ICDAR 2025 Handwritten Notes Understanding Challenge

Aniket Pal[§1], Sanket Biswas[§2], Alloy Das[§3], Ayush Lodh[§3], Priyanka Banerjee[§3], Soumitri Chattopadhyay[4], Ajoy Mondal[1], Dimosthenis Karatzas[2], Josep Lladós[2], C.V. Jawahar[1]

[1] CVIT Lab, IIIT Hyderabad, India
[2] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[3] Habitat Labs, Habitat Lens Pvt. Ltd., India
[4] UNC Chapel Hill, USA

Abstract. This report presents the results of the inaugural ICDAR 2025 Handwritten Notes Understanding Competition, centered on evidence-based question answering over complex scientific handwritten notes. The final competition test set included over 2,000 real academic note images and 1,000 curated questions across diverse STEM domains, characterized by varied layouts, diagrams, and dense equations. The overall challenge was structured as a single task with a multi-phase evaluation protocol designed to assess *not only answer correctness but also grounded reasoning ability*, requiring models to localize the exact visual evidence supporting their answers. Hosted on the Robust Reading Challenge (RRC) Portal, the competition ensured fair, private-set benchmarking, with 6 valid submissions out of 16 registered teams. The results underscore the obstacles that current state-of-the-art (SOTA) vision-language models (VLMs) face in multimodal reasoning over handwritten content. This first edition marks a promising step toward interpretable, layout-aware document understanding and sets the foundation for future iterations.

Keywords: Handwritten Document Understanding · Multimodal Reasoning · Evidence-Based Question Answering · Vision-Language Models

1 Introduction

Handwritten notes-those scribbled pages filled with equations, doodles, arrows, and underlined concepts, have long been silent witnesses of deep learning and discovery in classrooms, labs, and research desks. Whether outlining the trajectory of a falling apple or mapping a neural network on paper, these notes capture the cognitive process of learning in its rawest form. But as education and research increasingly go digital, the ability to read, understand, and reason over these informal and richly visual documents remains far beyond the reach of even today's most powerful AI systems.

The field of Document Analysis and Recognition (DAR) has evolved into a multimodal discipline, integrating applications from both Natural Language Processing (NLP) and Computer Vision (CV). This convergence has led to the development of comprehensive Document Understanding (DU) systems capable of interpreting Visually Rich Documents (VRDs), where both textual and layout information are pivotal. DU encompasses key subtasks such as key-value information extraction (KIE) [10], layout analysis (DLA) [21], document visual question answering (VQA) [13,18], scene text VQA [3], infographic understanding [12] table understanding, and mathematical expression recognition [1,9,15]. Previous benchmarks—like ICDAR 2019 Scene Text VQA and ICDAR

[§] These organizers contributed equally to the challenge.

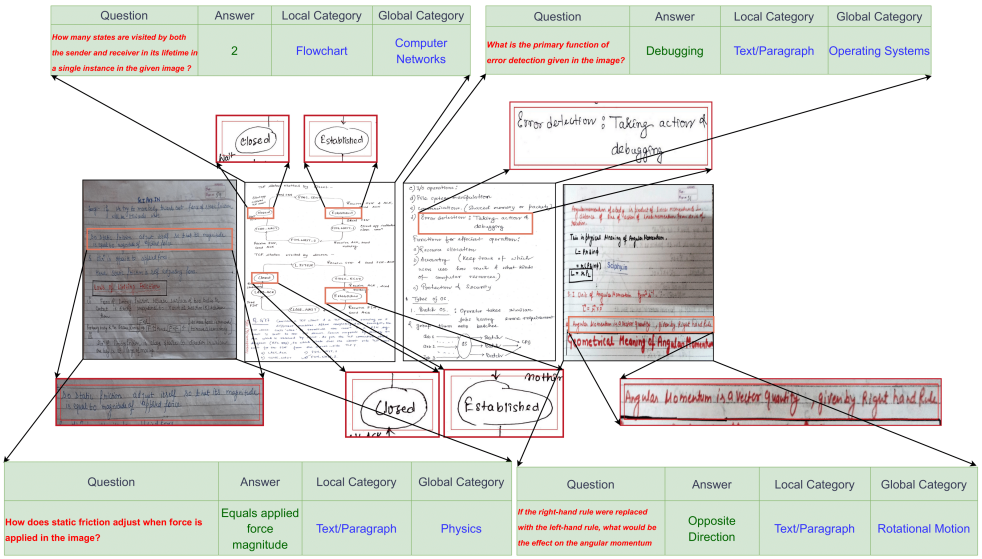

Fig. 1: **Representative examples from the ICDAR 2025 Handwritten Notes Understanding (HNU) test set.** Each example shows a natural language question grounded in a scientific note, alongside its corresponding answer, local evidence region (e.g., text fragments, flowchart, boxed formulas), and the annotated global category labels (e.g., subject domains like Rotational Motion). This highlights the complexity of the multimodal reasoning task, requiring models to localize specific visual evidence (e.g., boxed formulas, diagrams, text fragments) across diverse STEM domains.

2021 DocVQA—have driven significant progress in this domain [3,18,13]. Despite these advancements, handwritten scientific documents remain a largely unmet challenge. Their inherent variability, complex visual structures, and domain-specific multimodality demand more than OCR-based pipelines or template-driven approaches. The HNU Challenge confronts this gap directly, offering a high-impact testbed to develop and evaluate next-generation document AI models capable of fine-grained reasoning and layout-aware understanding inspired by the philosophy of ICDAR 2023 DUDE competition [19]. We believe this effort will catalyze progress in both academia and industry, fostering models better suited to low-resource, high-complexity settings that mirror real-world scientific workflows.

The ICDAR 2025 Handwritten Notes Understanding (HNU) Competition was conceived to directly confront this challenge. Organized as the first edition of its kind, this competition invited participants to develop systems capable of evidence-based question answering over real-world handwritten notes—a significantly harder task than existing document understanding (DU) challenges related to document visual question answering [13,12,18,19], majorly due to unstructured layouts, diverse visual modalities, and domain-specific shorthand that characterize such documents. At the core of this challenge lies a novel benchmark evaluation "golden dataset" of 1,000 carefully curated question-answer (QA) pairs grounded in 2,000 pages (images) of authentic handwritten scientific notes. These notes span a wide spectrum of STEM domains, including mathematics (e.g., differential equations, calculus, geometry), physics (e.g., fluid mechanics, rotational motion), chemistry (e.g., group theory, organometallics), computer science (e.g., DBMS, compiler design), and various branches of engineering. The dataset is distinguished by its high complexity as shown by representative examples in Figure 1—featuring dense mathematical equations, tables, graphs, charts, and rich visual

annotations such as strikethroughs—demanding robust multimodal reasoning from participating models. Other representative domain-specific datasets launched in the past include chemical structures [4], handwritten tables [8] and the Handwritten VQA competition for extractive-only questions [14].

1.1 Challenge Objectives

We aim to foster the development of models with strong reasoning abilities tailored to the unique challenges of handwritten scientific notes. The ICDAR 2025 Handwritten Notes Understanding Challenge is the first to introduce evidence-grounded visual question answering in this domain, requiring models not only to answer questions but to also localize the exact visual content—such as equations, diagrams, or tables—that supports their responses. Our objectives are:

- **Advance multimodal reasoning over unstructured handwritten layouts** Stimulate research into understanding complex handwritten documents where text, mathematical expressions, diagrams, and tables are intermingled without a fixed reading order. This includes integrating visual information from flowcharts, graphs, and figures with natural language for accurate question answering.
- **Improve fine-grained recognition and grounding of scientific visual elements** Drive progress in identifying and interpreting handwritten formulas, scientific symbols, and structured visual components in their layout context. Models are expected to output bounding boxes of relevant evidence and classify both local (e.g., table, formula) and global (e.g., physics, chemistry) content categories.
- **Benchmark model performance on real-world handwritten artifacts** Evaluate systems on challenging handwritten data that includes strikethroughs, shorthand, variable writing styles, and layout inconsistencies—typical features in authentic academic notes.
- **Promote generalization across domains and document styles** Assess model robustness and adaptability using a diverse dataset covering multiple STEM disciplines (e.g., physics, computer science, mathematics) and a wide variety of writing conventions and formats.

By structuring the task as a DocVQA-style competition with an emphasis on **visual evidence localization**, this challenge provides a realistic and interpretable benchmark for future research in document intelligence. It aims to attract interest from the broader Document AI, NLP, and Vision-Language communities, fostering innovation in layout-aware learning, handwritten content analysis, and explainable reasoning in scientific contexts.

1.2 Challenge Contributions

The ICDAR 2025 Handwritten Notes Understanding Challenge tackles a critical frontier in Document AI: `the interpretation of complex, real-world handwritten scientific material`. By introducing a task that requires not only answering questions but also localizing the supporting visual evidence within dense and unstructured notes, the challenge pushes participants beyond traditional document understanding pipelines. The dataset's intrinsic difficulty—featuring handwritten equations, diagrams, tables, and text interwoven without a predictable structure—necessitated significant innovation in both architectural design and reasoning strategies. Several key trends emerged from the submitted solutions:

- **Custom Pipelines for Handwritten Layout Parsing** Given the irregular structure of handwritten notes, participants designed pipelines that could segment and relate visual elements such as equations, diagrams, and text blocks. These pipelines often replaced standard document models that assume cleaner layouts.
- **Visual Grounding for Evidence Localization** To meet the requirement of evidence-backed answers, many approaches incorporated visual grounding techniques. Some teams used segmentation models like Grounded SAM [16], combined with prompts or region proposals, to detect and highlight the source of the answer on the page.
- **Prompt-Based Answer Categorization and Extraction** Teams employed structured prompts to guide models in extracting specific types of information, such as formulas, definitions, or values from diagrams. These prompt strategies helped models handle a wide range of question types across scientific domains.

Overall, the challenge encouraged targeted methods for layout parsing, visual grounding, and multimodal reasoning in handwritten contexts. These approaches reflect progress in making document understanding systems more accurate, explainable, and better suited to real-world handwritten content.

2 Novelty

Advancing Document Understanding (DU) requires more than incremental gains in computer vision and natural language processing. It calls for tackling underrepresented, real-world document types that expose current system limitations. Handwritten scientific notes are a prime example: structurally complex, multimodal, and rarely addressed in standard benchmarks. Most existing datasets cater to printed documents with regular layouts. In contrast, handwritten academic notes often feature:

- **Highly unstructured layouts** mixing free-form text, mathematical derivations, diagrams, tables, and symbolic notations.
- **Multi-domain coverage** spanning mathematics (e.g., differential geometry), physics (e.g., fluid mechanics), chemistry, computer science, and engineering.
- **Visual irregularities** including varied handwriting styles, strikethroughs, annotations, and dense visual content distributed across multi-page formats.

Such material presents challenges that go beyond text extraction and OCR—they demand deep semantic comprehension grounded in visual context. To address this gap, the ICDAR 2025 Handwritten Notes Understanding Challenge introduces the following:

- A novel benchmark of 2,000 pages of handwritten notes and 1,000 expert-authored QA pairs focused on evidence-based question answering.
- A formulation as `Evidence-Based DocVQA`, requiring not just answers but also grounding them in the specific regions of visual content (equations, diagrams, etc.).
- A focus on generalization, encouraging models that adapt to unseen domains, layouts, and scientific content with limited supervision.

The competition promotes models that can: (1) Understand and navigate highly variable handwritten layouts. (2) Reason across multiple modalities (text + visual elements) to answer domain-specific questions. (3) Identify and localize visual evidence to support interpretability. (4) Operate effectively in few-shot or zero-shot settings, minimizing dependence on task-specific annotation. By setting these objectives, the challenge creates a structured and practical path toward robust handwritten DU systems. It offers a reproducible benchmark and invites the community to build solutions that reflect the real complexities of handwritten scientific content.

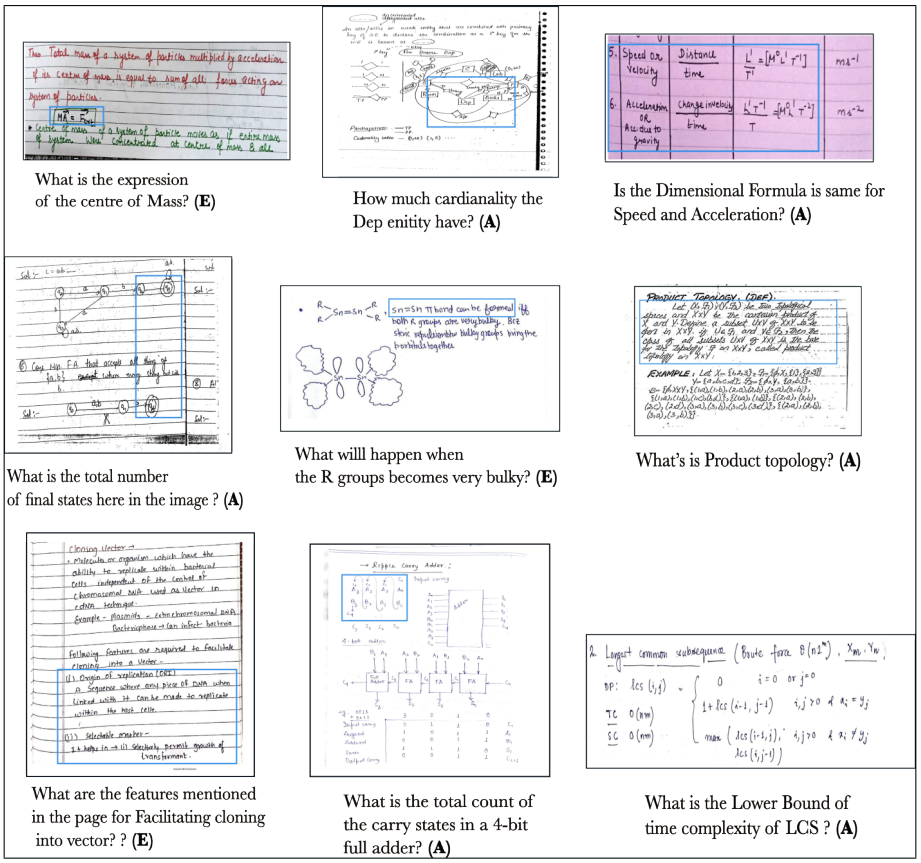

Fig. 2: **Example Questions with (E)xtractive and (A)bstractive answers for particular layout semantics under consideration.** Unlike prior benchmarks focused on extractive QA [14], our dataset is the first to introduce open-ended, multi-domain reasoning questions requiring interpretation of diagrams, equations, and layout elements, pushing beyond OCR-based extraction toward genuine understanding. The last example depicts the 'None' type question.

3 The ICDAR HNU Competition Dataset

The ICDAR 2025 Handwritten Notes Understanding (HNU) Challenge is grounded on a newly constructed dataset designed specifically to evaluate models on complex, multi-modal handwritten content. The dataset reflects a wide array of scientific and technical disciplines, including mathematics (e.g., differential equations, calculus), physics (e.g., fluid mechanics), chemistry (e.g., group theory), computer science (e.g., DBMS, compiler design), and core engineering domains as shown in Figure 3. To ensure real-world diversity, handwritten materials were sourced from public educational repositories (e.g., IIT-JEE, GATE, SSC prep in India, and GRE resources), along with notes collected from undergraduate and engineering courses. The documents exhibit significant variation in handwriting styles, ink types, page textures, and scanning quality—including cases with poor lighting and alignment—resulting in a dataset that captures the visual and structural complexity often absent in synthetic or printed document benchmarks.

The released dataset contains 1,000 expert-annotated question–answer (QA) pairs linked to over 2,000 pages of handwritten academic notes. Each QA pair is grounded

in specific regions of the page through bounding boxes, and includes metadata for fine-grained reasoning evaluation. Questions span a variety of information types—definitions, formula retrieval, diagram interpretation—and are designed to reflect both local content structure and broader scientific context. The annotation process was conducted in **four** structured stages:

- **Document Collection and Pre-filtering:** The dataset included handwritten notes sourced from publicly available educational repositories, including websites that host preparation materials for competitive examinations such as IIT-JEE, GATE, SSC, and GRE. Permission to use these notes was obtained from content providers prior to inclusion. Documents were filtered for quality, and low-resolution scans, multilingual content, or poorly structured notes were excluded. This curation ensures high-quality handwritten inputs with varied styles, layouts, and page formats (e.g., ruled, unruled, multi-color ink).
- **QA Generation and Evidence Grounding:** Domain experts mainly included STEM engineering students (JEE/GATE entrance exam qualified) formulated high-quality QA pairs across diverse domains. Each QA pair was grounded by annotating bounding boxes over the document region(s) containing the visual evidence. Questions were verified iteratively, with authors providing guidelines and supervision. Answers involving formulas were encoded in LaTeX to preserve semantic and typographic fidelity.
- **Validation and Metadata Annotation:** Local and global categories (e.g., 'Diagram', 'Math Equation', 'Physics') were assigned. Errors and inconsistencies were resolved in multiple verification rounds. Final annotations were exported to JSON format for distribution.
- **Final Quality Assurance Check:** The annotation team consisted of 15–20 trained STEM personnels, and all examples passed through multiple rounds of verification. A Quality-Assurance Team of 6 domain experts combined and generated the final valid annotations. All annotations underwent a final quality assurance pass by the ICDAR 2025 HNU Organizing Team. The dataset was initially maintained in CSV format and later converted into structured JSON files for public release, ensuring inter-operability with standard document AI pipelines.

4 Competition Evaluation Protocol

The ICDAR 2025 Handwritten Notes Understanding Competition was conducted between February and May 2025. The final test set, consisting of 1,000 questions across 2,000 handwritten documents, was made available to participants during a restricted window from April to May. Submissions were required in the form of results on a blind public test set, rather than executable models, though participants were encouraged to open-source their implementations. The evaluation was hosted on the `Robust Reading Competition (RRC) portal`[5], and participants were expected to comply with the competition rules in good faith, in line with the scientific integrity policy of the platform.

4.1 Task Formulation

The central task of this competition challenges participants to build systems capable of answering questions grounded in specific pages of real-world handwritten scientific

[5] https://rrc.cvc.uab.es/?ch=33&com=introduction

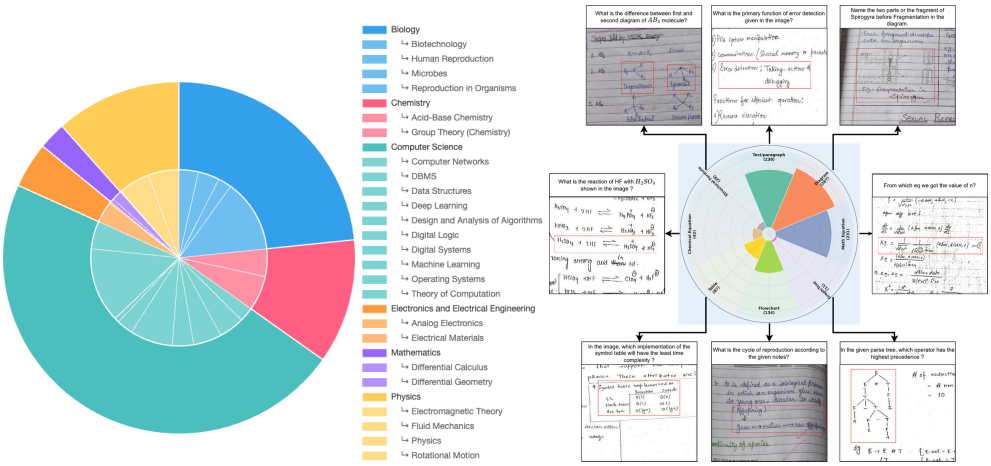

Fig. 3: **Distribution of domain coverage and content categories in the HNU dataset. (Left):** Hierarchical sunburst chart showing the global subject areas (e.g., Physics, Chemistry, Computer Science) and their subtopics (e.g., DBMS, Fluid Mechanics, Reproduction in Organisms), demonstrating the dataset's breadth across STEM fields. **(Right):** Visualization of local content types annotated in the dataset (e.g., flowcharts, mathematical expressions, tables), alongside example questions and bounding box-highlighted evidence regions that reflect the multimodal complexity of the task.

documents. Each document typically spans 3–4 pages and includes densely packed content—text, equations, diagrams, and tables—often requiring cross-referencing across sections. The task goes beyond simple information retrieval; it requires models to comprehend complex handwritten content, interpret contextual cues, and locate the exact visual evidence that justifies their answers.

Questions are crafted to evaluate multiple levels of understanding:

- **Information extraction** of factual or numerical content
- **Contextual synthesis** across multiple handwritten elements
- **Robustness to handwriting variability**, including diverse styles, layouts, and notational conventions

Answers may range from short spans (e.g., numeric values or terms) to more elaborate text synthesized from multiple visual cues. Crucially, all answers must be explicitly supported by evidence visible within the document pages—reinforcing the task's emphasis on traceable and explainable reasoning

4.2 Evaluation Metrics

The primary evaluation phase operates under the assumption of independently and identically distributed (i.i.d.) data. The training, validation, and test splits in the HNU dataset maintain a consistent distribution of document types, layout structures, and question-answer characteristics. Submissions were evaluated through the Robust Reading Challenge (RRC) platform, which offers real-time leaderboards and interactive visualization tools adapted to handle multi-page PDF inputs. To assess the correctness of

model-generated answers across diverse formats—ranging from plain text to symbolic expressions—we adopt the `Average Normalized Levenshtein Similarity (ANLS*)` as our primary evaluation metric. This variation of the standard ANLS is adapted to handle scenarios where the predicted or ground truth answer may be `None`. The Normalized Levenshtein Similarity (NLS) between a predicted answer s_{pred} and a ground truth answer s_{gt} is defined as:

$$\text{NLS}(s_{\text{pred}}, s_{\text{gt}}) = 1 - \frac{L(s_{\text{pred}}, s_{\text{gt}})}{\max(|s_{\text{pred}}|, |s_{\text{gt}}|)} \tag{1}$$

where $L(\cdot)$ denotes the Levenshtein edit distance and $|\cdot|$ is the string length. The ANLS* score modifies this definition as follows:

- If both s_{pred} and s_{gt} are `None`, then $\text{NLS}^* = 1$.
- If only one of s_{pred} or s_{gt} is `None`, then $\text{NLS}^* = 0$.
- Otherwise, $\text{NLS}^* = \text{NLS}(s_{\text{pred}}, s_{\text{gt}})$.

The final ANLS* score is computed as the mean of all NLS^* values across the test set. To evaluate **evidence localization**, we use the `Intersection over Union (IoU)` metric between the predicted and annotated bounding boxes:

$$\text{IoU}(B_{\text{pred}}, B_{\text{gt}}) = \frac{\text{Area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{Area}(B_{\text{pred}} \cup B_{\text{gt}})} \tag{2}$$

Higher IoU values indicate better alignment of the predicted bounding region with the ground truth evidence. For **category classification**, we report accuracy scores for both *local* (e.g., flowchart, diagram, text) and *global* (e.g., physics, chemistry, computer science) answer categories:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}} \tag{3}$$

Predictions are marked correct if the predicted category matches the annotated ground truth category. All submissions had to follow the prescribed JSON format strictly. Submissions not conforming to this format were automatically assigned a zero score across all metrics. To support transparency and ease of development, we provided the complete set of metric implementations and evaluation scripts in a public repository, enabling participants to validate their models locally before final submission.

5 Competition Results and Analysis

5.1 Submitted Methods

Among the sixteen teams that registered for the ICDAR 2025 Handwritten Notes Understanding Challenge, six submitted valid runs for the evidence-based VQA task. All participating systems leveraged large-scale VLMs or hybrid pipelines that combined OCR with LLMs. These approaches were characterized by carefully designed prompting strategies to produce structured outputs across the multiple sub-tasks of the competition. Table 1 provides a concise summary of the methods.

A common architectural choice across several submissions was the use of advanced VLMs such as Qwen-2.5-VL and InternVL, both of which have shown strong performance in multimodal reasoning tasks. Qwen-2.5-VL, developed by Alibaba, combines a visual encoder with a powerful language model, enabling end-to-end multimodal reasoning. Its architecture supports detailed analysis of both image regions and text, making it well-suited for evidence-grounded VQA in complex handwritten documents. Participants

Table 1: **Descriptions of methods submitted to the ICDAR 2025 HNU competition.** Only the latest submission per team is considered for final ranking.

	ORGANIZING TEAM
Qwen VL 2.5 (ours)	The Qwen2.5VL (7B Instruct) [2] model served as the foundation. A three-step prompting pipeline was used: (i) to predict the final answer, (ii) to generate bounding boxes representing the evidence span in the handwritten note, and (iii) to classify the local (e.g., paragraph, equation) and global (e.g., chemistry, CS) content categories. Outputs were converted into the RRC-compatible format using a custom script.
	VANTAGE LABS, UNC CHAPEL HILL
Textract + MiniCPM	MiniCPM-V [20] (8B) was used as the primary VLM, paired with Amazon Textract for OCR-based text extraction. The model was prompted using handcrafted templates to sequentially extract answer, evidence region, and categorical information. Four outputs were parsed and mapped to the required JSON submission using automated formatting tools.
	KALMAN LABS
Intern-VL	InternVL3 [5] (8B) was used in its vision-language setting with an engineered prompt for single-shot inference. Outputs include bounding box prediction and label identification. InternVL was selected for its alignment capability in dense handwritten layouts and symbol-rich regions.
	VESTIGE R&D
Kimi + CoT	Kimi-VL [17] (8B) was prompted using a Chain-of-Thought (CoT) format to encourage detailed reasoning over diagrams and formulae. Prompt engineering focused on guiding the model to explain its steps in deriving answers from contextual elements and highlight those regions. A formatting pipeline parsed the structured output for evaluation.
	WINDEN RESEARCH
Ovis + ICL	Ovis [11] (4B) was paired with Visual In-Context Learning (ICL) where few-shot visual examples of answer-evidence pairs were provided. The prompt setup encouraged region-aware reasoning by showing the model annotated exemplars within its context. A structured output wrapper was used to match RRC specifications.
	NEWTRON AI
Sail-VL	SAIL-VL-1.5 [7] was paired with AIM-V2 ViT vision encoder. Adaptive stream packing and longer sequence modeling during training made it apt for large page-level reasoning. Task-specific prompts were used to extract evidence-aware answers, and special decoding heads processed multi-modal spans from the image.
	CASTOR LABS
Molmo VL	Molmo VL [6] (Open Source) used a set of four dedicated prompts to extract answer spans, predict box locations, and classify hierarchical content. The model employed transformer-based decoders trained on multi-modal representations to jointly process image patches and spatial embeddings.

Table 2: **Final Leaderboard of the ICDAR 2025 Handwritten Notes Understanding Challenge.** Scores reflect performance on the evidence-based VQA task. Highest scores per column are highlighted.

Rank	Team + Model	ANLS* ↑	IoU ↑	Local ↑	Global ↑
1	**Qwen VL 2.5 (Baseline)**	**0.4468**	**0.0284**	24.58	**47.75**
2	Winden Research — Ovis + ICL	0.4365	0.0063	**26.27**	21.78
3	Vestige R&D — Kimi + CoT	0.4143	0.0127	24.08	22.68
4	Kalman Labs — Intern VL 3	0.3380	0.0145	23.08	25.87
5	Castor Labs — Molmo VL	0.2109	0.0274	24.48	12.19
6	Newtron AI — SAIL-VL	0.0611	0.0070	19.18	14.59
7	Vantage Labs — Textract + MiniCPM	0.0466	0.0053	23.28	22.88

Note. All models are evaluated on the hidden HNU test set via the RRC evaluation platform. Highlighted values indicate the top scores per metric.

using Qwen-2.5-VL often designed multi-stage prompting pipelines to independently extract answers, supporting bounding boxes, and answer type classifications.

Another frequently adopted model was InternVL, a high-capacity vision-language model pre-trained on diverse visual corpora. It is built to align vision and language tokens using a multi-level cross-attention mechanism, facilitating visual grounding and structured output generation. Teams built specialized prompts to better guide InternVL in navigating the noise and variability present in handwritten notes.

In addition to these `end-to-end VLMs`, other teams explored `OCR-first pipelines`, such as combining Amazon Textract with models like MiniCPM, a lightweight open-source language model optimized for constrained environments. These pipelines used OCR for text extraction and employed LLMs to infer answers and localize supporting content, though they generally showed limited robustness on visually grounded tasks.

To ensure structured and evaluable output, nearly all methods adopted prompt-based modular decoding, with dedicated prompts for answer generation, visual grounding (bounding box prediction), and classification of local/global content types. In particular, some submissions utilized `chain-of-thought (CoT) prompting` to elicit step-by-step reasoning, while others explored `visual in-context learning (ICL)` to better anchor outputs in visual evidence. Overall, the submitted systems reflect a growing maturity in leveraging multimodal models for complex document understanding, while also revealing current limitations in layout reasoning and visual localization within handwritten content.

5.2 Performance Analysis

Table 2 presents a comparative overview of the submitted methods evaluated on the HNU test set, using four key metrics: **Average Normalized Levenshtein Similarity (ANLS*)**, **Average Intersection over Union (Avg. IoU)**, **Local Category Accuracy**, and **Global Category Accuracy**. Higher scores across all metrics indicate better performance.

– **Top Performer — Qwen VL 2.5:**
  - Achieved the highest **ANLS* of 0.4468**, showcasing strong textual answer generation that closely aligns with the semantic and syntactic structure of the ground truth.

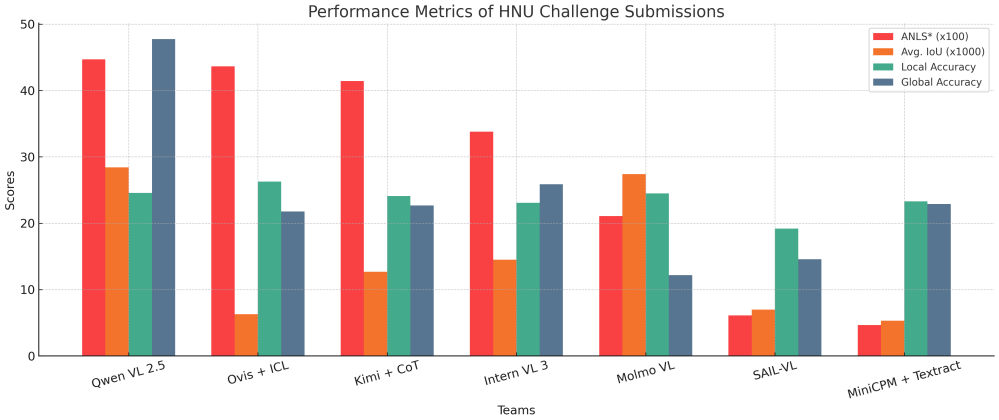

Fig. 4: Comparative performance of submitted methods on the HNU test set. ANLS* and IoU are scaled (×100 and ×1000, respectively) for visualization.

  - Led in **Avg. IoU (0.0284)**, reflecting relatively superior ability in evidence region localization, though all models showed limited performance in this area.
  - Demonstrated **best Global Category Accuracy (47.75%)**, suggesting superior high-level understanding of question intent.
  - Served as the official competition baseline, and set a challenging standard across multiple evaluation axes.
- **Specialized Strength — Ovis + ICL (Winden Research):**
  - Recorded the highest **Local Category Accuracy (26.27%)**, indicating effectiveness in recognizing fine-grained question types.
  - Utilized visual in-context learning (ICL) for strong context-aware grounding, though overall textual accuracy (ANLS) and IoU remained modest.
- **Balanced Yet Unexceptional — Kimi + CoT and Intern VL:**
  - **Kimi + CoT** displayed moderate performance across all metrics, aided by chain-of-thought prompting, but failed to lead in any specific category.
  - **Intern VL 3** had solid mid-range ANLS and category accuracy scores, showcasing robustness but lacking standout performance.
- **Challenging Scenarios — Lower-Performing Models:**
  - **Molmo VL**, **SAIL-VL**, and **Textract + MiniCPM** scored significantly lower, particularly in ANLS* and IoU, suggesting difficulties in visual-textual reasoning and precise evidence alignment.
  - OCR-first or hybrid pipelines such as Textract + MiniCPM particularly struggled with layout-dependent reasoning, as reflected in both localization and answer generation metrics.

Overall Insight: While the Qwen VL 2.5 model led in most dimensions as shown in Figure 4, the performance gap across metrics indicates that no single approach achieved comprehensive superiority across all aspects of the handwritten document understanding task. Notably, the generally low IoU values across the board point to spatial grounding in handwritten documents as a significant open challenge for current vision-language systems.

5.3 Prizes and Recognition

All teams that submitted valid entries will be acknowledged with official certificates of participation. Additionally, the top-performing teams will receive commemorative

mementoes and will be invited to attend the associated competition session, providing visibility within the document understanding research community.

6 Conclusion and Future Directions

The inaugural ICDAR 2025 Handwritten Notes Understanding (HNU) Challenge introduced a new benchmark focused on grounded question answering over complex handwritten scientific documents. By releasing a carefully curated dataset of 1,000 questions grounded in 2,000 pages of diverse handwritten notes, the competition pushed participants to tackle the challenges of unstructured layouts, multimodal evidence, and scientific reasoning. The submissions showcased a variety of VLMs, highlighting promising directions in prompt engineering and multimodal alignment. However, the consistently low IoU scores and variations in category accuracy underline the persistent difficulty of fine-grained evidence localization and contextual understanding in handwritten content. Future iterations of the challenge will expand into broader domains, introduce multi-page and multi-hop reasoning tasks, and promote end-to-end systems that unify recognition and understanding. New evaluation criteria focused on interpretability and partial semantic alignment will also be considered to better capture real-world model capabilities. We also plan to foster broader community involvement and increase participation in future editions through extended outreach and open development tools.

Acknowledgements

This work acknowledges the financial support of Department of Research and Universities of the Generalitat of Catalonia to the DocAI Research Group: Group on Document Intelligence (2021 SGR 01559), the Spanish project PID2021-126808OB-I00 funded by MCIN/AEI/10.13039/501100011033, and the PhD grant AGAUR (2023-FI-3-00223). The Computer Vision Center is part of the CERCA Program/Generalitat de Catalunya. The authors state that they have no competing interests.

References

1. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 993–1003 (2021) 1
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923 9
3. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019) 1, 2
4. Chen, M., Wu, H., Chang, Q., Cheng, H., Ma, J., Hu, P., Zhang, Z., Liu, C., Pi, C., Hu, J., et al.: Icdar 2024 competition on recognition of chemical structures. In: International Conference on Document Analysis and Recognition. pp. 397–409. Springer (2024) 3
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 9
6. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models (2024), https://arxiv.org/abs/2409.17146 9
7. Dong, H., Kang, Z., Yin, W., Liang, X., Feng, C., Ran, J.: Scalable vision language model training via high quality data curation (2025), https://arxiv.org/abs/2501.05952 9
8. Gao, L., Huang, Y., Déjean, H., Meunier, J.L., Yan, Q., Fang, Y., Kleber, F., Lang, E.: Icdar 2019 competition on table detection and recognition (ctdar). In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1510–1515. IEEE (2019) 3
9. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4083–4091 (2022) 1
10. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019) 1
11. Lu, S., Li, Y., Chen, Q.G., Xu, Z., Luo, W., Zhang, K., Ye, H.J.: Ovis: Structural embedding alignment for multimodal large language model. arXiv:2405.20797 (2024) 9
12. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022) 1, 2
13. Mathew, M., Tito, R., Karatzas, D., Manmatha, R., Jawahar, C.: Document visual question answering challenge 2020. arXiv preprint arXiv:2008.08899 (2020) 1, 2
14. Mondal, A., Mahadevan, V., Manmatha, R., Jawahar, C.: Icdar 2024 competition on recognition and vqa on handwritten documents. In: International Conference on Document Analysis and Recognition. pp. 426–442. Springer (2024) 3, 5
15. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. pp. 732–747. Springer (2021) 1
16. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) 4

17. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang, Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z., Lin, Z.: Kimi-vl technical report (2025), https://arxiv.org/abs/2504.07491 9
18. Tito, R., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D.: Icdar 2021 competition on document visual question answering. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16. pp. 635–649. Springer (2021) 1, 2
19. Van Landeghem, J., Tito, R., Borchmann, Ł., Pietruszka, M., Jurkiewicz, D., Powalski, R., Józiak, P., Biswas, S., Coustaty, M., Stanisławek, T.: Icdar 2023 competition on document understanding of everything (dude). In: International Conference on Document Analysis and Recognition. pp. 420–434. Springer (2023) 2
20. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 9
21. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1015–1022. IEEE (2019) 1