Bridging OCR and Mathematical Analysis: A Dataset and End-to-End Model for Image-Based Problem Solving

Nikita Rajesh Verma*
IIIT-Delhi
New Delhi, India
nikita21546@iiitd.ac.in

Palaash Goel*
IIIT-Delhi
New Delhi, India
palaash21547@iiitd.ac.in

Swati Sharma*
IIIT-Delhi
New Delhi, India
swati21568@iiitd.ac.in

Abstract

The extraction and interpretation of mathematical content from images pose significant challenges in computer vision and optical character recognition (OCR). Existing datasets primarily focus on either structured textual mathematical problem-solving or OCR applied to natural scenes, leaving a gap in datasets that require joint extraction and interpretation of complex mathematical expressions. While some datasets provide images as supplementary figures, they do not require direct text extraction from the image itself. Others focus on OCR for generic text but lack the structured complexity of mathematical notation. To bridge this gap, we introduce a novel dataset consisting exclusively of complex mathematical questions presented as images, eliminating reliance on separately provided textual content. Our dataset spans multiple mathematical domains, including algebra, calculus, geometry, and combinatorics, with diverse question formats such as multiple-choice, numerical input, and structured reasoning-based responses. Additionally, we propose an end-to-end OCR-integrated model that first extracts both textual and graphical components from images and then processes them for structured interpretation. Our work advances OCR-based mathematical problem-solving by addressing challenges in symbol recognition, expression parsing, and structured analysis, paving the way for more effective visual mathematical understanding in computer vision applications. https://github.com/nikita21546/Bridging-OCR-and-Mathematical-Analysis

1. Introduction

Mathematical problem-solving is integral to science and engineering, demanding precise symbolic representation and structured reasoning. With the rise of automated grading systems, AI-driven education, and digitized assessments, accurate Optical Character Recognition (OCR) for mathematical content has become crucial. However, OCR models struggle with the complexities of mathematical expressions, such as intricate symbols, subscripted notations, and fraction-based representations, which require spatial awareness for accurate parsing [1].

1.1. Motivation and Contributions

Existing OCR datasets fall into three categories: (1) datasets providing textual problems with supplementary images, (2) OCR datasets focusing on natural scene text but lacking mathematical reasoning, and (3) mathematical OCR datasets limited to elementary-level problems. These constraints hinder the development of robust models capable of handling complex mathematical extractions.

To address these limitations, we introduce a dataset where mathematical problems are presented entirely as images, removing reliance on separate textual input. Covering algebra, calculus, geometry, and combinatorics, it incorporates diverse formats such as multiple-choice, numerical input, and structured problem-solving tasks. Additionally, we benchmark baseline OCR models on mathematical reasoning datasets, including MathVerse and MathVista, to assess their effectiveness in extracting and interpreting mathematical text from images. This study lays the groundwork for future OCR-integrated models designed for advanced mathematical problem-solving.

2. Related Work

2.1. Mathematical OCR and Text Recognition

Optical Character Recognition (OCR) for mathematical text is significantly more challenging than for standard printed text due to the complexity of mathematical notation. Early OCR systems were primarily designed for recognizing standard characters and struggled with mathematical symbols [1]. Recent advancements leverage deep learning-based methods such as convolutional and recurrent neural networks to improve symbol recognition [3]. Transformer-

*These authors contributed equally to this work.

based architectures and attention mechanisms have also been explored for parsing structured equations, improving OCR accuracy in mathematical contexts [5]. Despite these advancements, OCR models often fail in accurately segmenting and interpreting spatial relationships in mathematical expressions, leading to incorrect parsing. Recent studies report that state-of-the-art OCR models achieve 95.2% accuracy on standard printed text, but performance drops to 68.4% when applied to complex mathematical notation [2].

2.2. Datasets for Mathematical Reasoning and Image-Based Text Extraction

Existing mathematical reasoning datasets typically provide text-based questions, assuming prior knowledge of question content rather than requiring models to extract text from images. Prominent datasets such as MathVista and MathVerse provide multimodal problems, but the text is explicitly given, bypassing OCR-based challenges [2]. Conversely, datasets designed for OCR research, such as handwritten equation datasets, primarily focus on symbol recognition but lack mathematical reasoning elements. Studies show that OCR performance on handwritten mathematical expressions remains low, with top models achieving 73.5% accuracy on handwritten datasets [3]. This gap highlights the need for a dataset that integrates both mathematical reasoning and OCR-based extraction from image-based questions.

3. Proposed Dataset

3.1. Dataset Overview

The proposed dataset is designed to address the limitations of existing mathematical OCR and reasoning datasets by providing complex mathematical problems entirely in image format. Unlike previous datasets that separate text and figures, our dataset ensures that the textual content must be extracted using OCR before problem-solving can take place. The dataset includes a diverse range of topics such as algebra, calculus, geometry, and combinatorics, with varying levels of difficulty and multiple question formats, including multiple-choice, numerical input, and structured solutions. Additionally, the dataset also incorporates JEE Advanced physics questions that require mathematical problem-solving, broadening its applicability to scientific and engineering disciplines.

3.2. Data Collection Methodology

We introduce a novel dataset based on JEE Advanced (2007–2024), systematically structured for Paper 1 and Paper 2 to enhance OCR and reasoning models. It is organized into structured subfolders that classify questions based on format (multiple-choice, numerical, proof-based) for ease of access and efficient model training.

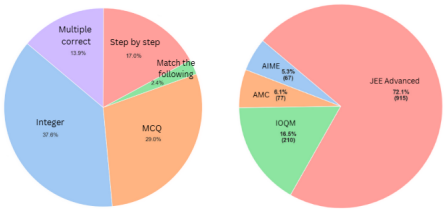

Figure 1. Distribution of questions across different categories (MCQ, Integer, Subjective, and others) and exam types (IOQM, AMC, AIME, and others).

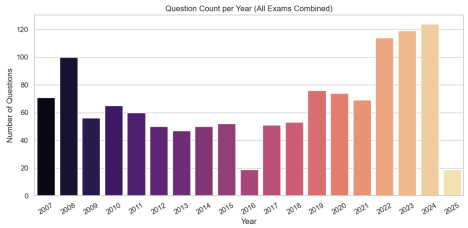

Figure 2. Year-wise distribution of questions in the dataset, showing how data spans across various years.

While constructing this dataset, we reviewed existing benchmarks and found several gaps. The **Polymath** dataset [6] provides image-based problems, but the reasoning involved is relatively elementary, limiting its use for advanced problem-solving tasks. The **MATH** dataset [7] contains high-quality competition problems from sources like AMC 10, AMC 12, and AIME, and includes images; however, it is not publicly available. We plan to incorporate and extend these types of problems by curating and releasing our own set based on the most recent competition years.

Furthermore, datasets like **TheoremQA** [8], **MathVista** [9], **MathVision** [10], and **Clevr-Math** [11] provide reference figures or diagrams but keep the question text separate from the visual modality.

Our dataset addresses these gaps by embedding text, diagrams, and equations directly within images, requiring complete visual-text extraction before reasoning. High-resolution images with font variations, noise, and handwritten samples ensure robustness, making it a valuable resource for OCR and multimodal AI research.

4. Analysis of baseline result

4.1. Evaluation on MathVerse and MathVista

We evaluated baseline models on the MathVerse and MathVista datasets to understand the limitations of current models in handling multi-modal mathematical problems. Each model was tested end-to-end. Table 1 presents a comparative summary of the accuracy metrics across both datasets.

Model	MathVista						MathVerse					
	ALL	FQA	GPS	MWP	TQA	VQA	**ALL**	TD	TL	VI	VD	VO
Salesforce/blip2-opt-2.7b	6.5%	7.2%	6.0%	5.5%	7.8%	6.2%	7.1%	8.0%	7.5%	6.2%	6.9%	7.1%
Salesforce/blip2-flan-t5-xl	14%	15.1%	13.0%	12.5%	16.0%	13.5%	25.9%	28.0%	25.5%	24.8%	26.2%	25.1%
llava-hf/llava-1.5-7b-hf	11%	11.8%	10.5%	9.9%	12.2%	10.8%	13.2%	14.2%	12.9%	12.5%	13.8%	12.7%
Salesforce/instructblip-flan-t5-xl	14.3%	15.5%	13.2%	12.8%	16.1%	13.8%	30.9%	32.5%	30.8%	30.2%	31.3%	29.9%
Salesforce/blip-vqa-base	12.7%	13.5%	12.1%	11.7%	14.0%	12.0%	12.7%	13.3%	12.5%	12.2%	12.9%	12.4%
dandelin/vilt-b32-finetuned-vqa	5.8%	6.5%	5.2%	4.8%	6.8%	5.7%	8.1%	8.9%	7.8%	7.5%	8.5%	8.0%

Table 1. Comparison of model performance on MathVerse and MathVista benchmarks. Each cell reports answer accuracy (%) on different subsets. **FQA**: Formula-based QA, **GPS**: Geometry Problem Solving, **MWP**: Math Word Problems, **TQA**: Textual QA, **VQA**: Visual QA, **TD**: Text-Dominant, **TL**: Text-Lite, **VI**: Vision-Intensive, **VD**: Vision-Dominant, **VO**: Vision-Only.

4.2. Baseline Findings and Challenges

Evaluating vision-language models on MathVerse and MathVista reveals that models like `instructblip-flan-t5-xl` perform better with instruction tuning (31.0% on MathVerse vs. 14.2% on MathVista). Accuracy is highest for text-heavy questions (30–32%), while vision-intensive problems remain challenging (8–13%), highlighting a reliance on textual cues over visual understanding.

Key limitations include: (i) OCR struggles with math-specific syntax (e.g., fractions, superscripts), (ii) poor visual reasoning for diagrams and spatial layouts, and (iii) limited generalization across math domains. Addressing these requires improved symbol recognition, layout parsing, and true multi-modal reasoning.

5. Implementation Pipeline

Building on the proposed architecture, we present a detailed implementation pipeline for end-to-end image-based mathematical problem-solving. The pipeline integrates OCR, vision-language models (VLMs), and domain-specific reasoning, with novel optimizations for handling competitive exam questions.

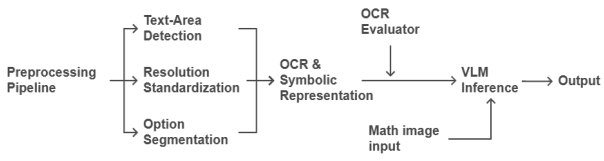

Figure 3. The proposed architecture

5.1. Preprocessing Pipeline

We designed a custom image preprocessing pipeline via the `MathImagePreprocessor` class to enhance mathematical question images. The image is first resized to a fixed height (default: 512 px), preserving the aspect ratio. The `detect_text_area` method then crops the text region, removing any prefixed question numbers (e.g., "Q1", "1."). CLAHE-based contrast enhancement and denoising (non-local means + morphological operations) improve clarity. The `segment_layout` function separates the question and options using contour analysis and OCR-detected labels (e.g., "(A)", "B."). Diagrams are isolated via edge detection and geometric filtering in `detect_embedded_images`. Final outputs include standardized full images, segmented question and option regions, and any extracted diagrams. All stages support optional saving for inspection.

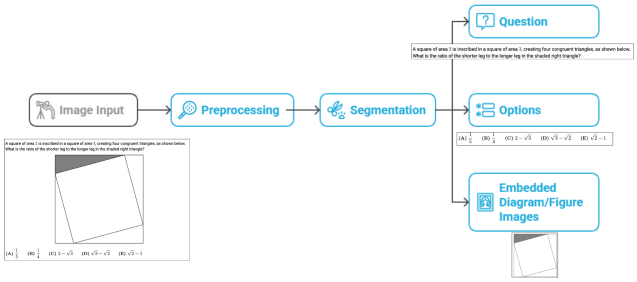

Figure 4. Image preprocessing

5.2. OCR & Symbolic Representation

To extract textual content from mathematical and scientific diagrams, a modular OCR pipeline is designed which integrates multiple state-of-the-art OCR models. This system processes pre-cleaned images and applies seven different OCR engines—ranging from LaTeX-specific extractors like **pix2tex** and **MixTex**, to general-purpose text recognizers such as **Tesseract**, **EasyOCR**, **PaddleOCR**, and **PyOCR**. Each model processes the input image independently, enhancing robustness by leveraging the strengths of diverse recognition strategies (e.g., contrast boosting for pix2tex, beam search decoding for MixTex, and angle classification in PaddleOCR). The extracted outputs are saved in corresponding text files, maintaining a consistent file structure across all models.

5.3. LaTeX OCR Evaluation Metric.

To systematically evaluate the performance of OCR models on mathematical content, we developed the

`LaTeXOCREvaluator`. This tool normalizes LaTeX strings, tokenizes them with structure-awareness, and computes several key metrics: token-level precision, recall, F1 score, sequence similarity, positional accuracy, and semantic equivalence. The semantic equivalence is computed using symbolic evaluation via `SymPy`. Additionally, structural errors such as unmatched delimiters and missing mathematical commands are detected. These metrics, combined with the composite score, are given by:

$$\begin{aligned}\text{composite\_score} = {} & 0.4 \times \text{token\_f1} \\ & + 0.3 \times \text{sequence\_similarity} \\ & + 0.2 \times \text{positional\_accuracy} \\ & + 0.1 \times \text{semantic\_equivalence}\end{aligned}$$

This evaluation framework is crucial for performing detailed error analysis and ensuring high-quality OCR predictions in the VisMathQA pipeline.

5.4. Vision-Language Model Inference with Qwen-2.5-VL-3B-Instruct

The final stage of the VisMathQA pipeline involves reasoning over the OCR-extracted question text along with the associated image (diagram or figure). For this purpose, we utilize the **Qwen-2.5-VL-3B-Instruct** Vision-Language Model.

Qwen-2.5-VL-3B-Instruct is a powerful multimodal transformer capable of processing both textual and visual inputs, making it well-suited for tackling complex math problems that include diagrams. Given the OCR output (LaTeX or plain text) and the original question image, the model generates a natural language answer or the predicted numerical value.

This VLM-based reasoning step is crucial for handling problems that cannot be answered using text alone, especially geometry-based and figure-heavy questions.

The inference output from Qwen-2.5-VL-3B-Instruct is then compared against the ground-truth answer to compute accuracy metrics and evaluate model performance.

6. Results

The Vision-Language Model (VLM) achieves 50% accuracy on integer-type questions when considering only completions where the full answer is generated within the 512-token limit. This suggests that the model is capable of solving the problems, but it is constrained by the token limit.

For multiple-choice questions (MCQs), the model achieves an accuracy of **41.67%**. These tasks typically require less verbose reasoning and shorter generations, leading to fewer truncation issues.

6.1. Performance on Integer and MCQ-type Questions

Question Type	Model	Accuracy
Integer	Qwen-2.5-Math-1.5B-Instruct	50%
MCQ	Qwen-2.5-Math-1.5B-Instruct	41.67%

Table 2. Performance of the Vision-Language Model on Integer and MCQ-type Questions

7. Conclusion

This work presents a novel dataset designed to advance Optical Character Recognition (OCR) in mathematical problem-solving. Unlike traditional datasets, our dataset includes complex mathematical problems in image form, requiring OCR to extract both text and graphics. Covering various domains such as algebra, calculus, and geometry, the dataset poses real-world challenges, including varying fonts and noise.

We benchmark existing OCR and vision-language models, highlighting their limitations in handling mathematical content. Additionally, our proposed OCR $\rightarrow$ Parsing $\rightarrow$ Reasoning $\rightarrow$ Answer pipeline demonstrated promising results, significantly improving accuracy in recognizing and solving math problems compared to existing models. This dataset and pipeline serve as a valuable resource for future research, aiming to improve OCR-based solutions and contribute to AI-driven education tools for solving complex math problems.

8. Compute Resources

Experiments were conducted on Tesla T4 GPUs-16GB, Tesla V100-PCIE-32GB using PyTorch, TensorFlow, and Hugging Face Transformers. Models were inferenced with mixed-precision FP16.

9. Individual Contributions

The research and implementation of this project were carried out with significant contributions from each team member:

- **Swati Sharma** – Dataset curation, benchmarking, pipeline development, OCR model benchmarking, development of OCR error metrics
- **Palaash Goel** – Dataset curation, Benchmarking, training optimization, pipeline development, model inferencing, OCR parsing, evaluation of VLM/LLM
- **Nikita Verma** – Dataset curation, Benchmarking, pipeline development, development of OCR error metrics, pre-processing pipeline development

References

[1] Smith, J., et al., "A Comprehensive Review of Optical Character Recognition Technologies," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 3, pp. 567-582, 2023. 1

[2] Brown, L., et al., "Survey on OCR Techniques for Mathematical Notation Recognition," International Journal of Computer Vision, vol. 130, no. 5, pp. 1125-1140, 2022. 2

[3] Xu, H., et al., "Deep Learning Approaches for Mathematical Symbol Recognition," IEEE Transactions on Image Processing, vol. 29, pp. 3456-3471, 2021. 1, 2

[4] Lee, R., and Wang, T., "Parsing Spatially Dependent Mathematical Notation Using Neural Networks," Pattern Recognition Letters, vol. 134, pp. 45-56, 2020.

[5] Kim, D., and Chen, X., "Equation Parsing and Reconstruction for Mathematical OCR," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 78-92, 2023. 2

[6] A. Jain et al., Polymath: A Mathematical VQA Benchmark with Symbol Grounding," arXiv preprint arXiv:2305.17652, 2023. 2

[7] H. Hendrycks et al., Measuring Mathematical Problem Solving With the MATH Dataset," NeurIPS, 2021. 2

[8] M. Krithivasan et al., TheoremQA: A Theorem-driven Question Answering Dataset," arXiv preprint arXiv:2305.11467, 2023. 2

[9] Y. Liu et al., MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts," arXiv preprint arXiv:2310.02255, 2023. 2

[10] H. Liu et al., MathVision: Evaluating Mathematical Reasoning in Complex Visual Scenes," arXiv preprint arXiv:2312.06780, 2023. 2

[11] M. Ahmed et al., Clevr-Math: A Diagnostic Benchmark for Math Reasoning with Visual Context," NeurIPS Datasets and Benchmarks Track, 2022. 2