Beyond Edit Distance: Reference-Free Structural-Reconstruction Fidelity for Document OCR

Ward Vandepitte — VDP STRUCT&GEO BV, Belgium

Abstract

Modern document-OCR systems report near-perfect text accuracy: on OmniDocBench, normalized edit distance and character error rate (CER) have largely saturated. This saturation reflects a blind spot rather than a solved problem: edit-distance-family metrics score the *content* a parser emits against hand-annotated references; they do not test whether that output is sufficient to **reconstruct the page**. We introduce **Visual Roundtrip Fidelity (VRTF)**, a **reference-free** measure of **structural-reconstruction fidelity**: from a parser's output alone we re-typeset the page at a single, corpus-consistent typography, place each block at its detected bounding box, and score the ink overlap of the reconstruction against the source image as a pixel confusion matrix. The metric needs no ground truth — only the source image the parser already consumed. Applied to a state-of-the-art parser (dots.ocr, 1.7B) on 2 351 scored pages of real engineering literature, character accuracy is saturated (CER 0.0045 on a transcribed subset) while structural reconstruction is only 54–73 %, with a heavy tail concentrated on the tables, charts and figures that carry the design information. Scored on a second parser, VRTF reproduces — with no annotation — the ordering the annotated benchmark assigns. Because the metric is reference-free, it also **decomposes** the gap: each capability that enriches a parser's output (e.g. formula cleanup or chart-from-data recreation) moves VRTF by a measurable amount, so ΔVRTF ranks where reconstruction fidelity is won or lost — and an expert spot-check shows such gains can be partly coarse placement, which the protocol reports rather than hides. This makes VRTF a **reference-free training signal**: a capability roadmap for a reconstruction-aware vision-language model, and a reward to train it against. We position reconstructability as the evaluation axis that remains open after edit distance saturates, and release the metric code and per-page corpus scores as a harder, annotation-free benchmark.

1. Introduction

Document parsing has been reshaped by vision-language models (VLMs) that unify layout detection and content recognition in a single network — dots.ocr (Li et al., 1.7B) [arXiv:2512.02498], MinerU [Wang et al., arXiv:2409.18839], and a rapid succession of OCR-VLM technical reports. On OmniDocBench [Ouyang et al., CVPR 2025], the benchmark on which current parsers are ranked, the leading systems sit within a point or two of each other on the text metrics — the current leader reports text edit distance 0.032 while its formula edit distance on the same pages is 0.329 [Li et al., their Table 1]. The text component, at least, is *saturating*.

Saturation of an edit-distance metric is easy to read as "OCR is solved." We argue it instead marks the limit of what that family of metrics can see. Character/word error rate (CER/WER), the tree-edit-distance similarity TEDS for tables, and character-detection matching CDM for formulas all compare the *emitted content* — a string, an

HTML table, a LaTeX expression — against a human annotation. They answer "did the parser recognize the right symbols?" They do not answer "is the parser's output enough to put those symbols back where they belong on the page?" For a great many downstream uses — retrieval over technical corpora, cross-referencing a value to the figure it depends on, automated checking of a design against a code — the second question is the operative one, and it is invisible to edit distance.

We make the second question measurable. Our metric is **reconstruction-based and reference-free**. From the parser's output we render a source-blind re-typesetting of the page — one calibrated body size, equations at that body scale, tables as plain grids — placing every block at its detected bounding-box top-left, *without ever consulting the source image to resize, shift, or align the rendering*. We then overlay the reconstruction on the binarized source and score it as a pixel confusion matrix over "ink" — the binarized foreground pixels (matched ink / missed source ink / spurious ink), at a tolerance tied to the text height so the score is comparable across scan resolutions. Because the only input besides the parser output is the source image the parser already consumed, **no annotation is required**, and the metric applies to any document the parser can ingest.

Our contributions: 1. **A reference-free structural-reconstruction-fidelity metric (VRTF)** for document OCR that tests reconstructability rather than content-match, complementing the annotation-based edit-distance family. The reconstruction protocol is *source-blind by construction*: it forbids the auto-compensation mechanisms (per-block font resizing, cross-correlation alignment, source-guided layout) with which a renderer could game the score by consulting the very image being scored. 2. **A resolution-invariant scoring tolerance** and its sensitivity analysis, making scores comparable across a corpus spanning ~110–300 dots per inch (DPI). 3. **Evidence that the axis is open**: on 2 351 scored pages of real engineering literature a state-of-the-art parser is character-saturated (CER 0.0045) yet only 54–73 % structurally faithful, with failures concentrated exactly on the structure-dense content. 4. **A decomposition into improvement avenues**: because VRTF is reference-free, toggling each output-side post-processing capability and re-scoring yields a ΔVRTF that quantifies that capability's contribution — a prioritized roadmap of what to improve, and (since no annotation is needed) a candidate **reward for reconstruction-aware VLM training**. The same experiments separate such legitimate, output-side capabilities from source-peeking rendering tricks — the auto-compensation family above — and illustrate the score inflation a peeking mechanism buys (+0.34 pp from a single line-graph mechanism), grounding why the protocol excludes them.

2. Related work

Document parsers. Modern document parsing has converged on vision-language models that emit a structured representation — per-block boxes, types, reading order, table HTML, formula LaTeX: Nougat for academic PDFs [Blecher et al., arXiv:2308.13418], GOT-OCR2.0's unified "OCR-2.0" model [Wei et al., arXiv:2409.01704], olmOCR for corpus-scale PDF linearization [Poznanski et al., arXiv:2502.18443], MinerU2.5's decoupled layout/recognition design [Niu et al., arXiv:2509.22186], and dots.ocr, a single 1.7B VLM jointly trained for layout, recognition and reading order [Li et al., arXiv:2512.02498]. We treat such a parser as the

system under test and ask whether its output suffices to *reconstruct* the page.

Reference-based metrics and their saturation. These parsers are ranked by reference-based metrics: character/word edit distance for text, tree-edit-distance similarity (TEDS) for table structure [Zhong et al., arXiv:1911.10683], and CDM for formulas [Wang et al., arXiv:2409.03643]. OmniDocBench [Ouyang et al., CVPR 2025, arXiv:2412.07626] assembles exactly this suite — edit distance (text, reading order) + TEDS + CDM — over ~1 000 pages with >100 000 manually curated annotations; olmOCR-Bench replaces string matching with 7 000 hand-authored binary unit tests, trading one form of annotation for another. Three properties matter here. First, every component is reference-based, so coverage is bounded by annotation effort, and each content type needs a *different* metric — there is no unified structural-fidelity number. Second, the text component is saturating: the current OmniDocBench leader reports text edit distance 0.032 against 0.329 for formulas on the same pages [arXiv:2512.02498, their Table 1], and our corpus shows CER 0.0045 coexisting with 54–73 % structural fidelity (§5). Third, layout is the least-measured axis: a 2006–2025 survey of OCR evaluation finds layout and typography errors rarely reported at all, and documents the mismatch between "CER good" and "usable" when structure or reading order is scrambled [Beyene & Dancy, arXiv:2603.25761].

Render-and-compare evaluation. Scoring markup by rendering it is not new: Im2Latex evaluated predicted LaTeX by compiling it and comparing the rendered image to the ground-truth render, already with a small pixel-misalignment tolerance [Deng et al., ICML 2017, arXiv:1609.04938]; CDM modernizes this for formulas by rendering *both* prediction and reference with per-character color tags and matching characters in image space [arXiv:2409.03643]; recent reconstruction systems score rebuilt artefacts with bounding-box intersection-over-union (IoU) against (synthetic) ground truth [Gonzalez, arXiv:2602.07645]. All of these require a reference to render or compare against, and the formula-level ones are binary or content-only. Our metric inherits the render-and-compare idea — including the tolerance, which we make resolution-invariant — but compares against the *source scan itself*, at document level, with no reference annotation.

Reference-free quality estimation. Where ground truth is unavailable, OCR/HTR quality has been estimated from the output text alone: lexicon ratios and (pseudo-) perplexity rank handwritten-text-recognition (HTR) models with $\rho \geq 0.8$ against CER-based rankings [Ströbel et al., LREC 2022, arXiv:2201.06170]; unsupervised signals (semantic coherence, region entropy, textual redundancy) assess OCR of historical newspaper archives where transcription ground truth will never exist [Beyene & Dancy, arXiv:2509.13236]. These are reference-free but purely *linguistic* — none measures whether the page's structure survived. VRTF occupies the empty cell in this grid: render-based like Im2Latex/CDM, reference-free like lexicon/LM estimation, and aimed at exactly the axis the survey finds unmeasured — structural reconstruction at document level.

3. Method

3.1 Honest reconstruction

Given a parser's per-block output (bounding box, type, text/LaTeX/table-HTML) and the source image, we render a reconstruction and score its ink overlap with the source. Crucially the rendering is **source-blind**: it never reads the source pixels to decide size, position, or alignment. Each block is rendered at a single corpus-consistent typography and placed at its detected bounding-box top-left. We call this protocol *honest*; the rest of the paper uses the neutral term source-blind for the mechanism and reserves "honest" for the protocol as a whole.

A reconstruction metric is only as honest as its renderer, because a renderer has many opportunities to *auto-compensate* — to consult the source image and quietly absorb the parser's errors. Three such mechanisms are tempting and must be explicitly forbidden: - **Per-block font resizing** — sizing the font to fill the detected box, so a too-tall box is silently absorbed by larger text; - **Per-line cross-correlation alignment** — shifting rendered lines onto the source ink; - **Source-guided layout** — taking line positions, pitch, and margins from the source crop. Each lets the reconstruction "snap to" the source and so hides detection error — the score rises while the parser's output is unchanged. The protocol therefore renders without any of them, and any bounding-box or layout error surfaces as a recall loss. (The same rules apply to equations — rendered at body scale, not stretched to the box, no alignment — and tables — rendered as plain grids, no source grid detection.) §6 returns to this family and measures the score inflation such a mechanism buys.

Figures. The block-level outputs of today's document parsers carry no figure content: they record that a figure exists (a box and a type), but nothing from which the chart, diagram or drawing could be rebuilt. (Chart-specialist modes exist — GOT-OCR2.0 can emit chart markup [Wei et al., arXiv:2409.01704] — but the document-parsing outputs evaluated here, like the interchange schema of §4, record only the box.) This is a genuine, large hole in current OCR output — on structure-dense documents figures carry much of the information, and §6 shows they account for a quarter of all missed ink on Deel II — and closing it (e.g. emitting chart data or vector geometry alongside the box) is one of the principal improvement avenues the metric is built to measure. The baseline protocol therefore scores figures as what they are in the output: missed. Their source ink is left unmatched — not masked out, and not recreated in the baseline — so every book is scored under the same rule and cross-book numbers stay comparable regardless of which optional extraction stages a pipeline happens to run.

3.2 Per-book typography calibration

The one global quantity the reconstruction needs is the document's typography, and measuring it is the protocol's single, deliberate exception to source-blindness. We calibrate **once per book** — the dominant single-line text-band height plus per-level heading sizes, measured over a page sample of the book itself, stored in resolution-independent units (per-mille of page height) and converted to pixels per page. The carve-out is safe because the consulted statistic is *book-global*: an aggregate over

hundreds of pages cannot absorb any individual block's detection error, which is precisely what per-block source access would do. Calibrating globally rather than per block keeps the renderer blind at the level where errors live.

3.3 Pixel-confusion scoring

The reconstruction is binarized and overlaid on the binarized source. Every source-ink pixel is a **hit** if rendered ink lies within tolerance τ (green) and a **miss** otherwise (red); rendered ink with no source within τ is a **false positive** (blue). We report

recall = G/(G+R), precision = G/(G+B), F1 = 2G/(2G+R+B),

so the reported score is exactly what the green/red/blue overlay displays (Figure 1). We define the **VRTF of a page** as this F1 — the fidelity of the parse-render roundtrip from the source image, through the parser's output, back to pixel space — and a document's VRTF as the mean over its scored pages.

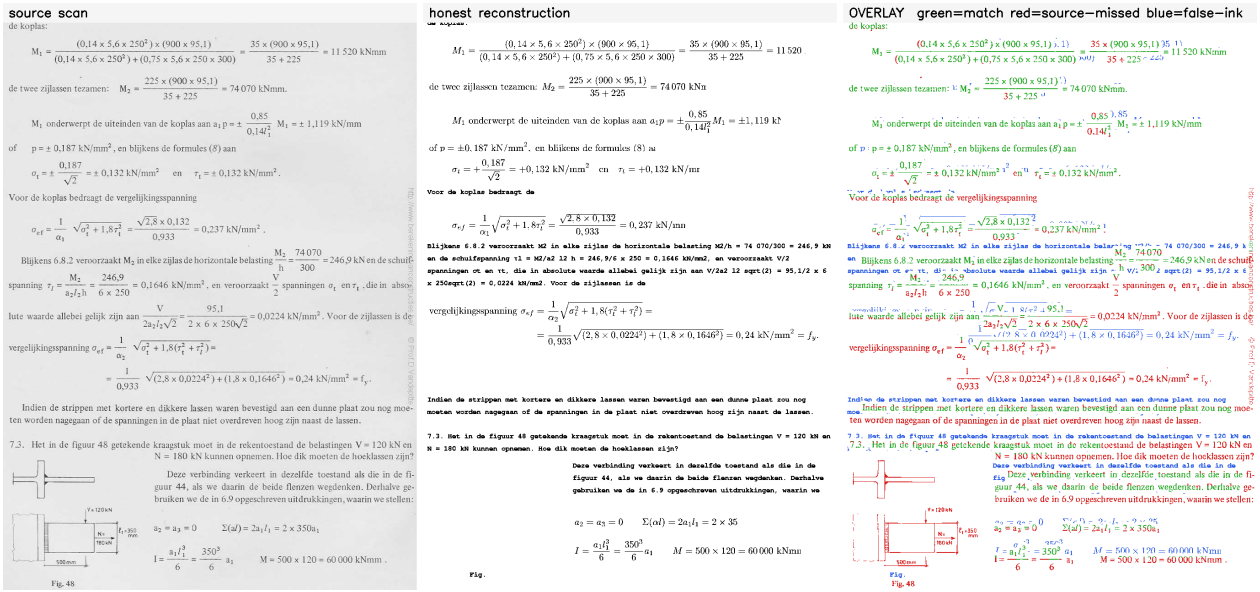

Figure 1 — the source-blind protocol on an equation-dense page (Vandepitte Deel II, p. 44; F1 0.57). Left: source scan. Middle: source-blind reconstruction at corpus-consistent typography. Right: pixel-confusion overlay — green = source ink matched within τ, red = source ink missed, blue = spurious rendered ink. Text lines reconstruct well; display equations rendered at body scale do not sit on the source's larger centered math, and the mismatch is visible as red/blue, not hidden by auto-compensation; the technical drawing (bottom left) is absent from the reconstruction and counts entirely as missed — the parser's output contains nothing to rebuild it from (§3.1).

3.4 Resolution-invariant tolerance

A tolerance is required because typeset-and-scanned text is inherently variable: the source carries print and scan variance (inking, pressure, paper distortion, digitization) and the reconstruction is set in a font that is not the document's own, so at

the pixel level the two can never coincide exactly. This variance cannot be avoided and is not the construct we want to measure — without a tolerance the metric scores glyph-shape agreement rather than structure, so τ is what filters that noise out of the placement signal. (Glyph fidelity itself is not discarded as uninteresting: a parser that captured the document's typography — font, weight, size per block — would be measurably better under this same protocol, and font capture is one of the improvement avenues of §6. The tolerance separates the two axes; it does not deny the second one.) A fixed-pixel τ, however, is a different fraction of the text size at different scan resolutions. We therefore tie τ to the calibrated body line-height, $\tau = f \cdot h\_body$, making it the same fraction of text size at any DPI. A sensitivity sweep (Table 1; $f \in \{0.25, 0.5, 1.0\}$; samples of up to 30 pages per book, the full document where shorter) shows $f = 0.25$ (a 4–8 px window) collapses into glyph-shape noise — 97–100 % of even clean text pages fall below 70 % — while $f = 1.0$ forgives nearly a full line. We adopt **$f = 0.5$** (≈ one x-height — the height of a lowercase letter), chosen a priori as the natural typographic placement scale and applied consistently across resolutions; Table 1 reports the sensitivity, and the qualitative conclusions — geotech design guides worst, tails differentiating — hold at all three values of f.

Table 1 — τ-sensitivity (mean F1 / % pages < 70 %; τ in px shows the resolution spread the body-fraction normalizes away):

Document	N	body px	f = 0.25	f = 0.5	f = 1.0
Vandepitte I	30	19	48.5 % / 97 %	76.0 % / 13 %	90.2 % / 3 %
Vandepitte II	30	16	38.0 % / 97 %	64.5 % / 57 %	82.7 % / 17 %
Vandepitte III	30	16	27.2 % / 100 %	60.4 % / 80 %	83.2 % / 10 %
CUR 198	30	16	49.7 % / 100 %	75.0 % / 20 %	85.0 % / 10 %
CUR 226	24	32	36.8 % / 100 %	59.6 % / 79 %	81.3 % / 17 %
Abbs 1988	5	26	44.8 % / 100 %	64.7 % / 80 %	77.9 % / 0 %
Bruce 1986	16	25	41.1 % / 100 %	60.5 % / 69 %	68.9 % / 38 %

Why pixel-confusion F1, and not SSIM or raw overlap? We scored the same reconstructions with two natural alternatives on a 70-page Deel II stride sample: the structural similarity index (SSIM) [Wang et al., IEEE TIP 2004], the standard perceptual image-quality measure, which compares local statistics of intensity — luminance, contrast and structure — between two images over a sliding window; and raw pixel overlap, the Jaccard index [Jaccard, 1901] of the two ink masks (intersection over union, IoU) — exact pixel agreement with no placement tolerance. Full-page SSIM *anti-correlates* with VRTF (Pearson $r = -0.50$, Spearman -0.28; range 0.34–0.88): on binarized text imagery, SSIM's local-statistics windows reward smooth regions and penalize the high-frequency structure of correctly placed type, so it measures neither placement nor content — consistent with the block-level anti-correlation that led us to drop SSIM during development. Strict pixel overlap (IoU at $\tau = 0$, the naive limit below even $f = 0.25$) preserves rank order (Spearman 0.87 vs VRTF) but collapses every page onto a glyph-mismatch floor — values 0.0005–0.13, mean 0.06 — because a reconstruction in a different font shares almost no exact pixels with the scan however correct its layout. The tolerance is what turns that compressed ordering into an interpretable, discriminating score.

4. Experimental setup

Parser. dots.ocr (1.7B VLM, rednote-hilab), the public checkpoint. Its native output is converted, without loss, into the block-level content-list representation (per-block bounding box on a normalized 0–1000 scale, type, text/LaTeX/HTML, page order) that the evaluator consumes — the de-facto interchange schema popularized by MinerU [Wang et al., arXiv:2409.18839]. The conversion is a format mapping only; no MinerU component is involved. Decoupling the evaluator from any one parser's output format is deliberate: VRTF scores whatever block-level representation a parser emits, so any system in §2 can be evaluated by the same code path. We demonstrate this with a second parser: an archival MinerU pass over Deel II (MinerU 1.x [Wang et al., arXiv:2409.18839], run on 2×-upsampled input in an earlier stage of this project), scored under the identical protocol with its own per-book calibration.

Corpus. ≈2 400 scanned pages of real engineering literature, of which 2 351 are scored (pages with no parseable blocks are excluded), in two domains: structural (Vandepitte, *Berekening van Constructies*, I/II/III; 1 943 scored pages) and geotechnical (CUR 198 retaining structures; CUR 226 sheet-pile design; Abbs 1988 and Bruce 1986 on grouted piles). Scans span ~110–300 DPI — itself a stress test for resolution-invariance.

Reference text. Character-level ground truth was transcribed for one 50-page subset (Vandepitte Deel II) to relate structural fidelity to CER; the structural metric itself needs no such reference.

Decomposition & avenue runs. The per-content-type decomposition and the capability ("avenue") runs of §6 are measured on Deel II (691 scored pages), the volume that also carries the CER ground truth; the corpus-wide numbers of §5 are unaffected by them.

5. Results

Text is saturated. On the transcribed 50-page Deel II subset, dots.ocr attains CER 0.0045 (page-mean; 99.55 % characters correct) — saturated even on these modest-quality scans. On the same 50 pages, per-page structural fidelity is *uncorrelated* with per-page CER (Pearson $r = -0.19$, $p = 0.18$; Spearman $\rho = -0.18$): per-page VRTF spans 0.24–0.92 while CER stays essentially flat. The two metrics measure different constructs.

Structure is not. At $\tau = 0.5 \cdot h_{body}$, structural-reconstruction fidelity per book (Table 2):

Table 2 — VRTF per book (source-blind protocol, full books):

Domain	Document	N	Mean VRTF	Median	Pages <70 %
Structural	Vandepitte Deel I	560	72.2 %	75.0 %	33 %
Structural	Vandepitte Deel II	691	67.5 %	70.0 %	50 %
Structural	Vandepitte Deel III	692	73.1 %	75.3 %	32 %
Geotech	CUR 198	258	68.6 %	79.4 %	32 %
Geotech	CUR 226	129	54.3 %	58.4 %	88 %
Geotech	Abbs 1988	5	65.2 %	67.0 %	60 %
Geotech	Bruce 1986	16	53.6 %	54.4 %	94 %

(Figures/images are not reconstructed and count as missed for *every* book — see §3.1 — so figure handling does not bias the cross-book comparison.)

VRTF ranks parsers. On Deel II the archival MinerU pass scores a mean VRTF of 55.4 % against dots.ocr's 68.8 % over the 645 pages both parsers produced — dots.ocr reconstructs better on 569 of those pages, MinerU on 75 (paired comparison; the MinerU pass is incomplete, lacking output for 33 of 693 pages, and ran on 2×-upsampled input, so the gap partly reflects preprocessing as well as the parser). The direction agrees with the reference-based OmniDocBench ranking of the two systems — reference-free VRTF reproduces, without any annotation, the ordering that the annotated benchmark assigns. The archival pass uses MinerU 1.x, the version in production use during the earlier project stage; the substantially redesigned MinerU2.5 [Niu et al., arXiv:2509.22186] is not evaluated here. The comparison therefore validates that VRTF recovers a known ordering reference-free rather than ranking current parsers. Scoring contemporary systems against each other — precisely where annotated benchmarks report closely tied, near-saturated numbers — is a direct application of the metric and is left to future work.

Structural reconstruction sits at 54–73 % corpus-wide against 99.6 % character accuracy — on Deel II, where both are measured, the reconstruction shortfall is a factor of ≈70 larger than the character error. It is worst on the table- and chart-dense geotechnical design guides (CUR 226 and Bruce average ~54 %, with 88–94 % of pages below 70 %).

τ-sensitivity. See §3.4 (Table 1): f = 0.25 collapses all books into the glyph-shape floor, f = 1.0 forgives nearly a full line; f = 0.5 is the discriminating mid-point.

Qualitative. Figures 2a and 2b contrast a clean text page (Deel II p19, VRTF 0.92, overlay near-all green) with a sheet-pile design page (CUR 226 p56, VRTF 0.37, the structured regions entirely red) — same engine, same τ; the difference is the density of structured content.

Analysis. Failures concentrate on equations (rendered at body scale, they do not sit on the source's larger centered display math), tables (grids and embedded sketches are not reconstructable from block-level output) and figures — these types lose the most *per page on which they occur*, while by raw pixel share the diffusely-spread text term is the largest single bucket (§6, Table 3, quantifies the split). The robust differentiator across domains is the *tail* (fraction of pages below 70 %), not the mean: struc-

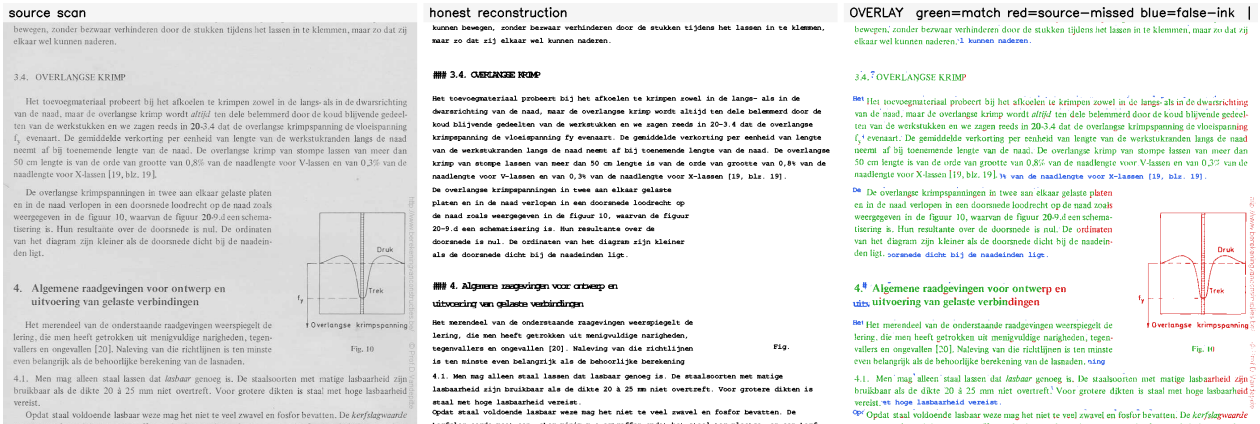

Figure 2a — a page the parser's output can rebuild (Vandepitte Deel II, p. 19; F1 0.92). Source, reconstruction, overlay: running text reconstructs almost entirely green.

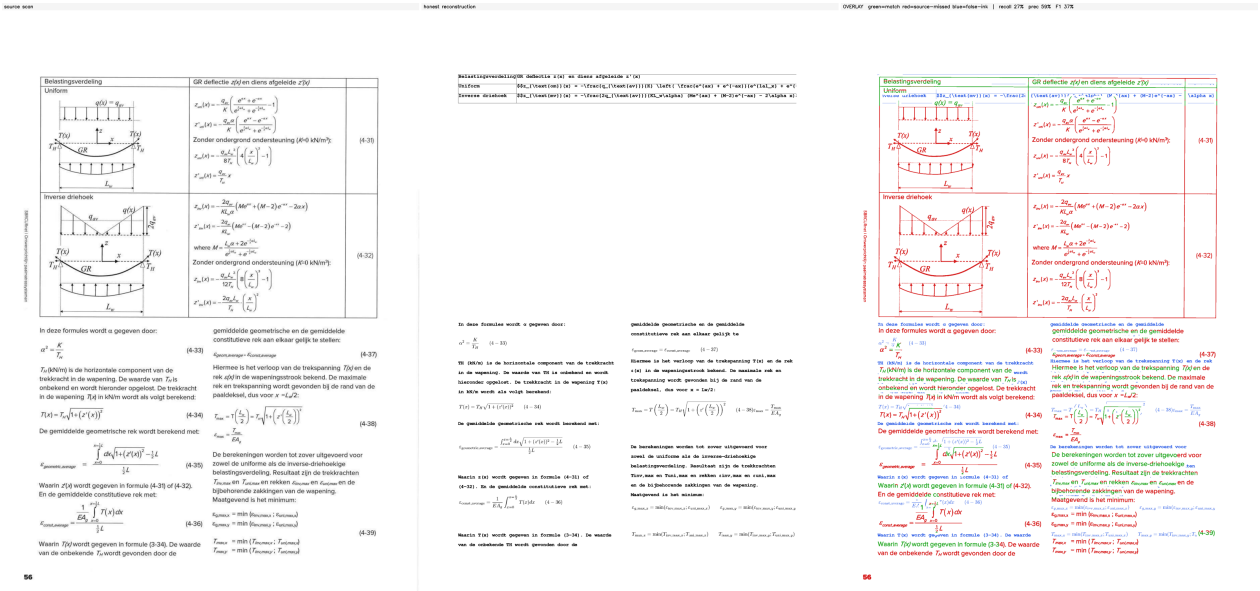

Figure 2b — a page it cannot (CUR 226, p. 56; F1 0.37). The diagram- and table-dense regions are red: the block-level output carries nothing from which they could be rebuilt.

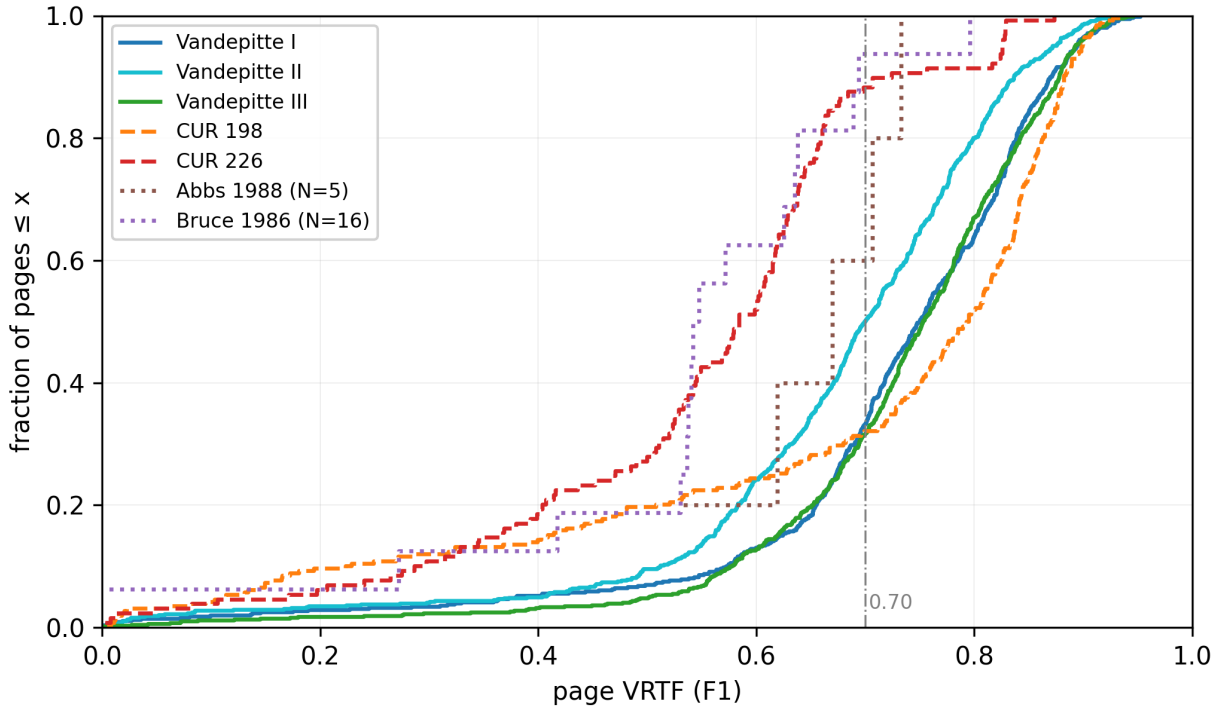

Figure 3 — per-book distribution of page VRTF (empirical CDFs). The structural volumes (solid) sit to the right of the chart- and table-dense geotech design guides (dashed/dotted), and the difference is carried by the left tail rather than the mode. The 0.70 line is the reading aid used for the "pages < 70 %" column of Table 2, not a calibrated threshold — the distributions make the tail claim without it.

tural texts 32–50 %, geotech design guides up to 94 %. Pages near zero (e.g. Bruce p. 0 at 0.006) are parse catastrophes — covers, or pages where the parser emitted almost nothing — and are left in the means, which affects the small-N books most. VRTF is lowest exactly where the engineering information lives.

6. VRTF as a roadmap for OCR improvement and VLM training

The baseline score is a floor — what a parser's *raw* output reconstructs. Its value as a research instrument is that the floor can be **decomposed**, because VRTF is reference-free and so can be recomputed under any change to the parser output at zero annotation cost. Everything in this section is measured on Deel II (§4); the corpus numbers of §5 remain exactly as reported there — figures-as-missed, no recreation, no cleanup.

6.1 Where the gap lives

Attributing every red (missed) and blue (spurious) pixel of the baseline reconstruction to the block bounding box it falls in gives a per-content-type decomposition of the structural gap (Table 3). Attribution paints bboxes in the priority order text < equation < table < figure (heading merged into text; bbox overlap is 1.2 % of painted area, bounding the misattribution this ordering can cause); source ink outside every bbox is *unparsed* — a mix of deliberately discarded furniture (reported separately, 0.4 % of the gap), undetected content (including page furniture such as running heads and page numbers), and spillover from undersized boxes, so the content-type shares are slight underestimates. The counterfactual column is first-order: F1* = 2(G+R_t) / (2(G+R_t) + (R−R_t) + (B−B_t)), i.e. the type's missed ink becomes hit and its spurious ink disappears, ignoring cross-type interactions within τ.

Table 3 — Deel II structural gap by content type. Shares are pixel-summed over the book; "ΔF1 if perfect" is the page-mean of the first-order counterfactual (and, in parentheses, the mean over only the pages where the type occurs).

Type	Share of missed ink (R)	Share of spurious ink (B)	Share of gap (R+B)	ΔF1 if reconstructed perfectly
Text (incl. headings)	58.5 %	87.7 %	68.2 %	+21.2 pp (691 pages)
Figure[1]	25.4 %	2.2 %	17.7 %	+7.9 pp (+11.3 pp on 484 pages)
Equation	6.5 %	8.6 %	7.2 %	+2.7 pp (+4.2 pp on 438 pages)
Table	0.7 %	1.5 %	1.0 %	+0.3 pp (+8.0 pp on 27 pages)
Unparsed / discarded	8.9 %	0.0 %	5.9 %	— (not an avenue)

[1] The figure category covers all image blocks, including the few photos/unclassified images (3 of 759 on Deel II) that no recreation capability can address.

Text dominates the gap by pixel share — partly genuine layout error, partly the renderer-drift term acknowledged in §7 — but it is structurally *shallow*: its red ink is

spread thinly across every page. The structured types are the opposite: figures alone are a quarter of all missed ink, and equations/tables concentrate their deficit on the pages where they occur (+4.2 pp and +8.0 pp on their own pages).

6.2 Avenue runs: capabilities as measurable deltas

Every capability that enriches the output is a candidate improvement: cleaning a malformed LaTeX expression so it compiles, recovering a table's cell structure, re-plotting a chart from extracted data, recovering within-block line geometry, capturing the document's typography (font, weight, per-block size — the axis the tolerance of §3.4 deliberately sets aside for the baseline score). Enabling such a capability and re-scoring gives a **ΔVRTF** that quantifies, in a single number and with no new annotation, how much of the structural gap that capability closes. We measure two such avenue runs (Table 4).

Table 4 — avenue runs on Deel II (ΔVRTF vs. the baseline, mean 67.5 %). ΔVRTF is the paired page-mean over the 691 common pages, with 95 % confidence interval (CI) from a paired page-level bootstrap (10 000 percentile resamples); "affected" restricts to pages whose score changed. The capability rows are demonstrated with auxiliary models — parse-time extraction by Claude Sonnet 4.6 (SVG/series output, iteratively refined against the reconstruction objective — i.e. **metric-tuned**: the extraction was optimized against the very score reported here, with no held-out split, so its ΔVRTF is an upper estimate of what a generic chart extractor would recover), formula re-recognition by UniMERNet-small [Wang et al., arXiv:2404.15254] on the source crops of the 71 equations whose raw LaTeX fails pdflatex — so they quantify the **value of the capability**, not a score of dots.ocr alone.

Avenue run	ΔVRTF	95 % CI	Affected pages	ΔVRTF on affected
Figure recreation from parsed data (source-blind render)	+2.3 pp	[+1.8, +2.7]	462	+3.4 pp
Formula compile-failure recovery (UniMERNet-small)	+0.07 pp	[+0.04, +0.10]	30	+1.6 pp
Both	+2.3 pp	[+1.9, +2.8]	473	+3.4 pp
Contrast: figure recreation, line-graph renderer peeking enabled	*+2.6 pp*	*[+2.1, +3.1]*	*462*	*+3.9 pp*

Figure recreation realizes just under a third of the first-order figure ceiling (+2.3 of +7.9 pp, 29 %); the remainder is extraction and rendering imperfection — 41 of 756 extractions fail to render at all, and 54 pages score *lower* with recreation than without (a bad re-plot adds spurious ink where missing ink merely subtracted), which the protocol reports rather than hides. Formula compile-failure recovery is real but small on this book: only 71 of 1 697 equations fail to compile, 48 are recovered, and the effect is +1.6 pp on the 30 pages it touches. The two avenues are independent to within noise (the combined run is additive). The ranking — figures first, then equations, tables mattering only on table pages — is the §6 failure analysis made quantitative.

6.3 Human spot-check (n = 20 + 20)

A review by the author — a practising structural/geotechnical engineer, hence a domain expert for this corpus, but also the metric's designer and a single rater (§7) — of 20 random figure recreations and 20 cleaned formulas qualifies what the deltas mean. The figure recreations are *not* faithful vectorizations: the drawing elements are generally present, but their position, orientation and direction are frequently wrong — load arrows inverted, moment diagrams and sub-figures mirrored about the horizontal axis (a systematic SVG y-convention error), truss web members at wrong directions and density, and occasional invented conventions (hatching absent from the source). Much of the recovered ink overlap is therefore coarse mass placement within τ rather than correct geometry. Of the 20 cleaned formulas, 12 are faithful re-recognitions, 5 contain minor errors and 3 compile but are wrong, garbled or partly hallucinated — compile success is not semantic correctness. Most formula defects trace to the *input crop*, not the re-recognizer: the crop is the parser's detected bounding box, so an undersized box truncates the formula before re-recognition — the same bbox-truncation phenomenon the decomposition surfaces as unparsed spillover, and an equally actionable fix (padded crops at parse time). The avenue rows must accordingly be read as the metric headroom these capabilities *as currently demonstrated* recover, not as evidence that faithful recreation has been achieved; the gap between +2.3 pp and the +7.9 pp ceiling is in large part exactly this unfaithfulness, and the systematic error patterns (orientation/direction) are themselves actionable roadmap items.

6.4 Improvement vs. peeking

Not every manipulation that raises a reconstruction score is an improvement. The litmus test: *could the enriched output be shipped as the parser's output and re-scored on a fresh render with no access to the source pixels?* Reading the source at **parse time** is what parsing is — a chart extractor or a formula re-recognizer consumes the source crop and emits a structured, source-detached result (data series, an SVG, compilable LaTeX) that passes the test. Source access inside the **scoring loop** does not: cross-correlating rendered lines onto the source ink, taking line positions or pitch from the source crop, detecting a plot area or table grid from the source image while rendering — these raise the score only by consulting the very image being scored and encode no transferable capability. A third category sits between the two and must be disclosed rather than forbidden: **metric-tuning at parse time**, as in our figure-extraction loop, passes the litmus test (the output is source-detached) but couples the parser's output to the evaluation signal — the spot-check's "coarse mass placement within τ" (§6.3) is its visible signature. The protocol (§3) excludes them by construction, and the roadmap admits only capabilities that pass the test. The contrast row of Table 4 measures the violation directly: letting the line-graph renderer detect the plot area, copy the source's grid ink into the render at exact pixel positions, and cross-correlate onto the source inflates ΔVRTF by a further +0.34 pp (paired per-page bootstrap 95 % CI [+0.22, +0.48], 59 pages affected) — with no change whatsoever to what the parser emits. That figure understates legacy peeking (it covers the line-graph family only, 74 blocks; it does not measure e.g. masking photo regions out of the source), but it makes the point empirically: ΔVRTF from a source-peeking mechanism measures evaluation inflation, not progress.

The avenue ranking is measured on one structural-engineering volume; the decomposition itself is cheap to recompute per book, and the figure/table shares — hence the ranking — shift with document type (the geotech design guides of §5 are far more table- and chart-dense).

6.5 Toward reconstruction-aware training

Two properties make VRTF unusual as a target. It needs no ground truth — only the source image the model already consumes — so it can be computed over arbitrarily large unlabelled corpora. And it scores *reconstructability*, a property that current next-token objectives never see. Together these suggest using VRTF (or its per-block decomposition) as a **reference-free reward**: rank candidate outputs by how well they rebuild the page, and train the VLM toward the capabilities the ΔVRTF roadmap shows are most valuable — internalizing as native behavior what is today bolted on in post-processing. The metric thus closes a loop: it diagnoses the current state, ranks the directions, and supplies the signal to move along them. The spot-check of §6.3 is also the warning label for this proposal: at a fixed τ a reward-maximizing model can learn coarse mass placement rather than correct geometry, so a VRTF reward needs a fidelity guardrail — periodic human or oracle spot-checks, or a τ-annealing curriculum that tightens the tolerance as training progresses.

7. Discussion

Why reference-free matters. Edit-distance, TEDS and CDM each require hand annotation; our metric needs only the source image, so it scales to arbitrary corpora and stays discriminative after content metrics saturate. It measures a different, complementary construct — reconstructability — and the two should be reported together.

What OCR should optimize next. A parse that records *block-level* structure (boxes + types + reading order) is not the same as one that records enough to *reconstruct* the page (within-block line positions; equation size/placement; table cell geometry). The results suggest the next gains lie in representations rich enough to reconstruct, not only label, structure.

Limitations. The reconstruction renderer assumes uniform leading (line spacing), so some text-page shortfall reflects renderer drift rather than parser error; figure and photo blocks count as missed (not reconstructable from block-level OCR); Abbs (N=5) and Bruce (N=16) are small; per-book calibration introduces a global term whose sensitivity we bound (τ tied to body size moves with it). The tolerance is also a granularity floor: at $\tau = 0.5 \cdot h\_body$, ink that lands in roughly the right place earns credit even when the geometry below τ is wrong — the §6 spot-check shows recreated figures collecting overlap from coarse mass placement while orientation and direction are incorrect, so ΔVRTF gains should be read together with such a fidelity check. Reading order is invisible to a placement metric — blocks render at their detected boxes regardless of emission order — so VRTF complements, rather than replaces, reading-order metrics. The corpus-wide numbers are for a single parser; the cross-parser comparison is demonstrated on one volume. The spot-check of §6.3 is by a single rater who is also the metric's designer. The metric scores *reconstructability of layout*, not semantic correctness — it complements, not replaces, content metrics.

Finally, external validation is planned, not yet done: a print-scan corpus of born-digital documents (printing and re-scanning a document whose true content and layout are known) will measure the print/scan noise floor — anchoring what a perfect parse can score under this protocol — and allow the reference-free score to be correlated against reference-based structural metrics; a blinded human page-rating study (30 stratified pages, already designed) will test agreement with human judgment of reconstruction quality.

8. Conclusion

Edit-distance saturation on document-OCR benchmarks marks the limit of content-match evaluation, not the end of the problem. A reference-free structural-reconstruction metric reveals a wide, open gap — 54–73 % on real engineering documents against near-perfect text — concentrated on the structured content that matters most. We offer reconstructability as the next evaluation axis, and the metric code and per-page scores as an annotation-free benchmark for it.

Acknowledgments and disclosure

The author shares a surname with Prof. D. Vandepitte, the author of the *Berekening van Constructies* volumes used in the corpus. The relation is a distant relative (at least 4 generations removed). The volumes were chosen as corpus material for their engineering content and structural density.

The metric implementation, experiments and manuscript drafting were carried out with the assistance of Claude Code (Anthropic), using the Claude Opus 4.5–4.8 and Claude Fable 5 models; all design decisions, the review of generated code and results, and the final text are the author's responsibility. This work received no external funding or sponsorship of any kind and was carried out entirely on the author's own time and resources.

Data availability

The metric implementation will be open-sourced on the author's GitHub (github.com/Ward-Vandepitte), together with the evaluation scripts and the complete per-page scores and decomposition/avenue-run data for all seven documents. The source documents themselves are copyrighted works (the Vandepitte volumes and the CUR design guides are commercial publications) and cannot be redistributed; the released per-page scores are keyed to page numbers so that results are verifiable against lawfully obtained copies.

Appendix A — Implementation and compute

Binarization. Otsu thresholding on background-normalized grayscale (morphological closing estimates the background when the page median is dark), with a fixed-threshold fallback when the result is degenerate (<2 % or >98 % ink).

Rendering. Text is set with a fixed monospace font (Nimbus Mono PS Bold) by a word-wrapping bitmap renderer at the calibrated body size; equations compile via pdflatex

(10 pt standalone, DPI tied to the body size as 7.2·body_px, 5 s timeout, falling back to text rendering on compile failure); tables render as plain grids; technical-drawing SVG rasterizes via cairosvg in the avenue runs. Tolerance uses an L2 distance transform.

Calibration. Per book, on a stride-5 page sample of the parser's own output: the dominant single-line text-band height sets the body size; heading sizes cluster by k-means with a fixed seed. Stored in per-mille of page height.

Parsing. dots.ocr public checkpoint, 50-page batches, single workstation GPU (NVIDIA RTX 2000 Ada). The archival MinerU pass predates this work's protocol and is described in §4/§5.

Compute and cost. The evaluator is CPU-only: a full book scores in roughly 10–20 minutes single-process; the complete corpus re-scores in under two hours on one workstation. The avenue-run figure extractions cost ≈ $97 of commercial API usage (Claude Sonnet 4.6, classification + extraction + refinement over the three Vandepitte volumes); UniMERNet-small re-recognition of 71 formula crops runs in minutes on the workstation GPU. Exact configurations ship with the released code.

References

- Beyene, F. S., & Dancy, C. L. (2025). Layout-aware OCR in Black digital archives: an unsupervised evaluation approach. arXiv:2509.13236.
- Beyene, F. S., & Dancy, C. L. (2026). A survey of OCR evaluation methods and metrics and the invisibility of historical documents. arXiv:2603.25761.
- Blecher, L., Cucurull, G., Scialom, T., & Stojnic, R. (2023). Nougat: neural optical understanding for academic documents. arXiv:2308.13418.
- Deng, Y., Kanervisto, A., Ling, J., & Rush, A. M. (2017). Image-to-markup generation with coarse-to-fine attention. ICML 2017. arXiv:1609.04938.
- Gonzalez, L. (2026). From dead pixels to editable slides: infographic reconstruction into native Google Slides via vision-language region understanding. WWW Companion 2026. arXiv:2602.07645.
- Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547–579.
- Li, Y., Yang, G., Liu, H., Wang, B., & Zhang, C. (2025). dots.ocr: multilingual document layout parsing in a single vision-language model. arXiv:2512.02498.
- Niu, J., Liu, Z., Gu, Z., et al. (2025). MinerU2.5: a decoupled vision-language model for efficient high-resolution document parsing. arXiv:2509.22186.
- Ouyang, L., Qu, Y., Zhou, H., et al. (2025). OmniDocBench: benchmarking diverse PDF document parsing with comprehensive annotations. CVPR 2025. arXiv:2412.07626.
- Poznanski, J., et al. (2025). olmOCR: unlocking trillions of tokens in PDFs with vision language models. arXiv:2502.18443.
- Ströbel, P. B., Clematide, S., Volk, M., Schwitter, R., Hodel, T., & Schoch, D. (2022). Evaluation of HTR models without ground truth material. LREC 2022. arXiv:2201.06170.
- Wang, B., Gu, Z., Liang, G., Xu, C., Zhang, B., Shi, B., & He, C. (2024). UniMERNet: a universal network for real-world mathematical expression recognition.

arXiv:2404.15254.
- Wang, B., Wu, F., Ouyang, L., et al. (2024). Image over text: transforming formula recognition evaluation with character detection matching (CDM). arXiv:2409.03643.
- Wang, B., Xu, C., Zhao, X., et al. (2024). MinerU: an open-source solution for precise document content extraction. arXiv:2409.18839.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600–612.
- Wei, H., et al. (2024). General OCR theory: towards OCR-2.0 via a unified end-to-end model. arXiv:2409.01704.
- Zhong, X., Shafiei Bavani, E., & Jimeno Yepes, A. (2020). Image-based table recognition: data, model, and evaluation. ECCV 2020. arXiv:1911.10683.