RENDERING AS SUPERVISION: EXACT-GEOMETRY SYNTHETIC DOCUMENTS FOR GROUNDED DOCUMENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Production document AI can read but cannot *point*: models answer questions yet emit no dependable evidence location, blocking auditable extraction and human review. On our probe, open vision–language bases at every scale (0.8B–9B) produce no parseable box at all, and the strongest frontier model is coordinate-frame-unreliable (0.157 acc@0.5 as instructed; 0.892 only under an oracle frame) — while the same models read text from *given* regions at up to 0.86 hit rate. The missing box supervision is unaffordable to annotate and unreliable to OCR precisely where degradation strikes. Our position: *when you synthesize the degradation, you own its geometry*. Our engine persists the exact, invertible correspondence field of every degraded view (882 B/page, bit-reproducible), turning 126K rendered pages into 233K zero-error grounding labels. A dosed two-stage recipe confers coordinate emission on a 4B model (0.821 mIoU, generalizing across document classes, model families, and languages), improves real-document extraction by +14.8 exact-EM, and regresses no tracked public benchmark (+5.4 F1 on replay-clean XFUND-zh). A token-level equivariance loss on the same field fails for a measured reason: the warp is sub-resolution for the vision tokenizer. Controlled studies map *how* label quality matters: i.i.d. label noise at OCR grade costs 3–11 mIoU (monotone in both domains; 3 seeds synthetic), whereas an equal-dose systematic bias is learned and fully correctable (0.526 raw, 0.841 after un-shifting — the exact-label anchor): exactness pays against the error structure OCR produces.

1 INTRODUCTION

Document understanding has been transformed by large vision–language models (VLMs) (Qwen Team, 2025; Chen et al., 2024), yet two failure modes persist in deployment. First, *grounding*: models answer document questions but cannot point to the evidence — and without an evidence box, extraction cannot be audited, highlighted, or routed for human review. On our pixel-JSON probe, open bases emit no parseable box at any scale and the strongest frontier model emits boxes in an unreliable coordinate frame (Tab. 3). Second, *acquisition robustness*: print, scan, photo, screen and screenshot channels move accuracy away from leaderboard conditions (Sec. 5.4).

Both failures trace to supervision economics: dense geometric ground truth is unaffordable to annotate at scale, OCR-derived pseudo-labels inherit estimation error precisely where degradation strikes, and synthetic pipelines — the field's label factories — discard the geometric transformation they themselves apply.

Our position is simple: **when you synthesize the degradation, you own its geometry**.

We render documents (or re-typeset real photographed substrates) and degrade them through five channels whose geometric component — perspective, rotation, non-rigid crumpling, compositing offset — is generated by us and therefore known *exactly*. The persisted field does two things: it transports every line box through every degradation with zero label error, and it doubles as a *measurement instrument* — we can inject controlled bias, variance, or smooth deformation into

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

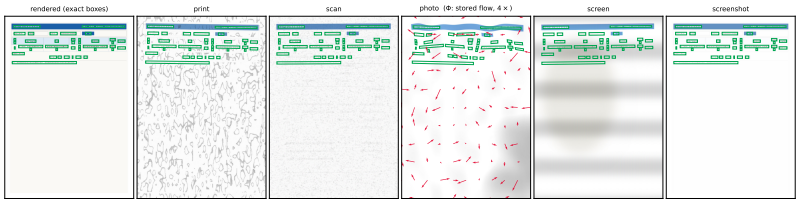


Figure 1: One rendered page, five acquisition channels, one exact geometry. Green boxes are line boxes transported by the stored field Φ (fixed-point inversion under crumpling) — zero label error in every view. Red arrows on the photo channel show the persisted non-rigid flow ($4\times$ magnified). All panels are real pipeline output; this page is rendered through all five channels for illustration (in the corpus, one traffic-weighted channel per page).

otherwise-perfect labels and measure which error structure a VLM absorbs (Sec. 5.2). The answer reshapes what such a pipeline should promise: exactness is nearly free, and its value concentrates against label *variance* — the error OCR actually produces.

Contributions:

- **Engine + corpus (C1).** A dual-path generation engine whose every output carries zero-error text, line boxes, field QA, and the exact, invertible, bit-reproducible correspondence field of each degraded view — to our knowledge the only corpus co-registering instruction-level semantics with its own degradation geometry. 126K pages, 827K refined QA, 233K grounding items across 38 document classes.
- **Recipe science (C2).** A controlled study of *how* exact synthetic supervision should be dosed: a two-stage recipe with a 500-sample replay sweet spot that confers coordinate emission (format floor 0.767 at 500 labels; 0.821 unified) without regressing any tracked public benchmark; an *anatomy of label-error structure* (variance: monotone dose–response in both domains, 3 seeds synthetic; equal-dose bias: fully correctable, 0.841 un-shifted vs. 0.526 raw; smooth warps: $\approx$0); the finding that forgetting hits the *output interface* before knowledge; and a negative result with a measured mechanism — token-level equivariance through the same field hurts because the warp is sub-resolution for the vision tokenizer.
- **Benchmark + scaling analysis (C3).** A five-channel degradation benchmark over public test sets — at production-calibrated doses, re-photographing is the only channel that materially degrades a strong base — and a model-scale study showing real-domain and grounding gains scale monotonically 0.8B$\rightarrow$9B while no further degraded-domain gain is observed beyond 4B (within seed noise).

2 Related Work

Document understanding models. OCR-free decoders read documents end-to-end (Kim et al., 2021; Lee et al., 2022); layout-aware backbones inject explicit structure (Huang et al., 2022; Wang et al., 2023a); KOSMOS-2.5 (Lv et al., 2023) couples spatially-aware text with generation. General VLMs (Wang et al., 2024a; Qwen Team, 2025; Chen et al., 2024) dominate benchmarks such as OCRBench (Liu et al., 2023); grounding-capable VLMs acquire coordinate emission through dedicated referring/grounding data — (Peng et al., 2024; Chen et al., 2023; You et al., 2024; Wang et al., 2024b; Bai et al., 2023; Cheng et al., 2024) — a conclusion our zero-shot probes confirm for documents: coordinate-conditioned reading is present, emission is absent until grounding data supplies it. Document specialists improve structure awareness (Hu et al., 2024; Liu et al., 2024; Baechler et al., 2024); LMDX (Perot et al., 2023) prompts LLMs for grounded extraction but inherits estimated coordinates. We change what supervises the reader rather than the reader itself — and where its grounding data comes from, at zero annotation cost.

Synthetic document data. Synthetic documents have long served as label factories (Gupta et al., 2016; Yim et al., 2021; Xia et al., 2024): SynthDoG feeds OCR-free readers (Kim et al., 2021), DocUNet built 100K warped pages for unwarping (Ma et al., 2018), Augraphy/ShabbyPages and

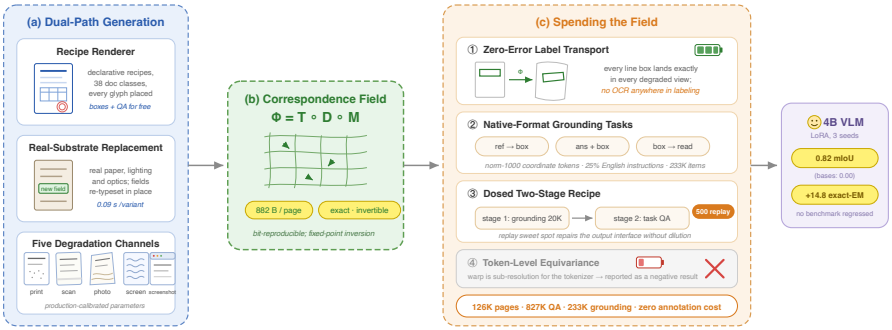


Figure 2: Architecture of the generation engine. The dual-path engine degrades every rendered page through five production-calibrated channels and *persists* the exact correspondence field Φ of each view. The field is spent three ways — zero-error label transport, native-format grounding tasks, and a dosed two-stage recipe; the one use that mechanically requires a dense field (token equivariance) is reported as a negative result.

Genalog make degradation *controllable* (Groleau et al., 2023a;b; Microsoft, 2021), and DocILE adds box supervision at scale (Šimsa et al., 2023). Across all of these, the applied geometric operators are appearance-level, parameter-logged but never exported as a dense correspondence, or emitted without semantic labels; none co-registers the degradation geometry itself with instruction-level supervision. That co-registration is what our engine adds.

Geometric self-supervision and rectification. Self-generated warps as exact supervision is an established recipe we inherit: known homographies for interest points (DeTone et al., 2017), known non-rigid warps for dense correspondence (Truong et al., 2020; 2021), stored flow/3D for dewarping rectifiers (Ma et al., 2018; Das et al., 2019; Feng et al., 2021; 2022; Verhoeven et al., 2023). Our delta is where the signal lands: the stored field transports *semantic* supervision into every degraded frame of the *VLM's* training data, under acquisition composites rather than paper curl alone.

Consistency and equivariance. Consistency regularization enforces output invariance under augmentation (Sohn et al., 2020a); semi-supervised detection transports pseudo-boxes through known augmentations (Sohn et al., 2020b; Liu et al., 2021; Xu et al., 2021; Wang et al., 2023b). Our label transport is mechanically STAC's operation with a constructively exact rather than teacher-estimated source — a difference our error anatomy prices at 3–11 mIoU for i.i.d. variance (bias is correctable). Feature-level equivariance under known warps (Thewlis et al., 2017; Dangovski et al., 2022; Devillers & Lefort, 2023) succeeds at pixel resolution from scratch (Wang et al., 2021; Xie et al., 2021) (stop-gradient per SimSiam (Chen & He, 2021)); we compare output invariance, known-field equivariance, and exactly-transported labels, and the third wins while the second fails measurably: our warp is sub-resolution for a pretrained VLM's token lattice (Sec. 5.5). Finally, fine-tuning erodes pretrained ability (Zhai et al., 2023) and replayed data mitigates (Yang et al., 2024); we map the replay dose–response curve explicitly.

3 Method

3.1 Dual-Path Generation Engine

Recipe rendering. Declarative recipes compile structural primitives (key–value blocks, tables, clauses, seals, multi-page layouts) into full pages; field values come from consistency-constrained generators (name ↔ gender ↔ ID-digit ↔ birth date agree). Every glyph is *placed*, not recognized: text, line boxes and field QA are byproducts with zero annotation error.

In-place replacement on real substrates. For classes whose realism hinges on material appearance (security patterns, lamination glare, wear), we re-typeset field content onto photographs of real documents: substrate, lighting and optics are real; replaced fields carry pixel-exact masks (0.09 s per variant after a one-time 0.39 s substrate bake).

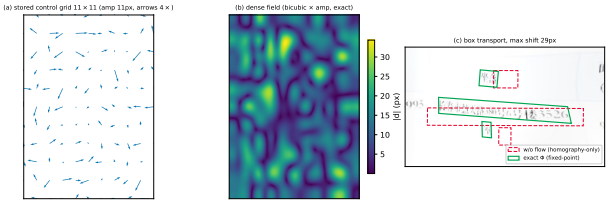


Figure 3: The stored crumple field and why it matters for labels. (a) persisted $s \times s$ control grid (unit-Gaussian $\times$ amplitude); (b) bicubic reconstruction — the exact dense field used by the pipeline; (c) a line box transported by the homography alone (red) versus the full field via fixed-point inversion (green): crumpling contributes up to $\sim$29 px of label shift here — error OCR-based labeling would silently absorb.

Five degradation channels. Print, scan, photo, screen, screenshot: compositions of appearance operators and geometric operators, with channel weights following production traffic.

3.2 The Exact Correspondence Field

For each degraded view $\tilde{x} = \mathcal{D}(x)$ we persist

$$\Phi \;=\; T_{\text{off}} \circ D_{\text{crumple}} \circ M, \tag{1}$$

where M is the composed perspective–rotation homography, D_{crumple} a non-rigid displacement field stored as an $s \times s$ control-point grid ($s \in [6, 12]$, bicubic upsampling, amplitude-scaled), and T_{off} the compositing translation (compositing never rescales). The crumple amplitude is 0.4–1.2% of the page's short side (5–20 px at our resolutions; local shifts reach $\sim$30 px). Storage: 882 B/page on average.

Three properties separate Φ from anything obtainable on real data or from homography-based self-supervision: **(i) dense and non-rigid** — crumpling maps every point, which homography-only pipelines cannot express (expressiveness, not measured performance gain: Sec. 5.2); **(ii) exact and invertible** — since `remap` is a reverse lookup ($q + d(q) = p$), forward correspondences follow from fixed-point iteration (3 iterations to sub-pixel at these amplitudes); we use this to correct all line boxes under crumpling (Fig. 3c), a 30 px-tail systematic label debt that we correct by construction (whether training could have absorbed it is itself measured — Sec. 5.2); **(iii) bit-exact reproducible** — the full chain is seed-deterministic (third-party libraries' global RNGs pinned; the single non-exportable pixel-displacing operator was removed and its semantics re-implemented parametrically).

3.3 Supervision Through the Field

Given paired views $(x, \tilde{x}, \Phi)$, the field admits a family of supervision signals; we study all three and report which ones earn their keep.

(a) Exact label transport (the workhorse). Every line box b^* of the canonical frame yields $\Phi(b^*)$ in the degraded frame with zero label error — standard LM cross-entropy on both views with *constructively correct* targets. This requires Φ (an OCR or homography pipeline cannot produce degraded-frame boxes at this fidelity) yet adds no new loss machinery.

(b) Grounding task forms. Exact line boxes yield three free task forms — refer$\rightarrow$box, box$\rightarrow$read, answer-with-evidence-box — in the model's native norm-1000 coordinate tokens; 25% of grounding items use English instructions to decouple capability from instruction language.

(c) Token-level equivariance (tested; negative). With visual tokens $z(\cdot)$, LayerNorm'd features, and bilinear grid-sampling GS over the token lattice,

$$\mathcal{L}_{\text{tok}} = 1 - \cos\big(\text{GS}_{\Phi^{-1}}\big(z(\tilde{x})\big),\, \text{sg}[z(x)]\big). \tag{2}$$

This is the operation that mechanically *requires* a known dense field. Across a $\lambda \times$ layer sweep it never helps and often hurts badly; we keep it as a documented negative result with its dose–damage landscape (Sec. 5.5) rather than as a contribution.

Document classes	38 (+ special renderers)
Documents / pages	95,200 / 125,826
Refined QA (tagged)	827,403
unanswerable negatives	11,335
multi-page questions	43,668
VQA train/val/test	343,009 / 18,760 / 18,453
Grounding train/val/test	209,993 / 11,505 / 11,243
Degraded views (one channel per page, traffic-weighted)	125,826
Correspondence fields	100% coverage, 882 B/page
Crumple-corrected boxes	21,045 pages

Table 1: Corpus statistics. Splits are by document-id hash; a page never crosses splits.

Dataset	semantic labels	degradation geometry	co-registered
SynthText (800K) (Gupta et al., 2016)	text + word/char boxes	✗	—
SynthDoG (2M) (Kim et al., 2021)	text sequence only	✗	—
Augraphy/ShabbyPages (Groleau et al., 2023a;b)	text (source-level)	params logged, no field	✗
DocGenome (6.8M pg) (Xia et al., 2024)	structure/boxes/QA	✗ (clean renders)	—
DocUNet/Doc3D (∼100K) (Ma et al., 2018; Das et al., 2019)	none	dense flow / 3D	✗
DocILE (∼940K) (Šimsa et al., 2023)	KIE boxes + text	✗ (appearance-level)	—
Ours (126K pg)	QA + boxes + class	Φ, 882 B/pg, bit-exact	✓

Table 2: The empty quadrant. Large synthetic corpora carry semantic labels but discard the geometry they apply; dewarping corpora persist geometry but carry no semantics. To our knowledge, ours is the only corpus in which a persisted degradation field and instruction-level semantic labels (QA, referring boxes) share one coordinate frame — a property required for exactness by construction; whether *training* needs the dense non-rigid component is itself measured (Tab. 5: homography-only .840 vs. full-Φ .838).

(d) Forgetting control. Two-stage training with an explicit replay dose: stage 1 trains grounding on 20K field-supervised items; stage 2 trains task QA with 500 grounding replays (sweet spot; Tab. 5) plus 3K public-data replays.

4 Data and Benchmark

Exact by construction: text, boxes, QA, class labels and Φ carry no OCR or human annotation anywhere in the labeling path. Questions whose answers are not visibly printed are tagged and excluded from training (447K filtered): answering the invisible teaches hallucination. **Outside the claim:** appearance parity with real photographs for recipe-rendered classes, and coverage of non-document imagery, are neither claimed nor evaluated.

Degradation benchmark (C3). The five-channel stack applied to public test sets yields per-channel drop curves (Sec. 5.4); generation parameters follow production traffic, and at these parameters only the re-photograph channel materially degrades a strong base — the benchmark measures *realistic* rather than adversarial acquisition.

Internal real-domain validation. 100K frozen real QA (7,849 source-document types); we evaluate a domain-matched subset (740 items) and an off-domain degraded control (500), reporting aggregate metrics only — providing sim2real evidence without releasing real financial documents.

5 Experiments

5.1 Setup

Base models: Qwen3.5 {0.8B, 2B, 4B, 9B} (native norm-1000 coordinate vocabulary); main line 4B. Cross-family rivals fine-tuned on identical data: Qwen2.5-VL-7B, Qwen3-VL-4B, each under its own coordinate convention and parser. Frontier zero-shot references: Kimi-K3 and Qwen3.8-27B (pixel-JSON, 300-item stratified probe). Main recipe (3 seeds, mean±std): LoRA $r{=}32/\alpha{=}64$, lr

Model	ref2box mIoU / acc@.5	ans+box mIoU / acc@.5	box2text hit
Frontier reference (zero-shot, pixel-JSON, 300-item stratified probe)			
Qwen3.8-27B	.016 / .012	.017 / .000	.000
Kimi-K3 (as instructed)	.142 / .157	.105 / .119	.435
Kimi-K3 (norm-1000 asked)	.740 / .434	.557 / .284	—
Kimi-K3 (heuristic frame†)	.621 / .795	.315 / .358	—
Kimi-K3 (oracle frame†)	.708 / .892	.391 / .450	—
Open bases, zero-shot (parse-failure rate in parentheses)			
Qwen3.5-0.8B/2B/4B	.000 (100%)	.000 (100%)	.70/.75/.45
Qwen3.5-9B	.000 (99%)	—	.792
Qwen2.5-VL-7B	.000 (points only‡)	.000	.623
Qwen3-VL-4B	.000 (verbal, no coords)	.000	.856
Fine-tuned on our corpus			
Ours 4B (grounding-only, 20K)	**.838±.004** / .968	.773±.001 / .91	.89
Ours 4B (unified, final)	.821±.004 / .970±.002	.765±.007 / .914±.008	.889±.006
Qwen3-VL-4B (same data)	.841 / —	—	—
Qwen2.5-VL-7B (same data)	.735 / .881	.664 / .792	—

Table 3: Grounding on the synthetic-domain test split. Zero-shot bases at *every* scale fail to emit parseable boxes ($\geq 99\%$ parse failure) while reading text from given boxes at up to 0.86 hit — localization ability exists but is format-locked. †K3 parses 83/83 but mixes coordinate conventions at temperature 0 (62/83 answers arrive norm-1000 despite the pixel instruction); "heuristic frame" applies a deployable no-ground-truth rule (any coordinate $>1000 \Rightarrow$ pixel, else norm-1000); "oracle frame" takes the better reading per item, an upper bound (App. H). ‡points, never boxes (102/278 queries; 0.259 point-in-box hit). Rival rows use each family's native coordinate convention. mIoU averages over parseable predictions while acc@0.5 counts unparseable as miss (App. A) — hence rows where the two move in opposite directions; "—" = not evaluated. The non-monotone zero-shot box2text sequence (.70/.75/.45/.79) reproduces across metrics; we report it as observed.

2×10^{-5} (probe-selected), effective batch 32; stage 1 = 20K grounding items (25% EN instructions); stage 2 = task QA + 500 grounding replay + 3K public replay (27% of stage 2; *composition: DocVQA-train 2K + InfographicVQA-train 1K — in-domain for two of the five tracked benchmarks, disclosed accordingly in Tab. 4*). Internal metrics: exact EM on matched740 and degraded500 (real); grounding mIoU/acc@0.5 on a held-out 1,000-item split (278 ref2box / 361 ans+box / 361 box2text). Public benchmarks are double-reported: lenient EM (reference matches exactly or is contained in the normalized prediction; direction reference $\subseteq$ prediction) and, for DocVQA/InfographicVQA, official ANLS; XFUND-zh uses value-level F1; OmniDocBench is excluded (parsing benchmark, official metric not yet run). Signs agree across all metric pairs (Tab. 4).

5.2 Grounding: Unlocking and Sharpening a Format-Locked Interface

No zero-shot base at any tested scale produces parseable boxes under our protocol — including grounding-advertised Qwen2.5-VL-7B (*points*, never boxes; 0.259 point-in-box hit) and Qwen3-VL-4B (verbal descriptions only) — while all read text from *given* boxes (Tab. 3): reading exists, emission is absent. The headline jump decomposes into two parts: formatting is nearly free (500 exact-label samples $\times$ 3 epochs already reach 0.767 mIoU, the *format-unlocking floor*, Tab. 5), and the corpus's net contribution is what the floor cannot buy — pushing precision to 0.838 ± 0.004 (grounding-only) and 0.821 (unified final — the headline, as the only arm carrying QA and boxes in one deployable model), generalizing to five held-out classes (0.783), transferring across families (Qwen3-VL-4B 0.841, Qwen2.5-VL-7B 0.735 under its own coordinate convention) and across a language boundary (below); acc@0.75 is 0.829 (Tab. 8). The σ-study shows the floor itself presupposes label exactness.

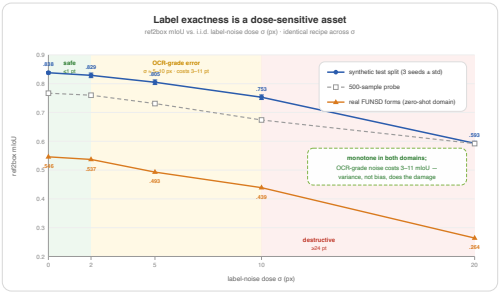


Figure 4: Exactness is a dose-sensitive asset: corrupting transported box labels with Gaussian noise σ degrades grounding monotonically in both the synthetic domain and on real FUNSD forms; OCR-grade error ($\sigma \approx 5$–10 px) already costs 3–11 mIoU points. Synthetic series: 3 seeds, mean±std; FUNSD and 500-sample series: seed 42. Identical recipe across σ.

Inference-time OCR baseline. The industrial default — OCR at query time + string matching — reaches 0.820 acc@0.5 (vs. ours 0.970; its 9.7% no-prediction queries already counted as misses) but collapses at tighter thresholds (0.536 @0.75 vs. ours 0.829: line boxes miss field boundaries) and cannot answer-and-ground (App. E).

What is exactness worth? An anatomy of label-error structure. Write a transported label as $\tilde{b} = \Phi(b^*) + \varepsilon$ and consider three error structures — constant offset (bias), i.i.d. fluctuation (variance), spatially smooth field — each probed by a dose-matched control (per-coordinate RMS); only one costs capability — bias is learnable because the target family is closed under translation, while i.i.d. noise admits no deterministic correction. *Variance*: Gaussian σ-corruption of transported boxes degrades grounding monotonically in two domains (Fig. 4) — OCR-grade i.i.d. error is not absorbable. *Bias*: a dose-matched control shifting every training box by a constant $+8$ px scores 0.526 raw but **0.841 after un-shifting its predictions** — equal to the exact-label anchor (0.838 ± 0.004): a systematic offset is learned *losslessly*, while the matched i.i.d. arm (σ=8, 0.792) pays the only irrecoverable cost. *Smooth self-generated warps*: dropping the crumple component of Φ during transport (16.9% of items, median 8.7 px) matches full-Φ training within noise (0.840 vs. 0.838) and costs ≈ 0.7 mIoU on the crumpled subset. Finally, a *real-OCR* arm grounds the abstraction: PaddleOCR (PP-OCRv5) labels on the same images retain 81% of samples with heavy-tailed error (median 3.8 px, p95 180 px) and train to 0.751 ± 0.002 (3 seeds), coinciding with the σ=10 point (0.753 ± 0.008, Tab. 6); a second system (PP-OCRv4 via RapidOCR; median 3.0 px, 79% kept) lands at 0.782 — a smaller dose, ordered with its smaller median error (version number $\neq$ label quality on this corpus). On real-document extraction the PP-OCRv5 arm reaches 0.577 exact-EM vs. 0.624 for its exact-label counterpart (base 0.549): OCR labels forfeit 62% of the real-domain gain. At both doses the Gaussian law prices real OCR labeling accurately rather than as a worst case; error shapes differ and sample loss is not separated from noise cost, so these remain empirical correspondences. Exact transport remains cheap insurance against a 30 px-tail label debt (Fig. 3c); the value of exactness concentrates against label *variance*, not against systematic or smooth deviation, which training absorbs.

Real-domain transfer across a language boundary. Trained purely on Chinese synthetic documents, the model grounds real English FUNSD forms: 0.553 ± 0.006 mIoU (grounding-only, Chinese prompts, 3 seeds) and 0.457 ± 0.007 under English prompts with zero parse failures ($137/150 \to 0/150$ via the 25% EN-instruction variant); an intermediate EN-variant recipe scores $\approx$0.51, splitting the gap into $\sim$5 points recipe cost and the rest prompt language. No FUNSD or English document image appears in training.

5.3 Real-Document QA: Interface Alignment and Net Accuracy

The final recipe reaches 0.697 ± 0.011 exact-EM on the domain-matched real set (base 0.549) and 0.488 ± 0.005 on the degraded control (base 0.282). We dissect the +14.8 gain with dual metrics because exact-EM conflates two effects. Under a *containment* metric the base already scores 0.824 (its verbose answers usually contain the right span) vs. 0.736 fine-tuned; a Qwen3.6-35B-A3B judge

Benchmark	Metric	Base	Synth-only	Final (3 seeds)
DocVQA (5,349)	lenient EM	.918	.908	**.936±.001**
	ANLS (official)	.931	.928	**.947±.002**
InfographicVQA (1,000)	lenient EM	.758	.720	**.805±.003**
	ANLS (official)	.776	.749	**.812±.002**
DocVQA-degraded (1,000)	lenient EM	.890	.883	**.912±.002**
OCRBenchV2 (1,000)[§]	lenient EM	.379	.370	.385±.003
XFUND-zh KIE	value F1	.640	.621	**.694±.012**
matched740 (real)	exact EM	.549	.676±.012	**.697±.011**
degraded500 (real)	exact EM	.282	**.492±.005**	.488±.005
ref2box	mIoU	.000	.818	**.821**

Table 4: Capability at bounded, measured cost; *signs agree across lenient and official metrics*. "Synth-only" = two-stage + 500 replay; "Final" adds 3K public replay from the train splits of DocVQA (2K) and InfographicVQA (1K) — in-domain for those two rows and the derived DocVQA-degraded row: their gains measure *replay repair*, not synthetic transfer; replay-clean evidence is XFUND-zh (+5.4 F1) and OCRBenchV2 (within noise). Mean output length is 32–33 characters for base and final alike, ruling out containment verbosity artifacts. [§]In-house reimplementation of the metric; the official OCRBenchV2 script is pending.

rates base 88/93 and fine-tuned 82/90 semantically correct on a 100-item subsample (cross-checked by an independent Kimi-K3 judge: 97.8% agreement, Cohen's κ=0.82), and a McNemar exact test on the 85 doubly-judged items finds no significant difference (discordant 1 vs. 4, p=0.375). The matched set is thus close to semantically saturated for the base: fine-tuning primarily converts latent, verbosely-expressed knowledge into the short exact extractions deployment requires (exact-EM $0.549 \to 0.697$; judged semantic correctness $88/93 \to 82/90$), at a small semantic cost within judge noise. The recipe buys an *extraction interface* and bounded net-accuracy movement, not a wholesale reading upgrade. The same dissection holds on the degraded control: its $+20.6$ exact-EM gain (3-seed mean) moves containment by only $+1.4$ (seed 42: $0.602 \to 0.616$) — interface alignment dominates on both real sets. The grounding-free ceiling of this effect is 0.738 ± 0.002 (pure task QA, Tab. 5); grounding-bearing recipes trade up to 4.1 EM for the capability.

Transfer is *dose-sensitive*: a light-touch 7.5K subset reaches 0.738 ± 0.002, while scaling the same mixture to 627K degrades real-domain EM monotonically ($0.738 \to 0.578 \to 0.477$) — heavy exposure overwrites what light exposure releases (steps/volume entangled; confound flagged). Single-stage mixing collects a dilution tax (-5.7 EM at just 3K grounding items; probe-scale arm, not in Tab. 5); two stages avoid it, but *bare* sequential training destroys the box format (275/278 parse failures) while leaving QA intact (0.708 EM) — forgetting hits the *output interface* before it hits knowledge. A 500-sample replay repairs the interface (none: format collapse; 1500: -3.1 EM re-dilution). The same recipe lifts Qwen2.5-VL-7B from 0.589 to 0.712: the effect is not family-specific.

5.4 Public Benchmarks and Degradation (C3)

Per-channel degradation. The degraded set assigns one channel per item at production traffic weights; by channel slice (98–299 items each, hence sampling noise), base / synth-only / final: print $.923/.909/.932$, scan $.919/.907/.942$, screen $.872/.872/.896$, screenshot $.959/.969/.959$, **photo** $.826/.819/.856$. Photo (perspective + lighting) is the only channel far below clean (.826 vs. .918); synth-only training does not close it, the final recipe recovers $+3.0$. Traffic-weighted totals reproduce Tab. 4 ($.890/.883/.910$; slices are seed 42, so the final total sits within 1σ of the 3-seed $.912 \pm .002$). Leave-one-channel-out *grounding* training shows the channels cover for each other (≤ 3 ref2box mIoU cost on the excluded channel).

Scale and steps. Across Qwen3.5 0.8B/2B/4B/9B under the identical synth-only recipe (Tab. 4, second column; the final-recipe 4B row is higher), real-domain EM scales monotonically ($.549/.580/.676/.730$)[1] and grounding likewise ($.768/.786/.820/.842$ mIoU; single-seed sweeps,

[1]The .549 tie with the 4B base is coincidence (verified on raw predictions).

Arm	matched EM	ref2box mIoU (fails)
Final recipe (3 seeds)	**.697±.011**	**.821** (0)
w/o public replay	.676±.012	.818 (0)
w/o any replay (bare 2-stage)	.708	format collapse (275/278)
replay 1500 (over-dose)	.661	.830 (0)
single-stage mix (+20K gr)	.658	.833 (0)
pure task QA (no grounding)	.738±.002	no box format
homography-only transport (3 seeds)	—	.840±.001 (0)
OCR pseudo-labels (PaddleOCR, 81% kept, 3 seeds)	.577	.751±.002 (0)
format probe: 500 exact labels	—	.767 / .928
class-holdout (5 unseen classes)	—	.783 / .932
blind image control	—	.003–.006
shuffled image control	—	.007–.011

Table 5: Recipe ablations and memorization controls. Pure-QA = grounding-free EM ceiling; grounding-bearing arms trade EM for capability. Blind/shuffled controls confirm image reading; class-holdout drops 5 points on unseen classes.

3-seed 4B mean .818), while the degraded-domain gain saturates at 4B ($.422/.470/.492/.490$): robustness to acquisition noise is capacity-limited earlier than in-domain reading. Grounding also grows with steps ($234/469/938 \rightarrow .807/.819/.834$), a quantity axis orthogonal to the σ quality axis (Fig. 4).

5.5 A Negative Result with a Mechanism: Token-Level Equivariance

The one loss that mechanically requires a dense field is the one that fails — for a measurable reason. On 7.5K paired views, $\mathcal{L}_{\text{tok}}$ at $\lambda = 0.5$ collapses real-domain EM $0.727 \rightarrow 0.527$; a $\lambda \times \text{layer}$ sweep ($\lambda \in \{0.05, 0.1\}$, layers 4–24) sits uniformly below the $\lambda = 0$ anchor (0.689–0.611, monotone in dose and depth) and replicates in the degraded domain (0.486 vs. 0.466/0.438). **Mechanism.** The stored crumple displacement (median 8.7 px, p95 21.9 px) sits below the vision tokenizer's 28 px effective stride (14 px patches, 2×2 merge): the warp moves content by a *fraction of one token* at the resolution where $\mathcal{L}_{\text{tok}}$ operates, so $\text{GS}_{\Phi^{-1}}$ is near-identity and the loss degenerates into pure collapse pressure toward $\text{sg}[z(x)]$, explaining the monotone dose–damage landscape. The design rule: *a deformation that is sub-resolution for the representation yields no equivariance signal — spend the field in label space*. The paired-view *data* wins (0.727 is the strongest arm within this probe-scale sweep); the equivariance *loss* only damages.

Ablations and controls. Table 5 adds the memorization controls: blind/shuffled images collapse grounding to ≤ 0.011; five held-out classes lose only 5 mIoU — the capability generalizes. The 500-sample format probe reaches 0.767 mIoU, yet its σ-sweep is as steep as the full-scale one (Tab. 6): sample efficiency *presupposes* label exactness.

6 Conclusion and Limitations

Rendering owns its geometry; persisting it exactly turns synthetic documents into dense supervisors: a dosed recipe unlocks coordinate emission no tested base expresses (0.821 mIoU), converts latent knowledge into deployable exact extraction (+14.8 exact-EM), regresses no tracked benchmark, and the error-structure controls price exactness in label variance alone. **Limitations:** Chinese-heavy corpus (English transfer only via FUNSD); internal sets unreleasable; acc@0.9 ≈0.26 caps redaction; photo remains the largest gap despite $+3.0$.

Reproducibility Statement

All training and evaluation numbers in this paper trace to a single audited ledger; every arm specifies base model, LoRA configuration, learning rate, effective batch, seeds (42/43/44), and step counts (Sec. 5.1). The generation engine is seed-deterministic: third-party RNGs are pinned and the single non-exportable pixel-displacing operator was re-implemented parametrically *before* the corpus used in all experiments was generated; determinism is verified by re-running the pipeline under a fixed environment. We will release the engine, recipe DSL, the synthetic corpus with the persisted field Φ (schema documented), the degraded public test sets, evaluation scripts including all prompts and

parsers, and the sampling manifests of every public-benchmark subset. The internal real-document sets cannot be released for compliance reasons; we document their selection protocol and provide per-item audit tuples (question_family, source, answer_length_bucket, per-arm correctness, judge verdicts; App. A.1) so that all statistical claims can be re-derived, and FUNSD plus the public benchmarks serve as the external reproduction path.

Ethics Statement

The corpus synthesizes documents with *fictitious* field values from consistency-constrained generators; no real personal data enters the labeling path. For the in-place replacement path, real document substrates are used only after removing or replacing all personally identifiable content, and face regions come from a licensed synthetic-face source. The internal real-document evaluation sets contain production data and are therefore not released; only aggregate metrics are reported. The primary LLM judge that audits 100 of these documents runs on internal infrastructure; the second judge (Kimi-K3) is accessed through an organization-controlled gateway and receives only question and answer text strings — no document images; all judge traffic remains under the organization's data-governance controls. Document synthesis capabilities can in principle be misused for forgery; our corpus targets degraded *understanding* rather than high-fidelity counterfeiting, and we release degradation channels rather than substrate-replacement tooling for identity documents.

The Use of Large Language Models

LLMs were used as writing and coding assistants throughout this project (drafting and revising text, generating analysis and plotting scripts under author review) and as components of the evaluation harness where explicitly stated (a primary LLM judge — Qwen3.6-35B-A3B — and a second, independent judge — Kimi-K3, which also serves as a frontier baseline in Tab. 3 — audit 100 items in Sec. 5.3; frontier LLM/VLM baselines in Tab. 3). All experimental designs, claims, and final text were reviewed and are the responsibility of the authors.

References

Gilles Baechler et al. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*, 2024.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. arXiv:2011.10566.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9313–9332, 2024.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. In *ICLR*, 2022. arXiv:2111.00899.

Sagnik Das, Ke Ma, Zhixin Shu, Dimitris Samaras, and Roy Shilkrot. Dewarpnet: Single-image document unwarping with stacked 3d and 2d regression networks. In *ICCV*, 2019.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *arXiv preprint arXiv:1712.07629*, 2017.

Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations (ICLR)*, 2023.

Hao Feng, Yuechen Wang, Wengang Zhou, Jiajun Deng, and Houqiang Li. Doctr: Document image transformer for geometric unwarping and illumination correction. *arXiv preprint arXiv:2110.12942*, 2021.

Hao Feng, Wengang Zhou, Jiajun Deng, Yuechen Wang, and Houqiang Li. Geometric representation learning for document image rectification. *arXiv preprint arXiv:2210.08161*, 2022.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. Augraphy: A data augmentation library for document images. In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 384–401, 2023a.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. Shabbypages: A reproducible document denoising and binarization dataset. *arXiv preprint arXiv:2303.09339*, 2023b.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016. arXiv:1604.06646.

Anwen Hu et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. *arXiv preprint arXiv:2111.15664*, 2021.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *arXiv preprint arXiv:2210.03347*, 2022.

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. arXiv:2102.09480.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Yuliang Liu et al. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023.

Ke Ma, Zhixin Shu, Xue Bai, Jue Wang, and Dimitris Samaras. Docunet: Document image unwarping via a stacked u-net. In *CVPR*, 2018.

Microsoft. Genalog: Synthetic document image generation with custom degradations and text alignment. `https://github.com/microsoft/genalog`, 2021. Open-source Python package.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*, 2024.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, et al. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*, 2023.

Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 147–166, 2023.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020a. arXiv:2001.07685.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020b.

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. arXiv:1705.02193.

Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *CVPR*, 2020. arXiv:1912.05524.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *ICCV*, 2021. arXiv:2104.03308.

Floor Verhoeven, Tanguy Magne, and Olga Sorkine-Hornung. Uvdoc: Neural grid-based document unwarping. *arXiv preprint arXiv:2302.02887*, 2023.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2023a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3240–3249, 2023b.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3024–3033, 2021.

Renqiu Xia et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16684–16693, 2021.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*, 2024.

Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 109–124, 2021.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

A Evaluation Protocol Details

A.1 Matched740 Selection Protocol

The matched740 real-document evaluation set is drawn from production data and cannot be released for compliance reasons. We document the selection protocol so that the statistical claims in the paper can be re-derived from the per-item audit tuples we will release.

Composition. The 740 items come from two production sources: contract element extraction (680 items) and complaint-management documents (60 items). All items are factual extraction questions (single verifiable answer grounded in the document image). Questions are organized into 12 template families (q04–q15) covering clause lookup, entity extraction, numeric-value extraction, table-cell lookup, and cross-reference questions.

Coverage matrix. Question-family $\times$ source distribution:

Question family	Contract	Complaint
q04 (retention/duration lookup)	107	7
q05 (cross-document reference)	109	7
q06 (clause-title extraction)	114	9
q07 (code$\rightarrow$value lookup)	108	10
q08 (obligation enumeration)	92	6
q09 (conditional-clause reading)	58	8
q10 (credential/ID extraction)	41	3
q11–q15 (long-tail: rates, contacts, table cells)	51	10
Total	680	60

Answer-length distribution. Short ($\leq$5 characters): 230; medium (6–20): 429; long ($>$20): 81. The predominance of short and medium answers reflects the extraction-oriented nature of the set and motivates the exact-EM metric.

Audit tuples. For each of the 740 items, we release a tuple (question_family, source, answer_length_bucket, base_correct, synth_only_correct, final_correct, judge_verdicts) where the correctness flags are binary and judge_verdicts carries both judges' YES/NO ratings for the audited subset. This allows all statistical claims (aggregate accuracy, per-stratum deltas, confidence intervals) to be re-derived without releasing the underlying documents or answers.

A.2 Grounding Evaluation Protocol

The grounding evaluation set comprises 1,000 items across three tasks:

- **ref2box** (278 items): given a field name, emit the bounding box coordinates.
- **ans+box** (361 items): answer a question and provide the box.
- **box2text** (361 items): given a box, read the text inside.

Prompt templates. All three tasks use a structured prompt with the image and a natural-language instruction. The instruction templates are:

- ref2box (EN): "Locate the text '{field_name}' in the document and return the bounding box in the format `(x1,y1),(x2,y2)`."
- ref2box (ZH): the Chinese equivalent ("draw the box around {field_name}"); exact strings are in the released evaluation scripts.
- ans+box (EN): "{question} Provide the answer and the bounding box."
- ans+box (ZH): the Chinese equivalent of the above.
- box2text (EN): "Read out the text from the box `(x1,y1),(x2,y2)`."
- box2text (ZH): the Chinese equivalent of the above.

25% of the prompts use English instructions; the remainder use Chinese.

Parsing and normalization. Model outputs are parsed with a regex-based extractor that looks for coordinate tuples in the forms `(x1,y1),(x2,y2)`, `[x1,y1,x2,y2]`, or JSON `{"box":[x1,y1,x2,y2]}`. Boxes are normalized to `[0,1000]` before computing IoU. For ans+box tasks, the text answer is normalized by stripping punctuation, lowercasing, and collapsing whitespace, then checked for exact match or containment (reference $\subseteq$ prediction).

Metrics. mIoU is averaged over *parseable* predictions only; acc@0.5 is computed over the *full* item count, with unparseable outputs counted as misses. Parse-failure rates are reported alongside so both quantities can be reconciled.

B Training Hyperparameters and Compute

All training uses the `ms-swift` library (version 4.4.2) with LoRA adapters. The base model is Qwen3.5-4B. Hyperparameters are consistent across all arms:

- **LoRA**: rank 32, alpha 64, target modules = all linear layers.
- **Learning rate**: 2e-5, warmup ratio 0.03.
- **Effective batch size**: 32 (bs=4 $\times$ grad-accum=8, or bs=8 $\times$ grad-accum=4 on larger-memory nodes; equivalent).
- **Optimizer**: AdamW (library defaults otherwise).
- **Mixed precision**: bf16; gradient checkpointing on.
- **Max sequence length**: 4,096 tokens.
- **Seeds**: 42, 43, 44 (both `seed` and `data_seed`; reported as mean $\pm$ std where applicable).

Two-stage recipe. Stage 1 trains on grounding data only (20,000 items, 625 steps per seed at effective batch 32). Stage 2 continues from the stage-1 checkpoint with the full mixture (task QA + 500 grounding replay + 3K public replay, 11,086 items, 347 steps per seed). Each seed's full chain (both stages plus five evaluations) takes $\sim$2.5 hours on a single H200. The 234/469/938-step S-curve in the main text refers to a separate grounding-only sweep, not to these stage lengths.

Compute budget. Each training arm in the experimental matrix is a single-GPU LoRA run of 0.5–3 hours; the full matrix (Tab. 5, the σ/scale/LOCO sweeps, and failed or superseded arms) consumed on the order of a few hundred H200 GPU-hours. Exact per-arm step counts and wall-clock are recorded in the released experiment ledger.

C Seed Determinism and Reproducibility

Random number generator pinning. All third-party RNGs are seeded: Python `random.seed(SEED)`, NumPy `np.random.seed(SEED)`, PyTorch `torch.manual_seed(SEED)`, and CUDA `torch.cuda.manual_seed_all(SEED)`. Training uses `torch.use_deterministic_algorithms(True)` where supported.

Non-exportable operators. The crumple operator (`_crumple` in the rendering engine) was originally implemented with a third-party library whose RNG state was not serializable. We re-implemented it parametrically (an $s \times s$ control-point displacement grid, bicubic-upsampled, $s \in [6, 12]$) *before* the corpus used in all experiments was generated, so that the correspondence field Φ could be persisted exactly. Determinism is verified by re-running the pipeline under a fixed environment (Python 3.10, NumPy 1.24, PyTorch 2.1, CUDA 12.1) and checking that the output images and Φ fields are bit-identical.

Data loading. The training loop uses `DataLoader` with `worker_init_fn` to seed each worker independently. Shuffling is deterministic per seed.

D Public Benchmark Evaluation Details

D.1 DocVQA and InfographicVQA

We evaluate on the validation splits: DocVQA-val (5,349 items) and InfographicVQA-val (1,000 items). Prompts follow the format:

```
<image> Question: {question} Answer:
```

Lenient EM. A prediction is correct if it exactly matches any reference answer *or* if the normalized reference is contained in the normalized prediction (direction: reference $\subseteq$ prediction). Normalization: strip `<think>.*?</think>` tags, apply NFKC Unicode normalization, collapse whitespace, strip trailing punctuation (ASCII and CJK variants of period, comma, semicolon, colon, and quotes), and lowercase.

ANLS (official). We also report the official ANLS metric for these two benchmarks. ANLS computes the best-matching reference for each prediction:

$$\text{ANLS} = \frac{1}{N}\sum_{i=1}^{N} \max_{r \in \text{refs}_i} \left(1 - \frac{\text{Levenshtein}(r, p_i)}{\max(|r|, |p_i|)}\right) \mathbb{1}_{\text{NL}<0.5}$$

where NL is the normalized Levenshtein distance. Normalization: strip, lowercase, and apply NFKC. The ANLS script is adapted from the official DocVQA evaluation.

D.2 OCRBenchV2

OCRBenchV2 (1,000 items) is a held-out slice of the full OCRBench. We report lenient EM with the same normalization as DocVQA. The official OCRBenchV2 script has not yet been run; we plan to update the camera-ready version with the official metric if it differs.

D.3 XFUND-zh KIE

XFUND-zh (Chinese) is a key-information extraction benchmark. The gold annotations are JSON objects with field-value pairs. We parse both the gold and predicted JSON, extract quoted values, normalize them (same normalization as lenient EM), and compute value-level precision, recall, and F1:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \cdot \text{P} \cdot \text{R}}{\text{P} + \text{R}}$$

where TP counts values that appear in both gold and prediction (after normalization), FP counts predicted values not in gold, and FN counts gold values not in the prediction.

D.4 Degraded Public Benchmarks

For each public benchmark, we apply the five degradation channels (print, scan, photo, screen, screenshot) at production-calibrated parameters. The degraded images are persisted with their manifest (channel, seed, parameter values) so that the degradation is not re-sampled at evaluation time. The headline degraded set (e.g. DocVQA-degraded, 1,000 items) assigns *one* channel per item, sampled at production traffic weights (photo 299 / scan 258 / print 220 / screen 125 / screenshot 98) — a traffic-weighted mixture, which is why its aggregate differs from the unweighted mean of the per-channel runs reported in the main text.

E Inference-Time OCR Baseline

The baseline runs PaddleOCR (PP-OCRv5 mobile) on each evaluation image at query time, then string-matches the requested field name against recognized lines (difflib ratio, threshold 0.5) and returns the best line's box. On ref2box it reaches 0.820 acc@0.5 with no training, but 0.536 @0.75 and 0.151 @0.9 (ours: 0.970/0.829/0.255) — OCR line boxes do not align with field boundaries, so precision collapses as the threshold tightens. 27/278 queries (9.7%) produce no prediction at all (OCR misses or no match above threshold); reading text from given boxes scores 0.765 vs. our 0.889; and the pipeline cannot answer-and-ground end-to-end, pays OCR latency at every query, and inherits OCR's degradation sensitivity precisely where robustness is needed.

F LLM-Judge Protocol

The judge audit (Sec. 5.3) samples 100 matched740 items (fixed seed) and asks the judge whether each arm's answer is semantically correct given the gold answer (YES/NO, temperature 0, three attempts). Primary judge: Qwen3.6-35B-A3B on internal infrastructure. An independent second judge (Kimi-K3, accessed through an organization-controlled gateway; inputs are question and answer text strings only, no document images) re-rated the same items: on 179 paired verdicts (of 2×100; 21 verdicts dropped where a judge returned no parseable YES/NO) the two agree 97.8% (Cohen's $\kappa = 0.82$, "almost perfect"); per arm, agreement is 98.9% (base) and 96.6% (fine-tuned). The judge verdicts are included in the released audit tuples.

G Numerical Tables

Series	σ=0	σ=2	σ=5	σ=10	σ=20
Synthetic ref2box (3 seeds)	.838±.004	.829±.008	.805±.008	.753±.008	.593±.007
FUNSD ref2box (s42)	.546	.537	.493	.439	.264
500-sample probe (s42)	.767	.760	.731	.674	.592

Table 6: Numerical values behind Fig. 4. Reference points: dose-matched bias arm (constant +8 px) 0.526 raw / 0.841 after un-shifting; dose-matched i.i.d. arm (σ=8) 0.792; real-OCR arms: PaddleOCR PP-OCRv5 (81% kept, 3 seeds) 0.751±0.002, PP-OCRv4 via RapidOCR (79% kept, s42) 0.782.

Qwen3.5 scale	0.8B	2B	4B	9B
Real-domain exact-EM	.549	.580	.676	.730
ref2box mIoU	.768	.786	.820	.842
Degraded-domain exact-EM	.422	.470	.492	.490

Table 7: Scale-sweep values (synth-only recipe, single-seed runs; Sec. 5.4).

Arm	acc@0.5	acc@0.75	acc@0.9
ref2box, final (3 seeds)	.970	.829	.255
ref2box, grounding-only σ_0	.970	.876	.315
ans+box, final	.914	.770	.203

Table 8: Multi-threshold accuracy (deployment view). acc@0.5 supports evidence highlighting and human review; acc@0.9 is not yet sufficient for automated redaction (Limitations).

H FRONTIER LLM/VLM BASELINES

The frontier baselines in Table 3 are evaluated on a 300-item stratified probe drawn from the 1,000-item grounding evaluation set (83 ref2box / 109 ans+box / 108 box2text). The models are **Kimi-K3** and **Qwen3.8-27B**, both accessed through OpenAI-compatible gateways.

Frontier models receive the image as base64 and an explicit instruction to return the box as JSON in *original-image pixel coordinates*; responses are taken from `message.content` with generous `max_tokens` (reasoning models return empty content otherwise). For box2text items the `<bbox>` placeholder is replaced by the ground-truth pixel region so the model knows which region to read. Failed or unparseable responses count as parse failures; 12 of 600 requests (2%) failed at the transport layer after retries and are counted as failures.

Coordinate-frame unreliability of Kimi-K3. K3 parses 83/83 ref2box probes, but at temperature 0 its answers mix coordinate conventions: 62/83 arrive normalized to $[0, 1000]$ despite the explicit original-pixel instruction, and repeated calls on the same item can switch convention. We therefore report three readings in Table 3: *as instructed* (boxes scored in the requested pixel frame), *heuristic frame* (a deployable no-ground-truth rule: any coordinate $>1000 \Rightarrow$ pixel frame, else norm-1000), and *oracle frame* (the better of the two interpretations per item, an upper bound). The heuristic recovers most of the oracle gain (0.795 vs. 0.892 acc@0.5), so the residual gap measures genuinely ambiguous items, not an unreachable ceiling; the as-instructed vs. oracle gap (0.157 vs. 0.892) locates the failure in protocol adherence rather than localization ability, mirroring the format-locked behavior of the open bases. Asking K3 directly for norm-1000 — the convention it spontaneously prefers — drops its compliance to 40/83 parseable (0.434 acc@0.5): the unreliability is in protocol adherence itself, not in which frame is requested. The box-parsing protocol is otherwise identical to the one used for our own models (Appendix A).