Do Optical Text Compressors Read or Guess? Disentangling Visual Recognition from Linguistic Priors

Anonymous ACL submission

Abstract

Optical text compression has recently been proposed as a way to reduce long textual contexts into compact visual tokens and reconstruct them with a language decoder. High OCR scores on natural text, however, may conflate faithful visual recognition with decoder-side linguistic completion. We introduce the *Semantic Vacuum Stress Test* (SVST), a prior-disentangled evaluation protocol that progressively removes sentence-level, word-level, and lexical-semantic cues while preserving visual rendering conditions. Using DeepSeek-OCR as a representative optical text compressor, we show that its strong performance on natural text is highly sensitive to linguistic priors: under zero-prior random text, Tiny-mode precision drops from approximately 90% to 19.84%, while larger visual-token budgets reduce but do not eliminate this dependence. A cross-system evaluation over 14 OCR/VLM settings further shows that many integrated E2E/VLM OCR/document systems degrade sharply under semantic-vacuum inputs, whereas systems with more explicit recognition pathways, especially the pipeline OCR reference, remain substantially more robust. Beyond character recovery, deliberately simple QA/VQA probes indicate that compressed representations often fail to support basic evidence-grounded understanding, even when the relevant information is explicit in the document. Finally, context-length stress tests reveal sharp collapse between 8,000 and 10,500 tokens. These findings suggest that compression-centric OCR/VLM systems should be evaluated not only by natural-text OCR accuracy, but also by prior-disentangled fidelity, semantic preservation, and context-boundary diagnostics. We release all data, results and scripts used in this study in an anonymous GitHub repository.

1 Introduction

Long-context processing remains a central bottleneck for Transformer-based language models because attention cost grows quadratically with sequence length. One emerging response is *optical text compression*: instead of feeding long documents as text tokens, a system renders text into images, encodes the image into a compact set of visual tokens, and reconstructs or consumes the content through a language decoder. DeepSeek-OCR (Wei et al., 2025) is a recent and influential example of this paradigm. It reports that pages can be reconstructed from as few as 64–400 vision tokens, with high precision under moderate compression and non-trivial precision even at more aggressive compression ratios. Such results have motivated the view that visual tokens may provide an efficient substrate for storing or transmitting long textual contexts.

Despite this promise, natural-text OCR accuracy is an ambiguous measure for compression-based OCR/VLM systems. When the decoder is a strong language model, accurate outputs may arise from two sources: visual evidence encoded in the image tokens, and linguistic priors learned from natural text. Under a tight visual-token budget, the decoder may normalize, repair, or substitute visually ambiguous strings with more probable words and phrases. This behavior can improve scores on ordinary natural text, but it can also mask whether the system has faithfully read the input. In fidelity-critical settings, such substitution errors differ from ordinary information loss: the system does not merely omit uncertain details, but may confidently replace visually present evidence with linguistically plausible alternatives.

This paper reframes the evaluation problem as a question of *prior-disentangled fidelity*: to what extent does an optical text compressor recover what is visually present, rather than what is linguistically expected? We introduce the *Semantic Vacuum Stress Test* (SVST), a diagnostic protocol that progressively removes linguistic cues while preserving visual rendering conditions. SVST evaluates a sys-

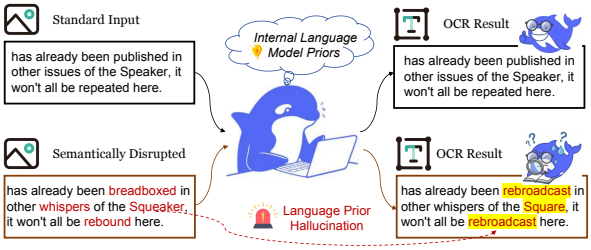

Figure 1: Example of prior-driven reconstruction under semantic disruption. When the input contains visually present but semantically unlikely strings, the decoder may substitute them with more plausible alternatives.

tem under five complementary probes:

- **Sentence-level disruption** tests whether global contextual coherence inflates OCR accuracy under compression.
- **Word-level disruption** tests whether local lexical and character-order priors repair visually present but malformed strings.
- **Zero-prior text** replaces natural words with random character sequences, creating a semantic vacuum in which recognition must rely primarily on visual evidence.
- **Simple QA/VQA probing** examines whether compressed representations retain the elementary semantic information needed for evidence-grounded answering.
- **Context-boundary stress testing** identifies the length at which a given optical-compression setup ceases to support reliable reconstruction.

We instantiate SVST primarily on DeepSeek-OCR because it is a representative and widely discussed optical text compression system, and compare its behavior with a broad set of OCR/VLM baselines to separate model-specific effects from broader architectural trends. Our results show that sentence-level semantic disruption reduces DeepSeek-OCR precision most strongly in high-compression settings (Tiny: -11.2% on average; Small: -3.6%; Base: -0.6%). Word-level perturbations reveal stronger dependence on lexical priors: 10% letter shuffling causes an average Tiny-mode drop of 11.3%, and zero-prior random text reduces Tiny-mode precision to 19.84%. Across the evaluated systems, many integrated E2E/VLM OCR/document systems degrade sharply on zero-prior inputs, while systems with more explicit recognition pathways remain comparatively robust. On deliberately simple QA/VQA probes, compressed representations often fail to support basic evidence-grounded answering, even though standard LLMs solve the corresponding text-QA setting with over 90% accuracy. Finally, long-context stress tests show that all DeepSeek-OCR modes collapse between 8,000 and 10,500 tokens.

Our contributions are threefold:

- We propose SVST, a reusable prior-disentangled evaluation protocol for measuring whether OCR/VLM systems read visual evidence or reconstruct linguistically plausible text.
- We provide a controlled empirical analysis of DeepSeek-OCR across sentence-level, word-level, zero-prior, semantic-preservation, and context-boundary probes.
- We conduct cross-system comparisons showing that prior dependence is especially pronounced in many integrated E2E/VLM OCR/document systems, while explicit-recognition systems provide useful robustness references.

We provide all code, data, and scripts in the anonymous supplementary material.

2 Related Work

2.1 Vision Encoders in Vision-Language Models

VLMs employ three main vision encoder designs: dual-tower for parallel high-resolution processing but incur heavy preprocessing and training overhead (Wei et al., 2024a); tile-based for memory efficiency but produce excessive tokens (Chen et al., 2024); adaptive-resolution for flexibility but suffer quadratic memory growth (Bai et al., 2025b). DeepSeek-OCR's DeepEncoder combines windowed SAM (Kirillov et al., 2023) and global CLIP (Radford et al., 2021) with a convolutional compressor, prioritizing low activations. But none of these designs isolate linguistic priors from visual recognition, leaving open how much OCR performance reflects true visual understanding.

2.2 End-to-End OCR and Document Understanding Models

End-to-end OCR has replaced traditional pipelines. Nougat (Blecher et al., 2023) pioneered paper parsing; GOT-OCR2.0 (Wei et al., 2024b) broadened

dense tasks; VLMs like Qwen-VL (Bai et al., 2025b), InternVL (Chen et al., 2024), Donut (Kim et al., 2022), and Pix2Struct (Lee et al., 2023) enhanced OCR; specialized models like UDOP (Tang et al., 2023) and DocLLM (Wang et al., 2024) target layouts. Despite high accuracies, evaluations focus on edit-distance or ANLS on natural text, leaving unanswered how few vision tokens are needed for meaningful text decoding.

2.3 Linguistic Priors in Multimodal Models

Multimodal models frequently exploit language priors over visual signals. Blind LLMs can outperform vision-enabled ones on some VQA tasks (Lin et al., 2023). CLIP shows textual biases (Luo et al., 2024; Materzyńska et al., 2022). In OCR, large models excel at printed text but struggle with handwritten or complex layouts, indicating reliance on linguistic context (Liu et al., 2024b). Probing studies suggest high compressed OCR scores often reflect decoder hallucination (Laurençon et al., 2024; Luo et al., 2024).

3 Semantic Vacuum Stress Test

SVST is designed to separate visual recognition from linguistic reconstruction in OCR-centric vision-language systems. Given a document image x and ground-truth text y, a standard OCR evaluation measures whether a model output $\hat{y}$ matches y. However, when y is natural language, high similarity between $\hat{y}$ and y can be supported by both visual evidence and learned language statistics. SVST therefore constructs paired inputs that preserve rendering conditions while reducing the usefulness of linguistic priors.

The protocol contains three levels of prior removal. First, **sentence-level disruption** replaces key words and phrases with semantically implausible alternatives that retain English-like orthographic patterns, testing reliance on global discourse coherence. Second, **word-level disruption** applies intra-word swaps and shuffles, testing whether a decoder normalizes malformed but visually present strings into likely vocabulary items. Third, **zero-prior generation** creates random character sequences with comparable length distributions, removing lexical and syntactic cues as much as possible. Together, these settings form a semantic vacuum in which robust OCR should continue to transcribe visible characters, whereas prior-dependent systems should lose accuracy or hallucinate plausible text.

Input	Average Precision (%)		
	Tiny	Small	Base
Fox original	83.88	93.89	96.62
Rendered text	88.00	95.23	97.93
Sentence-disrupted	$76.75^{-11.25}$	$91.56^{-3.67}$	$97.31^{-0.62}$

Table 1: Average DeepSeek-OCR precision under sentence-level semantic disruption. Superscripts show drops from rendered text; full token-bin results are in Appendix Table 7.

In addition to character-level fidelity, SVST includes two complementary diagnostics. The **semantic-preservation probe** uses deliberately simple QA/VQA questions whose answers are explicit in the document, testing whether the compressed representation preserves information usable for basic evidence-grounded answering. The **context-boundary probe** increases document length under fixed compression modes, identifying where visual-token bottlenecks induce repetition, hallucination, or total collapse. The following sections instantiate these diagnostics on DeepSeek-OCR and compare them with other OCR/VLM systems.

4 Probe 1: Sentence-Level Semantic Disruption

4.1 Experimental Setup

We base our evaluation on the Fox benchmark (Liu et al., 2024a), comprising 112 English document pages with ground-truth token lengths ranging from 600 to 2500. As a clean baseline, we render the ground-truth text into images, denoted as **Rendered text** in Table 1. For disruption, we apply targeted replacements guided by a controlled distortion process: key nouns, verbs, and phrases are substituted with absurd alternatives mimicking English patterns (e.g., "butterfly" $\rightarrow$ "breadflutter"), preserving character shapes and layout while eliminating meaningful context. The disrupted text is also rendered into images using the same pipeline, denoted as **Sentence-disrupted** in Table 1. Both Rendered text and Sentence-disrupted sets are evaluated on DeepSeek-OCR in Tiny, Small, and Base modes. We also report results on the original Fox images for reference, and performance is measured by OCR precision.

4.2 Results and Analysis

Table 1 summarizes average precision, with full token-bin results reported in Appendix Table 7. The summary shows that introducing sentence-level semantic disruption substantially degrades accuracy, particularly under more aggressive compression settings. In Tiny mode, the sentence-disrupted setting reduces average precision to 76.7% (-11.2%), while Small and Base modes see smaller but still significant drops to 91.5% (-3.6%) and 97.3% (-0.6%), respectively.

These patterns indicate that sentence-level semantic priors substantially support reconstruction when visual tokens are limited, especially in Tiny mode, where the decoder appears to rely heavily on global linguistic context to reconstruct plausible text. In contrast, ample vision tokens (Base) enable near-perfect recovery regardless of semantic validity, indicating that sufficient visual resolution reduces dependence on higher-order linguistic priors. Overall, sentence-level semantic disruption reveals a clear trade-off: OCR accuracy under compression can be overestimated when evaluation inputs contain strong linguistic regularities, because the decoder can exploit global context in addition to visual evidence.

4.3 Case Study

Figure 1 illustrates a concrete example of prior-induced hallucination under sentence-level disruption. The original text is: "has already been published in other issues of the Speaker, it won't all be repeated here.", and we replace "published" with "breadboxed", "issues" with "whispers", "Speaker" with "Squeaker", and "repeated" with "rebound" to create the disrupted text: "has already been breadboxed in other whispers of the Squeaker, it won't all be rebound here." For the original text, DeepSeek-OCR largely preserves the input, suggesting that natural linguistic context supports stable reconstruction. For the disrupted text, when faced with the visually clear but semantically unlikely word "Squeaker", DeepSeek-OCR fails to transcribe the visual input faithfully. Instead, it outputs the more plausible word "Square" and normalizes "rebound" into "rebroadcast". This behavior confirms that the model prioritizes linguistic probability over visual evidence.

Perturbation	Average Precision (%)		
	Tiny	Small	Base
Swap	$83.69^{-4.31}$	$91.51^{-3.72}$	$95.45^{-2.48}$
Shuffle	$81.83^{-6.17}$	$91.24^{-3.99}$	$95.84^{-2.09}$

Table 2: Average DeepSeek-OCR precision under 5% word-level corruption. Superscripts show drops from rendered text; full token-bin results are in Appendix Table 8.

Perturbation	Average Precision (%)		
	Tiny	Small	Base
Swap	$78.11^{-9.89}$	$88.07^{-7.16}$	$93.75^{-4.18}$
Shuffle	$76.70^{-11.30}$	$86.55^{-8.68}$	$93.99^{-3.94}$

Table 3: Average DeepSeek-OCR precision under 10% word-level corruption. Superscripts show drops from rendered text; full token-bin results are in Appendix Table 9.

5 Probe 2: Word-Level and Zero-Prior Disruption

Having established the role of sentence-level priors, we now turn to finer-grained disruptions at the word level.

5.1 Experimental Setup

We continue using the Fox benchmark as the baseline with three word-level perturbation strategies:

- **Swap**: Randomly select 5% or 10% of words and swap two letters within each selected word, creating minor spelling distortions that preserve most word structure but introduce errors repairable by linguistic priors.
- **Shuffle**: Randomly select 5% or 10% of words and fully shuffle the letters within each selected word, destroying internal word structure while keeping character distributions similar.
- **Zero-Prior Random Text**: Generate entirely new "words" (2–10 random letters, mixed case) to form documents with identical length distribution of the original Fox instances, but devoid of any lexical or syntactic structure.

5.2 Results and Analysis

The performance degradation across these settings (Tables 2, 3, and 4) exposes a heavy dependency on lexical priors; full token-bin word-corruption results are reported in Appendix Tables 8 and 9.

Text-Token Bin	Precision (%)		
	tiny	small	base
600-700	26.81	45.37	63.54
700-800	24.29	45.25	61.93
800-900	22.77	46.53	63.47
900-1000	23.44	42.44	60.94
1000-1100	15.44	43.15	61.07
1100-1200	15.21	40.35	61.60
1200-1300	11.20	33.48	60.18
1300-1400	8.40	33.02	62.36
1400-1500	5.23	25.87	58.28
1500-1600	2.48	23.21	57.83
1600-1700	4.43	24.00	56.05
1700-1800	2.86	18.91	56.64
2400-2500	0.00	6.97	45.60
Average	19.84	42.12	61.70

Table 4: OCR precision (%) of DeepSeek-OCR on zero-prior random text across text-token bins. Zero-prior inputs replace natural words with random character sequences while preserving approximate length distributions, thereby reducing lexical and semantic cues.

First, DeepSeek-OCR is sensitive to letter order: even minor disruptions cause disproportionate failures in high-compression modes. At 10% disruption, Tiny-mode precision drops by 9.89% under Swap and 11.30% under Shuffle on average. Although bin-level results vary, the average drops under Shuffle are larger than Swap in Tiny and Small modes, suggesting that the model benefits from standard character-order regularities during reconstruction. When these regularities are disrupted, visually grounded recovery becomes less reliable, especially under aggressive compression.

The zero-prior setting produces the largest degradation: Tiny-mode precision falls to 19.84%, while Small and Base reach 42.12% and 61.70%, respectively. The large gap between natural-text performance (~90% in Tiny mode) and zero-prior performance suggests that linguistic predictability substantially contributes to OCR accuracy under aggressive compression.

Scenario	Comparison (Input → OCR Result)
Swap	**Input:** ...has ayreadl ... **Output:** ...has **already** ...
Shuffle	**Input:** ...it won't all be eepetadr here. **Output:** ...I won't be all **expected** here.
Zero-Prior	**Input:** EYuoV qUtjpy pWxZCks vUQnwh K qCuYCXmor **Output:** E'vuol u'qtippy piwZckus u'Quwnh kq'CuYcKorom

Table 5: Examples of Model Behavior under Different Word-Level Disruptions

5.3 Case Study

As shown in Table 5, we continue with the example in Probe 1 to illustrate word-level disruptions. In the Swap scenario ("ayreadl" instead of "already"), the model acts as an auto-corrector, outputting the correct English word "already" despite the visual mismatch. This shows reliance on lexical priors to fix minor errors. In the Shuffle scenario ("eepetadr" instead of "repeated"), the model again leverages context to produce "expected", a plausible English word, rather than attempting to decode the nonsensical input. Finally, in the Zero-Prior Random Text scenario, the model largely fails to recover the ground-truth random strings. It attempts to force-fit the random visual patterns into quasi-syllabic structures, resulting in output that is neither the ground truth nor a valid word, but a manifestation of the decoder struggling without priors.

6 Probe 3: Cross-System Generalization

To determine whether linguistic prior dependence is specific to DeepSeek-OCR or reflects a broader pattern across OCR/VLM systems, we conduct a comparative analysis across diverse systems.

6.1 Experimental Setup

We compare DeepSeek-OCR (Tiny/Small modes) against a diverse set of OCR and VLM models spanning 125M–72B parameters on clean natural text and zero-prior random text. The evaluated systems include integrated end-to-end or VLM-based OCR/document systems and a traditional OCR pipeline. The type labels in Table 6 are intentionally coarse: they distinguish E2E/VLM systems from a pipeline OCR reference, while avoiding a strict architectural taxonomy over models with different design goals. The zero-prior random text serves as a prior-removal probe for OCR systems, forcing them to rely primarily on visual recognition capabilities.

6.2 Results and Analysis

The average results in Table 6 reveal a clear disparity across system types, with full token-bin results reported in Appendix Tables 10 and 11. On natural text, most systems achieve high precision, often comparable to or higher than DeepSeek-OCR. However, under zero-prior random text, many E2E/VLM systems suffer large drops in precision. DeepSeek-OCR (Tiny) drops by 68.16 percentage points; similarly, HunyuanOCR (-59.79), Nougat

Model	Type	Natural Avg.	Zero-Prior Avg.	Drop
DeepSeek-OCR Tiny	E2E/VLM	88.00	19.84	-68.16
DeepSeek-OCR Small	E2E/VLM	95.23	42.12	-53.11
dots.ocr 7B	E2E/VLM	97.40	46.31	-51.09
Qwen2.5-VL 7B	E2E/VLM	98.10	46.86	-51.24
Qwen2.5-VL 72B	E2E/VLM	98.69	56.78	-41.91
GOT-OCR 0.58B	E2E/VLM	98.00	57.71	-40.29
MonkeyOCR 1.2B	E2E/VLM	96.60	71.55	-25.05
MonkeyOCR 3B	E2E/VLM	97.37	79.97	-17.40
SmolDocling 0.125B	E2E/VLM	94.30	56.77	-37.53
Nougat 0.35B	E2E/VLM	96.82	37.94	-58.88
HunyuanOCR 1B	E2E/VLM	97.83	38.04	-59.79
GLM-OCR 0.9B	E2E/VLM	97.21	47.46	-49.75
PaddleOCR-VL-1.5 0.9B	E2E/VLM	98.51	86.96	-11.55
PaddleOCR-v5 0.07B	Pipeline	94.44	89.53	-4.91

Table 6: Average OCR precision (%) on natural and zero-prior random text. Type distinguishes integrated end-to-end or VLM-based OCR/document systems (E2E/VLM) from a traditional OCR pipeline (Pipeline). This coarse grouping is used only to contextualize robustness patterns, not to define a strict architectural taxonomy. Full token-bin results are in Appendix Tables 10 and 11.

(-58.88), and Qwen2.5-VL 7B (-51.24) show substantial degradation. GLM-OCR follows the same trend, dropping from 97.21% on natural text to 47.46% on random text (-49.75), suggesting substantial prior dependence in this VLM-based OCR setting.

The trend is not uniform across all E2E/VLM systems. PaddleOCR-VL-1.5 remains comparatively robust, dropping by 11.55 points, while the traditional pipeline OCR system PaddleOCR-v5 drops by only 4.91 points. These results suggest that systems with more explicit visual-recognition pathways are less dependent on linguistic priors, whereas many integrated E2E/VLM OCR systems are more vulnerable when lexical and semantic cues are removed.

6.3 Case Study

As shown in Figure 4, we present an example of OCR results on natural text and random text across an optical-compression model (DeepSeek-OCR Small), a general VLM (Qwen2.5-VL 72B), and a traditional pipeline OCR model (PaddleOCR-v5). For natural text, all three models perform perfectly, achieving 100% precision. However, on random text, both DeepSeek-OCR Small and Qwen2.5-VL 72B struggle to recognize the visually present strings and tend to force-fit random character patterns into more familiar token sequences. In contrast, PaddleOCR-v5 maintains high precision and correctly identifies most of the semantically meaningless strings.

This comparison illustrates that strong natural-text OCR accuracy in E2E/VLM systems does not necessarily imply robustness under low-prior inputs. Rather than treating OCR accuracy on natural text as pure visual recognition, these results suggest that such outputs may partly reflect reconstruction through a linguistic lens, which becomes a liability when the text is unstructured or nonsensical.

7 Probe 4: Simple QA/VQA Semantic Preservation

High OCR accuracy does not guarantee that optically compressed representations preserve the basic semantic information needed to answer simple factual questions. We therefore use document QA and VQA as lightweight probes of semantic preservation, rather than as a broad evaluation of advanced reasoning ability.

7.1 Experimental Setup

We extend the Fox benchmark by annotating each document page with three fact-based question-answer pairs. The questions are intentionally simple: each can be answered from explicit evidence in the page and requires only basic reading comprehension and evidence selection, rather than domain expertise or multi-hop reasoning. VQA includes strong multimodal baselines: Qwen2.5VL-3B/7B, Qwen3VL-4B/8B (Bai et al., 2025a), and MiniCPM-V 4.5 (Hu et al., 2024). QA baselines include Qwen2.5-3B, Qwen3-4B, and Llama3.2-3B (all ~3–4B scale, similar to DeepSeek-OCR's activated parameters).

7.2 Results and Analysis

The performance gap illustrated in Figure 2 suggests that character-level reconstruction and elementary semantic accessibility should be evaluated

separately. Because the questions are intentionally simple and evidence-local, this probe is closer to a basic sanity check than a difficult reasoning benchmark. Strong text-only LLMs solve the corresponding text-QA setting with over 90% accuracy when given the ground-truth text, confirming that the required evidence is explicit in the document and that the language-understanding burden is minimal.

In contrast, DeepSeek-OCR remains near chance in the VQA setting (~20% accuracy). This result should not be interpreted as a broad criticism of DeepSeek-OCR as a reasoning-optimized VLM, since the model was primarily introduced as an optical text-compression and OCR system. Rather, it indicates that the compressed visual pathway, as evaluated here, does not reliably expose the document evidence needed for even simple question answering.

Even when DeepSeek-OCR's decoder is provided with uncompressed ground-truth text, it reaches only 27.7% accuracy. This result suggests that the DeepSeek-OCR decoding pathway does not reliably demonstrate the elementary language understanding and evidence-grounded reasoning needed to answer these simple document questions. We therefore interpret the QA/VQA results as a diagnostic of semantic accessibility: high surface-level OCR precision can coexist with weak support for basic document understanding.

7.3 Case Study

Figure 5 illustrates this evidence-accessibility gap with a simple question about legal consequences. The answer is explicitly stated in the document, and standard text-only LLMs correctly select answer B when given the ground-truth text. DeepSeek-OCR, however, selects different incorrect answers in QA and VQA modes. In the QA setting, where its decoder is given the uncompressed text, it selects an incorrect option; in the VQA setting, where the model must answer from the document image, it selects another incorrect option. This inconsistency suggests that the relevant evidence is not reliably accessed through either the DeepSeek-OCR decoding pathway or the compressed visual pathway, even though the answer is explicit in the source document. The case therefore illustrates why surface-level OCR fidelity and elementary evidence-grounded document understanding should be evaluated separately.

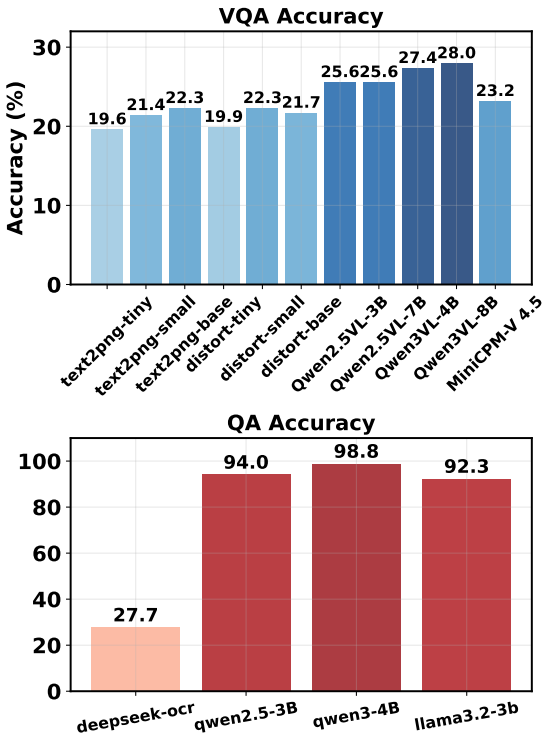

Figure 2: VQA and QA Performance.

8 Probe 5: Context-Length Boundary

The core premise of DeepSeek-OCR is that optical compression can bypass the quadratic complexity of standard LLMs, motivating the possibility of longer effective contexts. In this final analysis, we stress-test this claim. We investigate whether the "optical context" is truly scalable, or if the fixed resolution of vision encoders imposes a density-dependent visual information bottleneck that triggers catastrophic model collapse.

8.1 Experimental Setup

To evaluate the limits of optical compression, we construct a controlled long-form narrative benchmark. We prompt GPT-5.1 to generate five English stories with 5k words each. To achieve long contexts, each story is repeated until reaching approximately 20,000 tokens. Each story is then segmented into 40 spans (500–20,000 tokens, 500-token steps) and rendered as document images. Evaluation is performed in Tiny, Small, Base, and Large modes of DeepSeek-OCR to determine if scaling the visual encoder mitigates context length limitations.

8.2 Results and Analysis.

Figure 3 plots OCR precision against context length for each compression mode. For presentation clar-

ity, we truncate the x-axis at 12,000 tokens, since all subsequent precision values are zero in our experiments. In contrast to the expectation that optical compression can support long effective contexts, all DeepSeek-OCR modes exhibit a sharp performance cliff under our controlled long-form rendering setup.

Context-length boundary. We observe a consistent collapse region between approximately 8,000 and 10,500 tokens. Tiny mode begins to degrade substantially after around 6,000 tokens and reaches zero precision by approximately 8,500 tokens. Base and Large modes also collapse around 8,500 tokens, indicating that increasing the visual-token budget does not eliminate the density-induced bottleneck. Interestingly, Small mode remains marginally usable for longer inputs and collapses around 10,500 tokens, slightly later than the other modes. This non-monotonic behavior suggests that simply scaling the visual encoder does not guarantee better robustness under extreme text density, and may reflect instability in how different compression modes allocate visual capacity.

Density-fidelity trade-off. These results suggest a fundamental limitation of the current optical-compression paradigm. A fixed-grid visual encoder can preserve only a finite amount of textual information. Once the rendered text density exceeds this capacity, the signal-to-noise ratio of the visual representation drops below the decoder's recovery threshold, leading to repetition, hallucination, or complete reconstruction failure. Under our controlled stress-test setting, DeepSeek-OCR reduces textual token cost but encounters a restrictive visual information bottleneck, and therefore does not reliably scale to long rendered contexts beyond roughly 10,000 tokens.

8.3 Case Study

Table 12 shows the progressive degradation of OCR results as context length increases from 2,500 to 8,500 tokens in DeepSeek-OCR Base mode. At 2,500 tokens, the result almost perfectly reconstructs the original text, only with a minor error ("Xingfu" misrecognized as "Yingxiu"). However, after increasing the input by only 500 tokens to 3,000, significant errors emerge: the first half of the sentence becomes unrelated to the original text. At 3,500 and 4,000 tokens, the OCR result is totally hallucinated by the decoder, with repeated phrases "The first time in the past" in 3,500 tokens and "narrow hallway with a long" in 4,000 tokens. By 8,000 tokens, the OCR result is simply an irrelevant sentence: "The following text is a placeholder for an image. It does not contain any relevant information for the article." After 8,500 tokens, the model completely collapses and fails to produce any meaningful text, and the OCR output consists entirely of meaningless HTML tags: "<table><tr><td></td><td></td>..." This case vividly illustrates the severe semantic degradation incurred through optical compression as context length increases.

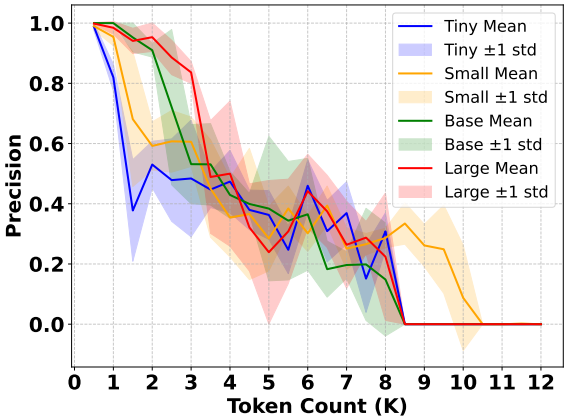

Figure 3: OCR precision under different DeepSeek-OCR compression modes as rendered context length increases.

9 Conclusion

We studied how optical text compression systems should be evaluated when natural-text OCR accuracy may conflate visual recognition with decoder-side linguistic reconstruction. We introduced SVST, a prior-disentangled protocol that removes sentence-level, word-level, and lexical-semantic cues while preserving rendering conditions. Using DeepSeek-OCR as a representative case and extending the analysis across OCR/VLM systems, we found that many integrated E2E/VLM OCR/document systems are substantially more sensitive to the removal of linguistic priors than systems with more explicit recognition pathways. We further showed that surface-level OCR precision does not necessarily imply semantic usability, and that long-form rendering exhibits a density-dependent failure boundary. Overall, these results suggest that optical text compression should be evaluated using prior-disentangled fidelity, semantic accessibility, and context-boundary diagnostics, rather than natural-text OCR accuracy alone.

Limitations

First, our evaluation focuses primarily on English dense textual documents; prior dependence may differ across languages, scripts, and document layouts. Second, although we evaluate a broad set of OCR/VLM systems, new long-context OCR systems may adopt different compression, recognition, and decoding mechanisms, so our conclusions should be interpreted primarily for the system types evaluated here. Third, our perturbations are synthetic and are intended to isolate linguistic priors; they may not fully capture naturally occurring low-prior inputs such as names, identifiers, code, formulas, and tables. Fourth, our QA/VQA experiments are lightweight semantic-preservation diagnostics rather than comprehensive reasoning benchmarks.

References

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

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

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*.

Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

Joanna Materzyńska, Antonio Torralba, and David Bau. 2022. Disentangling visual and written concepts in clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16410–16419.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong

Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024b. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

A Appendix

A.1 Full Token-Bin OCR Results

The following tables provide the complete token-bin results corresponding to the compact main-text summaries.

A.2 Cases Illustration for Probes 3, 4 and 5

Text-Token Bin	Precision (%)								
	Fox tiny	Fox small	Fox base	Rendered tiny	Rendered small	Rendered base	Sentence-disrupted tiny	Sentence-disrupted small	Sentence-disrupted base
600-700	95.93	98.30	98.51	98.64	99.00	99.56	$96.23^{-2.41}$	$98.78^{-0.22}$	$99.60^{+0.04}$
700-800	93.81	96.98	97.65	96.70	98.48	98.82	$87.94^{-8.76}$	$94.08^{-4.40}$	$98.12^{-0.70}$
800-900	91.91	96.65	97.74	94.46	97.49	98.75	$88.46^{-6.00}$	$96.91^{-0.58}$	$99.20^{+0.45}$
900-1000	84.19	96.68	98.80	87.73	96.94	98.94	$70.06^{-17.67}$	$90.11^{-6.83}$	$98.20^{-0.74}$
1000-1100	79.24	91.25	95.27	86.18	95.24	96.67	$74.54^{-11.64}$	$92.05^{-3.19}$	$96.32^{-0.35}$
1100-1200	74.34	89.21	93.67	80.72	94.21	95.46	$57.76^{-22.96}$	$87.11^{-7.10}$	$95.03^{-0.43}$
1200-1300	58.73	86.44	89.65	73.97	91.85	93.39	$54.26^{-19.71}$	$90.23^{-1.62}$	$97.62^{+4.23}$
1300-1400	69.34	90.98	96.05	64.22	87.61	98.16	$50.49^{-13.73}$	$81.22^{-6.39}$	$96.05^{-2.11}$
1400-1500	76.70	96.03	99.47	64.48	90.86	98.68	$1.94^{-62.54}$	$73.33^{-17.53}$	$96.75^{-1.93}$
1500-1600	34.41	76.01	92.23	53.59	84.41	95.83	$46.75^{-6.84}$	$80.46^{-3.95}$	$98.18^{+2.35}$
1600-1700	58.14	86.18	94.02	64.50	87.01	97.13	$44.12^{-20.38}$	$79.54^{-7.47}$	$96.97^{-0.16}$
1700-1800	34.43	76.29	95.78	60.00	74.34	94.29	$25.93^{-34.07}$	$20.00^{-54.34}$	$43.14^{-51.15}$
2400-2500	0.43	37.24	73.60	13.66	13.61	84.00	$40.59^{+26.93}$	$63.92^{+50.31}$	$81.05^{-2.95}$
Average	83.88	93.89	96.62	88.00	95.23	97.93	$76.75^{-11.25}$	$91.56^{-3.67}$	$97.31^{-0.62}$

Table 7: Full token-bin precision (%) for DeepSeek-OCR under sentence-level semantic disruption, corresponding to the summary in Table 1. Fox denotes the original benchmark images, Rendered denotes ground-truth text rendered with the same pipeline, and Sentence-disrupted denotes rendered text after semantic-disruption replacements.

Text-Token Bin	Swap (%)			Shuffle (%)		
	tiny	small	base	tiny	small	base
600-700	$94.19^{-4.45}$	$95.21^{-3.79}$	$96.55^{-3.01}$	$87.37^{-11.27}$	$91.83^{-7.17}$	$96.72^{-2.84}$
700-800	$92.02^{-4.68}$	$94.94^{-3.54}$	$95.94^{-2.88}$	$85.81^{-10.89}$	$88.53^{-9.95}$	$94.16^{-4.66}$
800-900	$90.82^{-3.64}$	$93.92^{-3.57}$	$96.16^{-2.59}$	$83.17^{-11.29}$	$89.55^{-7.94}$	$95.56^{-3.19}$
900-1000	$84.21^{-3.52}$	$92.31^{-4.63}$	$95.72^{-3.22}$	$77.48^{-10.25}$	$87.32^{-9.62}$	$93.02^{-5.92}$
1000-1100	$81.42^{-4.76}$	$91.73^{-3.51}$	$94.49^{-2.18}$	$73.93^{-12.25}$	$85.40^{-9.84}$	$92.63^{-4.04}$
1100-1200	$78.13^{-2.59}$	$89.98^{-4.23}$	$94.08^{-1.38}$	$72.87^{-7.85}$	$84.76^{-9.45}$	$93.26^{-2.20}$
1200-1300	$69.70^{-4.27}$	$87.08^{-4.77}$	$94.56^{+1.17}$	$62.56^{-11.41}$	$83.05^{-8.80}$	$92.85^{-0.54}$
1300-1400	$61.07^{-3.15}$	$85.31^{-2.30}$	$95.42^{-2.74}$	$56.58^{-7.64}$	$78.39^{-9.22}$	$92.84^{-5.32}$
1400-1500	$67.63^{+3.15}$	$91.58^{+0.72}$	$95.42^{-3.26}$	$25.23^{-39.25}$	$82.66^{-8.20}$	$90.65^{-8.03}$
1500-1600	$37.38^{-16.21}$	$78.82^{-5.59}$	$91.92^{-3.91}$	$40.32^{-13.27}$	$71.56^{-12.85}$	$92.69^{-3.14}$
1600-1700	$55.71^{-8.79}$	$83.86^{-3.15}$	$95.51^{-1.62}$	$43.50^{-21.00}$	$78.12^{-8.89}$	$93.42^{-3.71}$
1700-1800	$25.45^{-34.55}$	$21.05^{-53.29}$	$89.66^{-4.63}$	$37.93^{-22.07}$	$61.54^{-12.80}$	$84.69^{-9.60}$
2400-2500	$31.11^{+17.45}$	$60.50^{+46.89}$	$88.18^{+4.18}$	$1.94^{-11.72}$	$56.54^{+42.93}$	$87.76^{+3.76}$
Average	$83.69^{-4.31}$	$91.51^{-3.72}$	$95.45^{-2.48}$	$81.83^{-6.17}$	$91.24^{-3.99}$	$95.84^{-2.09}$

Table 8: Full token-bin precision (%) for DeepSeek-OCR under 5% word-level corruption, corresponding to the summary in Table 2.

Text-Token Bin	Swap (%)			Shuffle (%)		
	tiny	small	base	tiny	small	base
600-700	$90.14^{-8.50}$	$92.60^{-6.40}$	$95.51^{-4.05}$	$93.06^{-5.58}$	$95.19^{-3.81}$	$96.84^{-2.72}$
700-800	$87.82^{-8.88}$	$89.82^{-8.66}$	$94.03^{-4.79}$	$90.70^{-6.00}$	$93.76^{-4.72}$	$96.49^{-2.33}$
800-900	$86.29^{-8.17}$	$90.84^{-6.65}$	$94.45^{-4.30}$	$87.45^{-7.01}$	$93.86^{-3.63}$	$96.98^{-1.77}$
900-1000	$73.06^{-14.67}$	$88.91^{-8.03}$	$93.90^{-5.04}$	$82.85^{-4.88}$	$92.26^{-4.68}$	$96.03^{-2.91}$
1000-1100	$77.16^{-9.02}$	$88.09^{-7.15}$	$92.07^{-4.60}$	$79.41^{-6.77}$	$90.68^{-4.56}$	$94.77^{-1.90}$
1100-1200	$74.62^{-6.10}$	$87.09^{-7.12}$	$92.93^{-2.53}$	$73.31^{-7.41}$	$89.08^{-5.13}$	$94.44^{-1.02}$
1200-1300	$48.84^{-25.13}$	$85.85^{-6.00}$	$94.93^{+1.54}$	$71.08^{-2.89}$	$86.52^{-5.33}$	$94.82^{+1.43}$
1300-1400	$58.17^{-6.05}$	$80.99^{-6.62}$	$93.28^{-4.88}$	$60.87^{-3.35}$	$84.64^{-2.97}$	$94.64^{-3.52}$
1400-1500	$64.94^{+0.46}$	$86.89^{-3.97}$	$93.87^{-4.81}$	$66.04^{+1.56}$	$89.03^{-1.83}$	$94.71^{-3.97}$
1500-1600	$51.10^{-2.49}$	$73.37^{-11.04}$	$91.61^{-4.22}$	$32.55^{-21.04}$	$77.29^{-7.12}$	$92.58^{-3.25}$
1600-1700	$41.52^{-22.98}$	$79.64^{-7.37}$	$92.27^{-4.86}$	$59.63^{-4.87}$	$81.88^{-5.13}$	$95.63^{-1.50}$
1700-1800	$31.87^{-28.13}$	$59.15^{-15.19}$	$86.09^{-8.20}$	$35.94^{-24.06}$	$66.07^{-8.27}$	$89.47^{-4.82}$
2400-2500	$5.30^{-8.36}$	$46.86^{+33.25}$	$89.19^{+5.19}$	$30.43^{+16.77}$	$54.75^{+41.14}$	$83.49^{-0.51}$
Average	$78.11^{-9.89}$	$88.07^{-7.16}$	$93.75^{-4.18}$	$76.70^{-11.30}$	$86.55^{-8.68}$	$93.99^{-3.94}$

Table 9: Full token-bin precision (%) for DeepSeek-OCR under 10% word-level corruption, corresponding to the summary in Table 3.

Text-Token Bin	Precision (%)													
	E2E/VLM OCR-Document Systems													Pipeline OCR
	DeepSeek-OCR		dots.ocr	Qwen2.5vl		GOT-OCR	MonkeyOCR		SmolDocling	Nougat	HunyuanOCR	GLM-OCR	PaddleOCR-VL-1.5	PaddleOCR-v5
	tiny	small	7B	7B	72B	0.58B	1.2B	3B	0.125B	0.35B	1B	0.9B	0.9B	0.07B
600-700	98.64	99.00	98.99	99.52	99.73	99.56	98.48	99.46	98.90	99.02	99.50	99.30	99.77	97.82
700-800	96.70	98.48	97.12	98.02	98.33	98.16	97.34	97.53	90.34	97.17	98.23	97.60	98.85	95.73
800-900	94.46	97.49	97.51	98.23	98.76	98.75	97.00	97.28	93.60	96.85	97.53	97.50	99.22	95.39
900-1000	87.73	96.94	98.12	98.98	99.04	99.11	97.37	97.90	98.17	98.09	98.61	98.26	99.16	96.17
1000-1100	86.18	95.24	97.65	96.97	98.20	97.67	94.42	96.97	96.84	96.49	97.57	95.96	98.26	93.68
1100-1200	80.72	94.21	95.98	96.73	98.21	96.30	93.97	95.39	93.42	93.20	96.09	94.68	94.56	91.18
1200-1300	73.97	91.85	97.71	98.29	99.61	98.08	93.74	95.27	95.33	95.43	97.79	94.64	97.81	88.46
1300-1400	64.22	87.61	97.44	98.05	99.68	98.22	97.26	97.97	97.52	98.39	97.32	98.85	98.94	94.63
1400-1500	64.48	90.86	99.47	1.00	99.74	98.15	99.47	99.21	98.95	97.61	99.74	99.47	99.74	95.23
1500-1600	53.59	84.41	92.95	97.85	97.84	95.87	93.93	95.23	91.28	92.62	95.68	92.92	95.78	90.23
1600-1700	64.50	87.01	95.56	97.29	97.63	95.42	96.66	96.51	95.88	97.07	95.80	94.47	95.96	84.13
1700-1800	60.00	74.34	97.32	96.49	96.49	90.43	100.00	100.00	94.77	92.87	96.98	96.17	95.32	93.61
2400-2500	13.66	13.61	98.61	99.43	100.00	78.47	92.71	100.00	94.92	94.27	99.43	97.40	99.38	70.37
Average	88.00	95.23	97.40	98.10	98.69	98.00	96.60	97.37	94.30	96.82	97.83	97.21	98.51	94.44

Table 10: Full token-bin precision (%) on natural text across OCR/VLM systems, corresponding to the natural-text column in Table 6. The first 13 model columns correspond to E2E/VLM systems, and PaddleOCR-v5 is the pipeline OCR reference.

Text-Token Bin	Precision (%)													
	E2E/VLM OCR-Document Systems													Pipeline OCR
	DeepSeek-OCR		dots.ocr	Qwen2.5vl		GOT-OCR	MonkeyOCR		SmolDocling	Nougat	HunyuanOCR	GLM-OCR	PaddleOCR-VL-1.5	PaddleOCR-v5
	tiny	small	7B	7B	72B	0.58B	1.2B	3B	0.125B	0.35B	1B	0.9B	0.9B	0.07B
600-700	26.81	45.37	46.80	51.20	58.61	70.13	71.77	80.39	66.45	45.16	47.80	48.43	87.32	88.94
700-800	24.29	45.25	47.77	50.49	56.14	59.63	71.63	79.64	61.10	36.50	40.76	46.04	87.25	89.93
800-900	22.77	46.53	48.83	50.54	58.11	63.43	72.53	80.30	58.28	40.50	35.98	50.34	87.19	89.70
900-1000	23.44	42.44	43.97	44.11	54.55	57.50	71.17	79.34	54.29	42.36	35.95	47.08	85.94	89.59
1000-1100	15.44	43.15	46.69	38.83	59.35	55.23	72.07	80.86	45.34	38.58	38.44	47.02	87.17	89.00
1100-1200	15.21	40.35	45.65	49.88	58.25	55.87	71.71	80.68	54.21	39.59	36.45	46.25	86.97	89.66
1200-1300	11.20	33.48	43.98	38.00	55.57	49.34	70.16	79.44	59.58	15.51	38.67	48.04	85.81	89.43
1300-1400	8.40	33.02	40.26	37.78	54.10	47.15	70.91	79.97	58.03	39.06	34.21	44.51	86.78	89.59
1400-1500	5.23	25.87	44.16	43.45	53.61	45.33	70.84	80.38	56.95	36.57	34.29	44.89	88.95	99.96
1500-1600	2.48	23.21	40.95	48.66	53.30	40.58	68.64	79.56	49.05	21.83	31.51	41.35	85.87	88.94
1600-1700	4.43	24.00	39.35	24.22	51.74	35.60	67.74	76.64	46.03	37.74	36.12	50.53	84.91	87.55
1700-1800	2.86	18.91	38.88	41.89	56.18	29.68	65.39	77.78	54.39	0.00	30.45	39.06	88.84	86.98
2400-2500	0.00	6.97	32.89	44.87	50.16	10.69	66.72	80.13	38.57	36.06	31.11	48.91	85.83	88.98
Average	19.84	42.12	46.31	46.86	56.78	57.71	71.55	79.97	56.77	37.94	38.04	47.46	86.96	89.53

Table 11: Full token-bin precision (%) on zero-prior random text across OCR/VLM systems, corresponding to the zero-prior column in Table 6. The first 13 model columns correspond to E2E/VLM systems, and PaddleOCR-v5 is the pipeline OCR reference.

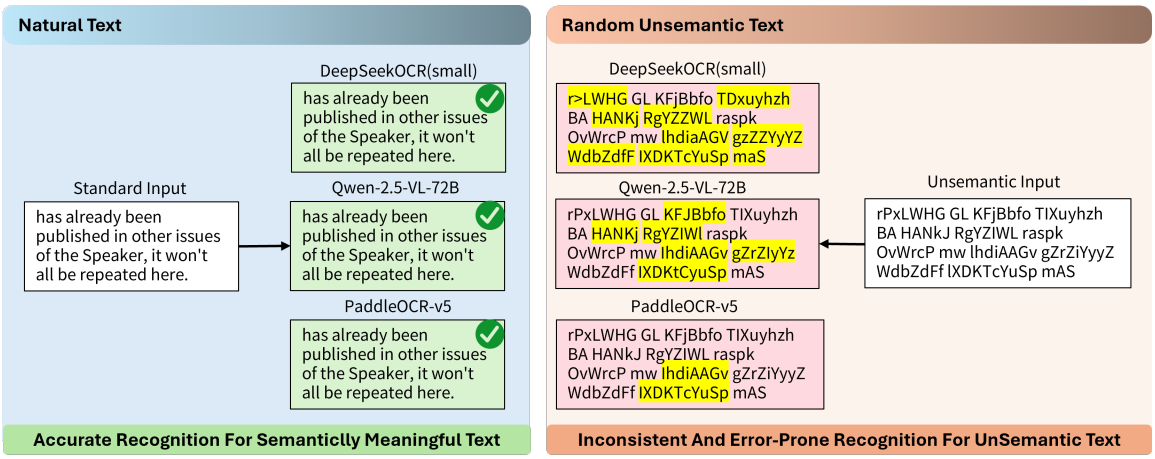

Figure 4: Case of OCR results on natural text and unsemantic text across different models.

Question: What action can a court take if a public body fails to comply with a disclosure order under the Freedom of Information Act?

Options:

A: Suspend the public body's operations
B: Punish the violation as contempt of court
C: Require the public body to hold a public hearing
D: Impose a criminal fine

Correct Answer: B

Reference Text: Failure to comply with an order of the court may be punished as contempt of court

QA Answers		VQA Answers	
Deepseek-OCR Answer:	A	Deepseek-OCR Answer:	C
Llama3.2-3B Answer:	B	Qwen2.5-VL-3B Answer:	D
Qwen2.5-3B Answer:	B	Qwen2.5-VL-7B Answer:	D
Qwen3-4B Answer:	B	Qwen3-VL-4B Answer:	C
		Qwen3-VL-4B Answer:	A

Figure 5: Case of QA and VQA results across different models.

Token Count	Category	Model Output Example & Analysis
Input Text	Ground Truth	The Old Phonograph. The first time Lin Mian saw the phonograph, it was tucked in the shadowy corner of the bookstore on Xingfu Road.
2.5k	Almost Perfect	The Old Phonograph. The first time Lin Man saw the phonograph, it was tucked in the shadowy corner of the second-hand bookstore on Yingxiu Road.
3k	Start Hallucination	The growth of the Internet has run like a sawmilling machine. It was launched by the discoverer of the second-hand bookstore on Kingfisher Road.
3.5k	Repetitive	The world of the past. The first time in the past. The first time in the past. The first time in the past... (Repeated loops)
4k	Repetitive	The ground floor of the train in San Mateo is the ground floor, it is located in the hallway of the car. The ground floor is a long, narrow hallway with a long, narrow hallway with a long... (Repeated loops)
8k	Irrelevant	*"The following text is a placeholder for an image. It does not contain any relevant information for the article."*
8.5k+	**Model Collapse**	`<table><tr><td></td><td></td><td>...` (Structural breakdown into raw HTML/noise)

Table 12: Qualitative Degradation of DeepSeek-OCR Base Outputs as Context Length Increases