HYBRID APPROACH FOR KEY-INFORMATION EXTRACTION FROM STRUCTURED DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated extraction of key information from structured document images is important for accounting, document management, and business automation, yet it remains difficult under blur, noise, mixed languages, and variable layouts. Although recent vision-language-model-based solutions have shown strong performance on document-understanding tasks, many of them rely on large models that require expensive hardware or cloud-based inference. In practical settings, however, document data may be sensitive, and some clients do not permit sending files to external servers, making fully local processing on affordable devices a necessary requirement. This work studies that problem through three extraction settings: a classical pipeline based on field detection, OCR, and post-processing; an OCR-free vision-language pipeline that generates JSON directly from the full document; and a hybrid pipeline that combines field localization with vision-language recognition on cropped regions. The proposed workflow includes orientation correction, field detection, recognition, normalization, and validation for producing machine-readable JSON outputs. Experiments were conducted under a unified schema and included comparisons at the detector, OCR, vision-language-model, and end-to-end levels. On the in-house receipt evaluation, the hybrid pipeline achieved the lowest CER and WER and the highest item exact match and slot F1 among the evaluated pipelines, at higher latency than full-document VLM extraction. On CORD-v2, its performance depended on the recognizer: the Qwen2-VL-2B hybrid underperformed its OCR-free counterpart, whereas the Qwen2.5-VL-3B hybrid improved several field-level metrics and JSON validity but had lower document exact match and no statistically significant item-exact-match gain. These findings provide practical guidance for selecting locally executable extraction pipelines according to the required balance of accuracy, output validity, and processing time.

1 INTRODUCTION

Automated key-information extraction supports accounting and document management, but sensitive documents may need to remain on local infrastructure. Data-protection requirements (European Parliament and Council, 2016) and security-control frameworks (Joint Task Force, 2020) motivate on-premise processing, while hardware and licensing costs constrain practical deployment. Given a document image and a fixed schema, the task is to return normalized JSON containing dates, totals, identifiers, and receipt line items despite layout variation and image degradation.

Classical detector-and-OCR pipelines offer explicit processing stages but propagate localization and recognition errors into structured outputs. OCR-free vision-language models generate outputs directly from document images, but may produce formatting errors or inconsistent field values. This study compares these alternatives with a hybrid pipeline that combines field localization, compact VLM recognition on crops, and field-specific normalization.

The contributions are an on-premise extraction workflow, component-level and end-to-end comparisons, and an analysis of when detector-guided recognition is useful. Evaluation covers an in-house receipt collection and CORD-v2. The results distinguish field accuracy, complete-document correctness, JSON validity, and latency, rather than treating them as interchangeable measures of system quality.

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.

2 RELATED WORK

OCR-first document extraction combines preprocessing, recognition, and rule-based parsing. Thresholding methods such as Otsu (Otsu, 1979) and engines such as Tesseract (Smith, 2007) remain practical components, while SROIE (Huang et al., 2021) and CORD (Park et al., 2019) provide receipt benchmarks. Field detectors such as Faster R-CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016) reduce the recognition area, but inaccurate boxes can truncate characters or introduce background clutter.

Layout-aware models combine textual, visual, and spatial information. DocFormer, LiLT, and TILT (Appalaraju et al., 2021; Wang et al., 2022; Powalski et al., 2021) explore complementary document representations; LayoutLMv3 (Huang et al., 2022) unifies text and image masking; and DiT (Li et al., 2022) studies document-image pretraining. OCR-free Donut (Kim et al., 2022) and image-to-text Pix2Struct (Lee et al., 2023) reduce reliance on separate recognition stages. These approaches motivate comparisons between OCR-based and direct structured generation under local deployment constraints.

More recent work moves toward region-aware and multi-stage extraction, combining localization or OCR with compact vision-language models rather than a single end-to-end pass — e.g., VDInstruct (Nguyen et al., 2025), OmniParser V2 (Yu et al., 2025), and multi-stage field extraction with OCR and compact VLMs (Jin et al., 2025)— which directly motivates the hybrid design pursued here.

Our study evaluates a modular detector-plus-recognizer design with compact, locally executable components. The comparison focuses on the measured trade-offs among extraction quality, output validity, and processing cost in the evaluated settings.

3 MATERIALS AND METHODS

3.1 SYSTEM OVERVIEW

The workflow contains three pipelines. The first pipeline (A) is a modular detector-and-OCR approach consisting of preprocessing, field localization, OCR recognition, normalization and validation, and finally JSON formatting. The second pipeline (B) is an end-to-end approach in which a vision-language model (VLM) generates the JSON directly from the document image under a fixed schema and controlled decoding configurations. The third is the proposed hybrid pipeline, in which YOLO11s first localizes individual fields and Qwen2-VL-2B-Instruct subsequently recognizes the content of each detected crop using field-specific prompts. The final output is formatted to follow a fixed set of keys; missing or illegible fields are set to null.

Figure 1 summarizes the three processing paths. Field normalization checks dates, cleans numeric values and identifiers, and standardizes decimal separators before JSON assembly. Missing or unreadable fields are represented as null.

3.2 DATA AND EVALUATION SUBSETS

For object-detection training, the receipt dataset uses eight target classes: DATE, FB, SUM, ORDER, FD, NAME, QUANTITY, and PRICE, split approximately into 180 training, 20 validation, and 20 test images. For isolated OCR evaluation, a fixed set of 100 cropped text regions with ground-truth labels was used.

For document-level VLM evaluation, a fixed set of 66 Dedkho receipt images was used. Each image was paired with a corresponding ground-truth JSON annotation, resulting in 66 image–annotation pairs. The same evaluation set was used for both VLM configurations to ensure a direct and consistent comparison. This document-level VLM evaluation is reported separately from the object-detection train/validation/test split.

The collection also includes 80 Emirates ID images, with a 60/10/10 training/validation/test split, and 100 rotation-training images, split 80/10/10. Receipt data contain Russian text; the ID data include Arabic and English. The end-to-end results reported below focus on receipts.

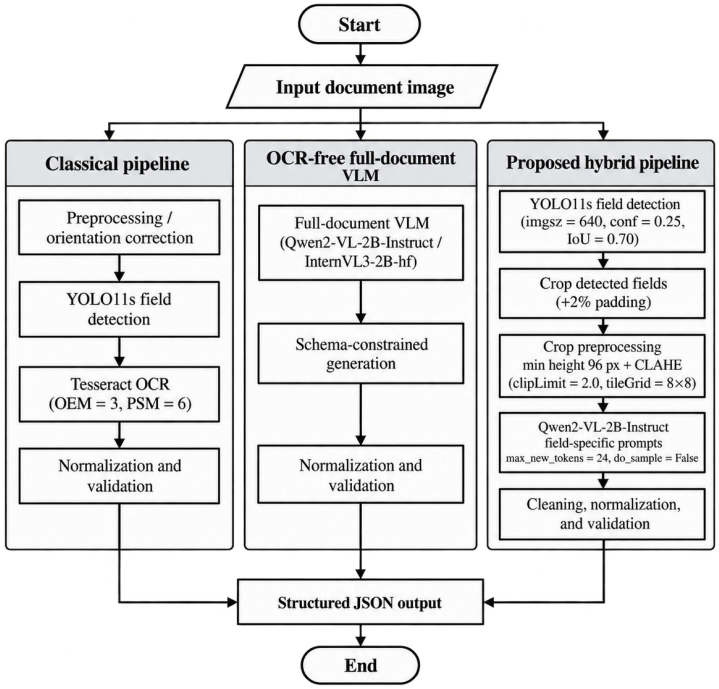

Figure 1: Project workflow.

To verify that the approach generalizes beyond the in-house corpus, the detection and end-to-end pipelines were additionally evaluated on the public CORD-v2 receipt dataset (Park et al., 2019) (public release: 1,000 receipts; 800 train / 100 validation / 100 test), spanning many store layouts. CORD categories were mapped to four shared fields — NAME (menu.nm), QUANTITY (menu.cnt), PRICE (menu.price), and TOTAL (total.total_price); the Russian fiscal identifiers remain specific to the in-house set. The architecture and training protocol were identical to those used on the in-house data.

3.3 Classical and OCR-free baselines

The classical pipeline applies preprocessing, detector-based field localization, Tesseract or EasyOCR recognition, and normalization. Evaluated preprocessing alternatives include resizing, denoising, CLAHE, adaptive binarization, and morphological closing. Resizing gave the lowest reported average CER, while CLAHE was fastest and improved OCR confidence without improving CER over the baseline. Tesseract uses OEM = 3 and PSM = 6; both OCR engines are evaluated on the same cropped regions.

To ensure a fair comparison and stable output, the same rules are applied to both models: produce JSON-only outputs, use a fixed key schema (e.g., DATE, SUM, PRICE, QUANTITY, ORDER, FD, FB, NAME) based on the document, return null when a field is missing or unreadable, and apply basic normalization rules for date format and numeric decimal separators. Generation was deterministic, using max_new_tokens = 256–512, do_sample = False, and use_cache = True. Qwen2-VL-2B-Instruct was loaded with torch_dtype = "auto" and device_map = "auto"; no model quantization was enabled in the main configuration. The processor used min_pixels = $256 \times 28 \times 28$ and max_pixels = $512 \times 28 \times 28$.

Both VLM models, Qwen2-VL-2B-Instruct and InternVL3-2B-hf were used without any task-specific fine-tuning on the datasets. The pretrained instruction-tuned checkpoints were evaluated through prompting only, and no gradient-based parameter updates were performed.

3.4 HYBRID RECOGNITION

For the hybrid pipeline, YOLO11s inference used an input size of 640 pixels, a confidence threshold of 0.25, and an IoU threshold of 0.70. Each detected bounding box was expanded by 2% before cropping in order to reduce the risk of truncating edge characters. These settings were fixed for all reported hybrid experiments.

Each preprocessed crop is passed independently to Qwen2-VL-2B-Instruct together with a field-specific textual prompt. The VLM input therefore consists of the cropped field image and a textual instruction selected according to the detected field class. No OCR-recognized text is supplied to the VLM at this stage.

Before recognition, crops shorter than 96 pixels are enlarged using bicubic interpolation while preserving aspect ratio. Each crop is converted to grayscale, processed using CLAHE with clipLimit = 2.0 and tileGridSize = (8, 8), and converted back to a three-channel image. No additional pixel-value normalization is applied at this stage. This preprocessing is fixed across the reported hybrid experiments.

The input message is formatted using the Qwen chat template and contains one image followed by the corresponding field-specific prompt. Generation is deterministic, with max_new_tokens = 24, do_sample = False, and use_cache = True. No stochastic sampling parameters are used during inference.

Field-specific prompts request the original item name, numeric price or total, quantity, date, or fiscal identifier. The exact prompts are retained in Appendix A. Only newly generated tokens are decoded. Cleaning removes Markdown fences, surrounding quotes, redundant whitespace, and unwanted labels; field-specific normalization and validation precede JSON assembly.

$$\hat{v}_f = \mathrm{Clean}(\mathrm{VLM}(\mathrm{Preprocess}(I[b_f]), P_f)) \,. \tag{1}$$

Here, $I[b_f]$ is the detected crop, P_f is its field-specific prompt, and $\hat{v}_f$ is the cleaned predicted value.

3.5 DETECTOR LEARNING OBJECTIVES

The detectors use standard supervised object-detection losses. Faster R-CNN combines RPN objectness and box regression with RoI classification and localization, including Smooth-L1 regression (Ren et al., 2015; Girshick, 2015). YOLO11s uses the default Ultralytics combination of box regression, classification, and distribution-based refinement. These follow established detection-loss formulations (Terven et al., 2025; Jiang et al., 2020; Zheng et al., 2019); no new loss function is introduced.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 EVALUATION PROTOCOL

The evaluation separates detector, cropped-field OCR, full-document VLM, and end-to-end pipeline comparisons. Detectors are compared using detection quality, latency, parameter count, and peak VRAM. OCR is evaluated before and after postprocessing using CER, exact match, and valid-output rate. The VLM component comparison uses the same 66 Dedkho receipts for both models. End-to-end evaluation uses a common test subset and JSON schema within each comparison, reporting CER, WER, ORDER item exact match, slot F1, document exact match, JSON validity, and latency. Lower CER and WER are better; higher exact match and F1 are better. Syntactically valid JSON does not establish semantic correctness.

4.2 COMPONENT COMPARISONS

YOLO11s has the more favorable reported localization-quality and resource trade-off. Tesseract gives better cropped-field recognition accuracy than EasyOCR, while EasyOCR has a slightly higher valid-output rate. Qwen2-VL-2B-Instruct gives stronger structured extraction scores and lower la-

tency than InternVL3-2B-hf; the latter uses slightly less peak VRAM. Table 4, Table 5, and Table 6 in Appendix B retain all component measurements.

Table 1: Selected component-level results. OCR CER and EM are after postprocessing; VLM scores are from the 66-image component evaluation.

Comparison / metric	First component	Second component
Detector	YOLO11s	Faster R-CNN
Reported mAP@0.5 / F1	0.981	0.926
Inference time (ms)	60.92	112.89
Peak VRAM (GB)	0.51	2.05
OCR	Tesseract	EasyOCR
Average CER	0.419	0.509
Exact match (%)	31.31	25.25
Valid output (%)	82.82	86.86
VLM	InternVL3-2B	Qwen2-VL-2B
Field EM (%)	23.11	26.89
ORDER item EM (%)	2.07	4.57
Inference time (ms)	10968.51	4429.37

4.3 In-house end-to-end results

The three end-to-end pipelines were compared on the same test subset using a unified JSON schema. The classical pipeline used YOLO11s for field localization followed by Tesseract OCR; the OCR-free pipeline used Qwen2-VL-2B-Instruct directly on the full document image; and the hybrid pipeline used YOLO11s for field localization followed by Qwen2-VL-2B-Instruct on the detected field crops. The comparison reports CER, WER, ORDER-item exact match, slot F1, document exact match, JSON validity, and average inference time.

Table 2: In-house end-to-end comparison. Classical: YOLO11s + Tesseract; OCR-free: full-document Qwen2-VL-2B; hybrid: YOLO11s + crop-level Qwen2-VL-2B.

Metric	Classical	OCR-free	Hybrid
JSON valid (%)	100.00	96.97	100.00
CER (%)	28.27	41.97	22.73
WER (%)	43.40	46.61	28.81
ORDER item EM (%)	4.00	0.00	29.33
Slot F1 (%)	44.27	30.55	74.11
Document EM (%)	0.00	0.00	4.65
Avg inference time (s)	N/A	13.42	17.74

The hybrid achieves the lowest CER and WER and the highest item EM and slot F1 in this comparison. Complete-document success remains limited to 4.65%. Its average latency is 17.74 s versus 13.42 s for OCR-free extraction, so the accuracy improvement comes with additional processing time.

4.4 CORD-v2 results

On CORD-v2, YOLO11s achieves mAP@0.5 of 0.961 and mAP@0.5:0.95 of 0.846, versus 0.897 and 0.658 for Faster R-CNN. Table 7 in Appendix B retains the per-field detector scores. End-to-end comparisons additionally include Qwen2.5-VL-3B-Instruct.

Using 10,000 bootstrap resamples over per-receipt scores, 95% confidence intervals for ORDER item EM are: Hybrid (Qwen2.5-VL-3B) 50.9 [42.1, 59.7], OCR-free (Qwen2.5-VL-3B) 47.2 [37.9, 56.2], Hybrid (Qwen2-VL-2B) 24.4 [16.9, 32.3], and Classical (Tesseract) 15.3 [9.6, 21.7]. Paired two-sided permutation tests confirm that the hybrid significantly outperforms the classical baseline (Δ = +35.6, $p < 0.001$) and that the newer compact model significantly outperforms the older one (Δ = +26.5, $p < 0.001$). For the newer model, the difference between hybrid and OCR-free

Table 3: CORD-v2 end-to-end results. All values are percentages. Full denotes OCR-free full-document extraction; Item EM and Doc EM denote item and document exact match.

Pipeline	JSON	CER	WER	Item EM	Slot F1	Doc EM
Tesseract	100.00	39.42	58.88	15.29	46.16	6.00
EasyOCR	100.00	38.80	52.74	10.88	37.32	0.00
Full Qwen2 2B	98.00	30.52	37.23	33.76	61.06	14.00
Hybrid Qwen2 2B	100.00	42.63	50.05	24.38	44.48	8.00
Full Qwen2.5 3B	77.00	37.46	45.52	47.16	60.92	29.00
Hybrid Qwen2.5 3B	100.00	26.77	36.37	50.91	65.97	18.00

is not statistically significant (Δ = +3.7, p = 0.50); there the hybrid's advantage is JSON validity, with valid JSON on 100% of evaluated inputs versus 77%, although document exact match is lower (18% versus 29%). Overall, the CORD-v2 results show model-dependent hybrid performance. With Qwen2-VL-2B, the hybrid is less accurate than OCR-free extraction; with Qwen2.5-VL-3B, it improves several field-level metrics but has lower document exact match. Both hybrid configurations produced valid JSON on all evaluated CORD-v2 inputs.

4.5 Field-wise analysis

Figure 2 compares exact match for FB, FD, DATE, SUM, NAME, PRICE, and QUANTITY. The hybrid is strongest on identifiers and item names in the displayed comparison, while OCR-free extraction is strongest on DATE. Both achieve 100% on SUM. These differences reinforce the need to distinguish field-level performance from whole-document correctness.

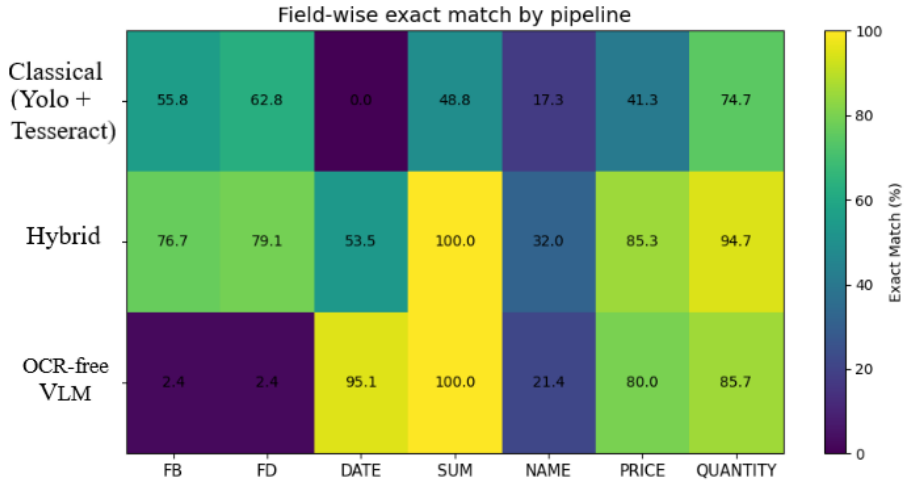

Figure 2: Field-wise exact-match rates by pipeline.

5 Discussion

The results support detector-guided recognition as a model- and dataset-dependent design choice. The in-house comparison favors the hybrid, but CORD-v2 shows that Qwen2-VL-2B benefits more from full-document extraction. With Qwen2.5-VL-3B, the hybrid improves several field-level measures and JSON validity, but not document exact match; its item-EM gain is not statistically significant (p = 0.50). Thus, field cropping should be selected jointly with the recognizer rather than assumed to improve every compact VLM.

Localization constrains recognition to relevant regions, while modular processing exposes detections and recognized values for inspection. This provides a practical basis for locating failures in individual stages. However, missed or poorly cropped fields cannot be fully recovered downstream.

Even strong individual-field scores can coexist with low document EM because the latter requires every field to be correct simultaneously.

The JSON-validity improvement has practical value for downstream software by reducing syntactic parsing failures. This benefit is distinct from accurate field values and correct row associations. Likewise, local execution addresses the requirement to keep document processing on-premise, but the measured accuracy-latency trade-off remains important: slot F1 rises from 30.55% to 74.11% on the in-house data while average latency increases from 13.42 s to 17.74 s. These comparisons support pipeline selection according to extraction quality, output validity, and available processing time.

The evaluation is limited by the relatively small in-house receipt collection, sensitivity to row alignment and normalization, and dependence on field detection. Rankings may change with document type, language, layout, and image quality. Future work includes broader evaluation, improved row grouping and date normalization, and selective VLM recognition of difficult fields.

6 Conclusion

This study compares classical OCR, full-document compact VLMs, and detector-guided VLM recognition for locally processed receipt extraction. The hybrid gives the strongest in-house end-to-end extraction scores, while CORD-v2 establishes a model-dependent trade-off: Qwen2-VL-2B favors OCR-free extraction, and Qwen2.5-VL-3B improves several hybrid field-level scores and JSON validity without a statistically significant item-EM gain or improved document EM. The practical value lies in identifying when localization, recognition, and normalization form a useful combination, with accuracy assessed alongside output validity and latency.

AI USE STATEMENT

Generative AI (ChatGPT/Codex) assisted with manuscript review, interpretation of reported results, feedback on methodological descriptions, editing and shortening the text, and preparation of the LaTeX format. This assistance did not execute the reported experiments or recompute their numerical results. The vision-language models described in the paper are components of the evaluated extraction systems. The authors are responsible for the final manuscript and reported results.

DATA AVAILABILITY

The in-house datasets used and analysed during the current study are not publicly available due to confidentiality and proprietary restrictions related to the industrial facility from which the data were obtained, but are available from the corresponding author on reasonable request and with permission of the data provider. CORD-v2 is publicly available (Park et al., 2019).

DECLARATIONS

Ethics approval and consent to participate: not applicable. Consent for publication: not applicable.

The authors declare that they have no competing interests.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The document data were provided by an industrial partner under a data-use agreement.

REFERENCES

Appalaraju, S., B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha. (2021) "DocFormer: End-to-End Transformer for Document Understanding." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 993–1003.

European Parliament and Council of the European Union. (2016) "Regulation (EU) 2016/679 (General Data Protection Regulation)." Official Journal of the European Union OJ L 119: 1–88.

Girshick, R. (2015) "Fast R-CNN." in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448.

Huang, Z., K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar. (2021) "ICDAR2019 competition on scanned receipt OCR and information extraction." arXiv 2103.10213.

Huang Y, Lv T, Cui L, Lu Y, Wei F (2022) LayoutLMv3: Pre-training for Document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia.

Jiang, S., H. Qin, B. Zhang, and J. Zheng. (2020) "Optimized loss functions for object detection: a case study on nighttime vehicle detection." arXiv 2011.05523.

Y. Jin, K. Chiu Jin-Chun, Y. Wang, K. Zhu Ke, Q. Zhong, and D. MacDonald, "Multi-Stage Field Extraction of Financial Documents with OCR and Compact Vision-Language Models," arXiv:2510.23066, 2025.

Joint Task Force. (2020) Security and privacy controls for information systems and organizations. NIST Special Publication 800-53, Rev. 5, Gaithersburg, MD, National Institute of Standards and Technology.

Kim, G. et al. (2022) "OCR-free document understanding transformer." in Computer Vision – ECCV 2022, Lecture Notes in Computer Science, vol. 13688, edited by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Cham, Springer, pp. 498–517.

Lee, K., M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova. (2023) "Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding." In Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202, pp. 18893–18912.

Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., and Wei, F. (2022). “DiT: Self-supervised Pre-training for Document Image Transformer.” In Proceedings of the 30th ACM International Conference on Multimedia, pp. 3530–3539.

S. Nguyen, G. Nguyen, H. Dao, T. Do, and D. Kim, "VDInstruct: Zero-Shot Key Information Extraction via Content-Aware Vision Tokenization," arXiv:2507.09531, 2025.

Otsu, N. (1979) “A threshold selection method from gray-level histograms.” IEEE Transactions on Systems, Man, and Cybernetics 9 (1): 62–66.

S. Park, S. Shin, B. Lee, J. Lee, J. Surh, M. Seo, and H. Lee, "CORD: A Consolidated Receipt Dataset for Post-OCR Parsing," in Workshop on Document Intelligence at NeurIPS, 2019.

Powalski, R., Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka. (2021) “Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer.” In Document Analysis and Recognition – ICDAR 2021, Lecture Notes in Computer Science, vol. 12822, pp. 732–747.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. (2016) “You only look once: unified, real-time object detection.” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788.

Ren, S., K. He, R. Girshick, and J. Sun. (2015) “Faster R-CNN: Towards real-time object detection with region proposal networks.” in Advances in Neural Information Processing Systems, vol. 28, pp. 91–99.

Smith, R. (2007) “An overview of the Tesseract OCR engine.” in Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR), vol. 2, pp. 629–633.

Terven, J. R., D. M. Cordova-Esparza, A. Ramirez-Pedraza, E. A. Chavez-Urbiola, and J. A. Romero-Gonzalez. (2025) “A comprehensive survey of loss functions and metrics in deep learning.” Artificial Intelligence Review 58: 195.

Wang, J., L. Jin, and K. Ding. (2022) “LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding.” In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 7747–7757.

W. Yu et al., "OmniParser V2: Structured-Points-of-Thought for Unified Visual Text Parsing," arXiv:2502.16161, 2025.

Zheng, Z., P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. (2019) “Distance-IoU loss: faster and better learning for bounding box regression.” arXiv 1911.08287.

A FIELD-SPECIFIC INFERENCE PROMPTS

NAME "This is a crop from a Russian receipt. Read the text exactly as printed in the ORIGINAL language and ORIGINAL script. DO NOT translate. DO NOT explain. DO NOT add labels. DO NOT write phrases like 'product name', 'the product name is', or quotes. Return ONLY the item name text."

PRICE "This is a crop from a Russian receipt. Return ONLY the price exactly as printed. DO NOT translate. DO NOT explain. DO NOT add extra words. Keep digits and decimal separator only. Example: 241.00"

QUANTITY "This is a crop from a Russian receipt. Return ONLY the quantity exactly as printed. DO NOT translate. DO NOT explain. DO NOT add extra words. If it looks like '*1', 'x1', 'x1', return only 1. Digits only."

DATE "This is a crop from a Russian receipt. Return ONLY the date exactly as printed. DO NOT translate. DO NOT explain. Keep only date and optional time."

SUM "This is a crop from a Russian receipt. Return ONLY the total amount exactly as printed. DO NOT translate. DO NOT explain. Keep digits and decimal separator only."

FD "This is a crop from a Russian receipt. Return ONLY the FD number. Digits only. DO NOT explain."

FB "This is a crop from a Russian receipt. Return ONLY the FB/FP number. Digits only. DO NOT explain."

B DETAILED COMPONENT RESULTS

Table 4: Detector comparison.

Criterion	YOLO11s	Faster R-CNN
Self-hosted	Yes	Yes
Parameters	9.42M	41.34M
Peak VRAM	0.51 GB	2.05 GB
Inference time	60.92 ms	112.89 ms
Main metric (mAP@0.5 / F1)	0.981	0.926

Table 5: OCR comparison.

Criterion	Tesseract	EasyOCR
Self-hosted	Yes	Yes
Language used	RU + EN + AR	RU + EN + AR
Configuration	OEM=3, PSM=6	defaults
Avg CER (before / after)	0.540 / 0.419	0.597 / 0.509
Median CER (before / after)	0.333 / 0.250	0.666 / 0.600
Exact match % (before / after)	22.22 / 31.31	9.09 / 25.25
Valid output (%)	82.82	86.86

Table 6: VLM component comparison on 66 Dedkho receipts.

Criterion	InternVL3-2B-hf	Qwen2-VL-2B-Instruct
Self-hosted	Yes	Yes
Evaluation images	66	66
JSON valid (%)	92.42	95.45
Field EM (%)	23.11	26.89
ORDER Item EM (%)	2.07	4.57
Peak VRAM (GB)	4.25 GB	4.32 GB
Average inference time (ms)	10968.51 ms	4429.37 ms

Table 7: CORD-v2 detector comparison.

Criterion	YOLO11s	Faster R-CNN
mAP@0.5	0.961	0.897
mAP@0.5:0.95	0.846	0.658
AP@0.5 NAME	0.952	0.859
AP@0.5 QUANTITY	0.968	0.915
AP@0.5 PRICE	0.978	0.928
AP@0.5 TOTAL	0.947	0.885
Params	9.7 M	41.31 M