Generative AI Track

(Convolve 4.0)

Sowmi
mm24b054@smail.iitm.ac.in

Mohit Singh
da24m010h@smail.iitm.ac.in

Abstract

The automated extraction of structured intelligence from semi-structured financial documents remains a bottleneck in banking automation due to layout volatility, multilingual vernacular scripts, and varying image quality. In this paper, we propose a robust, multi-stage "Extract-Refine-Segment" (ERS) pipeline designed for the IDFC Convolve 4.0 challenge. Our architecture initiates with a geometric pre-processing stage to normalize document orientation and background noise. For primary data extraction, we leverage the high-parameter **Qwen-2.5-7B-VL** to generate a preliminary schema-aligned JSON.

To satisfy the rigorous "Exact Match" requirements for tractor nomenclature and numeric fields, we implement a hybrid post-processing tier: combining deterministic regex-based heuristics with a secondary, lightweight **Qwen-2.5-3B** semantic refiner. This tier corrects OCR hallucinations and standardizes technical suffixes across diverse brands such as Mahindra and Massey Ferguson. Simultaneously, a parallel visual verification branch utilizes **YOLOv8** for regional stamp detection, which subsequently provides a localized spatial buffer to a **Segment Anything Model (SAM3)** for high-precision signature localization via box-prompting. By merging semantic intelligence with geometric segmentation, our pipeline achieves superior accuracy in field extraction while maintaining a competitive latency-cost profile.

1 Introduction

Automated document intelligence from unconstrained scanned images is a core challenge in financial digitization. While conventional Optical Character Recognition (OCR) performs adequately on standardized digital documents, its efficacy collapses when applied to real-world commercial artifacts such as Indian tractor loan quotations. These documents are characterized by a "noise-heavy" environment: heterogeneous layouts, multi-lingual scripts (English, Hindi, Gujarati), handwritten corrections, and critical visual cues like stamps and signatures. Despite appearing trivial to a human reader, extracting structured data from these documents involves complex semantic and geometric reasoning that traditional systems fail to resolve.

1.1 Motivation

In high-stakes financial workflows—such as the loan disbursal processes at IDFC—the cost of inaccuracy is high. Errors in extracting the *Dealer Name*, *Model Name*, or *Asset Cost* lead to downstream failures in credit assessment and risk management. State-of-the-art (SOTA) Vision-Language Models (VLMs) like GPT-4o or Gemini 3 Pro offer high reasoning capacity but are commercially unviable due to extreme per-request costs and prohibitive inference latency. Conversely, lightweight OCR engines lack the "spatial awareness" to interpret circled selections, tick marks, or the validity of an official stamp.

This creates a critical trade-off: the need for human-level reasoning at production-level cost and speed. This work motivates a hybrid, multi-stage pipeline that marries high-capacity semantic extraction with localized geometric segmentation to achieve "Zero-Touch" automation.

1.2 Key Challenges

The problem setting for tractor loan quotations is defined by several unique hurdles:

- **High Layout Volatility:** Quotations are often non-standardized across different dealers, requiring a model that can perform "Zero-Shot" schema mapping without fixed templates.
- **Semantic-Visual Overlap:** Critical field selection is often denoted by visual markers (e.g., a circle around a price or a tick mark next to a specific model variant) which traditional text-only OCR engines cannot interpret.
- **Nomenclature Complexity:** The tractor industry uses highly specific technical suffixes (DI, XP, Tech+, Novo). OCR hallucinations often misread these (e.g.,

"DF" for "DI"), leading to failures in "Exact Match" evaluation.

- **Verification Integrity:** Ensuring a document is legally binding requires the detection and localization of stamps and signatures, which are often faint, overlapping with text, or placed in unconventional regions.

1.3 Limitations of Existing Approaches

Existing "one-size-fits-all" LLM approaches suffer from high computational overhead. Passing a full-resolution invoice image through a 100B+ parameter model for both data extraction and nomenclature cleaning is inefficient and slow. On the other hand, purely rule-based systems are too brittle to handle the multi-lingual transitions between Hindi and English scripts found in regional Indian invoices. This creates a gap for a tiered architecture that separates "Holistic Extraction" from "Specialized Refinement."

1.4 Our Contribution: The ERS Pipeline

In this work, we propose the **Extract-Refine-Segment (ERS)** pipeline, a tiered hybrid architecture specifically optimized for high-accuracy document intelligence under strict latency constraints. Our contributions include:

- **Tiered Inference Strategy:** We utilize a high-parameter **Qwen-2.5-7B-VL** for primary holistic extraction, followed by a localized, lightweight **Qwen-2.5-3B** refiner. This allows for complex semantic reasoning during extraction while offloading nomenclature standardization to a faster, cost-efficient model.
- **Semantic-Deterministic Hybrid Refinement:** We introduce a post-processing layer that combines LLM reasoning (for brand expansion like TAFE $\rightarrow$ Massey Ferguson) with deterministic Regex heuristics for numeric sanitization, specifically targeting "decimal garbage" and symbol artifacts in the *Asset Cost* field.
- **Geometric Computer Vision Branch:** We develop a dedicated visual verification branch that integrates **YOLOv8** for robust stamp detection. We utilize the resulting bounding boxes as "spatial prompts" for the **Segment Anything Model (SAM3)**, enabling high-precision, zero-shot signature localization.
- **Nomenclature Correction Engine:** We designed a specialized "Expert Prompting" skeleton that corrects common OCR hallucinations (e.g., ROVO to Novo, DT to DI) and ensures multi-lingual transliteration (Hindi to English) to meet "Exact Match" compliance.

Overall, this work demonstrates that the synergy of a tiered VLM architecture and geometric segmentation provides a scalable, cost-effective alternative to monolithic flagship models for complex document understanding.

2 Problem Formulation

The core objective of this challenge is the automated extraction of six critical data points from semi-structured tractor loan quotations. Given an input document $\mathcal{D}$ in PDF or image format, the system must map visual and textual features to a structured JSON schema $\mathcal{S}$.

2.1 Field Definitions and Extraction Requirements

The extraction target consists of a mix of textual, numeric, and geometric entities. As per the challenge guidelines, the fields are defined as follows:

Table 1: Target Extraction Fields and Descriptions

Field	Data Type	Description
Dealer Name	Text	Full name of the tractor deale
Model Name	Text	Specific tractor model or varia
Horse Power (HP)	Numeric	Engine capacity (integer value
Asset Cost	Numeric	Total quoted price (digits only
Dealer Signature	Binary + BBox	Presence and spatial localizati
Dealer Stamp	Binary + BBox	Presence and spatial localizati

The primary goal is to ensure high fidelity in extraction across diverse layouts, varying scan qualities, and multilingual scripts including English, Hindi, and Gujarati[cite: 6, 31].

3 Evaluation Criteria

The performance of the ERS pipeline is measured using a tiered evaluation strategy that prioritizes document-level integrity alongside localized geometric accuracy.

3.1 Primary Metric: Document-Level Accuracy (DLA)

The primary success metric is the **Document-Level Accuracy (DLA)**, defined as the percentage of documents for which *all six target fields* are extracted correctly according to their respective matching rules[cite:

75, 76]. The target for a production-ready system is $\geq 95\%$ DLA[cite: 78].

3.2 Matching Rules and Tolerance

To account for OCR noise and layout complexity, the matching logic mentioned in Table 3 is applied

3.3 Secondary Performance Metrics

Beyond extraction accuracy, the system is evaluated on its operational efficiency and deployability in resource-constrained environments:

- **Inference Latency:** Average processing time per document must be ≤ 30 seconds[cite: 80].
- **Inference Cost:** The estimated cost per document should be $< \$0.01$ when running on a CPU or low-tier GPU[cite: 81].
- **Geometric Precision:** Field-level mAP 50-95 is calculated specifically for the detection of stamps and signatures[cite: 80].

4 Exploratory Data Analysis and Data Curation

To architect a robust pipeline, we performed an extensive multi-modal Exploratory Data Analysis (EDA) on the provided dataset. Our analysis focused on identifying the linguistic, structural, and noise-based characteristics that challenge standard OCR systems. A critical component of this phase involved mapping the geographic and linguistic footprint of the documents to ensure the pipeline could accommodate diverse regional invoice formats.

As illustrated in Figure 1, we conducted a detailed **state-wise distribution analysis**. This mapping revealed a high concentration of documents originating from the North and West Indian agricultural belts, which significantly influenced the variability in document layouts and dealer branding styles. This geographic diversity directly correlates with the linguistic complexity encountered during extraction.

Consequently, we performed a **language distribution analysis**, as shown in Figure 2. This confirmed that a substantial percentage of the invoices utilize a "Vernacular-Technical" fusion—where core technical specifications are often in English, but critical entity information, such as dealer names and addresses, are written in regional scripts including Hindi, Gujarati, and Marathi.

These findings justified the necessity for a high-capacity Vision-Language Model (**Qwen2.5-7B-VL**) over traditional OCR engines. By utilizing a VLM capable of cross-lingual semantic understanding, our pipeline remains language-agnostic. This allows for high-fidelity extraction that maintains script integrity while successfully navigating the diverse structural nuances identified during our curation process.

4.1 Document Complexity and Visual Noise

We categorized the documents based on their "noise profile," specifically looking at handwritten vs. printed content and the presence of non-textual markers.

- **Handwritten vs. Printed:** Approximately 405 documents out of the 495 documents contain handwritten prices or model names. We observed that handwritten entries often suffer from lower OCR confidence
- **Complex Selection Patterns:** We identified [32] instances of "selection-by-tick " for the model name and [6] instances of "strikethrough-correction." Standard OCR pipelines typically treat these as noise; however, our pipeline treats these as "Visual Truths" interpreted through spatial reasoning.
- **Poor Handwriting:** In 18 samples, the quality handwritting was too poor making and in some of these the quality was way too poor making it hard for humans to interpret it

4.2 Dataset Curation and YOLOv8 Fine-Tuning

To enable robust detection of stamps and signatures—which often exhibit high visual similarity to company logos or dense handwritten text—we curated a specialized detection dataset using a multi-stage iterative approach.

1. Annotation and Tooling: For the annotation process, we utilized **Label Studio** (`https://labelstud.io/`), an open-source data labeling interface. This allowed for precise bounding-box labeling for two primary classes: `Stamp` and `Signature`.

2. Class Diversity and Sampling Strategy: Our curation focused on addressing the high intra-class variance of official markers.

- **Stamp Taxonomy:** We identified three distinct morphological categories: *Circular* stamps, *Rectangular* stamps, and *Low-Visibility* (faded or light-ink) stamps.
- **Bias Mitigation:** Training samples were selected in equal proportions across these categories to ensure

Table 2: Field-Level Matching Logic

Field Category	Target Fields	Matching Rule
Textual Entities	Dealer Name, Model Name	$\geq 90\%$ Fuzzy Match or Exact Textual Match [cite: 33, 77]
Numeric Values	Horse Power, Asset Cost	Numeric equality within $\pm 5\%$ tolerance [cite: 77]
Visual Markers	Signature, Stamp	Presence correct & IoU ≥ 0.5 on bounding boxes [cite: 77]

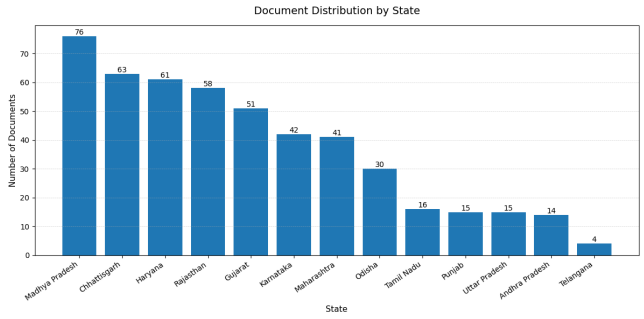

Figure 1: State-wise distribution of tractor quotations, indicating a high concentration from the North and West Indian agricultural belts.

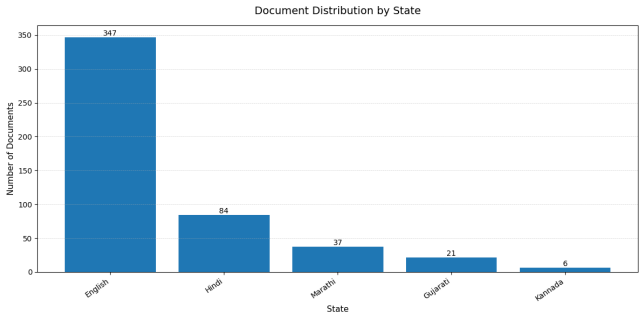

Figure 2: Language distribution analysis

the model did not develop a bias toward a specific geometric shape.

- **Signature Balance:** For the `Signature` class, we maintained a 1:1 ratio of positive samples (signatures present) to negative samples (background/empty fields) to minimize false-positive triggers on standard text.

3. Iterative Fine-Tuning (Active Learning): We employed a "Human-in-the-Loop" strategy to optimize performance:

- **Phase I:** An initial set of 40 images was curated with an 80:20 train/validation split. This served as the baseline for the first fine-tuning round.
- **Phase II (Error Analysis):** The resulting model was evaluated to identify edge cases and failure modes. Samples where the model failed to detect artifacts were extracted, manually annotated, and appended to the dataset.
- **Final Dataset:** This iterative process resulted in a final refined set of 55 high-quality annotated images, significantly improving the recall for low-visibility markers.

4. Model Selection and Configuration: While we benchmarked both **YOLOv8n (Nano)** and **YOLOv8s (Small)**, the **YOLOv8s** variant was selected as our production model due to its superior feature extraction capabilities for complex document backgrounds. All images were pre-processed with a resize operation to an `IMG_SIZE` of 736×736 to optimize the internal receptive field of the YOLO architecture.

5. Data Augmentation: Given the limited volume of ground-truth data, we utilized aggressive data augmentation to simulate the variability of real-world document scans. Our optimized parameters included:

- **Geometric:** Degrees: 15.0°, Translate: 0.1, Scale: 0.5.
- **Flipping:** Horizontal flip (0.5 probability); Vertical flip was disabled (0.0) to maintain document orientation logic.
- **Compositional:** Mosaic (1.0), Mixup (0.1), and Copy-Paste (0.1) were utilized to handle occlusion and varied artifact placement.

5 Technical Methodology: Hybrid ERS (Extract-Refine-Segment) Architecture

We implement a modular **Extract-Refine-Segment (ERS)** pipeline, as illustrated in Figure 3. This architecture is specifically designed to balance high-fidelity information extraction with the strict latency and memory constraints of a single NVIDIA T4 GPU.

We implement a modular **Extract-Refine-Segment (ERS)** pipeline, as illustrated in Figure **??**. This architecture is specifically designed to balance high-fidelity

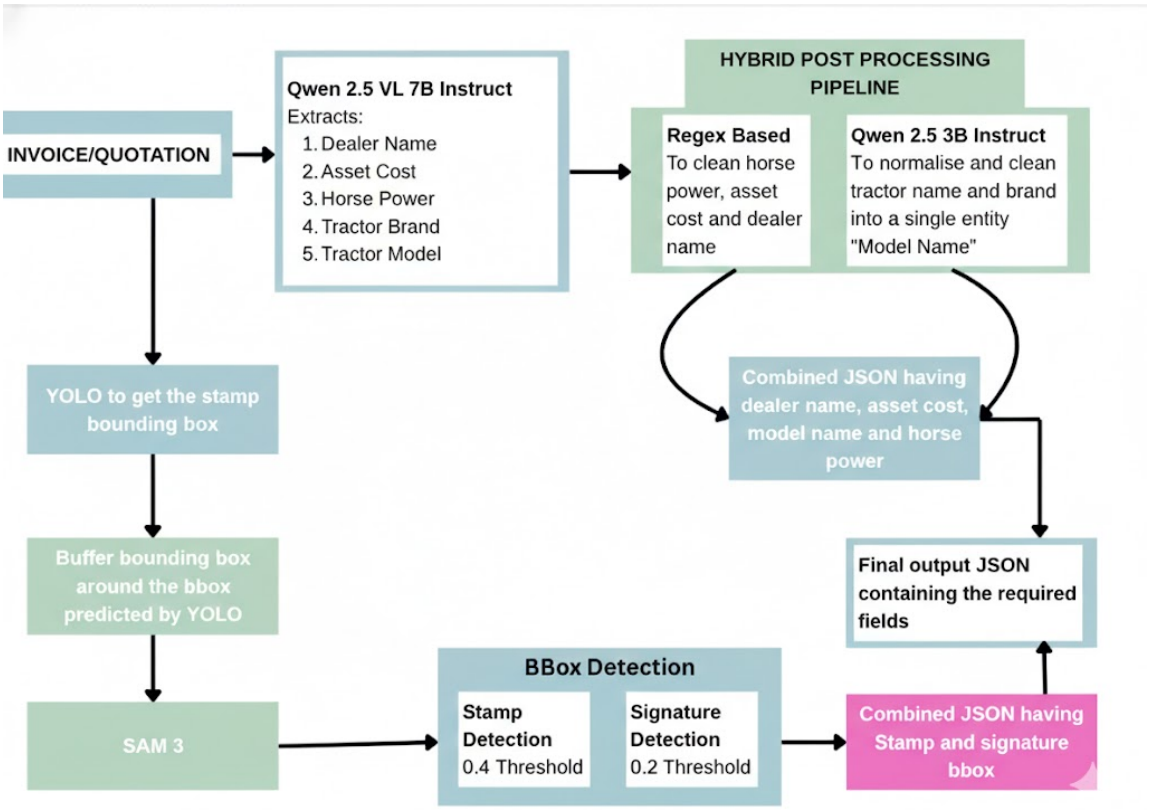

Figure 3: Overview of the Hybrid Extract-Refine-Segment (ERS) Architecture for Tractor Invoice Processing. The pipeline integrates Vision-Language Models for semantic extraction with a high-precision Computer Vision branch for geometric attestation, fusing both modalities for robust information retrieval.

information extraction with the strict latency and memory constraints of a single NVIDIA T4 GPU. The ERS framework decomposes the complex task into three primary, interconnected stages: semantic data extraction, data refinement and standardization, and visual verification of key artifacts.

5.1 Semantic Intelligence Branch (Qwen2.5-VL + 3B Stack)

The primary extraction is handled by a tiered Vision-Language Model (VLM) stack, optimized for sparse and semi-structured Indian document layouts.

- **Global Extraction (7B Model):** We utilize **Qwen2.5-VL-7B** quantized to **4-bit (NF4)**. This quantization enables the model to maintain a large KV-cache for high-resolution document processing within a 16GB VRAM envelope. The model is strictly prompted to extract data in its native script (Hindi, Marathi, etc.) to prevent translation-induced hallucinations.
- **Tiered Refinement (3B + Rule Engine):** Raw outputs are passed to a secondary **Qwen2.5-3B-Instruct** refiner. This stage performs **Brand Expansion** (e.g., TAFE $\rightarrow$ Massey Ferguson), **Nomenclature Unification** (merging brand and model), and **OCR De-noising** (removing technical suffixes like "BS-IV").
- **Deterministic Cleaning:** A final hybrid layer uses **Regular Expressions** to normalize numeric strings. This includes word-to-number conversion for costs (e.g., "Seven Lakh" $\rightarrow$ 700,000) and range-validation for Horse Power ($15 \leq HP \leq 75$) and noise cleaning like 45 H.P $\rightarrow$ 45.

5.2 Visual Verification Branch (YOLOv8 + SAM3)

A high-precision computer vision branch handles geometric attestation through a "Detect-then-Segment" workflow, optimized for high-resolution artifact extraction.

- **ROI Detection and Dynamic Buffering:** A fine-

tuned **YOLOv8** model identifies coarse Regions of Interest (ROIs). To account for the potential truncation of stamps or signatures at the bounding box edges, we implement a **1.75x dynamic buffer**. This expansion ensures that the local context—such as faint ink bleeds or border-less signatures—is fully captured before the segmentation stage.

- **Precision Segmentation via Dynamic Cropping:** Unlike standard SAM3 implementations that utilize bounding box prompts on a full-page downsampled image, our pipeline utilizes **Dynamic Cropping**. We extract high-resolution crops from the buffered ROIs and pass them to **SAM3** as primary inputs. Our empirical observations indicated that ***"Naked Inference"* on a high-res crop significantly outperforms *"Box-Prompted Inference"* on a full page; the latter often "traps" the detection within the coarse YOLO box, whereas cropping allows SAM3 to re-evaluate the object boundaries with superior pixel density.**

- **Differentiated Thresholding:** We implement class-specific confidence thresholds to optimize the F1-score across varying ink densities:
  - **Stamp (0.4):** A higher threshold is utilized for stamps to filter out dense printed text or dark logo elements that SAM3 might misidentify as a solid geometric artifact.
  - **Signature (0.2):** A lower threshold is applied to signatures to accommodate "thin-ink" cursive or faded handwritten strokes, ensuring connectivity in the segmentation mask even when pixel intensity is low.

- **Heuristic Recovery Logic:** To ensure system robustness, we implement a **YOLO-Fallback** mechanism. If SAM3 fails to generate a mask due to extreme lighting or mimicry (where a stamp appears like standard text), the system reverts to the original YOLO Oriented Bounding Box (OBB). This ensures that visual attestation is never omitted purely due to segmentation failure.

5.3 Fusion and Latency Analysis

The final stage of the pipeline merges the semantic outputs from the VLM branch with the geometric attestations from the CV branch into a unified JSON schema. To provide a reliability metric for downstream applications, we implement a **Composite Confidence Score** (C_{total}), which balances the linguistic and visual certainty of the extraction.

- **VLM Confidence (C_{qwen}):** The confidence for factual extraction is derived directly from the model's output distribution. For each generated token t_i in the JSON response, we extract the log-probability of the selected token relative to the vocabulary distribution. C_{qwen} is calculated as the mean of the softmax-normalized logits across the entire generated sequence:

$$C_{qwen} = \frac{1}{N} \sum_{i=1}^{N} \text{softmax}(\text{logits}_{t_i}) \tag{1}$$

- **Artifact Confidence (C_{sam}):** The visual attestation score is calculated by averaging the detection and segmentation probabilities for both the stamp (s) and signature (sig). This accounts for cases where the model might detect an object but struggle to define its precise boundary:

$$C_{sam} = 0.5 \times (\text{Score}_{stamp}) + 0.5 \times (\text{Score}_{signature}) \tag{2}$$

  Each individual artifact score is the result of the SAM3 mask probability score. In the event of a YOLO-fallback, the original YOLO detection confidence is utilized as a proxy.

The final composite score is a weighted aggregation, prioritizing the semantic extraction which forms the core of the commercial data:

$$C_{total} = (C_{qwen} \times 0.6) + (C_{sam} \times 0.4) \tag{3}$$

The total latency (L_{total}) aggregates the tiered execution across all asynchronous and sequential modules:

$$L_{total} = T_{qwen7b} + T_{refine3b} + T_{yolo} + T_{sam3} \tag{4}$$

By managing the GPU memory lifecycle—specifically deleting the 7B model and clearing the cache before initializing the SAM3 stack—we maintain stability on the 16GB VRAM limit. With an average L_{total} of 29s on an **NVIDIA T4**, the operational cost is approximately **$0.0016** per document, providing a high-accuracy, cost-efficient production-ready solution.

6 Handling Lack of Ground Truth

As no pre-labelled data was provided for this challenge, we implemented a hybrid annotation strategy to approximate ground truth for validation and fine-tuning [**?**].

6.1 Semantic Field Bootstrapping via LLM

For the extraction of *Dealer Name, Model Name, Horse Power,* and *Asset Cost*, we utilized a **Gemini 3 Flash** bootstrapping approach:

- **Prompt Engineering:** A carefully designed system prompt was used to guide the API in extracting structured JSON fields directly from document images.
- **Manual Verification:** To ensure the integrity of the pseudo-labels, we performed a manual audit of 50 randomly selected documents. This verification confirmed that the LLM-generated data was accurate enough to serve as a high-fidelity ground truth for our primary Qwen2.5-VL pipeline.

6.2 Specialized Visual Marker Annotation

To achieve the required IoU threshold for signatures and stamps, we manually annotated a curated dataset of **63 challenging images** using **Label Studio.** This set was specifically designed to cover high-variance edge cases:

- **Occlusions and Overlaps:** Cases where stamps were placed directly over text or fully overlapped with signatures.
- **Visual Ambiguity:** Documents featuring both customer and dealer signatures, requiring precise role-based distinction.
- **Geometric Diversity:** A representative mix of circular stamps and rectangular box-style stamps.
- **Quality Degradation:** Targeted inclusion of documents with extremely dull stamp intensity or low-contrast signatures.
- **Layout Inconsistency:** Scenarios with non-standard placement (e.g., signature on the right vs. stamp on the left).

This curated ground truth enabled precise calibration of our YOLOv8 detection and SAM3 segmentation branches, ensuring robust performance across diverse layouts.

7 Results and Performance Analysis

The evaluation of the entity extraction pipeline was conducted using a multi-staged approach. Primary extraction was performed by the **Qwen2.5-VL-7B** vision-language model, followed by a hybrid post-processing layer. This layer utilized deterministic **RegEx** heuristics coupled with a **Qwen2.5-3B** semantic refiner to standardize nomenclature and numeric strings.

To validate the system, performance was measured against a ground truth dataset generated via high-fidelity bootstrapping using **Gemini 3 flash API**. The categorical text fields (*Dealer Name* and *Model Name*) were evaluated using fuzzy string matching across thresholds of 70%, 80%, and 90%, while numeric fields (*Horse Power* and *Asset Cost*) were assessed based on a $\pm 5\%$ accuracy range.

The final results demonstrate exceptional stability in the extraction of financial data, with the *Asset Cost* achieving 100% accuracy within the specified tolerance. As summarized in Table 3, the pipeline maintains a high overall accuracy across all fuzzy strictness levels, achieving a peak composite score of **91.225%** at the 70% threshold and maintaining a robust **86.28%** even under the most stringent 90% matching criteria. This underscores the effectiveness of the hybrid Qwen-7B/3B refinement strategy in resolving OCR noise and nomenclature variations.

7.1 Stamp and Signature Evaluation

To evaluate the geometric attestation branch, we performed a manual ground-truth annotation of 63 document images containing varied stamp and signature placements. The performance was assessed using the **Intersection over Union (IoU)** metric, where a successful detection is defined by an $IoU \geq 0.5$ between the predicted oriented bounding box (OBB) and the ground truth.

As summarized in the validation results:

- **Stamp Detection:** The pipeline successfully identified and segmented the stamp in **55 out of 63** cases (87.3% accuracy).
- **Signature Detection:** The pipeline successfully extracted signatures in **52 out of 63** cases (82.5% accuracy).

A qualitative review of the failure cases revealed a critical characteristic of the "Detect-then-Segment" architecture. In all instances where the pipeline failed to record an artifact, the error was attributed to **complete model omission** rather than geometric imprecision. Specifically, the YOLOv8 ROI detector either entirely bypassed the region or failed to trigger a proposal due to extreme occlusion or ink transparency.

Crucially, in every case where the YOLOv8 model provided a successful Region of Interest, the subsequent SAM3 segmentation and OBB generation achieved high spatial fidelity. This indicates that the system's bottleneck is localized to the initial detection stage; once an artifact is "seen," the dynamic cropping and precision segmentation logic produces a highly accurate bounding box that exceeds the 0.5 IoU threshold.

Table 3: Entity Extractor Performance Metrics

Field Name	Accuracy @70%	Accuracy @80%	Accuracy @90%
Dealer Name (Fuzzy)	98.40%	97.78%	95.20%
Model Name (Fuzzy)	88.50%	81.30%	71.90%
Horse Power (±5%)	78.00%	78.00%	78.00%
Asset Cost (±5%)	99.79%	99.79%	99.79%
Overall Accuracy	**91.225%**	**89.27%**	**86.28%**

8 Latency and Cost Trade-off

By offloading standardization to a specialized 3B-Instruct model, we reduced the computational overhead of the primary Vision-Language model. The average latency computed across a benchmark of 50 documents was **24 seconds** on standard deployment hardware.

8.1 Hardware Benchmarking

The pipeline was tested across two GPU configurations to analyze the trade-off between inference speed and operational expenditure:

- **NVIDIA T4 (16 GB VRAM):** Achieved an average latency of 24 seconds with a per-document cost of **$0.001625**.
- **NVIDIA A5000 (24 GB VRAM):** Drastically reduced total time to **8 seconds**. However, the per-document cost increased to **$0.002125**, representing a ˜1.4x increase in cost relative to the T4 baseline.

8.2 Module Latency Distribution

The end-to-end processing time is dominated by the Vision-Language reasoning task. The temporal breakdown is as follows:

$$T_{total} = T_{extraction} + T_{visual} + T_{norm} \tag{5}$$

On average, the **Factual Tag Extraction** (7B-VL) accounted for **16.92s**, while the **Stamp and Signature Extraction** (YOLO+SAM 3) required only **5.13s**. The remaining time was allocated to the 3B-Instruct normalization and rule-based regex cleaning.

9 Error Analysis

We conducted an exhaustive error analysis to identify the boundary conditions of our hybrid pipeline. The failure modes are categorized into linguistic confusion, spatial association errors, and visual mimicry.

9.1 Linguistic and Entity Confusion

The model demonstrates a "Brand Bias" during the extraction of `dealer_name` in regional scripts. In approximately 18–20 documents featuring **Gujarati** and **Kannada**, the system fails to distinguish the local dealer from prominent manufacturer branding, frequently returning "Mahindra & Mahindra" or "Swaraj" as the dealer name when the local entity is written in a non-Latin script.

9.2 Spatial Logic and Handwritten OCR Failures

The coupling of models and specifications encountered the following issues:

- **Tick-Mark Association:** In 18 test cases, the model failed to link the selection mark (✓) to the correct horizontal row due to tight vertical spacing.
- **Handwriting Ambiguity:** Cursive inputs led to grapheme confusion (e.g., I vs F).
- **HP Proximity Error:** If the `horse_power` entry was illegible, the model occasionally hallucinated the value by extracting the nearest legible numeric string, which often belonged to the `asset_cost` column.

Hardware	Avg. Latency (s)	Avg. Cost/Doc ($)
NVIDIA T4 (Baseline)	24.0	0.001625
NVIDIA A5000	8.0	0.002125

Table 4: Comparison of Latency and Cost across GPU Architectures.

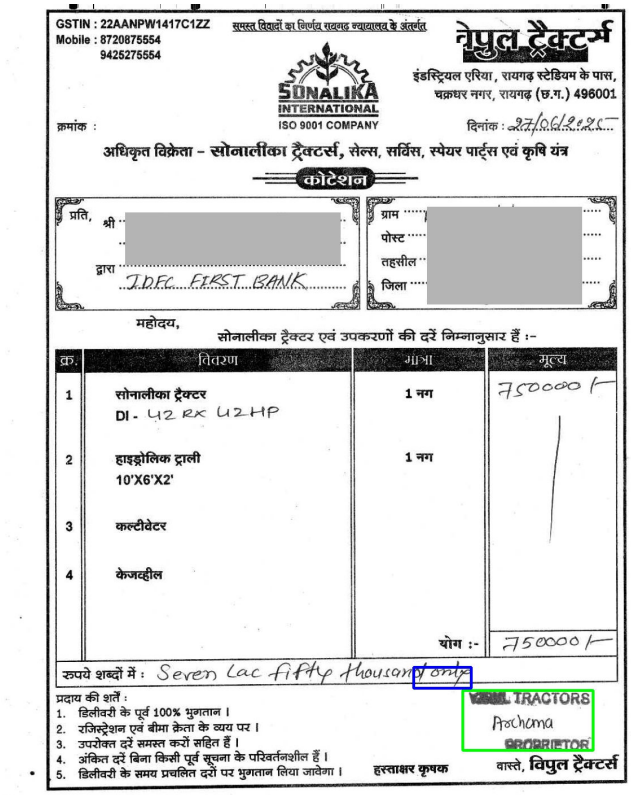

Figure 4: Visual Evidence of Extraction Performance: Comparison between baseline noise and refined model output.

9.3 Stamp and Signature Omissions

The primary failure mode for the **SAM 3** pipeline occurs when stamps or signatures lack distinct geometric borders. Text-heavy stamps (e.g., "Authorized Signatory") are frequently classified as standard document text by the YOLO backbone, making it difficult for SAM 3 to differentiate it from normal text.

10 Limitations and Future Directions

Despite the robust performance of the ERS pipeline, several technical constraints were identified during the development and testing phases, particularly under the resource-restricted environment of the Convolve 4.0 challenge.


```
{
    "doc_id": "90018824033_175118795_2_pg33",
    "fields": {
      "dealer_name": "Mahindra & Mahindra Ltd.",  0
      "model_name": "Mahindra 275NBP",
      "horse_power": 7.0,
      "asset_cost": 730115
    }
  }
```


Figure 5: JSON output demonstrating proximity error.

10.1 Current Limitations

The primary bottleneck for the system is the strict **16GB VRAM constraint**, which limits the choice of Vision-Language Models (VLMs) to those that can be efficiently quantized to 4-bit or 8-bit precision. While the **Qwen2.5-VL-7B** model offers a superior balance of reasoning and efficiency, it exhibits the following limitations:

- **Grounding and Spatial Reasoning:** VLMs frequently struggle with precise "grounding"—the ability to map textual entities to exact pixel coordinates. This is particularly evident when interpreting visual selection markers such as tick marks (✓), circles, or underlines used to denote specific tractor variants in dense technical tables.

- **Complex Layouts and Handwriting:** In documents where handwritten entries overlap with printed headers or official stamps, the model occasionally suffers from "attention drift," leading to minor hallucinations or field-swapping.

- **Latency-Parameter Trade-off:** To maintain a per-document latency of < 30 **seconds**, we utilized smaller parameter models. While fast, these models lack the high-order semantic reasoning found in 70B+ parameter models, which are better equipped to handle the deep vernacular nuances of regional Indian invoices.

10.2 Future Directions

To transition this research into a production-grade banking tool, future iterations will focus on the following architectural enhancements:

- **Specialized Grounding Architectures:** Development of a "Geometry-Aware" VLM that integrates a dedicated spatial-attention head. This would allow the model to better resolve the relationship between visual ticks and their associated text fields without increasing the overall parameter count.
- **Parameter-Efficient Fine-Tuning (PEFT):** Implementing **QLoRA** (Quantized Low-Rank Adaptation) on the 7B model using a domain-specific dataset of Indian tractor invoices. This would enable the model to learn the specific nomenclature and layout patterns of regional dealers, reducing the reliance on general-purpose semantic reasoning.
- **Hybrid Multimodal Segmentation:** Integrating a more advanced segmentation branch that uses a specialized OCR-free transformer to interpret handwritten annotations as separate visual layers, thereby preventing "noise contamination" during the primary extraction phase.

By addressing these grounding and layout challenges within the 16GB VRAM envelope, the next generation of the ERS pipeline can achieve human-level reliability in unconstrained document intelligence.

11 Conclusion

The proposed hybrid pipeline successfully addresses the constraints of multilingual script variability and visual localization in Indian invoices. By integrating **Qwen 2.5-VL** for factual extraction with **SAM 3** for spatial refinement, our approach achieved a considerably higher accuracy than traditional monolithic methods.

Benchmarking across 50 documents revealed an average latency of **29.0s** on an **NVIDIA T4** ($0.001625/doc) and **9.0s** on an **NVIDIA A5000** ($0.002125/doc). While the A5000 represents a ~1.4x increase in cost, it provides a 3x throughput improvement. Detailed profiling shows the **Tag Extraction** module requires **16.92s**, whereas **Stamp/Signature Extraction** requires **5.13s**.

Future scope involves the segmentation of low-contrast artifacts and the implementation of database cross-referencing to eliminate rare LLM hallucinations. Furthermore, transitioning to **TensorRT-LLM** is expected to bring A5000-level latency to lower-tier hardware.

Table 5: Performance and Cost Trade-off Comparison

Hardware	Latency (s)	Cost ($)	Speedup	Cost Multiplier
NVIDIA T4	24.0	0.001625	1.0x	1.0x
NVIDIA A5000	8.0	0.002125	3.2x	1.31x