the full-vocabulary argmax. We therefore do not view the overlap heads as sufficient on their own to emit the copied token. Instead, the two interventions show that they selectively transmit source-identity information from the target region under fixed attention, while unrelated visual locations contribute essentially no effect.

Table 1: Counterfactual value patching in Qwen2-VL-2B. Values are donor-directed next-token logit shifts. Target uses OCR–retrieval overlap heads at the target evidence; Random uses matched random heads; Non-target uses the same overlap heads at unrelated visual positions.

Task	#Pairs	Target source	Matched heads	Non-target source
Text	184	4.476	0.658	0.004
Table	105	2.425	0.812	-0.003
Formula	128	2.498	0.168	-0.003

6 How Does OCR Specialization Change the Circuit?

6.1 Matched specialization comparison

We next ask how OCR specialization modifies this shared retrieval-and-copy substrate: does specialization introduce a new OCR-specific head organization, or does it primarily reorganize an existing one? To answer this question, we perform matched general-to-specialized comparisons on Qwen2-VL-2B and Qwen3-VL-2B. We denote the general model before OCR specialization by G and the specialized checkpoint by S. Within each backbone, we apply the same OCR-head discovery procedure to G and S using the same evaluation samples, enabling a direct comparison of head identity and organization.

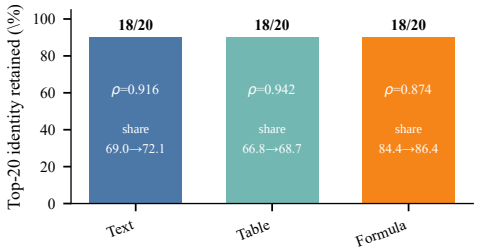

Figure 4: Qwen2-VL OCR specialization preserves 18/20 top heads in each task. Bar height shows identity retention; ρ is the all-head Spearman correlation. The percentages inside each bar show the share of total OCR-head score carried by the top 20, from the general to the specialized model.

To distinguish between circuit replacement and circuit reweighting, we examine specialization-induced changes from three complementary perspectives. We first ask whether OCR specialization changes the identity and organization of the dominant heads. We then examine whether specialization alters the distribution of functional importance within this largely preserved head population. Finally, we test whether these organizational changes correspond to shifts in causal contribution through cross-ranked ablations. This decomposition separates changes in which heads are recruited from changes in how strongly they contribute to OCR behavior.

Each configuration uses its own deployment prompt to ensure that the model produces outputs in a fixed and consistent format before and after fine-tuning (Tables 2 and 3). The matched comparison therefore characterizes the deployed general-to-specialized change, incorporating both parameter adaptation and prompt differences rather than isolating parameter updates alone.

6.2 OCR-head organization before and after specialization

We first examine whether OCR specialization changes which heads dominate OCR and how they are organized across layers. In Qwen2-VL, the general and specialized configurations share 18 of their top 20 OCR heads for each of text, table, and formula recognition (Figure 4 and Table 11). This agreement extends beyond the leading heads: the all-head rankings remain strongly correlated, with Spearman correlations of 0.9164, 0.9423, and 0.8736 for text, tables, and formulas, respectively. The score-weighted layer positions are also stable, shifting by at most 0.34 layers across the three tasks (full statistics in Appendix G).

In Qwen3-VL, the general and specialized configurations share 17/20, 20/20, and 19/20 leading heads for text, tables, and formulas, respectively.

6.3 OCR-head score redistribution

Although the dominant OCR-head identities remain largely stable, their relative scores change under specialization. In Qwen2-VL, the effective head count decreases across all three tasks, while the share of the summed OCR scores across all heads accounted for by the top 20 increases from 69.0%

to 72.1% for text, 66.8% to 68.7% for tables, and 84.4% to 86.4% for formulas. The decrease in effective head count and the increase in top-20 score share indicate a more concentrated OCR-score distribution after specialization. Cross-task sharing also increases: the three-way intersection of the text, table, and formula top-20 sets grows from 12 to 14 heads, indicating a larger common core across OCR content types. At the individual-head level, some conserved heads increase in OCR score while others decrease, indicating redistribution rather than uniform amplification.

6.4 Cross-ranked causal effects

While OCR-score redistribution reveals how specialization changes the functional emphasis of different heads, we further examine whether these changes correspond to shifts in causal contribution. We rank heads independently before and after specialization, then ablate each ranking's top-k set separately in both the general and specialized models. Each ablation is compared with five random head sets matched to its head count and per-layer allocation (Table 10).

In Qwen2-VL, ablating either top-20 set impairs OCR in both model configurations, but their relative effects differ by task. For text, the specialized-ranked set causes a larger NED increase than the general-ranked set in each model; for formulas, the general-ranked set causes a larger CDM decrease. For tables, the two sets cause similar TEDS decreases within the general model (0.460 and 0.452 for the general- and specialized-ranked sets) and larger decreases within the specialized model (0.667 and 0.696, respectively). Thus, the effects change with specialization even though most top-ranked head identities are retained.

In Qwen3-VL text OCR, ablating the general-ranked top-20 increases NED by 0.3968 in the general model and 0.3149 in the specialized model. Ablating the specialized-ranked top-20 increases NED by 0.2904 and 0.4201, respectively. Each increase exceeds the mean effect of its own layer-matched random controls. For tables, the two rankings contain the same top-20 heads. Ablating this set decreases TEDS by 0.2796 in the general model and 0.4358 in the specialized model. The separately sampled random controls have mean TEDS decreases of 0.2272 and 0.2834 for the two rankings in the general model, and 0.0655 and 0.0391 in the specialized model. The general-model table effect therefore does not exceed both control means. For formulas in the specialized model, ablating the specialized-ranked set causes a larger CDM decrease than ablating the general-ranked set (0.3917 versus 0.1752); their respective random-control means are 0.0128 and 0.0705.

Overall, the cross-ranked ablations suggest circuit conservation with causal reweighting, where OCR specialization largely preserves the underlying head population while altering the distribution of causal contribution within that population.

7 Conclusion

We studied the attention-head mechanisms underlying full-sequence OCR. Evidence-grounded discovery and held-out ablations identify sparse, causally relevant head sets for text, table, and formula transcription across three OCR-specialized models. In Qwen2-VL and Qwen3-VL, independently discovered OCR and textual retrieval/copy heads overlap substantially and exhibit cross-task causal effects. In Qwen2-VL, patching values at the target visual location shifts next-token logits toward a donor glyph while attention remains fixed. Matched comparisons before and after OCR specialization show that most leading OCR heads are retained, although their scores and causal effects change. Together, these results suggest that specialization changes how existing retrieval/copy-related heads contribute to visual transcription.

REFERENCES

Ingeol Baek, Hwan Chang, Sunghyun Ryu, and Hwanhee Lee. How do large vision-language models see text in image? unveiling the distinctive role of OCR heads. In *Proceedings of EMNLP*, pp. 20441–20453, 2025. URL https://aclanthology.org/2025.emnlp-main.1032/.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. 2025. URL https://arxiv.org/abs/2511.21631.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. 2023. URL https://arxiv.org/abs/2308.13418.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, Dexin Jin, Yuxuan Zhang, Guohong Wen, Yanfeng Wang, Yutao Zhang, Xiaohan Zhang, Wenyi Hong, Yukuo Cen, Da Yin, Bin Chen, Wenmeng Yu, Xiaotao Gu, and Jie Tang. GLM-OCR technical report. *arXiv preprint arXiv:2603.10910*, 2026. URL https://arxiv.org/abs/2603.10910.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *Computer Vision ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, pp. 498–517, Berlin, Heidelberg, 2022. Springer-Verlag. URL https://doi.org/10.1007/978-3-031-19815-1_29.

Aaron Branson Cigres Li, Zhaowei Wang, Yu Zhao, Yiming Du, Haobo Li, Xiyu Ren, Ginny Wong, Simon See, Lishu Luo, Haodong Duan, Pasquale Minervini, and Yangqiu Song. Can retrieval heads see images? multimodal retrieval heads in long-context vision-language models. *arXiv preprint arXiv:2605.27243*, 2026. URL https://arxiv.org/abs/2605.27243.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2409.09951.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Zirui Tang, Boyu Niu, Yuanhong Zheng, Dongsheng Ma, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Fangdong Wang, Jingzhou Chen, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, RuiLiang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Dechen Lin, Shenguanlin, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, Lei Bai, Pei Chu, Weijia Li, Jiang Wu, Lijun

Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. In *Proceedings of ACL: Industry Track*, pp. 13–42, 2026. URL https://aclanthology.org/2026.acl-industry.3/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. 2022. URL https://arxiv.org/abs/2209.11895.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of CVPR*, 2025. URL https://arxiv.org/abs/2412.07626.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in pdfs with vision language models. 2025. URL https://arxiv.org/abs/2502.18443.

Ruoxi Sun, Quantong Qiu, Juntao Li, Zecheng Tang, Yihang Lou, and Min Zhang. Mechanistic insights into functional sparsity in multimodal llms via core heads, 2026. URL https://arxiv.org/abs/2606.05843.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. Image over text: Transforming formula recognition evaluation with character detection matching. 2025. URL https://arxiv.org/abs/2409.03643.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. 2024. URL https://arxiv.org/abs/2409.12191.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. 2024. URL https://arxiv.org/abs/2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. 2025. URL https://arxiv.org/abs/2510.18234.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024. URL https://arxiv.org/abs/2404.15574.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026. URL https://arxiv.org/abs/2606.03264.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. 2020. URL https://arxiv.org/abs/1911.10683.

A AI USE STATEMENT

Large language models were used only to polish the language and improve the readability of this manuscript. The authors reviewed all AI-assisted edits and take full responsibility for the scientific content and final text.

B PROTOCOL AND EVIDENCE CONSTRUCTION

Visible evidence. For text, we re-render official annotations and retain a box for each character; a BPE token maps to the union of its constituent character evidence rather than one enclosing word rectangle. For tables, the complete structured output is preserved, but only visible cell-content tokens enter discovery; HTML/OTSL tags are excluded. For formulas, rendered glyphs, operators, and composite objects are aligned to source spans, while braces and non-visible control syntax are excluded. A visual-token position is included in $E(z)$ when its patch rectangle has IoU above 0.1 with at least one visible region for span z.

Free-generation replay. Greedy generations are replayed using their exact token IDs. We audit replay by checking whether every replayed position predicts the originally generated next token. In the GLM-OCR discovery runs, mean agreement across folds A and B is 100% for text, 99.99% for tables, and 99.96% for formulas. A generated BPE token is scoreable only when its entire decoded span belongs to a maximal exact reference–prediction match block. This avoids assigning a visual target to partially correct or mismatched output.

Fold transfer and interventions. Heads discovered on A are intervened on B, heads discovered on B on A, and the A/B consensus on C. Principal free-generation interventions zero head output only at answer positions.

C MODELS, COHORTS, AND METRICS

The evaluated models differ in decoder depth and query heads per layer: GLM-OCR has 16 layers with 16 query heads per layer (256 total); MinerU2.5 has 24 layers with 14 per layer (336 total); PaddleOCR-VL-1.6 has 18 layers with 16 per layer (288 total); Qwen2-VL-2B has 28 layers with 12 per layer (336 total); and Qwen3-VL-2B has 28 layers with 16 per layer (448 total). Each specialist uses 400 instances per task, divided into 160/160/80 by source page.

Prompts and output contracts. The OCR specialists are evaluated with task instructions documented by their model providers. GLM-OCR and MinerU2.5 use `Text Recognition:`, `Table Recognition:`, and `Formula Recognition:`; PaddleOCR-VL-1.6 uses `OCR:` for text and the same table and formula instructions. Text targets preserve the visible reading order, formula targets are delimited LaTeX, and table targets are canonical HTML. For MinerU2.5 and PaddleOCR-VL-1.6, which generate tables in OTSL format, the evaluation adapter converts HTML references to OTSL before token alignment.

The specialized Qwen2-VL and Qwen3-VL checkpoints use the short training-time prompts `Text Recognition:`, `Table Recognition:`, and `Formula Recognition:`. The untuned checkpoints do not reliably follow the required output formats with these short prompts, so prompt calibration selects the complete deployment prompts in Table 2 for Qwen2-VL and Table 3 for Qwen3-VL, whose text prompt is identical to Qwen2-VL's.

Generation budgets and cohort filtering. The specialist token budgets are 96, 128, and 2,048 new tokens for text, formula, and table, respectively; the matched Qwen comparisons use 128, 256, and 2,048. Discovery retains generations that satisfy the task output contract and do not reach the token budget.

For text transcription, we use normalized edit distance (NED) (Ouyang et al., 2025):

$$\mathrm{NED}(\hat{y}, y) = \frac{\mathrm{Levenshtein}(\hat{y}, y)}{\max(|\hat{y}|, |y|, 1)}.$$

Table 2: Complete deployment prompts for untuned Qwen2-VL, selected by prompt calibration.

Task	Prompt
Text	`Perform OCR on this image. Output only the visible text in natural reading order, verbatim. Preserve case, punctuation, symbols, spacing, and line breaks. No commentary, labels, Markdown, or code fences.`
Table	`Transcribe the table into compact HTML. Return only one <table>...</table> element with rows in reading order. Use <tr> and <td> or <th>; preserve every cell's text exactly and use rowspan or colspan only when visually required. Use no other attributes, prose, Markdown, or code fence.`
Formula	`Read only the mathematical expression shown. Respond with its exact LaTeX transcription between $$ delimiters. Keep fractions, roots, superscripts, subscripts, accents, relations, brackets, and punctuation. Add no explanation.`

Table 3: Deployment prompts for untuned Qwen3-VL. The text prompt is identical to that of Qwen2-VL (Table 2).

Task	Prompt
Text	`Perform OCR on this image. Output only the visible text in natural reading order, verbatim. Preserve case, punctuation, symbols, spacing, and line breaks. No commentary, labels, Markdown, or code fences.`
Table	`Perform table OCR and output only canonical HTML beginning with <table> and ending with </table>. Preserve row and column order and all visible cell text. Use only table, tr, td, and th tags plus necessary rowspan/colspan attributes. Do not add commentary or a code block.`
Formula	`Transcribe the displayed mathematical formula exactly into LaTeX. Return one LaTeX string only, enclosed in $$...$$. Preserve all symbols, operators, scripts, delimiters, and visible terminal punctuation. Do not explain and do not use a code fence.`

The matched general-to-specialized experiments fine-tune Qwen2-VL-2B and Qwen3-VL-2B with rank-32 LoRA (alpha 64) for two epochs on 12,000 mixed OCR examples.

For Qwen3-VL-2B, text NED is 0.0155 before and 0.0168 after specialization, while table TEDS rises from 0.7855 to 0.9192 on the full 400-example sets.

Following the retrieval-head protocol of Wu et al. (2024), we use three needle–question pairs, 20 context lengths, and 10 insertion depths for 600 trials per architecture. Our context lengths span 1K–32K tokens rather than the original 1K–50K. As in the original protocol, a retrieval hit requires the attended needle token to match the generated token; only generations with ROUGE-1 recall above 0.5 enter head scoring, and heads with mean retrieval score above 0.1 are selected. Qwen2-VL and Qwen3-VL yield 594 and 592 successful trials and identify 31 and 30 retrieval heads, respectively.

D Complete Discovery Results

We further characterize the discovered OCR-head populations through their head-wise score distributions, cross-fold stability, and cross-task overlap. Figure 5 shows the full OCR-score maps for text, table, and formula recognition across all three OCR specialists.

Cross-fold stability. The independently discovered head populations are highly stable across disjoint data folds. For GLM-OCR, the free-generation top-20 sets overlap by 20/20, 20/20, and 19/20 heads for text, tables, and formulas, respectively; MinerU2.5 retains 19/20, 19/20, and 20/20, and PaddleOCR-VL-1.6 retains 19/20, 20/20, and 20/20. Table 5 additionally reports the teacher-forced diagnostic for GLM-OCR, which yields similarly stable rankings across folds.

Table 4: Retrieval/copy reuse in untuned generalist VLMs. "OCR in R" is threshold-set containment; all top-20 overlaps exceed architecture-specific random expectation with hypergeometric $p < 10^{-11}$.

Model	OCR task	Retrieval heads	Top-20 overlap	Spearman	OCR in R
Qwen2-VL	Text	31	13/20	0.846	92.3%
	Table	31	17/20	0.886	93.8%
	Formula	31	15/20	0.677	85.7%
Qwen3-VL	Text	30	15/20	0.835	74.1%
	Table	30	15/20	0.859	84.6%
	Formula	30	13/20	0.806	83.3%

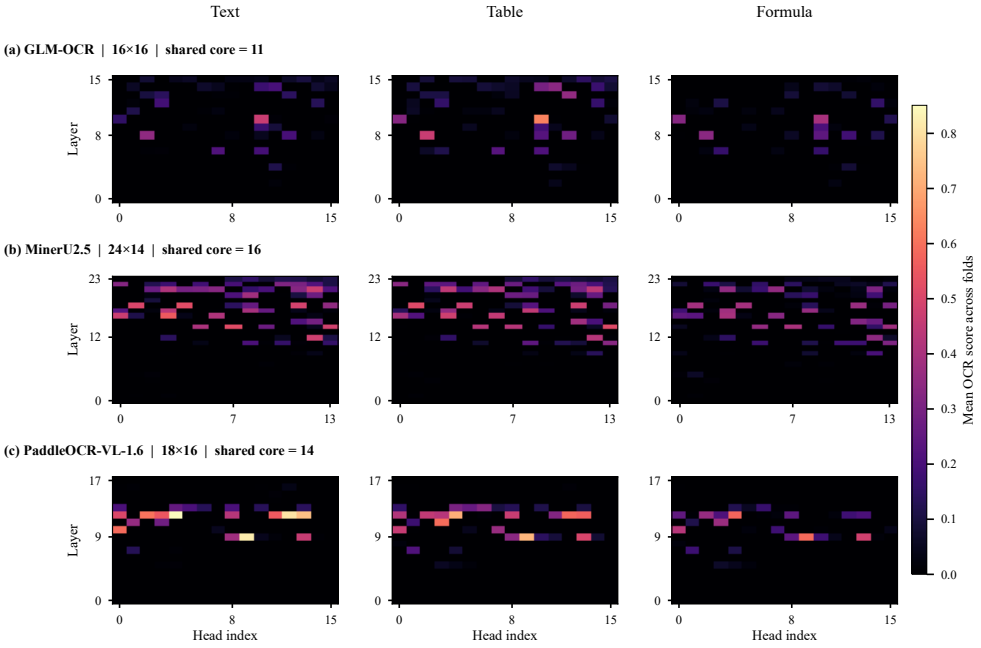

Figure 5: Head-wise OCR scores for text, tables, and formulas in three OCR specialists. High-scoring heads are sparse and architecture dependent, with task-dependent variation in their locations and strengths.

Cross-task sharing. The task-specific top-20 sets also overlap substantially within each specialist. Their three-way intersections contain 11 heads in GLM-OCR, 16 in MinerU2.5, and 14 in PaddleOCR-VL-1.6, forming the largest overlap category in all three models (Table 6).

Table 5: GLM-OCR cross-fold discovery stability. TF is the teacher-forced diagnostic and FG is free-generation replay; both top-set overlaps use 20 heads.

Task	TF ρ	TF top-20	FG top-20
Text	0.9931	19/20	20/20
Table content	0.9951	19/20	20/20
Formula	0.9944	20/20	19/20

Table 6: Mutually exclusive overlap categories of the task-specific top-20 OCR-head sets. Pairwise-only columns contain heads shared by exactly the named two tasks, excluding the third.

Model	Text only	Table only	Formula only	Text–Table only	Text–Formula only	Table–Formula only	All three
GLM-OCR	4	2	7	5	0	2	11
MinerU2.5	3	1	2	1	0	2	16
PaddleOCR-VL-1.6	3	0	3	3	0	3	14

E Detailed Causal Results and Controls

Held-out OCR ablation. Table 7 reports the measured NED, TEDS, and CDM under top-20 head ablation for the three OCR specialists, alongside the intact baseline and the mean over five matched random head sets.

Table 7: Held-out causal validation in specialized OCR models. Top-20 and Random denote the targeted and matched-random ablation conditions, respectively. Score is the metric value under each ablation, and Δ is the signed change from the intact baseline. Random results are averaged over five head sets exactly matched to the target set's per-layer head counts for all three tasks.

		Top-20 ablation		Random ablation	
Model	Base	Score	Δ	Score	Δ
Text (NED)					
GLM-OCR	0.033	0.337	$+0.303$	0.086	$+0.053$
MinerU2.5	0.027	0.070	$+0.043$	0.026	-0.001
PaddleOCR-VL-1.6	0.021	0.621	$+0.600$	0.023	$+0.002$
Table (TEDS)					
GLM-OCR	0.964	0.061	-0.903	0.932	-0.032
MinerU2.5	0.977	0.512	-0.465	0.862	-0.115
PaddleOCR-VL-1.6	0.978	0.111	-0.867	0.950	-0.027
Formula (CDM)					
GLM-OCR	0.978	0.132	-0.846	0.962	-0.016
MinerU2.5	0.982	0.528	-0.454	0.983	-0.001
PaddleOCR-VL-1.6	0.948	0.286	-0.662	0.925	-0.023

Cross-model and cross-task effects. Cross-task effects differ substantially across the three specialists. MinerU is asymmetric, especially on text targets; GLM-OCR is stronger and more uniform; and PaddleOCR-VL shows large effects across nearly every discovery-task and evaluation-task combination in Figure 6(b). This variation supports a shared routing pool with architecture- and task-dependent strength rather than one universal profile.

Specificity beyond OCR. To test whether the overlapping OCR and retrieval heads are broadly important for visual recognition rather than specifically involved in copying, we ablate the three task-specific overlap sets in Qwen2-VL-2B and evaluate object-presence judgments on POPE. Each intervention is compared with 20 random head sets matched in size and per-layer allocation. As shown in Table 8, targeted ablation reduces accuracy by only 0.22–0.56 percentage points, with no

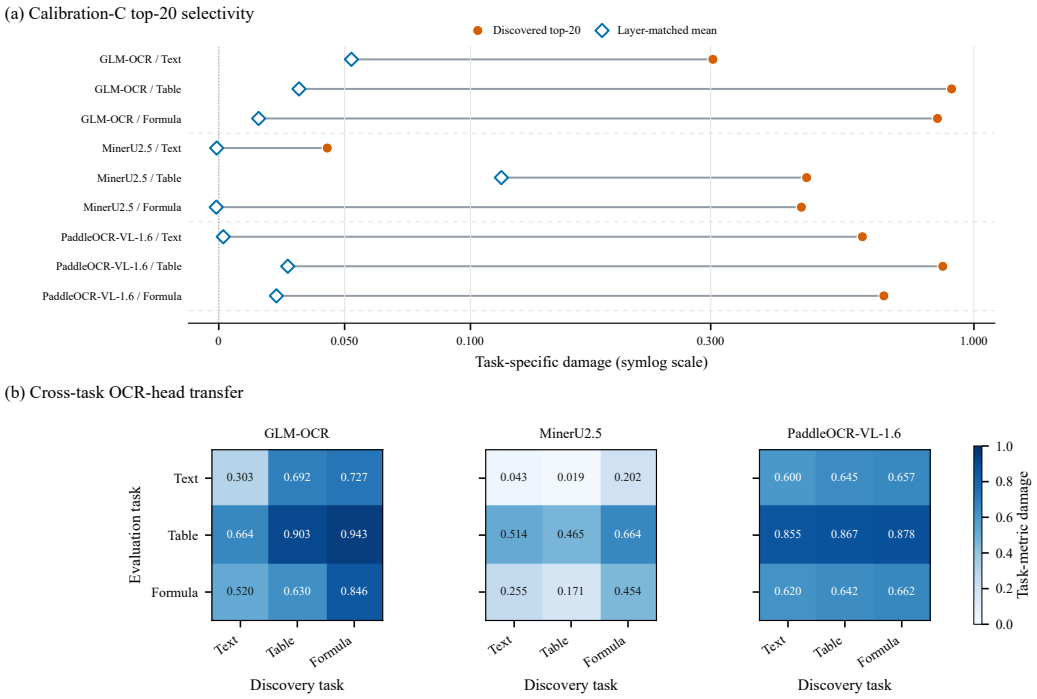

Figure 6: Causal effects of OCR-head ablation across three specialized OCR models. (a) Task-specific damage from ablating the discovered top-20 heads and layer-matched random heads. Text, table, and formula use ΔNED, TEDS drop, and CDM drop, respectively. (b) Cross-task transfer of the task-specific top-20 OCR heads. Columns indicate the discovery task and rows the evaluation task.

statistically significant excess over the random-control means. These results suggest that the effects observed in OCR and retrieval are not simply a consequence of disrupting generally important visual heads, although they do not establish exclusive involvement in copying.

Table 8: POPE specificity control. Values are accuracy drops in percentage points, pooled over Random, Popular, and Adversarial splits. Random denotes the mean of 20 layer-matched controls; Excess is the targeted drop minus the random mean.

Set	**#H**	**ΔAcc.**	**Random**	**Excess**	p
Text–Retrieval	12	0.56	0.16	0.39	.143
Table–Retrieval	15	0.22	0.21	0.01	.667
Formula–Retrieval	6	0.22	0.11	0.11	.381

F Token-Level Examples

To complement the aggregate ablation results in specialized OCR models, we examine where an identified OCR head attends during individual generation steps and how the corresponding outputs change under ablation. We use MinerU2.5 head L16H03 (layer index 16, head index 3) as a case study across text, table, and formula recognition. Figure 7 pairs the visual evidence for each selected output token with this head's attention over image patches and compares baseline generation with ablation of either L16H03 alone or the task-specific top-20 OCR heads.

L16H03 assigns 35.8%, 50.1%, and 75.6% of its visual attention to the evidence patches for `eleven`, `Forbidden`, and α, respectively. Its five most-attended visual patches account for 74.7%, 74.4%, and 94.3% of visual attention, indicating spatial concentration rather than necessarily target alignment. Ablating L16H03 alone produces only spacing or notation changes, with no output change in the table example. In contrast, top-20 ablation replaces `eleven` with `twelve` and omits the following option, changes the target table cell from `Forbidden` to `Allowed`, and rewrites the formula. These examples distinguish visual alignment by an individual head from the causal contribution of a jointly ablated head set; the content errors cannot be attributed to L16H03 alone.

G Matched General-to-Specialized Results

The three-task intersection grows from 12 to 14 heads, heads shared by exactly two tasks fall from 9 to 5, and the task-specific periphery changes from 6 to 8. L19H06 and L17H01 strengthen across all tasks, whereas L17H08 weakens across all three. These opposing changes rule out uniform amplification and instead indicate redistribution within a conserved circuit.

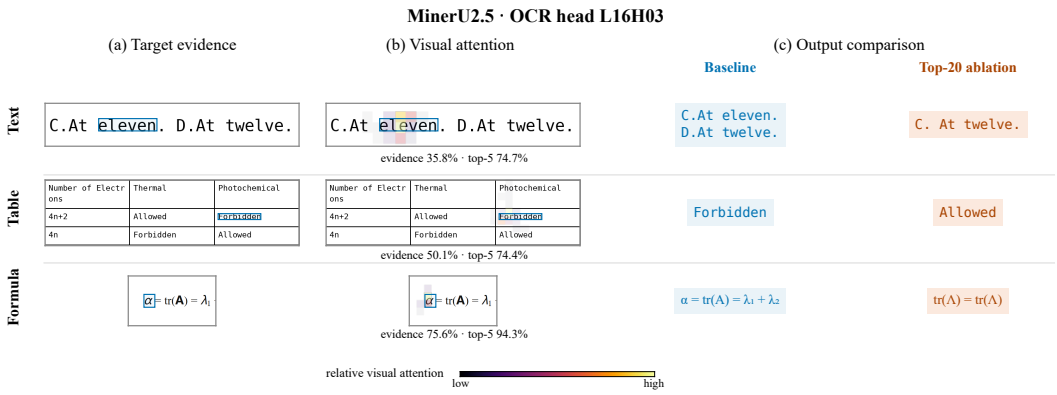

Figure 7: Token-level examples in MinerU2.5. Rows show text, table, and formula recognition. (a) Input crops with target evidence outlined in blue. (b) Visual attention of L16H03 at the selected generation step. Evidence and top-5 percentages are normalized by total attention to visual patches; argmax denotes the most-attended visual patch. (c) Corresponding output excerpts before and after task-specific top-20 ablation, with single-head effects reported separately. Input crops show local evidence and need not contain the full output excerpt.

Table 9: Matched specialization replicates across Qwen generations. Discovery is computed independently in G and S on common valid samples; G denotes the general model and S its OCR-specialized checkpoint.

Backbone	Task	Common N	Qualified G→S	Top-20 overlap	Spearman
Qwen2-VL	Text	319	17→20	18/20	0.9164
	Table	290	16→20	18/20	0.9423
	Formula	312	8→10	18/20	0.8736
Qwen3-VL	Text	319	27→33	17/20	0.9349
	Table	299	26→34	20/20	0.9372
	Formula	317	18→15	19/20	0.8946

Table 10: Held-out cross-ranked causal effects. Cells report performance after ablation, with signed change from base in parentheses. Text uses NED (↓), tables use TEDS (↑), and formulas use CDM (↑). G heads are ranked in the general model and S heads in its OCR-specialized checkpoint. Rnd mean averages ten layer-matched random control sets (five matched to each ranking in head count and per-layer allocation).

Model/task	Eval.	k	Base	G heads	S heads	Rnd mean
Qwen3-VL Text	General	5	0.012	0.022 (+0.010)	0.019 (+0.006)	0.018 (+0.005)
		10	0.012	0.028 (+0.015)	0.018 (+0.005)	0.019 (+0.006)
		20	0.012	0.409 (+0.397)	0.303 (+0.290)	0.020 (+0.007)
Qwen3-VL Text	Specialized	5	0.019	0.022 (+0.002)	0.024 (+0.005)	0.018 (−0.001)
		10	0.019	0.022 (+0.003)	0.022 (+0.002)	0.019 (0.000)
		20	0.019	0.334 (+0.315)	0.439 (+0.420)	0.020 (+0.001)
Qwen3-VL Table	General	5	0.773	0.645 (−0.128)	0.666 (−0.107)	0.603 (−0.170)
		10	0.773	0.673 (−0.100)	0.673 (−0.100)	0.579 (−0.194)
		20	0.773	0.493 (−0.280)	0.493 (−0.280)	0.518 (−0.255)
Qwen3-VL Table	Specialized	5	0.926	0.918 (−0.009)	0.931 (+0.005)	0.899 (−0.027)
		10	0.926	0.914 (−0.012)	0.914 (−0.012)	0.888 (−0.039)
		20	0.926	0.491 (−0.436)	0.491 (−0.436)	0.874 (−0.052)
Qwen3-VL Formula	General	5	0.969	0.967 (−0.002)	0.968 (−0.001)	0.950 (−0.019)
		10	0.969	0.963 (−0.007)	0.940 (−0.029)	0.956 (−0.013)
		20	0.969	0.880 (−0.089)	0.590 (−0.379)	0.933 (−0.036)
Qwen3-VL Formula	Specialized	5	0.976	0.973 (−0.003)	0.971 (−0.005)	0.967 (−0.010)
		10	0.976	0.975 (−0.002)	0.941 (−0.035)	0.961 (−0.015)
		20	0.976	0.801 (−0.175)	0.584 (−0.392)	0.934 (−0.042)
Qwen2-VL Text	General	5	0.080	0.140 (+0.060)	0.118 (+0.038)	0.092 (+0.012)
		10	0.080	0.153 (+0.073)	0.229 (+0.149)	0.099 (+0.019)
		20	0.080	0.415 (+0.335)	0.632 (+0.552)	0.149 (+0.068)
Qwen2-VL Text	Specialized	5	0.013	0.013 (+0.000)	0.012 (−0.001)	0.014 (+0.002)
		10	0.013	0.033 (+0.020)	0.103 (+0.090)	0.021 (+0.008)
		20	0.013	0.273 (+0.260)	0.508 (+0.495)	0.022 (+0.009)
Qwen2-VL Table	General	5	0.613	0.428 (−0.184)	0.462 (−0.151)	0.605 (−0.007)
		10	0.613	0.403 (−0.209)	0.298 (−0.315)	0.580 (−0.032)
		20	0.613	0.152 (−0.460)	0.161 (−0.452)	0.392 (−0.220)
Qwen2-VL Table	Specialized	5	0.842	0.579 (−0.263)	0.604 (−0.238)	0.819 (−0.023)
		10	0.842	0.535 (−0.307)	0.368 (−0.474)	0.788 (−0.054)
		20	0.842	0.175 (−0.667)	0.146 (−0.696)	0.576 (−0.266)
Qwen2-VL Formula	General	5	0.959	0.705 (−0.254)	0.678 (−0.281)	0.958 (−0.001)
		10	0.959	0.629 (−0.330)	0.477 (−0.482)	0.957 (−0.002)
		20	0.959	0.334 (−0.625)	0.497 (−0.462)	0.952 (−0.007)
Qwen2-VL Formula	Specialized	5	0.951	0.709 (−0.242)	0.683 (−0.268)	0.945 (−0.006)
		10	0.951	0.610 (−0.340)	0.431 (−0.520)	0.943 (−0.008)
		20	0.951	0.304 (−0.647)	0.402 (−0.549)	0.934 (−0.016)

Table 11: Matched OCR-head discovery in the general (G) and specialized (S) Qwen2-VL models. Top-head overlap, all-head rank correlation, and score-weighted layer position show little change in head organization.

Task	Common N	Qualified G→S	Top-20 overlap	Spearman	Weighted layer G→S
Text	319	17→20	18/20	0.9164	18.695→18.681
Table	290	16→20	18/20	0.9423	18.754→18.697
Formula	312	8→10	18/20	0.8736	17.227→17.566

H EXAMPLES OF SOURCE-IDENTITY REPLACEMENT

To illustrate how target-region value replacement changes the evidence for a specific output token, we examine matched Qwen2-VL-2B image pairs for text, tables, and formulas (Figure 8). Each pair shares the generated prefix before its target token, so both variants are compared at the same decoding step.

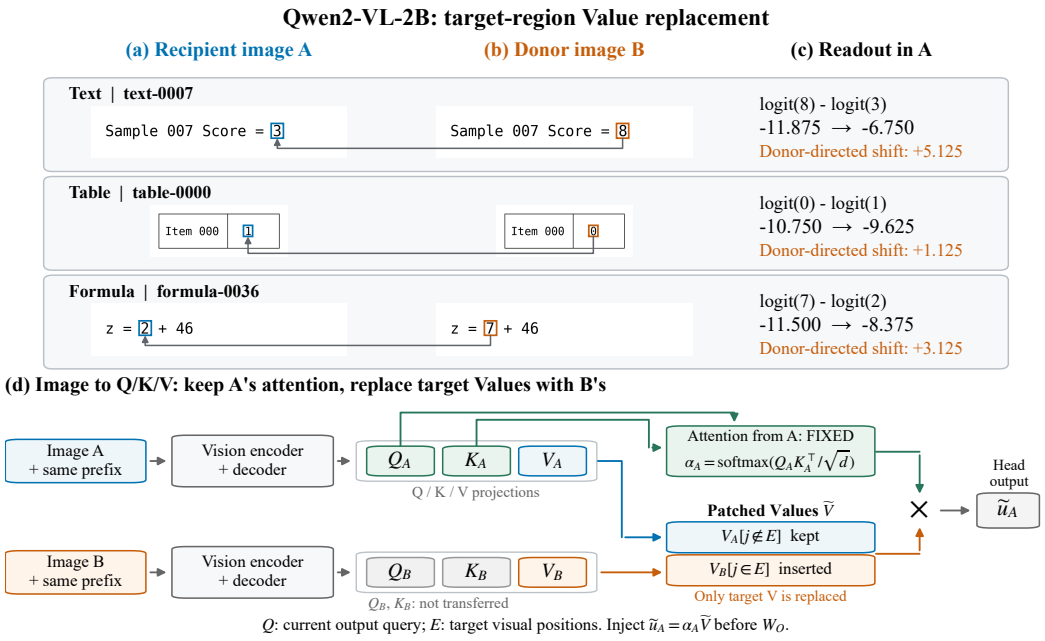

Figure 8: Examples of $A \leftarrow B$ source-identity replacement. Panel (d) traces image and prefix processing to Q/K/V: A's attention is fixed (green), non-target values are retained (blue), and B supplies target values only (orange). The overlap sets contain 13, 17, and 15 heads for text, tables, and formulas. Logits shift toward the donor without changing the next-token argmax.

1. Obtain Q/K/V from paired inputs. Image A reads `Sample 007 Score = 3`; B substitutes `8` for `3`. Both are recognized correctly. Processing each image and the same pre-digit output prefix through the vision encoder and decoder yields the current-output query Q and context keys K and values V. We cache A's attention row α_A, both runs' target-region values, and A's clean head output u_A. The glyph boxes identify matched visual positions E; table and formula pairs follow the same one-digit construction.

2. Fix A's attention and replace only target values. For each overlap head, we construct $\widetilde{u}_A = u_A + \sum_{j \in E} \alpha_{Aj}(V_{Bj} - V_{Aj})$. The cached attention weights determined by A's Q/K remain unchanged: B supplies only target-region values, not its Q/K. Contributions outside E also remain from A. We rerun A with the same prefix and inject $\widetilde{u}_A$ at the target step, before output projection W_O. This fixes attention in the constructed head outputs, not throughout the network: other heads are not overwritten, and downstream computation proceeds normally. Neither pixels nor output tokens are replaced.

3. Read the donor-directed logit shift. After injection, the text contrast $\mathrm{logit}(8) - \mathrm{logit}(3)$ changes from -11.875 to -6.750: a $+5.125$ shift, versus $+0.438$ averaged over 20 layer-matched controls. Table and formula shifts are $+1.125$ and $+3.125$, versus control means of $+0.188$ and $+0.163$. Wrong-location shifts are zero in all three cases. These are $A \leftarrow B$ results; the main analysis averages both directions. The shifts support donor-content transmission under fixed attention, but do not change the predicted token in these examples.