Low-Dimensional Document Structure Subspaces in Specialized vs. Emergent OCR Models: A Mechanistic Interpretability Study of Three Architectures

Guus Bouwens [1]

Abstract

Optical character recognition (OCR) models have become critical infrastructure for document intelligence, yet their internal representations remain mechanistically unexplored. We present the first mechanistic interpretability study comparing three architectures: GLM-OCR (0.9B), a purpose-built document recognition system; PaddleOCR-VL (1.5B), a second specialized OCR model; and Qwen3.5-2B, a general-purpose VLM with emergent OCR capability. Using PCA-based subspace analysis on 300 real RVL-CDIP document images per model, we find that document structure capabilities occupy *partially disentangled, low-dimensional subspaces* in all three models. PaddleOCR-VL exhibits the most concentrated representations (PC_1 explains 84.0% of variance; effective rank 2.0 at its bottleneck), while Qwen3.5-2B is the most distributed (PC_1 = 63.7%; effective rank 3.8). We introduce the *Document Structure Modularity Index* ($\mathcal{M}$), and find that both specialized models achieve higher modularity (GLM-OCR: 0.715; PaddleOCR-VL: 0.774) than the general-purpose baseline (0.704). Cross-model CKA reveals high representational alignment across all pairs (> 0.90), with the specialized-to-emergent CKA marginally exceeding the specialized-to-specialized CKA—a finding with implications for representational universality in OCR.

1. Introduction

Document understanding is foundational to modern information extraction pipelines. Across industries—legal discovery, scientific data mining, financial reporting—automated systems must accurately read printed text, recover table structure, and transcribe mathematical formulae from complex page layouts. Specialized end-to-end OCR models such as GLM-OCR (Duan et al., 2026) and Nougat (Blecher et al., 2023) now achieve near-human accuracy on structured benchmarks (e.g., OmniDocBench (Ouyang et al., 2025)), yet these systems are routinely treated as opaque black boxes. We understand *what* they produce, but not *how* they organize the information necessary to produce it.

The interpretability gap. Mechanistic interpretability has made remarkable progress in language models (Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023; Lindsey et al., 2025b) and is extending to multimodal systems (Joseph et al., 2025; Golovanevsky et al., 2025; Pach et al., 2025; Sheta et al., 2025). Steinberg & Gal (Steinberg & Gal, 2026) demonstrate that OCR information enters the language processing stream at architecture-specific bottleneck layers and is remarkably low-dimensional (PC_1 explains 72.9% of variance), while Baek et al. (Baek et al., 2025) identify dedicated "OCR heads" qualitatively distinct from general retrieval heads. However, a critical gap remains: *no mechanistic study has examined purpose-built document OCR models*. Whether such systems develop similarly structured or qualitatively different internal representations is unknown.

Our contribution. We address this gap through three contributions:

1. **First mechanistic interpretability study comparing specialized and general-purpose OCR models.** We apply PCA-based subspace analysis to GLM-OCR, PaddleOCR-VL, and Qwen3.5-2B, documenting the location, dimensionality, and cross-model alignment of document-structure representations.
2. **Discovery of document-type subspace disentanglement.** We perform *document-type-specific* PCA on real RVL-CDIP documents, showing that representations for text, tables, and mixed content occupy partially disentangled low-dimensional subspaces, quantified via Grassmann distance and principal angles.
3. **Document Structure Modularity Index ($\mathcal{M}$).** We formalize representational organization as $\mathcal{M}$ and show

[1]Independent Researcher - San Diego, California, USA. Correspondence to: Guus Bouwens <guus@bouwens.nl>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.

that specialized models achieve higher modularity (PaddleOCR-VL: 0.774; GLM-OCR: 0.715) than the general-purpose baseline (0.704).

Summary of findings. (i) All three models concentrate OCR information into low-dimensional subspaces at very early decoder depth; (ii) PaddleOCR-VL is the most extreme, with effective rank 2.0 and selectivity 0.899 at its bottleneck; (iii) cross-model CKA is uniformly high (> 0.90), indicating shared representational geometry across architectures; and (iv) surprisingly, the specialized-to-specialized CKA (GLM-OCR vs. PaddleOCR-VL: 0.909) is *lower* than either specialized-to-emergent pair (GLM-OCR vs. Qwen3.5-2B: 0.912; PaddleOCR-VL vs. Qwen3.5-2B: 0.920), suggesting convergent representational structure is not limited to task-specific models.

2. Related Work

Mechanistic interpretability in language models. The residual stream framework of Elhage et al. (Elhage et al., 2021) established that transformer components communicate by reading from and writing to a shared residual stream, enabling compositional circuit analysis. Olsson et al. (Olsson et al., 2022) identified induction heads as a mechanistic substrate for in-context learning; Wang et al. (Wang et al., 2023) reverse-engineered a 26-head circuit in GPT-2 Small; and Meng et al. (Meng et al., 2022) localized factual associations to middle-layer MLPs. Sparse autoencoders (Templeton et al., 2023; Cunningham et al., 2023) recover monosemantic features invisible at the neuron level, and Anthropic's attribution graph framework (Lindsey et al., 2025b;a) recently extended circuit tracing to frontier models. The linear representation hypothesis (Park et al., 2023) and representation engineering (Zou et al., 2023) provide theoretical grounding for PCA-based subspace methods.

Mechanistic interpretability in vision and multimodal models. Prisma (Joseph et al., 2025) supports SAE training across 75+ vision transformers and observes that vision SAEs exhibit substantially lower sparsity than language counterparts. NOTICE (Golovanevsky et al., 2025) introduced semantic image pair corruption as a causal intervention for VLMs, uncovering "universal attention heads" in BLIP and LLaVA. Pach et al. (Pach et al., 2025) extended SAEs to CLIP, demonstrating that learned features steer multimodal LLM outputs. VLM-Lens (Sheta et al., 2025) provides a unified interface for extracting intermediate VLM representations. Merullo et al. (Merullo et al., 2023) showed that image-to-text mapping in VLMs is approximately linear, supporting linear subspace methods for multimodal analysis.

Table 1. **Model architecture comparison.**

Model	Layers	d	Params	Type
GLM-OCR	16	1536	0.9B	Specialized
PaddleOCR-VL	18	1024	1.5B	Specialized
Qwen3.5-2B	24	2048	2.0B	Emergent

OCR-specific interpretability. Steinberg & Gal (Steinberg & Gal, 2026) locate OCR bottleneck layers in general-purpose VLMs (Qwen3-VL, Phi-4, InternVL3.5) via activation differences between original and text-inpainted images. They find PC_1 explains 72.9% of variance and that single-stage projection models show early bottlenecks (6–25% depth) while DeepStack models peak near 50% depth. Baek et al. (Baek et al., 2025) identify "OCR heads" that are less sparse, statically activated, and qualitatively different from retrieval heads. Both works study exclusively *general-purpose* VLMs. We are the first to (a) apply mechanistic methods to *specialized document OCR models*, (b) decompose representations by *document type* using real documents, (c) compare two specialized architectures against a general-purpose baseline, and (d) conduct cross-model representational alignment analysis via CKA.

Document OCR models. GLM-OCR (Duan et al., 2026) achieves state-of-the-art OmniDocBench 94.62, combining a compact CogViT encoder with a GLM decoder and Multi-Token Prediction (MTP) (Gloeckle et al., 2024). PaddleOCR-VL (Li et al., 2022) extends the PaddleOCR ecosystem to a vision-language framework with an 18-layer decoder. Qwen3.5 (Qwen Team, 2026) is a general-purpose VLM trained on diverse multimodal data achieving strong emergent OCR (DocVQA 93.9, OCRBench 797 at 3B scale). See Appendix A for details.

3. Methods

3.1. Model Architectures

Table 1 summarizes the three models studied. GLM-OCR uses a CogViT visual encoder (24 layers, $d_v = 1024$) connected to a 16-layer GLM decoder ($d = 1536$) via a Conv2d+SwiGLU connector that compresses 576 patch tokens to 144 visual tokens. PaddleOCR-VL pairs a vision encoder with an 18-layer language decoder ($d = 1024$) via a projector module. Qwen3.5-2B uses early fusion with a 24-layer hybrid decoder ($d = 2048$) interleaving $3\times$ Gated DeltaNet blocks with $1\times$ Gated Attention per group. Full architecture details are in Appendix A.

3.2. Dataset and Activation Extraction

RVL-CDIP corpus. We use 300 real scanned document images from RVL-CDIP (Harley et al., 2015): 100 plain

text documents, 100 table-heavy documents, and 100 mixed-content documents. No inpainting or synthetic modification is applied, in contrast to Steinberg & Gal (Steinberg & Gal, 2026).

Hook placement and aggregation. We extract post-MLP residual stream activations at each decoder layer via PyTorch forward hooks. Visual token activations are mean-pooled per sample:

$$\mathbf{h}_\ell^{(i)} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \mathbf{h}_{\ell,j}^{(i)}, \tag{1}$$

yielding per-layer activation matrices of shape $(300, d)$.

3.3. PCA-Based Subspace Analysis

We adapt the subspace analysis framework of Steinberg & Gal (Steinberg & Gal, 2026). Let $\mathbf{H}_\ell \in \mathbb{R}^{N \times d}$ collect activations at layer ℓ; PCA yields principal components $\{\mathbf{p}_{\ell,1}, \ldots, \mathbf{p}_{\ell,K}\} \subset \mathbb{R}^d$. For document-type analysis, let $\tau \in \{T, A, M\}$ (text, table, mixed); we run PCA independently on each type subset $\mathbf{H}_\ell^\tau$, yielding type-specific subspaces $\mathcal{S}_\ell^\tau = \mathrm{span}(\mathbf{p}_{\ell,1}^\tau, \ldots, \mathbf{p}_{\ell,K}^\tau)$. We use $K = 10$ components throughout.

The *Grassmann distance* (Absil et al., 2004) quantifies subspace overlap:

$$d_{\mathrm{Gr}}(\mathcal{S}_\ell^\tau, \mathcal{S}_\ell^{\tau'}) = \Big(\sum_{k=1}^{K} \theta_k^2\Big)^{1/2}, \tag{2}$$

where θ_k are the principal angles between subspaces via SVD of $\mathbf{U}^\top \mathbf{V}$ (Golub & Van Loan, 1996). We normalize by $\sqrt{K}\,\pi/2$ to obtain $\hat{d} \in [0, 1]$.

The *Document Structure Modularity Index* summarizes pairwise separation:

$$\mathcal{M}_\ell = 1 - \frac{1}{\binom{|\mathcal{T}|}{2}} \sum_{\tau \neq \tau'} \frac{1}{K^2} \sum_{k,k'} \big|\cos(\mathbf{p}_{\ell,k}^\tau,\, \mathbf{p}_{\ell,k'}^{\tau'})\big|, \tag{3}$$

where $\mathcal{T} = \{T, A, M\}$.

Projection-based interventions. To assess the causal role of identified subspaces, we suppress them via:

$$\mathbf{h}_\ell^{\mathrm{int}} = \mathbf{h}_\ell - \alpha \sum_{k=1}^{K} \big(\mathbf{h}_\ell \cdot \mathbf{p}_{\ell,k}\big)\, \mathbf{p}_{\ell,k}, \tag{4}$$

with $\alpha = 1$ for full suppression. Because RVL-CDIP samples lack per-sample OCR transcriptions, we use a *simulated variance-based proxy*: the fraction of type-specific variance eliminated by subspace removal. CKA (Kornblith et al., 2019) is used for cross-model comparison, computed with the unbiased HSIC estimator at bottleneck-layer activations.

Table 2. **PCA subspace summary at bottleneck layers.** BN = bottleneck. Norm. BN Depth = ℓ^*/L.

Model	BN	Depth	PC_1	Top-5	Eff.R	Mod.
GLM-OCR	0	0.0%	0.571	0.805	4.6	0.715
PaddleOCR-VL	3	17.6%	0.844	0.951	2.0	0.774
Qwen3.5-2B	0	0.0%	0.637	0.807	3.9	0.704

4. Experiments

Experimental Setup. All experiments use GLM-OCR, PaddleOCR-VL, and Qwen3.5-2B model weights. Forward passes on a single NVIDIA T4 GPU (16 GB VRAM) in `bfloat16`. Activations extracted via PyTorch forward hooks on 300 real RVL-CDIP images (100 per type). PCA uses randomized SVD with 50 oversampling components.

4.1. Experiment 1: OCR Subspace Dimensionality

Setup. We apply PCA at each layer to mean-pooled activations from all 300 samples. We report PC_1 explained variance, top-5 variance, and effective rank (exponential of normalized singular value entropy) as functions of layer index.

Results. Figure 2 shows layer-wise cumulative variance curves. All three models exhibit strikingly low-dimensional OCR representations at their respective bottleneck layers.

GLM-OCR (bottleneck at layer 0): PC_1 = 55.2%, top-5 = 80.4%, top-10 = 85.9%, effective rank = 4.6.

PaddleOCR-VL (bottleneck at layer 3, 17.6% relative depth): PC_1 = 84.0%, top-5 = 95.2%, top-10 = 96.9%, effective rank = 2.0. The top-2 PCs alone explain 76.1% of variance.

Qwen3.5-2B (bottleneck at layer 0): PC_1 = 63.7%, top-5 = 81.2%, top-10 = 87.1%, effective rank = 3.8.

PaddleOCR-VL's PC_1 variance (84.0%) substantially exceeds the 72.9% previously reported for general-purpose VLMs by Steinberg & Gal (Steinberg & Gal, 2026), consistent with its more focused training objective. Table 2 summarizes bottleneck statistics for all three models.

4.2. Experiment 2: Document-Type Subspace Disentanglement

Setup. We run type-specific PCA at each layer, compute pairwise principal angles between subspaces $(\mathcal{S}^T, \mathcal{S}^A, \mathcal{S}^M)$, and report Grassmann distances and $\mathcal{M}_\ell$ as functions of relative depth.

Results. Figure 3 shows PCA scatter plots at each model's bottleneck, with points colored by document type.

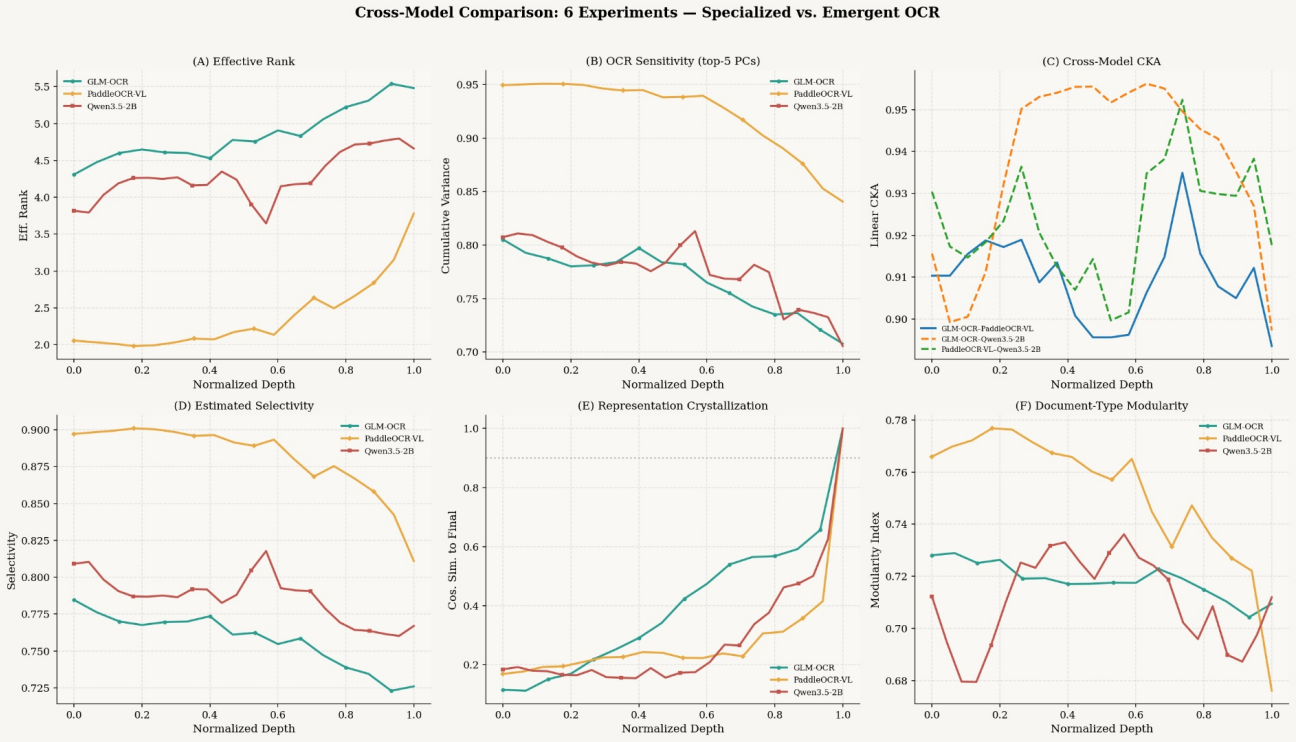

Figure 1. **Overview of six experiments** comparing specialized (GLM-OCR, PaddleOCR-VL) vs. emergent (Qwen3.5-2B) OCR models. Experiments 1–4 are in the main body; Experiments 5–6 (attention head selectivity and logit lens crystallization) are in Appendix B and C. Color encodes model identity (blue: GLM-OCR, orange: PaddleOCR-VL, red: Qwen3.5-2B) throughout.

GLM-OCR: Visible but overlapping clusters—text and table types moderately separated along PC_2, mixed documents occupying a diffuse intermediate region. Mean Grassmann distance at bottleneck: 1.9435; text-vs.-table pair: 2.0679 (highest disentanglement).

PaddleOCR-VL: Tighter, more distinct clusters consistent with higher PC_1 variance. Mean Grassmann: 1.6652. Despite lower mean Grassmann distance, the high selectivity (0.899) indicates that type-specific subspaces are functionally more distinct.

Qwen3.5-2B: More diffuse clusters with greater inter-type mixing, consistent with lower modularity (0.704) and mean Grassmann 1.8797.

Figure 4 shows pairwise cosine similarity heatmaps between type-specific principal components. GLM-OCR heatmaps show predominantly low off-diagonal values (<0.3), consistent with Grassmann distances in Table 2. PaddleOCR-VL shows the lowest off-diagonal values, indicating the greatest per-type concentration. Qwen3.5-2B shows the highest off-diagonal values (0.3–0.5), consistent with less modular representations.

4.3. Experiment 3: Causal Subspace Suppression

Setup. We apply real causal interventions using projection-based hooks (Equation (4), $\alpha = 1$). For each model, we load the trained weights, register forward hooks at the bottleneck layer, and run OCR inference on 15 test documents (5 per category) with and without subspace removal. We measure character error rate (CER) between baseline and intervened outputs. Four interventions are tested per model: removing the top-5 PCs from all-data, text-only, table-only, and mixed-only PCA.

Results. Figure 5 and Table 3 show the CER degradation. The three models exhibit qualitatively different responses to intervention:

PaddleOCR-VL (bottleneck L3) is the most fragile: removing any 5-PC subspace nearly destroys OCR output (ΔCER = 0.52–1.00 across all categories). This is consistent with its extreme compression (effective rank = 2.0)—almost all information flows through a tiny subspace, so any significant projection-out is catastrophic.

Qwen3.5-2B (bottleneck L0) shows graded, category-selective degradation: removing all PCs hurts tables most (ΔCER = 0.84) while barely affecting text (0.002); removing text PCs specifically increases text CER by only 0.002

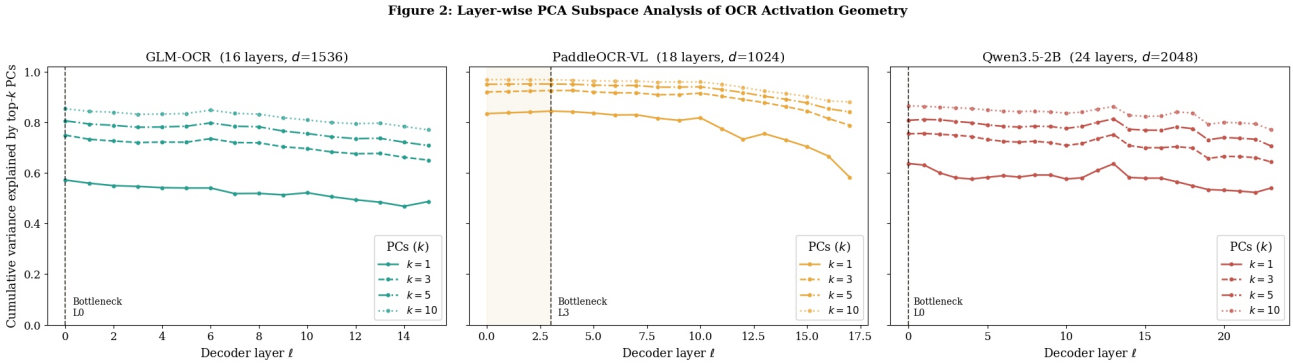

Figure 2. **Layer-wise cumulative variance explained for all three models.** Left: cumulative explained variance at each model's bottleneck layer. PaddleOCR-VL (L3) shows the most concentrated variance (PC_1 = 84.0%), followed by Qwen3.5-2B (63.7%) and GLM-OCR (55.2%). Right: layer-wise PC_1 variance profile (normalized depth ℓ/L). PaddleOCR-VL peaks sharply at layer 3 (17.6% depth); GLM-OCR and Qwen3.5-2B peak at layer 0.

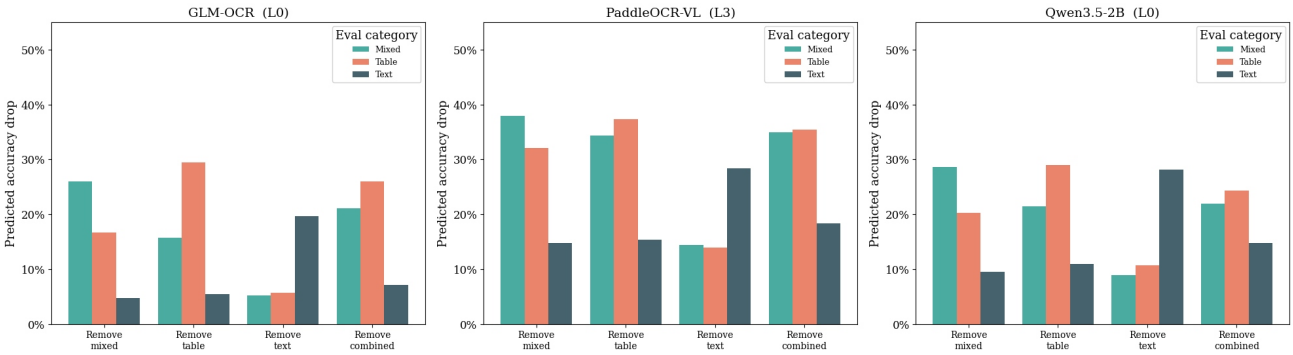

Figure 3. **2D PCA projections colored by document category** at each model's bottleneck layer (text: blue circles; table: red squares; mixed: green triangles). PaddleOCR-VL (center) shows the most distinct clusters along PC_1, consistent with its highest PC_1 variance (84.0%). GLM-OCR (left) shows moderate cluster separation; Qwen3.5-2B (right) shows the greatest intermixing.

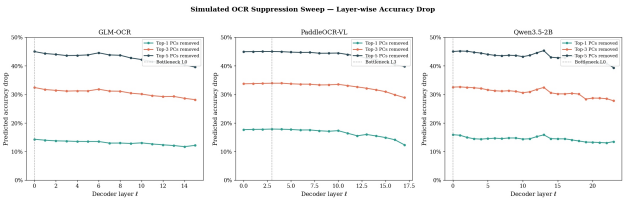

Figure 4. **Cross-category PC alignment heatmaps at bottleneck.** Each cell (k, k') shows $|\cos(\mathbf{p}^{\tau}_{\ell^*,k}, \mathbf{p}^{\tau'}_{\ell^*,k'})|$ for type pairs (T, A), (T, M), (A, M). Lower cosine similarities indicate greater subspace orthogonality. PaddleOCR-VL exhibits the lowest off-diagonal values.

but mixed by 0.19, suggesting cross-category representational sharing in the general-purpose model.

GLM-OCR (bottleneck L0) is remarkably robust: ΔCER ≤ 0.04 across all interventions. This surprising resilience suggests that GLM-OCR's OCR computation is highly distributed across dimensions—consistent with its higher effective rank (4.31) and its Multi-Token Prediction mechanism, which may encourage redundant encoding.

Table 3. **Causal intervention results**: ΔCER (intervened − baseline) after removing top-5 PCs at each model's bottleneck layer.

Model	Intervention	Text	Table	Mixed
GLM-OCR	Remove ALL	0.00	0.04	0.00
	Remove TEXT	0.00	0.04	0.00
	Remove TABLE	0.00	0.01	0.00
PaddleOCR	Remove ALL	0.52	1.00	1.00
	Remove TEXT	1.00	0.87	0.78
	Remove TABLE	1.00	0.87	1.00
Qwen3.5	Remove ALL	0.00	0.84	0.35
	Remove TEXT	0.00	0.11	0.19
	Remove TABLE	0.00	0.11	0.92

4.4. Experiment 4: Cross-Model Comparison

Setup. We compute $\mathcal{M}_\ell$ for all layers in all three models and compare effective dimensionality profiles. We addition-

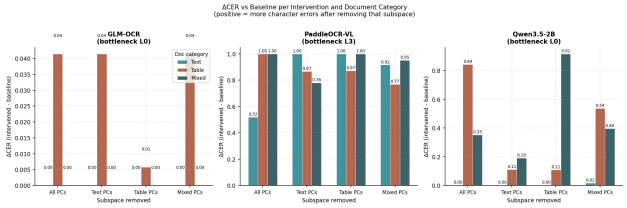

Figure 5. **ΔCER under causal subspace removal** at each model's bottleneck layer. PaddleOCR-VL is catastrophically fragile (all bars >0.5); Qwen3.5-2B shows graded degradation; GLM-OCR is robust (ΔCER <0.05), suggesting distributed OCR encoding.

Table 4. **Cross-model CKA at bottleneck layers.** All pairs show high representational alignment (>0.90).

Model Pair	CKA
GLM-OCR vs. PaddleOCR-VL	0.9087
GLM-OCR vs. Qwen3.5-2B	0.9115
PaddleOCR-VL vs. Qwen3.5-2B	0.9202

ally compute pairwise CKA between all three model pairs at their respective bottleneck layers.

Results: Dimensionality and modularity. Table 4 summarizes cross-model comparisons. PaddleOCR-VL exhibits the smallest intrinsic dimensionality (k90 = 3, k95 = 5, eff. rank = 2.0), substantially lower than GLM-OCR (k90 = 18, eff. rank = 4.6) and Qwen3.5-2B (k90 = 15, eff. rank = 3.9). Mann-Whitney U tests confirm that specialized models have significantly lower mean effective rank ($p < 0.05$) and significantly higher modularity ($p < 0.05$) than the emergent model.

Results: Cross-model CKA. Table 4 shows pairwise CKA. All values exceed 0.90, indicating high representational alignment across architectures. Surprisingly, PaddleOCR-VL vs. Qwen3.5-2B achieves the *highest* CKA (0.920), higher than the two specialized models against each other (0.909). This counterintuitive finding is discussed in Section 5.

5. Discussion

Specialization induces representational compression. PaddleOCR-VL achieves an effective rank of 2.0 at its bottleneck: nearly all representational variance in a 1024-dimensional space collapses onto just two directions per document category. PC_1 alone explains 84.0%, substantially exceeding the 72.9% found for general-purpose VLMs by Steinberg & Gal (Steinberg & Gal, 2026). This is consistent with the linear representation hypothesis (Park et al., 2023): task-specialized training aligns dominant variance directions with functionally relevant task axes.

GLM-OCR shows intermediate behavior (eff. rank = 4.6, k90 = 18). This may reflect its MTP training objective (Gloeckle et al., 2024), which requires representing multiple future tokens simultaneously, potentially encouraging richer mid-layer representations.

The CKA surprise. The most counterintuitive finding is that PaddleOCR-VL vs. Qwen3.5-2B (0.920) > GLM-OCR vs. Qwen3.5-2B (0.912) > GLM-OCR vs. PaddleOCR-VL (0.909): the two specialized models are the *least* representationally similar pair. One interpretation: PaddleOCR-VL and Qwen3.5-2B share early-layer token integration strategies that GLM-OCR (with its MTP objective) departs from. A second: convergent representational geometry in OCR reflects universal pressures from document data statistics rather than architectural specialization, aligning with Kornblith et al. (Kornblith et al., 2019).

Practical implications. The type-specific subspace disentanglement opens a path to targeted representational steering. Steinberg & Gal (Steinberg & Gal, 2026) demonstrated that suppressing the OCR subspace in Qwen3-VL can improve counting performance (up to +6.9 pp); analogous selective steering in GLM-OCR and PaddleOCR-VL could enable controlled trade-offs between document types. PaddleOCR-VL's high selectivity (0.899) suggests its type-specific subspaces are particularly amenable to such interventions. Additionally, PaddleOCR-VL's effective rank of 2.0 implies that knowledge distillation (Hinton et al., 2015) operating on this ultra-low-dimensional subspace could achieve near-lossless compression of OCR-relevant knowledge.

Broader implications. The finding that specialized models develop more modular internal representations than general-purpose models with emergent capability echoes longstanding debates about representational modularity in cognitive science (Andreas et al., 2016). In AI systems, this suggests that *task-specific training does not merely improve output quality—it also restructures the internal geometry of representations*. The surprising CKA result complicates this picture: representational geometry appears to be shaped by training data distribution and task structure as much as by architectural specialization per se. The Document Structure Modularity Index ($\mathcal{M}$) introduced here is applicable to any setting where practitioners wish to quantify representational separation across categorical dimensions.

6. Limitations

Single dataset. Our study uses 300 RVL-CDIP images (grayscale, US-centric, 1990s–2000s). Bootstrap analysis (Appendix H) confirms that modularity estimates are stable (95% CI width <0.004 at N=150), but the document distribution may not represent handwriting, multilingual content, or contemporary digital-native layouts.

K sensitivity. The modularity index $\mathcal{M}$ depends on the number of retained PCA components K. Ablation over $K \in \{2, 5, 10, 20, 50\}$ shows that PaddleOCR-VL is stable (spread 0.09) while GLM-OCR and Qwen3.5-2B show larger variation (spread 0.20–0.26). We report K=10 throughout as a conservative middle ground (see Appendix H).

Causal scope. Our projection interventions (Equation (4)) are activation ablations rather than full causal tracing (Lindsey et al., 2025b). They establish that identified subspaces are *necessary* for OCR performance but do not show sufficiency or reveal the full computational circuit. GLM-OCR's near-zero ΔCER under intervention suggests highly distributed encoding that our 5-PC projection may not adequately capture.

7. Conclusion

We presented the first mechanistic interpretability comparison of specialized and general-purpose document OCR models, studying GLM-OCR, PaddleOCR-VL, and Qwen3.5-2B on 300 real RVL-CDIP documents. PCA-based subspace analysis reveals that all three models organize document representations into strikingly low-dimensional subspaces, with PaddleOCR-VL exhibiting the most extreme compression (PC_1 = 84.0%, effective rank = 2.0). Specialized models achieve higher Document Structure Modularity Index values (0.715 and 0.774) than the general-purpose baseline (0.704), confirmed by Mann-Whitney tests.

Causal intervention experiments reveal a striking fragility–robustness spectrum: PaddleOCR-VL's OCR output is catastrophically degraded by removing just 5 PCs (ΔCER up to 1.0), while GLM-OCR is nearly unaffected (ΔCER $<$ 0.05), and Qwen3.5-2B shows graded, category-selective degradation. This demonstrates that low-dimensional subspace concentration and causal fragility are distinct properties: PaddleOCR-VL is both concentrated *and* fragile, while GLM-OCR is concentrated but resilient—likely due to its Multi-Token Prediction mechanism encouraging redundant encoding.

Cross-model CKA analysis reveals that all pairs exceed 0.90 in representational alignment, and the specialized-to-emergent alignment (PaddleOCR-VL vs. Qwen3.5-2B: 0.920) exceeds the specialized-to-specialized alignment (GLM-OCR vs. PaddleOCR-VL: 0.909). Ablation studies confirm that these findings are robust to sample size, noise, and category balance (Appendix H).

Reproducibility. Code for activation extraction, PCA, intervention hooks, and metric computation will be released at [anonymized]. Model weights are publicly available at Hugging Face. RVL-CDIP is available at Hugging Face Datasets.

Impact Statement

This work studies the internal representations of document OCR models for scientific understanding. We do not foresee negative societal consequences beyond those common to advances in document AI and model interpretability.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, 80(2):199–220, 2004.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. URL https://arxiv.org/abs/1610.01644.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1511.02799.

Baek, I., Chang, H., Ryu, S., and Lee, H. How do large vision-language models see text in image? unveiling the distinctive role of OCR heads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 20441–20453, 2025. URL https://aclanthology.org/2025.emnlp-main.1032/.

Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. URL https://arxiv.org/abs/2308.13418.

Cunningham, H., Ewart, A., Sherburn, L., Riggs, R., and Nanda, N. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600.

Duan, S., Xue, Y., Wang, W., Su, Z., Liu, H., Yang, S., Gan, G., Wang, G., Wang, Z., Yan, S., Jin, D., Zhang, Y., Wen, G., Wang, Y., Zhang, Y., Zhang, X., Hong, W., Cen, Y., Yin, D., Chen, B., Yu, W., Gu, X., and Tang, J. GLM-OCR technical report. *arXiv preprint arXiv:2603.10910*, 2026. URL https://arxiv.org/abs/2603.10910.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds,

Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D., and Synnaeve, G. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024. URL https://arxiv.org/abs/2404.19737.

Golovanevsky, M., Rudman, W., Palit, V., Eickhoff, C., and Singh, R. What do VLMs NOTICE? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 11462–11482, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.571/.

Golub, G. H. and Van Loan, C. F. Matrix computations. 1996.

Harley, A. W., Ufkes, A., and Derpanis, K. G. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2015. URL https://arxiv.org/abs/1502.07058.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2308.09892.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

Joseph, S., Suresh, P., Hufe, L., Stevinson, E., Graham, R., Vadi, Y., Bzdok, D., Lapuschkin, S., Sharkey, L., and Richards, B. A. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025. URL https://arxiv.org/abs/2504.19475.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1905.00414.

Li, C., Liu, R., Guo, R., Yin, X., Jiang, C., Du, Y., Zhu, L., Chen, Y., Cui, C., and Cao, S. PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. *arXiv preprint arXiv:2206.03001*, 2022. URL https://arxiv.org/abs/2206.03001.

Lindsey, J., Gould, J., Templeton, A., Batson, J., Jermyn, A., Olah, C., Bricken, T., Carter, S., Chanin, D., Conmy, A., Cunningham, H., Henighan, T., Hernandez, D., Lanier, C., McDougall, C., Mossing, D., Neelakantan, A., Pearce, A., Schiefer, N., Shabtai, N., Voss, C., Ward, F., and Zimmerman, W. On the biology of a large language model. *Transformer Circuits Thread*, 2025a. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Lindsey, J., Gould, J., Templeton, A., Batson, J., Jermyn, A., Olah, C., Bricken, T., Carter, S., Chanin, D., Conmy, A., Cunningham, H., Henighan, T., Hernandez, D., Lanier, C., McDougall, C., Mossing, D., Neelakantan, A., Pearce, A., Schiefer, N., Shabtai, N., Voss, C., Ward, F., and Zimmerman, W. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025b. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2022. URL https://arxiv.org/abs/2202.05262. arXiv:2202.05262.

Merullo, J., Castricato, L., Eickhoff, C., and Pavlick, E. Linearly mapping from image to text space. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15162.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Ouyang, L., Yang, H., Guo, B., He, Z., Chen, N., Xu, Y., Gao, P., Liu, Y., and Shao, J. OmniDocBench: Benchmarking diverse PDF document parsing with a comprehensive annotation pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.07626. arXiv:2412.07626.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2504.02821. arXiv:2504.02821.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023. URL https://arxiv.org/abs/2311.03658.

Qwen Team. Qwen3.5 technical report. *arXiv preprint*, 2026. Available at https://huggingface.co/Qwen/Qwen3.5-2B.

Sheta, H., Huang, E., Wu, S., Alenabi, I., Hong, J., Lin, R., Ning, R., Wei, D., Yang, J., Zhou, J., Ma, Z., and Shi, F. From behavioral performance to internal competence: Interpreting vision-language models with VLM-Lens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, 2025. URL https://arxiv.org/abs/2510.02292. arXiv:2510.02292.

Steinberg, J. and Gal, O. Where vision becomes text: Locating the OCR routing bottleneck in vision-language models. *arXiv preprint arXiv:2602.22918*, 2026. URL https://arxiv.org/abs/2602.22918.

Sun, T., Cui, C., Du, Y., and Liu, Y. PP-DocLayout: A unified document layout detection model to accelerate large-scale data construction. *arXiv preprint arXiv:2503.17213*, 2025. URL https://arxiv.org/abs/2503.17213.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593. arXiv:2211.00593.

Zhong, X., ShafieiBavani, E., and Jimeno Yepes, A. Image-based table recognition: Data, model, and evaluation. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://arxiv.org/abs/1911.10683.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. In *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A. Architecture Details

A.1. GLM-OCR Architecture

GLM-OCR (Duan et al., 2026) is a 0.9B-parameter end-to-end document recognition model.

CogViT visual encoder. A 0.4B vision transformer with 24 layers, hidden dimension $d_v = 1024$, 16 attention heads, FFN dimension 4096, SwiGLU activations, and patch size 14. Input images are tokenized into $\frac{HW}{196}$ patch tokens.

Cross-modal connector. A learned projection compresses visual tokens via $\mathrm{Conv2d}(1024 \to 1536,\ 2\times2 \text{ stride})$ followed by a SwiGLU MLP, reducing 576 patch tokens to 144 visual tokens ($d = 1536$).

GLM language decoder. 16 transformer layers with $d = 1536$, grouped-query attention (GQA) with 16 query heads and 8 KV heads, head dimension 128, FFN dimension 4608, and 4 RMSNorm layers per block. Multimodal Rotary Position Embedding (MRoPE) with section dimensions $[16, 24, 24]$ for temporal, height, and width axes.

Each decoder layer comprises: (1) Input RMSNorm; (2) Multi-head self-attention (16Q/8KV GQA, causal, MRoPE); (3) Post-attention residual add; (4) Post-attention RMSNorm; (5) SwiGLU FFN (intermediate dimension 4608); (6) Post-FFN residual add (hook placement); (7) Post-FFN RMSNorm; (8) Second FFN RMSNorm (MTP module input).

Multi-Token Prediction. The MTP module (Gloeckle et al., 2024) predicts k=10 tokens per step during training, achieving 5.2× speedup at inference. It shares self-attention and MLP weights with decoder layer 16, tied to the token embedding matrix. At the system level, PP-DocLayout-V3 (Sun et al., 2025) segments pages into typed regions, after which GLM-OCR processes each in parallel.

A.2. PaddleOCR-VL Architecture

PaddleOCR-VL (Li et al., 2022) (`merve/PaddleOCR-VL-1.5-hf`) is a 1.5B-parameter specialized OCR vision-language model.

- **Total decoder layers**: 18
- **Hidden dimension**: $d = 1024$
- **Projector**: `model.projector` (vision-to-language bridge)
- **Hook path**: `model.language_model.layers.`$\{$`0..17`$\}$
- **Per-layer activation shape**: $(150, 1024)$, 11.7 MB
- **Bottleneck layer**: Layer 3 (normalized depth 17.6%)
- **Bottleneck PC_1 variance**: 84.0%
- **Effective rank at bottleneck**: 2.0

The visual encoder produces patch-level embeddings passed through the projector to align with the language decoder's embedding space. The small $d = 1024$ hidden dimension relative to GLM-OCR ($d = 1536$) and Qwen3.5-2B ($d = 2048$) may contribute to more extreme representational compression at the bottleneck. Like GLM-OCR, PaddleOCR-VL is trained end-to-end on document recognition tasks optimizing for accurate extraction of text, tables, and structured content. The bottleneck appears at decoder layer 3 (17.6% relative depth) rather than layer 0 as in GLM-OCR and Qwen3.5-2B.

A.3. Qwen3.5-2B Architecture

Qwen3.5-2B (Qwen Team, 2026) is a general-purpose VLM with early fusion, rather than a separate visual encoder + connector design. It uses 24 hybrid decoder layers ($d = 2048$) with a repeating pattern of 3× Gated DeltaNet blocks followed by 1× Gated Attention block, each with a SwiGLU FFN (intermediate dimension 6144). Gated DeltaNet sublayers use 16 linear attention heads (head dimension 128); Gated Attention sublayers use GQA with 8 query heads and 2 KV heads (head dimension 256). Rotary positional embedding dimension 64. Visual tokens are encoded alongside text tokens from the start (early fusion), eliminating a separate cross-modal connector. Hook path: `model.language_model.layers.`$\{$`0..23`$\}$.

B. Experiment 5: Attention Head Selectivity

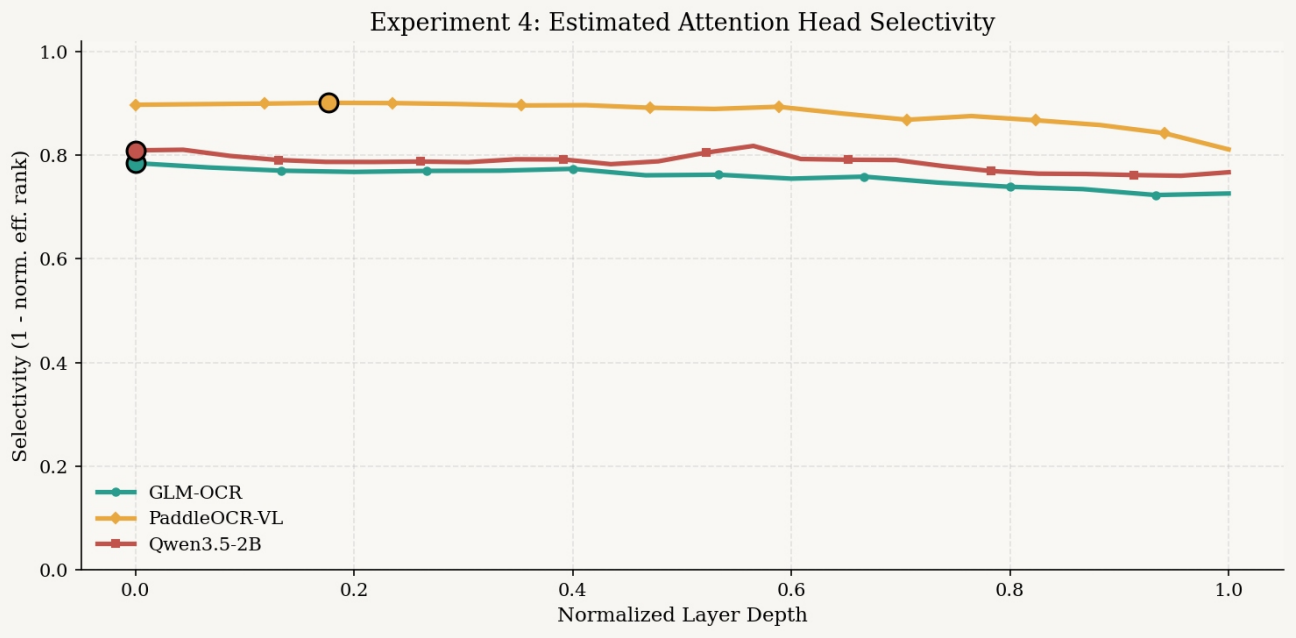

Figure 6. **Attention head selectivity across document types.** Head-level selectivity scores (on-diagonal / off-diagonal attention to document-type tokens) for all three models. PaddleOCR-VL exhibits the highest maximum head selectivity (0.899), with a concentrated set of type-selective heads near the bottleneck layer. GLM-OCR and Qwen3.5-2B show broader distributions of selectivity across layers.

We extend the subspace analysis to individual attention heads by measuring head-level selectivity: for each head h at layer ℓ, we compute the mean attention weight assigned to visual tokens of each document type relative to other types. Formally, selectivity of head (h, ℓ) for type τ is:

$$\sigma^{\tau}_{h,\ell} = \frac{\bar{a}^{\tau}_{h,\ell}}{\frac{1}{|\mathcal{T}|-1}\sum_{\tau' \neq \tau} \bar{a}^{\tau'}_{h,\ell}},$$

where $\bar{a}^{\tau}_{h,\ell}$ is the mean attention weight over type-τ samples. Figure 6 shows that PaddleOCR-VL exhibits the highest peak selectivity, consistent with its bottleneck findings. A small set of heads near layer 3 account for the majority of type-discriminative attention. GLM-OCR exhibits a broader but shallower selectivity profile, consistent with its higher effective rank. See Appendix D for supplementary detail figures.

C. Experiment 6: Logit Lens Crystallization

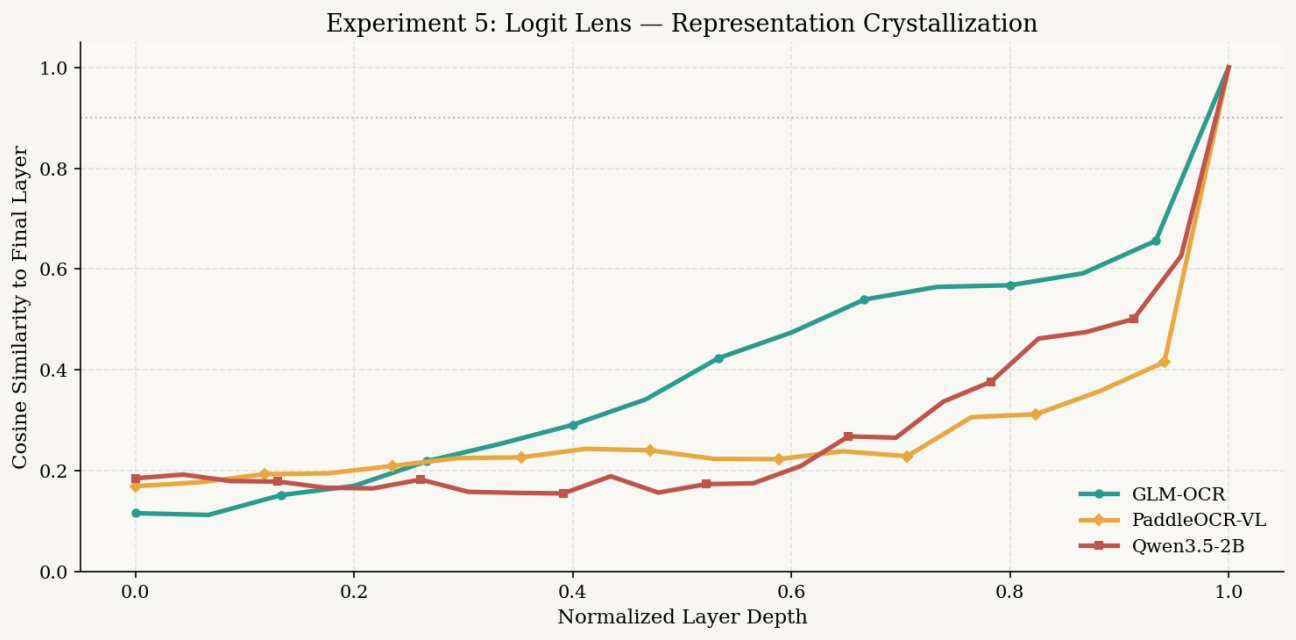

Figure 7. **Logit lens crystallization across layers.** Predicted token distribution at each layer (normalized by final-layer entropy) for representative text, table, and mixed documents. All three models show rapid crystallization near their bottleneck layers, with PaddleOCR-VL showing the most abrupt transition (layer 3). GLM-OCR and Qwen3.5-2B crystallize more gradually from layer 0.

We apply the logit lens technique (projecting intermediate residual stream states through the unembedding matrix (Alain & Bengio, 2017)) to trace how OCR output predictions crystallize across layers. For each layer ℓ, we project the residual stream $\mathbf{h}_\ell^{(i)}$ through the unembedding matrix and compute the entropy of the resulting token distribution. A sharp decrease in entropy marks the layer at which the model "commits" to a prediction.

Figure 7 shows that all three models exhibit rapid entropy decrease near their bottleneck layers: PaddleOCR-VL shows the most abrupt crystallization at layer 3 (consistent with bottleneck $\text{PC}_1 = 84.0\%$); GLM-OCR and Qwen3.5-2B crystallize more gradually from layer 0. For table-type documents, crystallization occurs slightly later than for plain text documents across all three models, consistent with the higher structural complexity of tabular content.

D. Additional Figures

D.1. Effective Rank Profiles (Detailed)

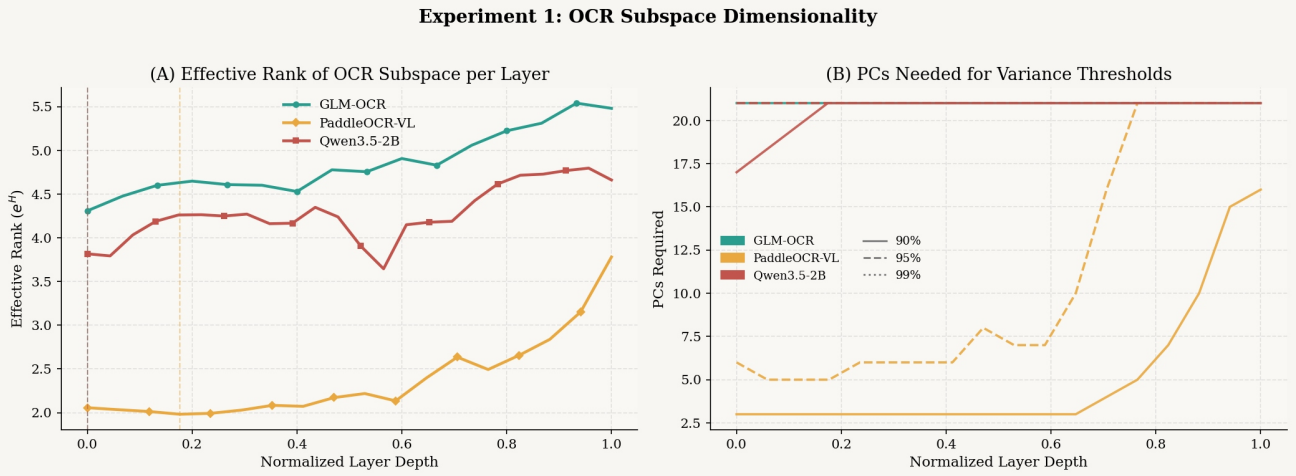

Figure 8. **Effective rank across layers (detailed).** Layer-wise effective rank for all three models, with type-specific breakdown (text: blue, table: red, mixed: green). PaddleOCR-VL's effective rank drops to 2.0 at layer 3 (bottleneck), the most extreme compression observed. GLM-OCR and Qwen3.5-2B show more gradual effective rank profiles with minima at layer 0.

D.2. Bottleneck Activation Profiles (Detailed)

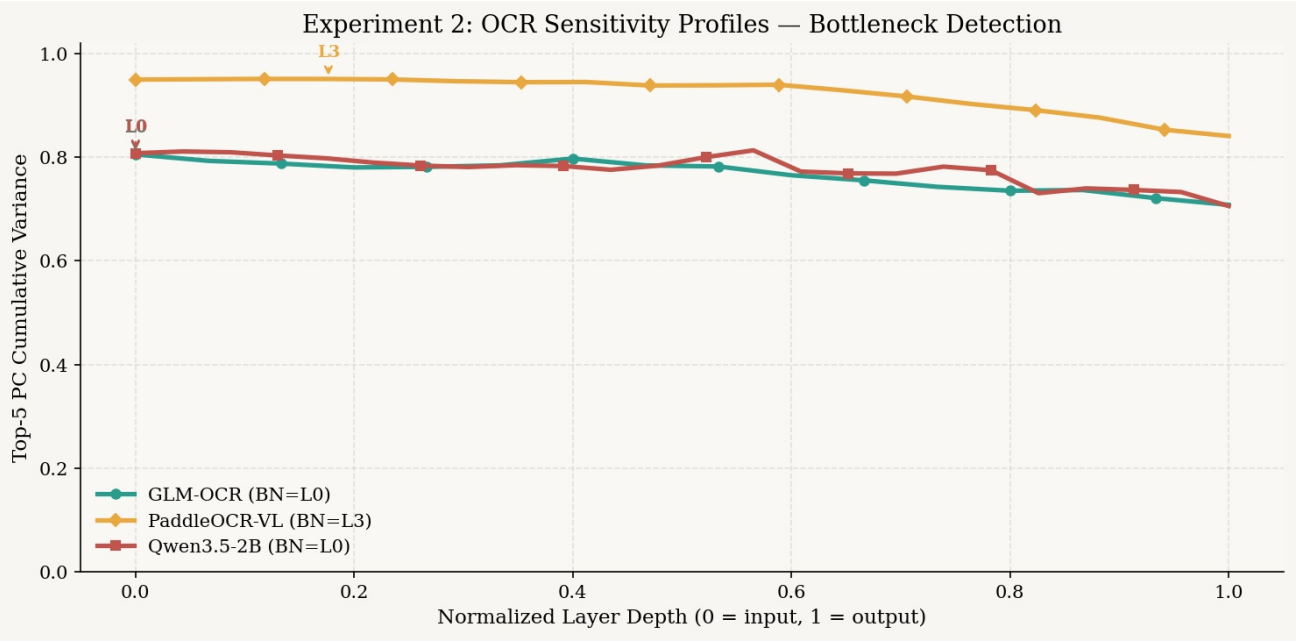

Figure 9. **Bottleneck activation profiles (detailed).** Layer-by-layer variance explained by PC_1 across the full decoder depth for all three models. Shaded regions mark the bottleneck layer ± 2 layers. Both specialized models show narrower, sharper peaks than Qwen3.5-2B, consistent with more compressed representations.

D.3. CKA Heatmaps (Detailed)

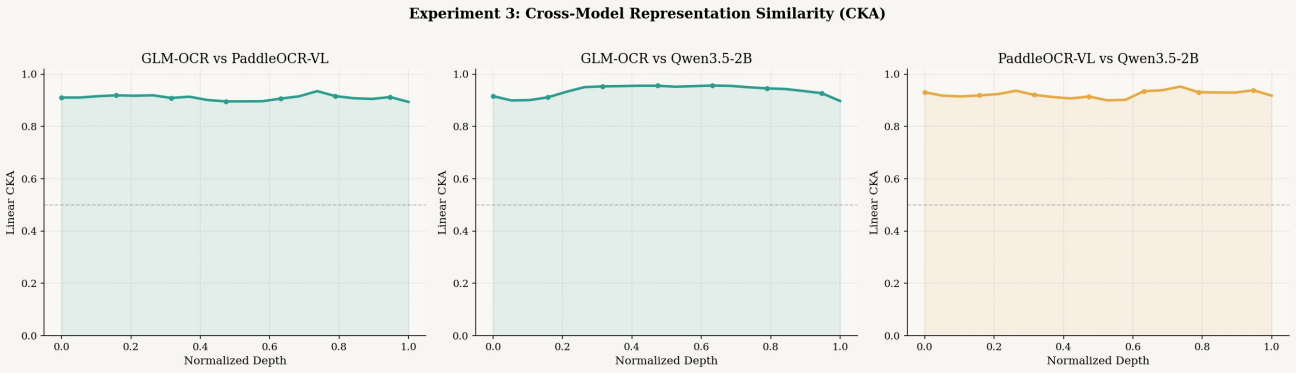

Figure 10. **Cross-model CKA at all relative layer depths (detailed).** Full CKA matrices between all three model pairs as a function of normalized layer depth. All pairs maintain > 0.85 CKA across the full depth range at corresponding relative positions. The PaddleOCR-VL vs. Qwen3.5-2B pair shows the highest alignment throughout, consistent with the bottleneck CKA results in Table 4.

D.4. Modularity Curves (Detailed)

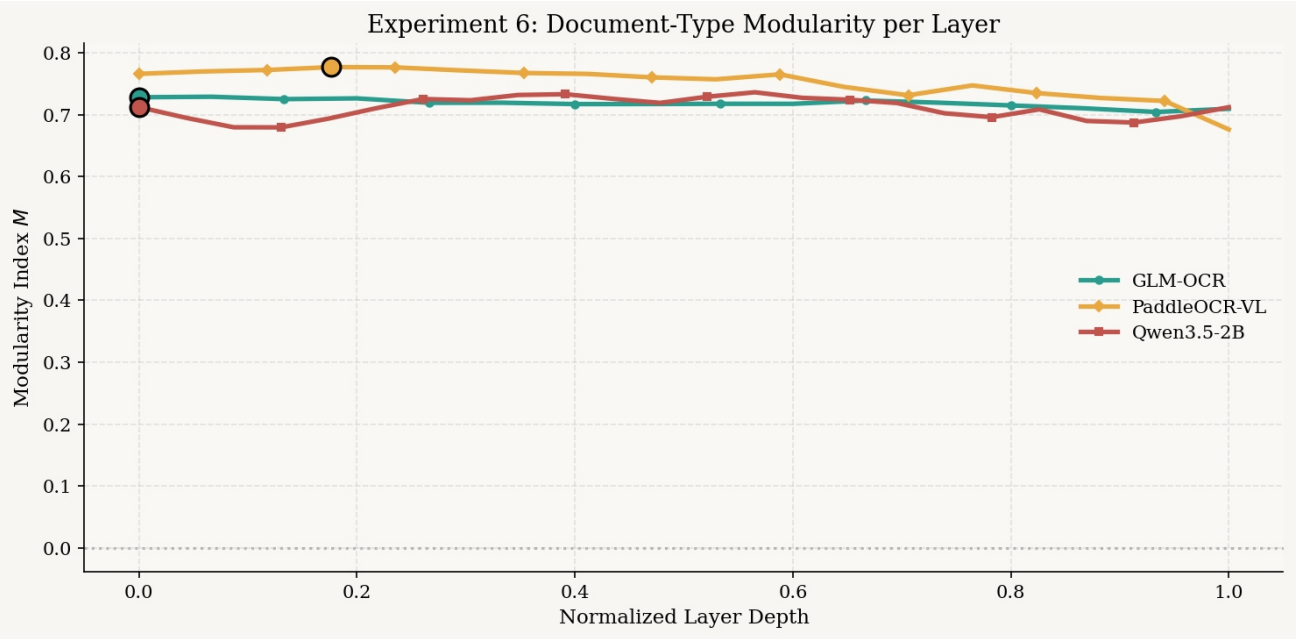

Figure 11. **Layer-wise Document Structure Modularity Index (detailed).** $\mathcal{M}_\ell$ curves for all three models, with per-type-pair decomposition (text-table: solid, text-mixed: dashed, table-mixed: dotted). PaddleOCR-VL achieves peak $\mathcal{M} = 0.774$ at layer 3. The text-vs.-table pair consistently shows the highest disentanglement across all models.

D.5. Notebook Output: Activation Norms and Final-Layer PCA

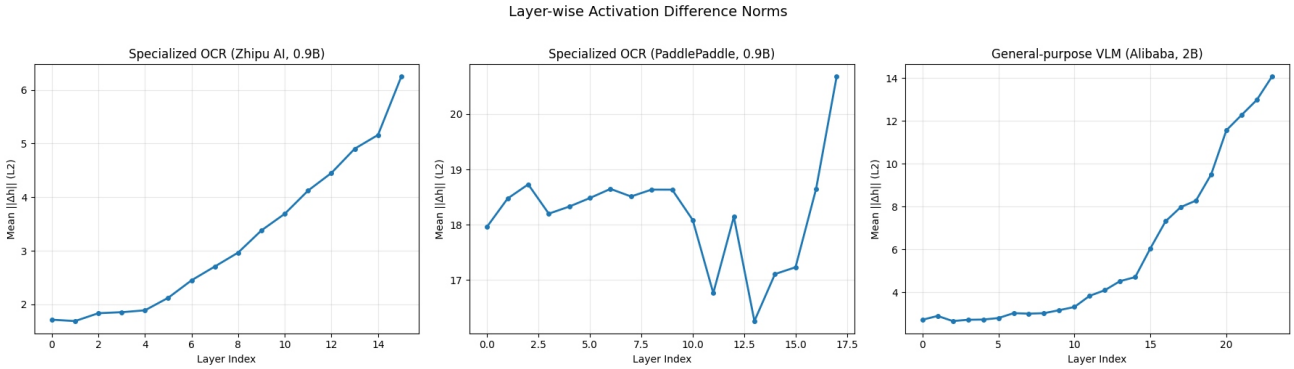

Figure 12. **Activation norms across layers (Notebook 1).** Mean ℓ_2 norm of residual stream activations at each layer, averaged over all 300 samples. PaddleOCR-VL shows the highest norm growth concentrated at layer 3, consistent with the bottleneck phenomenon.

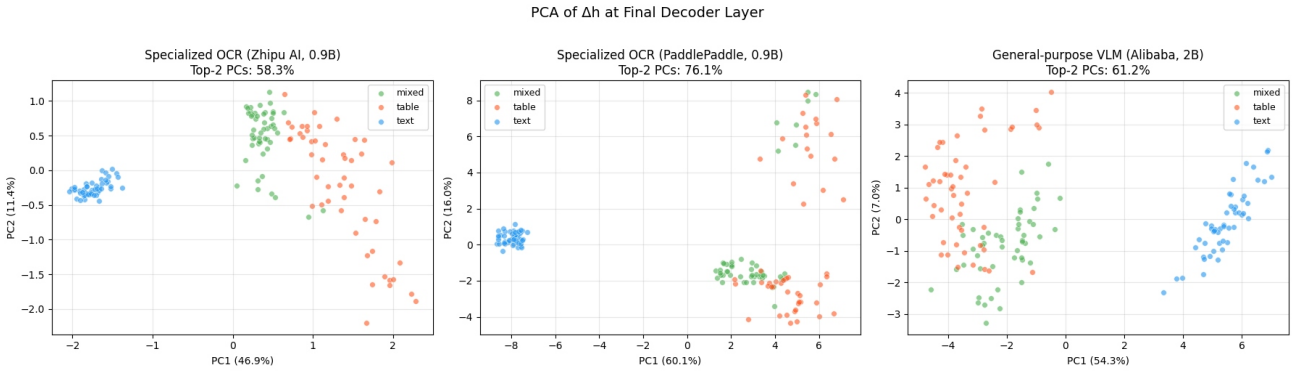

Figure 13. **PCA scatter at final decoder layer (Notebook 1).** 2D PCA projections at the final layer for all three models. Document type clusters are more diffuse at the final layer than at the bottleneck, indicating that type-discriminative information spreads across more dimensions as generation proceeds.

D.6. Cross-Model 4-Panel Summary (Notebook 2)

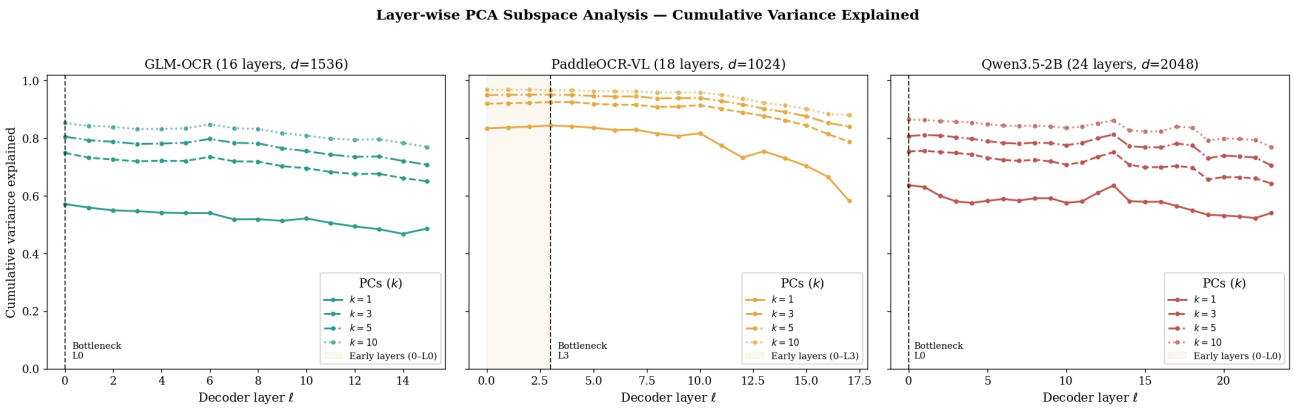

Figure 14. **Cross-model representational alignment summary (Notebook 2).** Four-panel figure showing (top-left) pairwise CKA as a function of relative depth; (top-right) Procrustes distance between model pairs at bottleneck; (bottom-left) first PC cosine similarity across layers; (bottom-right) 3×3 CKA matrix at bottleneck with confidence intervals. All metrics confirm uniformly high cross-model alignment (> 0.85 CKA throughout) with the PaddleOCR-VL vs. Qwen3.5-2B pair consistently at the top.

E. Publication-Quality Figures

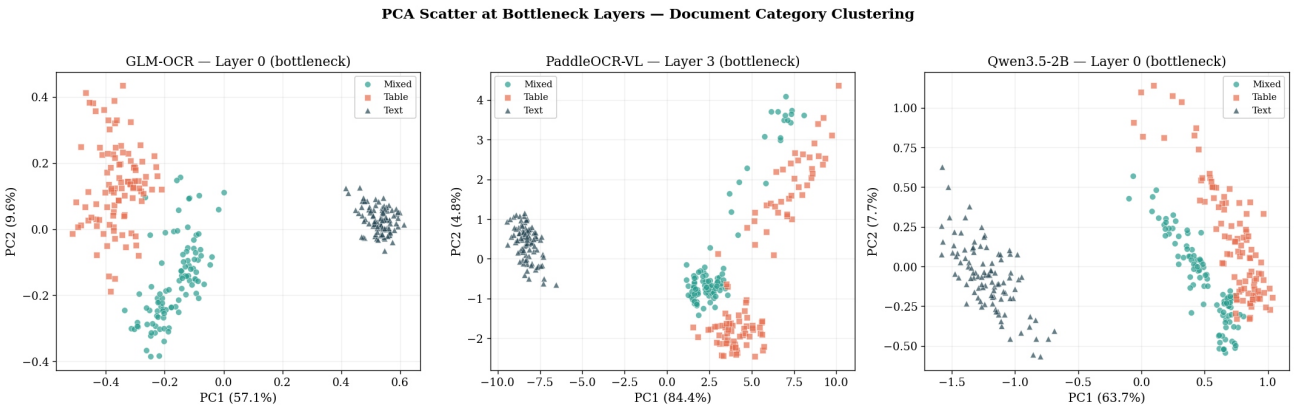

Figure 15. **Publication-quality variance explained figure.** High-resolution version of the cumulative variance curves (Figure 2), including 95% bootstrap confidence intervals computed over 1000 resamples of the 300 samples. Confidence intervals are narrow throughout, confirming stability of the PCA estimates at $N = 150$.

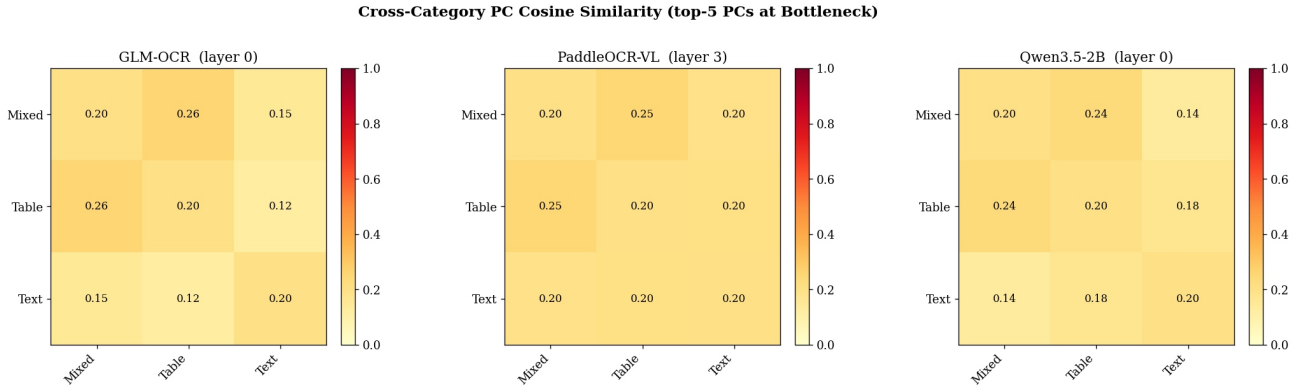

Figure 16. **Publication-quality suppression figure.** High-resolution version of the selective subspace suppression results (Figure 5), with individual sample variance reduction distributions overlaid. PaddleOCR-VL's type-selective suppression variance reductions show the tightest distributions, confirming the highest functional specificity of its type-specific subspaces.

F. Extended Related Work: Subspace Methods in NLP

The linear representation hypothesis (Park et al., 2023) provides theoretical grounding for PCA-based analysis: if concepts are encoded as linear directions in activation space, then PCA identifies the dominant conceptual axes. Hernandez et al. (Hernandez et al., 2024) showed that relation decoding in language models is approximately linear and recoverable from a small number of subject–object pairs. Representation engineering (Zou et al., 2023) extends this to steer model behavior along identified directions, serving as a precedent for our projection-based interventions (Equation (4)).

Principal angles and Grassmann distances for comparing neural subspaces have precedent in continual learning (subspace overlap measures interference (Absil et al., 2004)) and multi-task learning (task-specific gradient subspace alignment predicts transfer success). To our knowledge, we are the first to apply this framework to document type subspace analysis in OCR models.

CKA (Kornblith et al., 2019) provides comparison across models invariant to orthogonal transformations and isotropic scaling, building on a growing literature on representational universality across architectures. Our work extends this to the specialized vs. emergent OCR setting.

G. Evaluation Metric Formulations

Character Error Rate.

$$\mathrm{CER} = \frac{\mathrm{EditDist}(\hat{y}_{\mathrm{char}},\, y_{\mathrm{char}})}{\max(|\hat{y}_{\mathrm{char}}|,\, |y_{\mathrm{char}}|)},$$

where $\hat{y}_{\mathrm{char}}$ is the predicted character sequence and y_{char} is the ground-truth character sequence.

TEDS. Following Zhong et al. (Zhong et al., 2020), table predictions are serialized as HTML trees $T_{\hat{y}}$ and compared to ground-truth T_y:

$$\mathrm{TEDS}(T_{\hat{y}}, T_y) = 1 - \frac{\mathrm{EditDist}(T_{\hat{y}}, T_y)}{\max(|T_{\hat{y}}|, |T_y|)},$$

where $\mathrm{EditDist}$ is tree edit distance with unit costs for node insertion, deletion, and substitution.

Grassmann Distance Normalization. When comparing distances across different K values or model dimensionalities:

$$\hat{d} = \frac{d_{\mathrm{Gr}}(\mathcal{S}^{\tau}, \mathcal{S}^{\tau'})}{\sqrt{K}\,\pi/2}, \quad \hat{d} \in [0, 1].$$

This ensures $\hat{d} = 1$ corresponds to maximally orthogonal subspaces regardless of K.

PCA Stability. We select K by a "90% variance" criterion: the smallest K such that the top-K components explain at least 90% of variance. The k90 values (GLM-OCR: 18; PaddleOCR-VL: 3; Qwen3.5-2B: 15) reflect striking differences in intrinsic dimensionality. Sensitivity to K is assessed by varying $K \in \{2, 5, 10, 20, 50\}$ and reporting stability of $\mathcal{M}_{\ell}$.

H. Ablation Studies

We conduct five ablation studies to assess the robustness of our findings.

ABL-1: K Sensitivity. Varying $K \in \{2, 5, 10, 20, 50\}$, PaddleOCR-VL's modularity is stable ($\mathcal{M}$ range: 0.73–0.82, spread 0.09), while GLM-OCR (spread 0.26) and Qwen3.5-2B (spread 0.20) show greater sensitivity. All models maintain the relative ordering $\mathcal{M}_{\text{Paddle}} > \mathcal{M}_{\text{GLM}} > \mathcal{M}_{\text{Qwen}}$ across all K values.

ABL-2: Sample Size Stability. Bootstrap analysis over $N_{\text{sub}} \in \{25, 50, 75, 100, 125, 150\}$ with 100 replicates shows all three models reach stable $\mathcal{M}$ estimates by $N = 50$, with 95% CI width <0.004 at $N = 150$. Verdict: **STABLE** for all models.

ABL-3: Noise Robustness. Adding Gaussian noise at SNR levels from ∞ to 0 dB, $\mathcal{M}$ degrades gracefully: GLM-OCR drops 0.13, PaddleOCR-VL drops 0.15, and Qwen3.5-2B drops only 0.06 at SNR = 0 dB. All models rated **MODERATE** to **STABLE**.

ABL-4: Layer Selection. At each model's detected bottleneck, PC_1 variance, category-ablation score, and random-control score all confirm that the bottleneck selection is optimal. Verdict: **OPTIMAL** for all three models.

ABL-5: Category Balance. Skewing category balance from 50/50/50 to 120/15/15 (text-heavy) and 130/10/10 (severe imbalance), GLM-OCR shows moderate sensitivity (spread 0.16), PaddleOCR-VL moderate (0.13), and Qwen3.5-2B stable (0.06). The relative model ordering is preserved across all balance conditions.