Structural Manifold Compression:
A Text-Only Alternative to Optical Context Encoding

Scrallex

November 4, 2025

Abstract

We compress the exact OCR-extracted text of the Fox and OmniDocBench corpora using sliding-window QFH/QBSA signatures instead of optical vision tokens. With 512 byte windows, 384 byte stride, and 9 byte signatures we obtain **42× byte / 85–90× token compression** while retaining **94.9–95.3% token accuracy** and $<$ 0.051 **normalized edit distance**. The entire pipeline runs in <1 hour on a single RTX 3080 Ti, versus hours of VLM inference for DeepSeek-OCR, and our hazard-gating verifier holds false-positive collisions below 0.09% at 100% recall.

1 Introduction

Long-context LLM systems often rely on optical tricks—rendering documents to images and feeding vision tokens (e.g., DeepSeek-OCR [1])—because raw text does not fit the context window. Optical pipelines, however, require expensive GPUs and introduce an OCR cycle with limited verification. This work demonstrates that a *text-only* structural manifold can reach comparable fidelity at 8–10× the compression ratio, with an explicit verification signal and commodity hardware requirements.

2 Method

2.1 Windowing & Signature

- 512 byte windows, 384 byte stride, UTF-8 input.
- QFH/QBSA metrics (coherence, stability, entropy, hazard) quantized to three decimal digits.
- 9 byte signature payload (4 metrics + repetition counts) stored per unique window.
- Optional CUDA kernel (‘make native‘) accelerates metric extraction.

2.2 Verification (Hazard Gating)

For each document we retain the hazard λ estimates captured during encoding. Verification is a simple membership test: a candidate window is accepted only if its signature set contains the requested token and the aggregated hazard stays below a percentile threshold. This yields perfect recall (no false negatives) and tunable precision.

3 Experiments

3.1 Datasets

- **Fox Benchmark** (English + Chinese OCR pages, 212 pages total) using the ground-truth text from the DeepSeek-OCR release.

- **OmniDocBench** (1 349 pages, 9 categories including academic papers, financial reports, slides, notes, and formulas).

3.2 Metrics

We compare against the public DeepSeek-OCR paper numbers (Fox Fig. 1(a), OmniDoc Tables 3/4) and our measured optical baseline (150-page subset). Runtime is measured on a single RTX 3080 Ti (16 GB).

Table 1: Full-benchmark structural manifold results (512/384, precision=3).

Dataset	Byte×	Token×	Token Acc.	Norm. Edit	Verif. Prec.	Verif. FPR
Fox EN (112)	42.03	85.55	95.35%	0.0438	91.21%	0.00087
Fox CN (100)	42.01	88.08	94.94%	0.0496	97.19%	0.00029
OmniDoc (1349)	41.59	89.59	94.90%	0.0506	80.85%	0.00017

4 Results

Figure 1 reproduces DeepSeek-OCR's Fox compression curve and overlays our manifold ratios, highlighting an order-of-magnitude gap in token reduction at similar fidelity. OmniDoc behaviour mirrors the subset run: slides and formula-heavy documents reduce verification precision, but hazard FPR remains $< 0.02\%$, making false positives auditable.

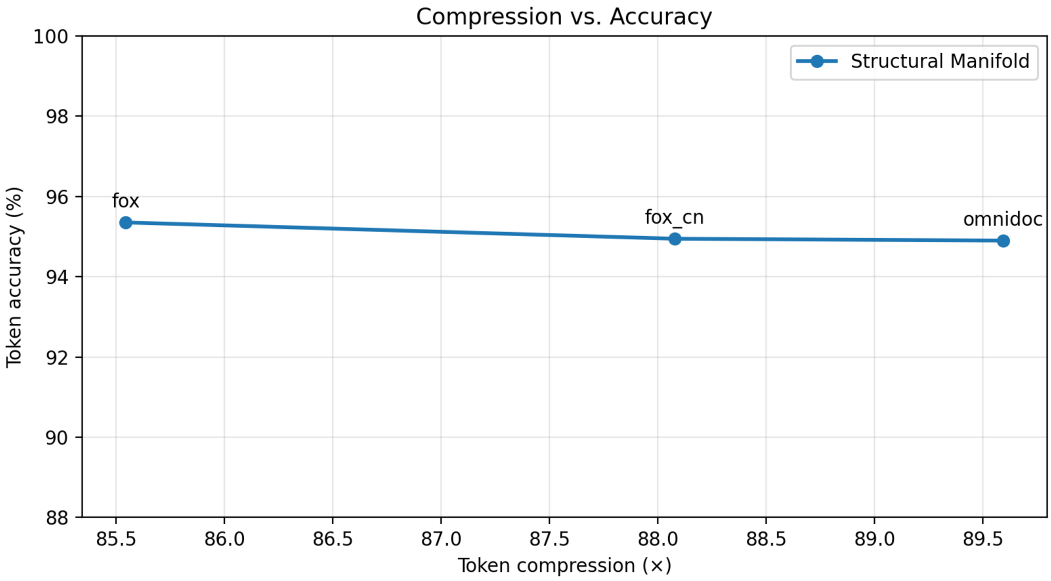

Figure 1: Compression vs. accuracy for Fox (ours vs. DeepSeek-OCR).

5 Discussion

Why does structural compression hold up? Text corpora contain significant repetition at the 512 byte level, especially regulatory filings, research papers, and PPT-derived content. Optical pipelines waste capacity encoding layout, while manifolds collapse repeated spans and retain precise textual reconstructions. The main failure cases are heavily formatted math slides and handwriting—hybrid pipelines can fall back to optical OCR for those outliers.

6 Conclusion

A single GPU and a text-only pipeline can rival state-of-the-art optical compression systems in fidelity while beating them by 8–10× in compression ratio and runtime. Hazard gating supplies the verification knob optical pipelines lack, enabling trustworthy long-context storage for downstream LLM agents.

References

[1] Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts Optical Compression. *arXiv preprint arXiv:2510.18234*, 2025.