ScienVisionQA: When Localization Matters for Scientific Document Understanding — Supplementary Material

1 Dataset Construction Details

This appendix provides the detailed prompts and procedures used in constructing ScienVisionQA.

1.1 Data Generation Prompts

1.1.1 Question Generation Prompt

```
You are a teacher who wants to help students find details in a research
    ↪ paper. Look at this image very carefully and give me a short
    ↪ question about some extremely tiny detail of this paper, focusing
    ↪ on figures and tables. If there is no table/figure on this page,
    ↪ just output empty. Please output the result in JSON format with key
    ↪  ‘question’. You should only output the JSON, nothing else. Like
    ↪ this: ‘‘‘json
{"question": "What is the specific notation used in the bottom-right
    ↪ corner of Table 2?"}
’’’
<image of a single page>
```

1.1.2 Answer Generation Prompt

This prompt is used for both Claude 3.7 (quality filter) and Qwen 2.5-VL (final answers):

```
Please give a concise answer to the following question based on the
    ↪ images provided: {question}
<image of a single page>
```

1.1.3 Answer Agreement Judging Prompt

We use o3-mini to judge answer agreement, with validation using QwQ 32B showing 85% inter-judge agreement:

```
Given the following reference answer:
{reference_answer}

Answer from the student:
{student_answer}

Please compare the student’s answer with the reference answer. Consider
    ↪ if the student’s answer is correct, concise, and closely matches
    ↪ the reference answer. Please output your result in JSON format with
    ↪  a key ‘student_correct’ to indicate if the student is correct, and
    ↪  nothing else. For example:‘‘‘json {"student_correct": true/false
    ↪ }’’’
```

1.1.4 QA Pair Categorization Prompt

```
You are a model that must categorize a single question-and-answer pair (
   ↪ with its associated image) along four predefined axes:

1. **Source Modality**: one of {Figure, Table, Equation, Caption}
2. **Answer Type**: one of {Numeric, Textual, Symbolic, Expression}
3. **Reasoning Complexity**: one of {Lookup, Arithmetic, Comparison,
   ↪ Multi-hop}
4. **Visual vs. Textual Dependency**: one of {Pure-Vision, Pure-Text,
   ↪ Mixed}

You will receive an input in this exact format:

Question: <question text>
Answer: <answer text>
<Image>

Your job is to examine the question, the answer, and the image, then
   ↪ output a **JSON** object with exactly these four fields (in this
   ↪ order) and their corresponding category tag. For example:

‘‘‘json
{
  "source_modality": "Figure",
  "answer_type": "Numeric",
  "reasoning_complexity": "Lookup",
  "visual_textual_dependency": "Mixed"
}
‘‘‘

Do not output any other text--only the JSON. Now process the following
   ↪ input:
Question: {question}
Answer: {answer}
<image of a single page>
```

2 Experimental Details

2.1 Model Versions

The following table summarizes the models and their corresponding versions used throughout our experiments.

Table 1: **Language Models and Version Identifiers.** Models evaluated in our experiments with their specific version identifiers.

Model Name	Provider	Version
GPT-4o	OpenAI	gpt-4o-2024-08-06
GPT-4o-mini	OpenAI	gpt-4o-mini-2024-07-18
o3-mini	OpenAI	o3-mini-2025-01-31
Claude 3.7 Sonnet	Anthropic	Claude 3.7 Sonnet 20250219

2.2 Training Details

Training hyperparameters for the three fine-tuned models:

Base Model	Prompt Template	k	Epochs	ϵ	GPU
Qwen2.5-VL 7B	Vanilla	3	3	10^{-8}	1x NVIDIA GH200 96GB
		5	3	10^{-8}	
		10	3	10^{-8}	
		20	5	10^{-8}	
	CoT	3	3	10^{-8}	
		5	3	10^{-8}	
		10	3	10^{-8}	
		20	5	10^{-8}	
InternVL3 8B	Vanilla	3	3	10^{-6}	4x NVIDIA A100 SXM 80GB
		5	3	10^{-6}	
		10	3	10^{-6}	
		20	5	10^{-6}	
	CoT	3	3	10^{-6}	
		5	3	10^{-6}	
		10	3	10^{-6}	
		20	5	10^{-6}	
Mistral Small 3.1 24B	Vanilla	3	3	10^{-6}	4x NVIDIA A100 SXM 80GB
		5	3	10^{-6}	
		10	3	10^{-6}	
		20	5	10^{-6}	
	CoT	3	3	10^{-6}	
		5	3	10^{-6}	
		10	3	10^{-6}	
		20	5	10^{-6}	

Table 2: **Training Setup for Qwen2.5-VL 7B, InternVL3 8B, and Mistral Small 3.1 24B.** Details of training configurations used for each model.

2.3 Data Augmentation Process

Our distractor-based training follows a systematic augmentation strategy:

For each QA pair and distractor count k:

1. Extract the page containing the answer
2. Randomly sample $k - 1$ other pages from the same document (or all available pages if the document has fewer than k pages)
3. Randomly shuffle all k pages to avoid positional bias
4. Create a training instance with the shuffled pages as input

This process is repeated for each epoch, ensuring diverse page combinations across training. The computational efficiency of smaller k values (e.g., $k = 3$ uses 85% fewer pages than full documents) makes localization training practical in resource-constrained settings.

2.4 MP-DocVQA Evaluation Setup

For transfer evaluation on MP-DocVQA [1], we use the validation split which contains 927 documents. To ensure clean evaluation without document-level overfitting, we retained exactly one question-answer pair from each document (specifically, the first listed pair per document). This results in our MP-DocVQA evaluation set of 927 QA pairs, with each question corresponding to a unique document. This setup ensures that transfer performance reflects genuine localization abilities rather than memorization of document-specific patterns.

3 Complete Experimental Results

3.1 Full Distractor Grid Analysis

Base Model	Prompt Template	*k*	Accuracy(%)	Page Predict Accuracy(%)
Qwen2.5-VL 7B	Vanilla	base	52.68	-
		3	57.45	
		5	57.67	
		10	57.97	
		20	63.57	
	CoT	3	55.93	86.32
		5	58.35	87.23
		10	59.64	88.51
		20	**64.25**	92.44
InternVL3 8B	Vanilla	base	54.50	-
		3	57.82	
		5	57.37	
		10	57.60	
		20	58.81	
	CoT	3	58.43	40.21
		5	**58.88**	41.87
		10	57.82	46.49
		20	57.45	45.65
Mistral Small 3.1 24B	Vanilla	base	74.68	-
		3	75.28	
		5	75.81	
		10	78.97	
		20	74.91	
	CoT	3	76.34	91.16
		5	75.06	90.70
		10	78.38	92.67
		20	**81.18**	94.63

Table 3: **Fine-tuned models evaluation results on ScienVisionQA.** Complete performance breakdown showing baseline (frozen model) accuracy and improvements across different prompt types (Vanilla/CoT) and distractor counts (k). Page prediction accuracy is reported for CoT models, which explicitly identify the relevant page before answering.

Base Model	Prompt Template	*k*	Accuracy(%)	Page Predict Accuracy(%)
Qwen2.5-VL 7B	Vanilla	base	77.11	-
		3	**81.55**	
		5	80.15	
		10	81.34	
		20	81.34	
	CoT	3	79.68	81.55
		5	81.01	81.23
		10	80.37	80.91
		20	80.58	81.66
InternVL3 8B	Vanilla	base	77.35	-
		3	80.47	
		5	80.15	
		10	80.80	
		20	81.23	
	CoT	3	**81.88**	65.05
		5	81.55	66.99
		10	80.47	67.31
		20	80.37	64.72
Mistral Small 3.1 24B	Vanilla	base	78.96	-
		3	80.56	
		5	80.04	
		10	81.23	
		20	79.07	
	CoT	3	**82.63**	81.98
		5	80.80	82.85
		10	81.34	81.98
		20	80.26	81.01

Table 4: **Fine-tuned models evaluation results on MP-DocVQA.** Transfer learning performance showing baseline (frozen model) accuracy and improvements when models trained on scientific documents are evaluated on diverse industry documents (letters, forms, reports, memos, newsletters, financial documents, etc.). Results demonstrate that localization skills learned on ScienVisionQA generalize across substantially different document types and visual layouts.

3.2 Per-Category Accuracy Analysis

To provide deeper insights into model performance, we analyze accuracy across the different categorization dimensions of ScienVisionQA. All results use the best-performing configuration for each model from our main experiments.

Accuracy by Evidence Source Modality

Modality	Qwen-7B (CoT, *k*=20)	InternVL-8B (CoT, *k*=5)	Mistral-24B (CoT, *k*=20)
Figure	66.1	60.1	80.3
Table	60.2	55.4	83.7
Equation	61.9	56.4	80.1
Caption	70.1	70.1	83.6

Table 5: **Performance across evidence source types.** Caption-based questions show highest accuracy, while table-based questions are most challenging for smaller models.

Accuracy by Answer Format

Answer Type	Qwen-7B	InternVL-8B	Mistral-24B
Numeric	65.8	59.2	83.1
Textual	66.0	61.4	80.5
Symbolic	57.3	53.1	77.1
Expression	54.9	52.9	80.4

Table 6: **Performance by answer format.** Expression and symbolic answers prove most challenging for smaller models.

Accuracy by Information Source

Dependency	Qwen-7B	InternVL-8B	Mistral-24B
Pure-Vision	59.4	55.7	78.5
Pure-Text	60.5	60.5	79.1
Mixed	67.2	60.6	82.8

Table 7: **Performance by information source.** Mixed vision-text questions show highest performance, suggesting complementary information aids extraction.

Key observations. Mistral-24B consistently outperforms smaller models across all categories, confirming that larger capacity remains beneficial even after localization training. Smaller models struggle most with symbolic and expression answers, while mixed vision-text questions—the most prevalent in our dataset—often show better performance than pure-vision tasks, suggesting that multimodal integration provides complementary information for accurate extraction.

References

[1] Rubén Pérez Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *Computing Research Repository (CoRR)*, abs/2212.05935, 2022.