Appendix

A. Analysis of Query-based Feature Extraction

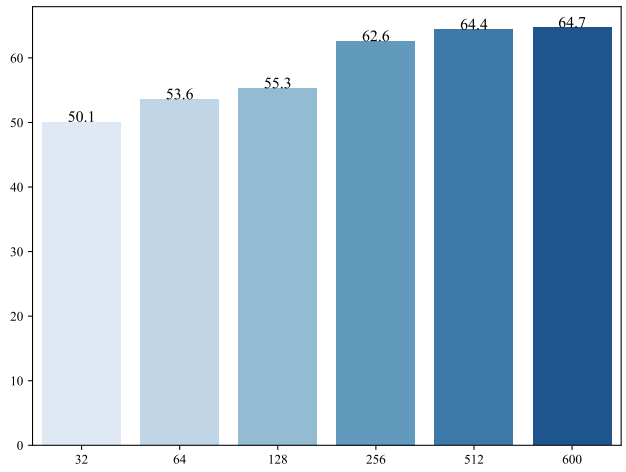


Figure 7. Performance with varying query numbers on the DocVQA dataset.

In this section, we offer an experimental analysis to clarify our reasoning behind not choosing a query-based feature fusion approach.

Building upon the Donut framework [33], we employ Q-Former [36] to extract image features and conduct cross-attention operations with Bart [41] using the extracted features. We fine-tune the model on the DocVQA dataset, and the experimental results are illustrated in Figure 7. When the image resolution is set to 1280, we observe that an insufficient number of query vectors can significantly degrade the model's performance. To mitigate this decline while maintaining the model's performance, 500 query vectors are required. However, this approach to information extraction is not highly efficient in practice. Consequently, we choose a direct fusion approach in the instruction filtering module to retain visual information to the greatest extent possible.

B. Visual Instruction Tuning

B.1. Instruction Templates

As shown in Table 4, we present additional instruction templates. A greater number of instruction templates can significantly enhance the model's generalization capabilities and improve its performance in real-world applications. It is worth noting that users' perspectives in posing questions are diverse; therefore, having an adequate number of templates allows the model to better understand and respond to real-world instructions.

Task	Format
IE	Human: What is the value of the {key}? AI: {value} Human: What is the {key}? AI: {value} Human: What is the content of {key}? AI: {value} Human: What is the essence of the {key}? AI: {value}
OCR	Human: Present all the text in the image. AI: {all text} Human: please output the OCR result AI: {all text} Human: What is the text content in this image? AI: {all text} Human: What is the textual context of this image? AI: {all text}
VG	Human: Where is the {obj}? AI: {x, y, x + w, y + h} Human: Where is the {obj} recorded? AI: {x, y, x + w, y + h} Human: Where is the {obj} located? AI: {x, y, x + w, y + h}
IC	Human: What is the abstract of the image? AI: {caption} Human: Can you describe the content of this picture? AI: {caption} Human: Could you put into words what's in this picture ? AI: {caption} Human: Can you summarize this picture in one sentence? AI: {caption}
TR	Human: What is the element in the table? AI: {element} Human: Please output the table in kv format? AI : {element}

Table 4. Additional examples of instruction tuning templates.

B.2. Details of Datasets

In this section, we provide a detailed introduction to the various datasets used in our experiments.

CORD The CORD [49] dataset comprises 800 training receipts, 100 validation receipts, and 100 test receipts. Each receipt is accompanied by a photo and a set of OCR annotations. The dataset identifies 30 fields across four categories, and the task's objective is to correctly assign each word to the appropriate field. The evaluation metric used is the entity-level F1 score, and official OCR annotations are utilized.

SROIE The SROIE [30] dataset is designed for extracting data from digitized receipts. It consists of 626 training samples and 347 testing samples. The objective is to re-

trieve information for up to four specific keys from each receipt: company, date, address, and total. The assessment metric used is the entity-level F1 score. Official OCR annotations are utilized, and the test set outcomes are supplied by the authorized evaluation platform.

DocVQA The DocVQA [45] dataset comprises 50,000 questions based on more than 12,000 pages from a diverse range of documents. The pages are divided into training, validation, and test sets at a ratio of approximately 8:1:1. The task's evaluation employs an edit distance-based metric called ANLS (average normalized Levenshtein similarity).

InfoVQA The InfographicVQA [46] dataset consists of 30,035 questions and 5,485 images, originating from 2,594 distinct web domains. This dataset employs the ANLS metric for evaluation, where higher scores are assigned if the predicted answer has a smaller difference from at least one of the ground-truth answers.

DeepForm DeepForm [7] is a socially important documents related to election spending with the objective of extracting contract numbers, advertiser names, payment amounts, and advertisement broadcast dates from advertisement disclosure forms. The dataset comprises 700 training samples, 100 validation samples, and 300 testing samples. The evaluation metric used is the F1 score.

KCL Kleister Charity [57] is a document understanding dataset designed for the extraction of information related to charitable organizations. It consists of 1,700 training samples, 400 validation samples, and 600 testing samples. The evaluation metric employed is the F1 Score.

WTQ WikiTableQuestions [5] is a question-answering dataset that comprises semi-structured HTML tables sourced from Wikipedia. It includes 1,400 training samples, 300 validation samples, and 400 testing samples. The evaluation metric employed is accuracy.

TabFact TabFact [16] is a dataset designed for investigating fact verification tasks in the context of semi-structured evidence. It consists of 13.2K training samples, 1.7K validation samples, and 1.7K testing samples. The evaluation metric employed is accuracy.

ChartQA ChartQA [15] is a question-answering dataset targeting data visualizations in the form of charts, involving both visual and logical reasoning. It comprises 9.6K manually curated questions and 23.1K questions automatically generated from manually curated chart summaries. The evaluation metric employed is relaxed accuracy.

TextVQA The TextVQA [56] dataset is constructed by extracting images and questions from the Open Images v3 dataset. It consists of 34,602 training samples, 5,000 validation samples, and 5,734 testing samples. The evaluation metric employed is accuracy.

VisualMRC The VisualMRC [59] dataset aims to enable machines to read and comprehend text in real-world documents and respond to natural language questions. This dataset comprises over 30,000 question and abstractive answer pairs derived from more than 10,000 document images spanning multiple web domains. The evaluation metric employed is CIDEr. The computation of CIDEr is based on syntactic consistency, content consistency, consistency metrics, and diversity evaluation, synthesizing the similarity and consistency scores between the generated image descriptions and multiple reference descriptions.

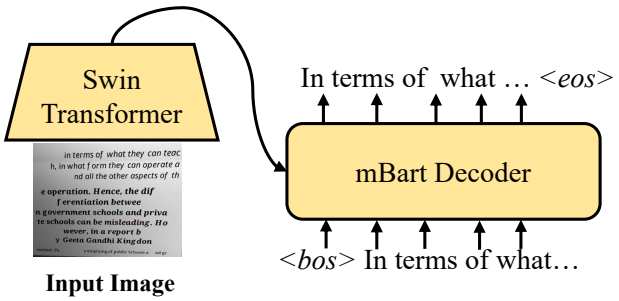


Figure 8. Schematic representation of the pretraining process for the image encoder.

TextCaps The TextCaps [55] dataset consists of 28,408 images and 142,040 captions, requiring models to read and comprehend textual information within the images and generate coherent descriptions. The evaluation metric employed is CIDEr.

C. Training

In this section, we primarily provide a detailed description of Stage 2 of our training strategy.

Stage 2 essentially involves the pretraining of the image encoder. Currently, open-source image encoders mainly focus on two aspects: one is performing image classification tasks using datasets like ImageNet, and the other is aligning image and text features based on contrastive learning. These two pretraining paradigms are not suitable for generative tasks such as text recognition, as there is a significant difference between the pretraining methods and downstream tasks.

To make the image encoder more suitable for text recognition and generation tasks, we employ a method similar to Donut for pretraining the image encoder, as illustrated in Figure 8. We primarily construct a temporary model to perform a pseudo-OCR task, which involves recognizing all text in the image in a top-to-bottom and left-to-right order. This pretraining task is more consistent with the downstream tasks, enabling our final HRVDA model to possess strong text recognition capabilities.

Human: What letter is on the button in the top left corner, and where is it located?
HRVDA: The top left button is Q, and it is located at [105, 440, 326, 522].

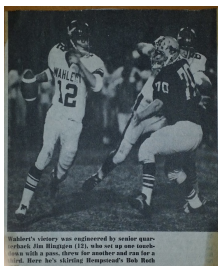
(a)

Human: What does the sign on the right say?
HRVDA: Collabora

(b)

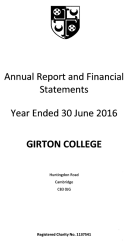
Human: what number is above the word "miglia"?
HRVDA: 1000

(c)



Human: What is the text under the photo?
HRVDA: Wahlert's victory was engineered by senior quarterback Jim Hingtgen (12), who set up one touchdown with a pass, threw for another and ran for a third. Here he's skirting Hempstead's Bob Roth

(d)

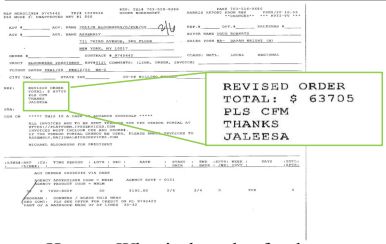


Human: Recognize all the text on the image
HRVDA:
Annual Report and Financial Statements
Year Ended 30 June 2016
GIRTON COLLEGE
Huntingdon Road
Cambridge …

(e)

Human: What is the abstract of the image?
HRVDA: A woman holding a phone which says the time is 12:10.

(f)



Human: What is the value for the gross_amount ?
HRVDA: 190

(g)

Human: What forms is the second least popular category of prizes?
HRVDA: female

(h)

Human: what is the number of times that the new york americans did not qualify for the playoffs?
HRVDA: 4

(i)

Figure 9. Additional qualitative examples generated by HRVDA. Green indicates that HRVDA answered correctly, while red represents incorrect answers.

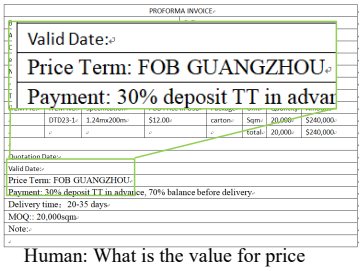
D. Qualitative Experimental Analysis

In this section, we provide some supplementary qualitative analysis.

As depicted in the first two rows of Figure 9, HRVDA can recognize colors, positions, and artistic fonts, which is primarily attributed to its visual pretraining. Furthermore, leveraging the semantic understanding capabilities of the LLM, HRVDA can also recognize text in complex regions, such as identifying the field above a particular field. Even when dealing with images containing long text, HRVDA demonstrates strong full-text OCR capabilities.

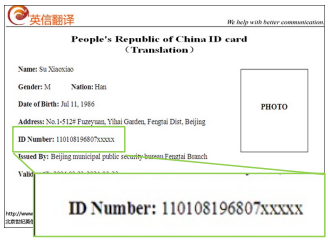
Nonetheless, HRVDA struggles with certain highly challenging examples, as illustrated in the last row of Figure 9. For example, the HRVDA model faces comprehension difficulties when processing images that have an exceptionally high density of text and exhibit intricate structural relationships. Moreover, HRVDA is not well-suited for images with extreme proportions. As demonstrated in Figure 9-(h), the model can only manage such images by performing multiple cropping operations, which inevitably compromises its grasp of the overall image structure. Furthermore, HRVDA is incapable of generating an adequate understanding for exceedingly complex instructions. To address these extremely challenging examples, we plan to further increase the resolution and employ a more powerful LLM in future iterations.

We also evaluate the performance of HRVDA using open-domain data, as shown in Figure 10. HRVDA performs exceptionally well in information extraction tasks for common fields, such as dates, amounts, fax numbers, *etc*. Overall, if the answer relies more on simple text recognition, HRVDA can perform very well, significantly advanc-



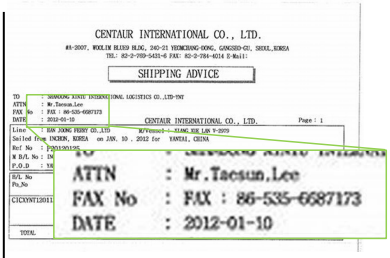
Human: What is the value for price term?
HRVDA: FOB GUANGZHOU

(a)



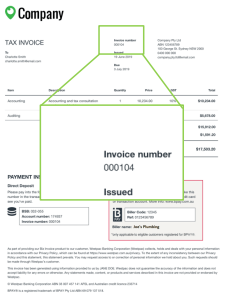
Human: What is the value of ID Number?
HRVDA: 110108196807xxxx

(b)



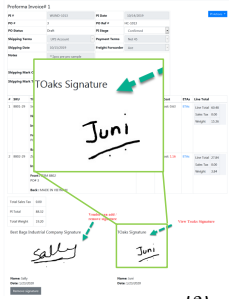
Human: what is the fax number?
HRVDA: 86-535-668713

(c)



Human: what is the invoice number?
HRVDA: 000104

(d)



Human: What is the Toaks Signature?
HRVDA: Juni

(e)

Human: What is the date?
HRVDA: 29-May-2016

(f)

Figure 10. Performance demonstration of HDVDA on open-world examples.

ing the practical application of MLLMs.