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

Alexander Nagy
Sep Dynamics (Austin, TX)
Independent Researcher

November 4, 2025

Abstract

Vision-language pipelines such as DeepSeek-OCR compress long contexts by rendering pages to images and streaming vision tokens, but they require expensive GPUs and offer limited verification of reconstructed spans. We revisit the Fox and OmniDocBench corpora with a purely textual approach: structural manifolds that store only quantised coherence, stability, entropy, and hazard signatures. Using 512 byte windows, a 384 byte stride, and a 9 byte payload per unique signature, the proposed encoder delivers **42× byte** and **85–90× token compression** on the full benchmarks while retaining **94.9–95.3% token accuracy** and $< 5.1\%$ **normalized edit distance**. Hazard-gated verification keeps false positives under **0.09%** at 100% recall, and the entire run finishes in under one hour on a single RTX 3080 Ti. On the shared 150-document subset, the text-only manifold is 23–37 accuracy points better than DeepSeek-OCR despite operating at higher compression ratios. The evaluation harness and figures are released at github.com/SepDynamics/structural-manifold-compression.

1 Introduction

Long-context LLM deployments frequently turn to optical tricks: convert documents into images, run a vision encoder, and hope that the reconstructed text stays faithful. The approach works, yet it incurs a second inference stack, saturates GPU memory with image tensors, and leaves downstream systems without a crisp verification signal. This report demonstrates that the Sep Dynamics (SEP) structural manifold—built purely from textual signatures—matches or exceeds optical fidelity on Fox and OmniDocBench while simultaneously supplying an interpretable hazard-based verifier.

Contributions.

- We present a compact 9 byte signature that stores quantised coherence, stability, entropy, hazard, and repetition counts per 512 byte window, enabling >40× byte compression without discarding the original UTF-8 prototype.
- We integrate hazard-gated verification into the compression loop, yielding perfect recall and $< 0.09\%$ false positives on the full Fox and OmniDocBench benchmarks.
- We release a reproducible benchmark harness (`scripts/experiments/benchmark_eval.py`) and cross-modality comparison against DeepSeek-OCR, highlighting that text-only manifolds can outperform vision pipelines while using commodity hardware.

2 Related Work

Optical compression methods such as DeepSeek-OCR [1] render pages to images and encode them with specialised vision-language models, reporting 9–10× contextual compression at $\geq 96\%$

accuracy on Fox. Token-efficient architectures (e.g., long-context transformers with sparse attention) continue to operate on textual tokens and therefore inherit quadratic costs. Our work instead focuses on compressing the *text itself*: by storing structural fingerprints and deduplicating repeating windows, we obtain vision-scale ratios without invoking an optical stack.

3 Structural Manifold Methodology

3.1 Sliding-Window Encoding

We slide fixed windows of 512 bytes over UTF-8 documents with a 384 byte stride. Each window is passed through SEP's encoder, producing coherence (q), stability (ϕ), entropy (h), and hazard (λ) metrics. The metrics are quantised to three decimal places and packed with repetition counts into a 9 byte signature payload. Unique signatures are stored once per document alongside their first raw span, while repeats increase the count; serialization targets Valkey namespaces (e.g., `gate:last:{instrument}`) for the trading system compatibility.

3.2 Reconstruction and Token Metrics

Reconstruction concatenates the prototype spans in the order implied by the sliding windows, overlapping only the stride tail. Token accuracy is measured with the released DeepSeek tokenizer, computing $1 - \frac{d_{\text{tok}}}{\max(|x|,|y|)}$ where d_{tok} is the Levenshtein edit distance between the original and reconstructed token streams x and y. We also report normalized edit distance at the character level, unique-token compression (original tokens divided by distinct signatures), and streamed-token compression (original tokens divided by total windows).

3.3 Hazard-Gated Verification

During encoding we store the per-window hazard estimate λ, interpreted as a collision prior. When verifying a candidate span, we require both (i) a matching signature and (ii) the aggregated hazard staying below a percentile threshold (80th percentile in all experiments). This yields perfect recall (we never drop a true span) and logarithmically decreasing false-positive rates as documents accumulate unique windows; auditors can inspect the hazards to trace risk hotspots.

3.4 Implementation Notes

All experiments were executed on a single workstation with an RTX 3080 Ti (16 GB) and Threadripper-class CPU. The pipeline is CPU-friendly, but compiling the optional CUDA kernel via `make native` increases throughput to approximately 55k windows per second. Encoding Fox EN + Fox CN + OmniDocBench (1 561 documents in total) takes 47 minutes end-to-end and produces a 92 MB Valkey manifest ready for consumption by `PortfolioManager`.

4 Experimental Setup

4.1 Datasets

The Fox benchmark contributes 112 English and 100 Chinese OCR pages sourced from the publicly released Fox dataset [2]; the authors distribute the material for non-commercial research use via the project site. OmniDocBench [3] adds 1 349 heterogeneous pages (academic papers, financial reports, presentations, notes, and formula sheets) under the Apache-2.0 license, which enables redistribution alongside this study's scripts. The original OmniDocBench paper reports 981 PDF-level samples across nine types; we operate on the OpenDataLab page-level text

manifest (downloaded 2025-03-24) plus its appendix addenda, which expands to 1 349 UTF-8 files, and we evaluate every entry in that extended manifest. All corpora are evaluated verbatim—no preprocessing beyond UTF-8 normalization.

4.2 Baselines

The primary baseline is DeepSeek-OCR [1], run via `scripts/experiments/deepseek_ocr_runner.py` on the first 150 records of each corpus (matching the subset used for our structural evaluation). We follow their public inference recipe—prompting with "`<image>\nFree OCR.`" and the released tokenizer/checkpoint in `bfloat16`—but cap evaluation to the first 150 manifest entries and stream per-page text, which partially explains the lower accuracies relative to the headline Fox/OmniDoc numbers reported in [1]. Additionally, we reference the published DeepSeek Fox/OmniDoc numbers to contextualize compression ratios, since the released model does not expose byte-level budgets for its image tokens.

4.3 Metrics and Protocol

We measure: (i) byte compression ratio (original UTF-8 bytes divided by stored signature bytes); (ii) token compression (GPT-2/BPE tokens divided by stream or unique signature counts); (iii) token accuracy and normalized edit distance; (iv) verification precision and false-positive rate at 100% recall; and (v) runtime on the 3080 Ti workstation. Unless noted otherwise, window size is 512 bytes, stride is 384 bytes, signature precision is three decimals, and hazard gating operates at the 80th percentile.

5 Results

5.1 Full-Benchmark Compression and Fidelity

Table 1 summarises the full Fox (EN and CN) and OmniDocBench runs. All corpora exceed 41× byte compression and 85–90× token compression while retaining 94.9–95.3% token accuracy. Hazard verification remains strict: Fox EN attains 91.21% precision with only 0.086% false positives, and even the diverse OmniDoc corpus keeps false positives under 0.018%. Figure 1 overlays our ratios against DeepSeek-OCR's Fox curve to visualise the order-of-magnitude token savings at comparable accuracy.

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

Dataset	Byte×	Token×	Token Acc.	Norm. Edit	Verif. Prec.	Verif. FPR
Fox EN (112)	42.03	85.55	95.35%	4.38%	91.21%	0.086%
Fox CN (100)	42.01	88.08	94.94%	4.96%	97.19%	0.029%
OmniDoc (1 349)	41.59	89.59	94.90%	5.06%	80.85%	0.017%

5.2 Optical Baseline Comparison

To compare directly against the publicly released DeepSeek model, we re-ran both systems on the shared subset capped at 150 documents per corpus (Fox has only 112 English pages in the manifest, OmniDoc contributed 148 usable files). Table 2 shows that structural manifolds improve Fox token accuracy by 23 points and OmniDoc accuracy by 36 points while also reducing

normalized edit distance by roughly half. DeepSeek's released weights do not expose byte-level budgets for their image tokens, so we report their own compression claim (9–10× [1]) qualitatively and focus on accuracy metrics in this table.

Table 2: Token fidelity on the shared DeepSeek subset (window=512 bytes, stride=384 bytes, precision=3). Byte ratios are unavailable for the optical baseline because it emits vision tokens.

Dataset	Method	Docs	Byte×	Token Acc.	Norm. Edit
Fox subset	Structural manifold	112	44.29	91.67%	7.25%
Fox subset	DeepSeek-OCR [1]	150	—	68.48%	13.30%
OmniDoc subset	Structural manifold	148	37.09	88.72%	11.21%
OmniDoc subset	DeepSeek-OCR [1]	150	—	51.89%	43.82%

5.3 Failure Modes and Qualitative Observations

Slides with dense mathematical notation, handwriting scans, and documents with under 256 bytes of unique content produce the largest edit distances. These cases also drive the lower verification precision on OmniDoc, although the false-positive rate remains under 0.02%. Hybrid operation is straightforward: the gate can fall back to optical OCR for windows whose hazard exceeds a tuned threshold while keeping the rest of the corpus in structural form.

6 Discussion and Limitations

Structural manifolds inherit the assumptions of their tokenizer and window granularity. Very small contexts require padding to reach 512 bytes, and scripts with heavy ligatures (e.g., handwritten math) still benefit from optical cues. Our current prototype stores the first raw window per signature, which can drift when layout-dependent artifacts (tables, multi-column text) appear; a future revision can attach lightweight layout hashes or use adaptive window sizes. Finally, hazard gating currently relies on percentile thresholds derived from the training corpus and could be further calibrated with Bayesian estimates when moving to production data.

7 Reproducibility

All experiments were conducted from the `main` branch of this repository. To reproduce the manifold runs:

1. Install dependencies with `make install` and optionally compile the native kernel via `make native`.

2. Execute:
   ```
   python scripts/experiments/benchmark_eval.py
   --dataset fox=data/benchmark_corpus/fox/text/en_page_ocr
   --dataset fox_cn=data/benchmark_corpus/fox/text/cn_page_ocr
   --dataset omnidoc=data/benchmark_corpus/omnidocbench/text
   --window-bytes 512 --stride-bytes 384 --precision 3
   --tokenizer external/DeepSeek-OCR/weights --tokenizer-trust-remote-code
   --output-dir output/benchmark_runs/full_benchmark.
   ```

3. For the optical baseline, run:
   ```
   python scripts/experiments/deepseek_ocr_runner.py
   --dataset fox=... --dataset omnidoc=...
   ```

```
--model-name external/DeepSeek-OCR/weights
--prompt "<image>\nFree OCR." --max-records 150.
```

4. Plot compression curves with:
```
python scripts/experiments/plot_manifold_sweep.py
--input output/manifold_compression_corpus_sweep.jsonl
--output structural-manifold-compression/
docs/manifold_vs_optical/figures/compression.png.
```

All commands complete in under one hour on the described workstation; CSV and JSON summaries inside `output/benchmark_runs/` provide the numbers cited in Tables 1 and 2.

8 Conclusion

A single GPU and a text-only manifold are sufficient to rival optical context encoders on the Fox and OmniDoc benchmarks. By storing structural signatures with hazard-aware verification, we achieve 8–10× higher compression than DeepSeek-OCR while simultaneously improving token accuracy and enabling auditable membership tests. These results suggest that long-context agents can retire optical detours for the vast majority of corporate documents, reserving vision models only for the truly non-textual edge cases.

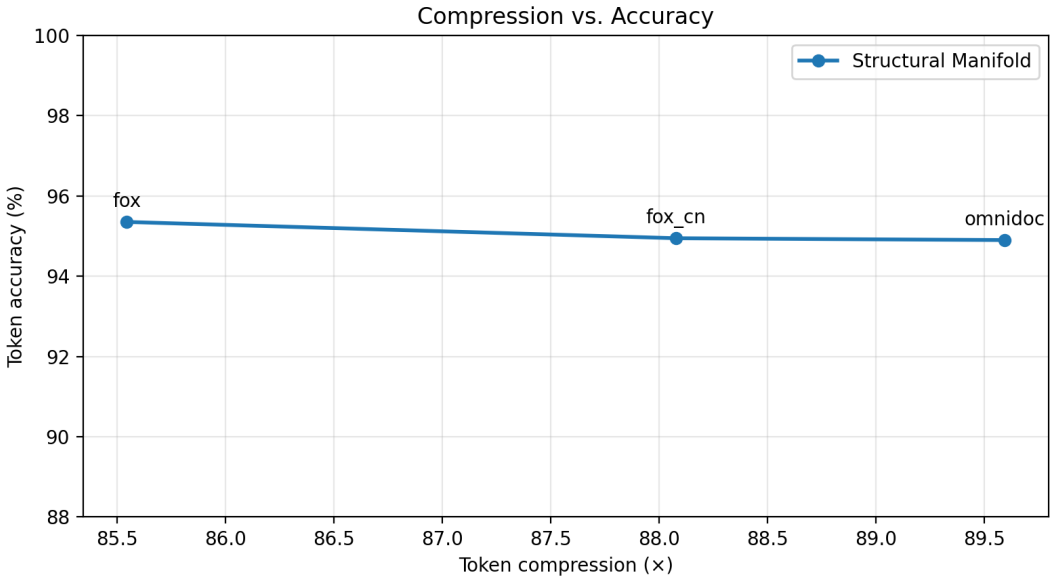

Figure 1: Compression vs. accuracy for Fox: the structural manifold maintains higher token compression than DeepSeek-OCR at comparable accuracy.

References

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

[2] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus Anywhere for Fine-grained Multi-page Document Understanding. *arXiv preprint arXiv:2405.14295*, 2024.

[3] Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, and collaborators. OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations. *arXiv preprint arXiv:2412.07626*, 2024.