

# GENERAL OCR THEORY: TOWARDS OCR-2.0 VIA A UNIFIED END-TO-END MODEL

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Traditional OCR systems (OCR-1.0) are increasingly unable to meet people’s usage due to the growing demand for intelligent processing of man-made optical characters. In this paper, we collectively refer to all artificial optical signals (e.g., plain texts, math/molecular formulas, tables, charts, sheet music, and even geometric shapes) as “characters” and propose the General OCR Theory along with an excellent model, namely GOT, to promote the arrival of OCR-2.0. The GOT, with 580M parameters, is a unified, elegant, and end-to-end model, consisting of a high-compression encoder and a long-contexts decoder. As an OCR-2.0 model, GOT can handle all the above “characters” under various OCR tasks. On the input side, the model supports commonly used scene- and document-style images in slice and whole-page styles. On the output side, GOT can generate plain or formatted results (markdown/tikz/smiles/kern) via an easy prompt. Besides, the model enjoys interactive OCR features, i.e., region-level recognition guided by coordinates or colors. Furthermore, we also adapt dynamic resolution and multi-page OCR technologies to GOT for better practicality. In experiments, we provide sufficient results to prove the superiority of our model.

## 1 INTRODUCTION

Optical Character Recognition (OCR) is a widely used technology that extracts the characters embedded in an optical image into an editable format. Typical OCR systems Du et al. (2021) in the OCR-1.0 era are mainly designed based on a multi-modular pipeline style, commonly including element detection, region cropping, and character recognition parts. Each module is prone to falling into local optima, making the whole system incur high maintenance costs. Moreover, traditional OCR methods have insufficient general ability, reflected as different OCR-1.0 networks usually designed for different sub-tasks. Nevertheless, choosing a suitable one from diverse OCR models for a special task is always inconvenient for users.

In the past year, Large Vision Language models (LVLMs) OpenAI (2023); Liu et al. (2023b); Ye et al. (2023a) have developed rapidly and showcased impressive performance. As a highly anticipated ability, the OCR performance of current LVLMs is continuously improving. Based on CLIP Radford et al. (2021), LLaVA Liu et al. (2023b) naturally acquires the English OCR ability after the instruct tuning phase. To lift the OCR accuracy and support other languages, e.g., Chinese, Qwen-VL Bai et al. (2023b) unfreezes its image encoder (a CLIP-G) and uses lots of OCR data in its stage-two training. Innovatively, Vary Wei et al. (2023) generates a new high-resolution OCR vision vocabulary paralleling the CLIP branch to deal with document-level dense OCR. By contrast, InternVL-1.5 Chen et al. (2024b) and other models Liu et al. (2024d); Ye et al. (2023b) utilize a sliding window manner to crop the whole image into multiple sub-patches for high-resolution OCR. Hence, a consensus is that optical character perception and recognition are the foundation of text-driven image understanding, drawing many researchers to pay more attention to LVLMs’ OCR booster.

However, the popular designs of LVLMs may not be suitable for diverse OCR tasks for the following reasons: 1) The conflicts between perception and reasoning. LVLMs mainly focus on visual reasoning performance, e.g., VQA Singh et al. (2019); Mathew et al. (2021), because that is what the LLM excels at. To quickly obtain the QA-gain benefits from LLMs, most LVLMs Liu et al. (2023b); Ye et al. (2023a); Li et al. (2023a) align image tokens to text ones. However, it is unreasonable to do this for pure perception OCR tasks, especially high-density text scenes, because each aligned vision

054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083  
084  
085  
086  
087  
088  
089  
090  
091  
092  
093  
094  
095  
096  
097  
098  
099  
100  
101  
102  
103  
104  
105  
106  
107

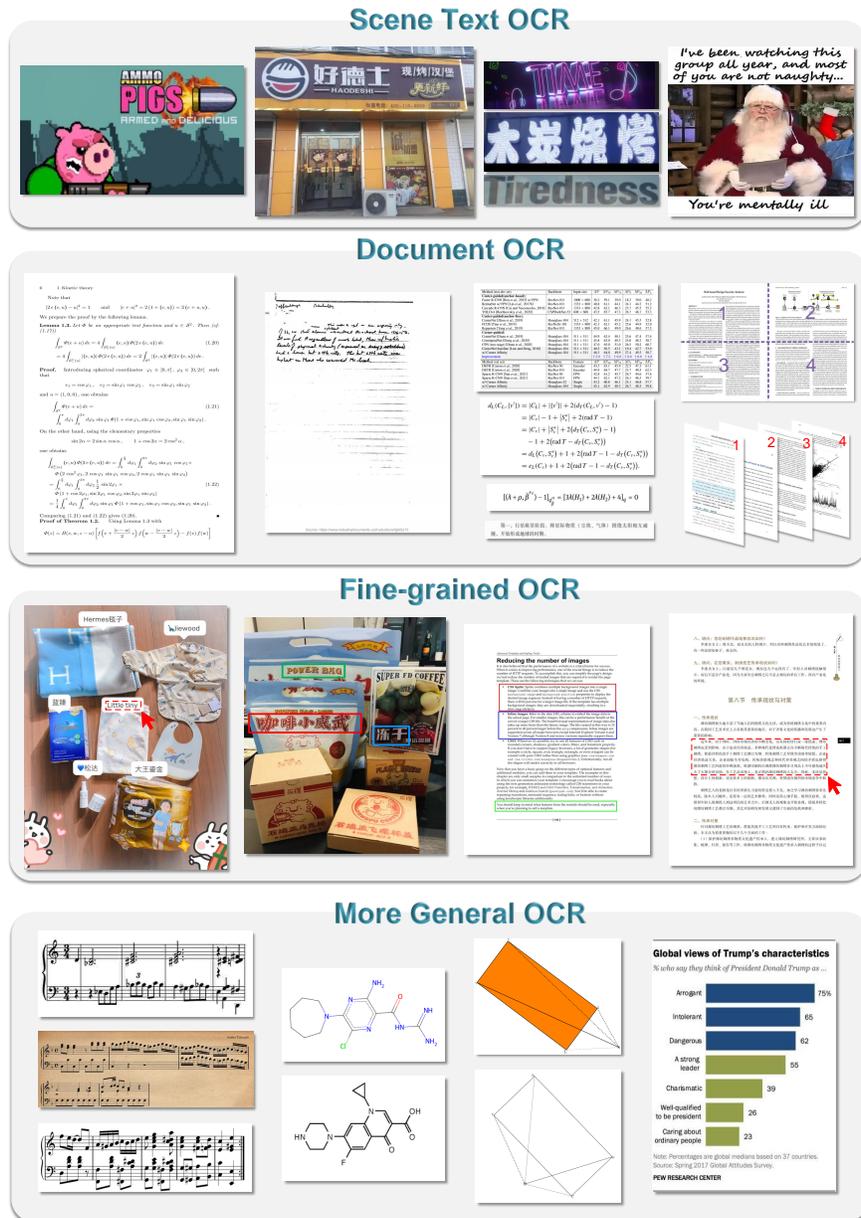


Figure 1: On the input side, GOT supports various optical image types, such as commonly used photographs and documents. Besides, as a general OCR-2.0 model, GOT can handle more tasks, e.g., sheet music, molecular formulas, easy geometric shapes, charts, etc. Moreover, the model can adapt to region-focus OCR, high-resolution OCR, and multiple-page OCR. GOT mainly supports English and Chinese and can control the structure results (Mathpix markdown/tikz/smiles/kern) via a prompt.

token (biased towards text token) cannot compress enough characters. Imagine how wasteful it is to use thousands of image tokens, e.g., the image-cropping manner Chen et al. (2024b); Liu et al. (2024c), to encode an equal amount of optical characters (e.g., texts within only an A4-PDF page). 2) High iteration and deployment costs. LVLm often enjoys billions of parameters, leading to the post-training and deployment costs being too high. Generally speaking, for LVLms, fine-tuning is not enough once we want to add a new OCR pattern, e.g., a new language, instead of enough GPU resources for pre-training. However, rerunning the pre-training with billions of parameters, only to introduce a new OCR feature, is also wasteful.

108 Accordingly, we propose the general OCR theory, i.e., OCR-2.0, to break the bottlenecks of both  
109 traditional and LVLM manners on OCR tasks. We think that a model of OCR 2.0 should have the  
110 following essential characteristics:

- 111
- 112 • **End-to-end.** Compared to OCR-1.0 models with complex procedures, the OCR-2.0 model should  
113 enjoy a unified and end-to-end architecture to ensure lower maintenance costs. It is cool that a  
114 beginner can quickly master the entire OCR system in the 2.0 era.
- 115 • **Low training and inference costs.** The OCR-2.0 model should not be a chatbot, like LVLM, that  
116 focuses on reasoning tasks. Its focus should be on strong perception and recognition of optical  
117 characters, so it needs a reasonable number of model parameters in exchange for lower training  
118 and inference costs.
- 119 • **Versatility.** The OCR-2.0 model’s other important point is versatility, including recognizing more  
120 general artificial optical “characters”, e.g., sheet music, charts, geometric shapes, etc. Besides, the  
121 model should support the output format with stronger readability, e.g.,  $\text{\LaTeX}$ /Markdown format for  
122 formulas and tables.

123  
124 Based on the proposed general OCR theory, we present a primary OCR-2.0 model (GOT) to bridge  
125 the gap between OCR-1.0 models and people’s higher optical character processing demands. In  
126 architecture, we adopt the unsophisticated encoder-decoder paradigm for the model. Specifically,  
127 GOT enjoys a high compression rate encoder to transfer the optical image to tokens as well as a  
128 long context length decoder to output the corresponding OCR results. The encoder has approx-  
129 imately 80M parameters posing  $1024 \times 1024$  input size which is enough to deal with commonly  
130 used photo/document input styles. Each input image will be compressed to tokens with  $256 \times 1024$   
131 dimensions. The decoder of GOT, with 0.5B parameters, supports 8K max length tokens to ensure  
132 it can tackle long-context scenarios. We devise an effective and efficient training strategy for GOT,  
133 which can be divided into three procedures, i.e., decoupled pre-training of the encoder, joint-training  
134 of the encoder with a new decoder, and further post-training of the decoder. Besides, to further lift  
135 the practicality of GOT, we additionally adapt the fine-grained OCR feature for better interactivity,  
136 dynamic resolution strategy for ultra-high-resolution images (e.g., over 2K), and the multi-page OCR  
137 technology to alleviate the problem of difficulty in breaking pages in PDF image-text pairs (e.g.,  
138 page breaks in *.tex* files). To support each training stage, we do many data engines for synthetic data  
139 production, which is the key to the success of GOT and will be described in detail in this paper. The  
main input data format supported by our model can be seen in Figure 1.

140 As a model for envisioning OCR-2.0, GOT demonstrates promising performance in our experiments  
141 in various OCR tasks. We hope the proposed simple and elegant GOT can draw more researchers to  
142 invest in the research of OCR-2.0. Of course, the path to OCR-2.0 is still long and GOT also enjoys  
143 much improvement room, such as supporting more languages, more general artificial signals, and  
144 more complex geometries. In this new era led by LVLMs, we are convinced that the pure OCR model  
145 is not over, it may even be a new beginning.

## 146 2 RELATED WORK

### 147 2.1 TRADITIONAL OCR

148  
149  
150 Optical Character Recognition (OCR) is a classic research topic that aims to convert the image’s  
151 optical contents into an editable format for further downstream processing. Traditional OCR systems,  
152 called OCR-1.0, typically use a framework that is assembled from multiple expert modules. For  
153 instance, to handle diverse optical characters, the OCR system Du et al. (2021) is usually developed  
154 by integrating several domain expert networks, such as layout analysis Zhong et al. (2019), text  
155 detection Liao et al. (2022); Liu et al. (2019b); Zhang et al. (2021), region extraction, and contents  
156 recognition Li et al. (2023b). The reason for using such a pipeline scheme is that the text recognition  
157 module (the OCR part) failed to scale up successfully, which can only deal with the image format of  
158 small slices, resulting in the entire OCR process being in the form of first detecting texts/cropping  
159 regions, and then recognizing the results within the slice. However, a system with complicated  
160 procedures may suffer potential systematic errors and high maintenance costs. Although some  
161 OCR-1.0 models, e.g., Nougat Blecher et al. (2023) can directly process documents at the whole page  
level, they are often designed and trained for a specific sub-task, leading to unsatisfactory general

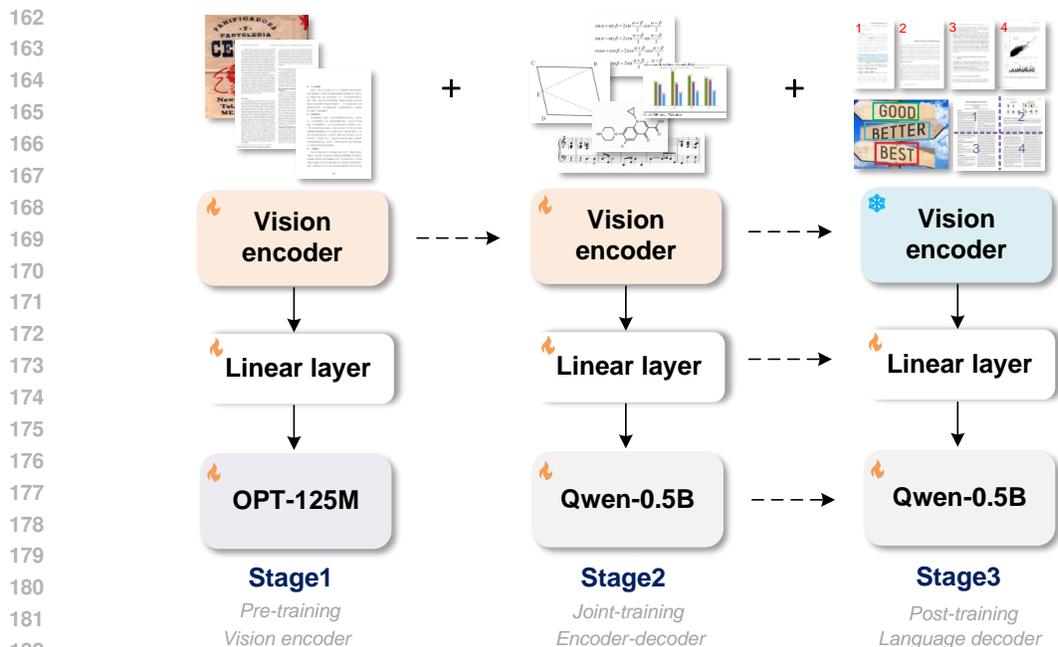


Figure 2: The framework of the proposed GOT. Stage 1: We pre-train the vision encoder using a tiny OPT-125M to adapt the OCR tasks efficiently. Stage 2: GOT is built by connecting the vision encoder to Qwen-0.5B and sufficient OCR-2.0 knowledge of more general optical characters is used in this stage. Stage 3: No modification of the vision encoder is required, and GOT is customized to new character recognition features.

ability. In the OCR-1.0 era, one inconvenient thing is that we usually need to switch different models according to various OCR needs.

## 2.2 LVLM-DRIVEN OCR

Large Vision-Language Models Liu et al. (2023b); Bai et al. (2023b); Wei et al. (2023); Ye et al. (2023a); Chen et al. (2024b); Liu et al. (2024d;a) have attracted lots of attention in the AI-community due to their powerful generalization capabilities. For the current LVLMs owning perception-reasoning comprehensive capacity, the OCR ability has become a hot spot with the increasing demand for text-driven visual understanding. Most LVLMs’ OCR capabilities come from the ready-made CLIP Radford et al. (2021), especially those that freeze CLIP encoder Liu et al. (2023b) to complete the entire LVLM training. For such models, the vanilla CLIP, mainly with English scene text knowledge, is the bottleneck for the OCR performance to out-of-domain tasks, such as other languages or documents. Some other LVLMs Ye et al. (2023a); Bai et al. (2023b) choose to unfreeze the encoder and freeze the LLM for training to enhance the CLIP-encoder and align the image tokens to text ones. These models will face the problem of low optical character compression rate, as it is difficult for frozen LLM to decode too much text from an aligned image token. To alleviate this problem, some models Chen et al. (2024b); Liu et al. (2024d); Ye et al. (2023b) adopt a sliding window manner to decompose input images into smaller patches. Although this dynamic resolution approach is highly effective in processing high-resolution input images, e.g., PDF, it will result in excessive image tokens and limit the max length of the generated OCR result to some extent.

## 3 GENERAL OCR THEORY

In this work, we propose the general OCR theory, i.e., OCR-2.0 (as expounded in Section 1) to promote the development of the OCR field. Based on the proposed new theory, we present a novel OCR model (GOT). In this section, we will introduce the technical details of our model.

### 3.1 FRAMEWORK

As illustrated in Figure 2, GOT comprises three modules, i.e., an image encoder, a linear layer, and an output decoder. The linear layer acts as the connector to map the channel dimension between the vision encoder and the language decoder. We utilize three main steps in optimizing the whole GOT model. First, we conduct the pure text recognition task to pre-train the vision encoder. To lift training efficiency and save GPU resources, we choose a tiny decoder to pass gradients to the encoder. In this stage, we feed images containing scene texts and manual images containing document-level characters into the model to allow the encoder to gather the two most commonly used characters’ encoding abilities. In the next stage, we form the architecture of GOT by connecting the trained vision encoder to a new larger decoder. We prepare lots of more general OCR data (*e.g.*, sheet music, math/molecular formulas, and geometric shapes) to scale up the OCR-2.0 knowledge for this stage. In the final stage, we intend to improve the generalization and applicability of GOT further. Specifically, fine-grained and multi-crop/page synthetic data are generated and added for GOT to support region prompt OCR Liu et al. (2024a), huge image OCR, and batched PDF OCR features.

### 3.2 PRE-TRAIN THE OCR-EARMARKED VISION ENCODER

As aforementioned, GOT enjoys the encoder-decoder structure. Inspired by the LVLMS design, the decoder can be initialized by a well-trained language model. However, we did not find a suitable pre-trained encoder for an OCR-2.0 model, so we must train one ourselves. We hope the new OCR encoder can work well on commonly used scene and document text recognition in various input shapes (both slices and whole pages).

#### 3.2.1 THE VISION ENCODER GENERATION.

The encoder structure we selected is VitDet Li et al. (2022) (base version with about 80M parameters) due to its local attention can greatly reduce the computational cost of high-resolution images. We follow the Vary-tiny setting Wei et al. (2023) to design the last two layers of the encoder, which will transfer a  $1024 \times 1024 \times 3$  input image to  $256 \times 1024$  image tokens. Then, these image tokens are projected into language model (OPT-125M Zhang et al. (2022)) dimension via a  $1024 \times 768$  linear layer. Unlike the Vary encoder which only focuses on a single document task under a relatively unitary input shape, we incorporated natural scenes and cropped slices during our pre-training. In the pre-processing stage, images of each shape are directly resized to  $1024 \times 1024$  squares, as square shapes can be used to adapt to images of various aspect ratios with a compromise.

#### 3.2.2 DATA ENGINE TOWARDS ENCODER PRE-TRAINING

In such an encoder pre-training stage, we use about 5M image-text pairs, including 3M scene text OCR data and 2M document OCR data. Their acquisition methods are as follows:

For the natural scene data, the English/Chinese images are sampled from Laion-2B Schuhmann et al. (2022) and Wukong Gu et al. (2022) datasets, respectively. Then, the pseudo ground truth in these diverse real scenes is captured using PaddleOCR Du et al. (2021) tools. Overall, we obtain 2M data with half in Chinese and half in English. For text ground truth, we perform two types of processing: 1) remove the bounding box and combine each text content in order from top to bottom and left to right. 2) crop the text region from the original image according to the bounding box and save it as image slices. The later method 2) allows us to obtain another 1M slice-type image-text pairs.

For the document-level data, we first collect open-source PDF-style files from the Common Crawl and employ the Fitz Python package to extract corresponding dense text content. In such a process, we gain 1.2M full-page PDF-style image-text pairs and 0.8M image slice data. The slice data, including line- and paragraph-level, is cropped from the PDF image via the parsed bounding box.

### 3.3 SCALING UP THE OCR-2.0 KNOWLEDGE VIA MULTI-TASK JOINT-TRAINING

#### 3.3.1 THE FINAL ARCHITECTURE OF GOT

After the pre-training step of the vision encoder, we connect it to a larger language model with more powerful capabilities to build the final architecture of GOT. Here, we adopt the Qwen Bai et al.

(2023a) with 500M parameters as the decoder because it has a relatively small number of parameters while incorporating prior knowledge of multiple languages. The dimension of the connector (i.e., the linear embedding layer) is adjusted into  $1024 \times 1024$  to align with the input channels of the Qwen-0.5B. Hence, GOT enjoys the seamless encoder-decoder paradigm with about 580M parameters in total, which is more computationally resource-friendly and easier to deploy on a consumer-grade GPU with 4G memory. The high compression rate ( $1024 \times 1024$  optical pixels to 256 image tokens) of the encoder saves a lot of token space for the decoder to generate new tokens. Meanwhile, the satisfactory decoding context length (we use about 8K max-length) of the decoder ensures that the GOT can effectively output OCR results under dense scenes.

### 3.3.2 DATA ENGINE FOR JOINT-TRAINING

To inject sufficient OCR-2.0 knowledge into GOT, instead of the above-mentioned plain OCR data, we carefully explore several synthesis methods and data engines in this stage, as shown in Figure 3. We will delve into the details of each type of synthetic data in the following paragraphs.

**Plain OCR data.** We use 80% of the data mentioned in Section 3.2.2 as plain OCR data. To further enhance the robustness of GOT, we also add the handwritten text recognition sub-task, which involves various styles of handwriting from letters and diaries in different languages. We collect the Chinese CASIA-HWDB2 Tek (2024a), English IAM Tek (2024b), and Norwegian NorHand-v3 Tek (2024c) datasets to meet our requirements. For the original image-text pairs with the line-level slice format, 6~8 pairs are grouped and randomly pasted into a blank document page to achieve longer-text handwriting recognition and improve training efficiency.

**Mathpix-markdown formatted data.** Preserving the optical content format is critical to maintaining strong readability for the output results, especially for mathematical formulas and tables. To this end, we use multiple approaches to gather as much formatted data as possible. The details of data collection and production are as follows:

- **Math formulas.** We crawl a large number of  $\LaTeX$  source *.tex* files on Arxiv and extract about 1M formula fragments from them. Next, we transfer the formula sources to Mathpix format and use the Chorme-driver to call Mathpix-markdown-it tool to render the sources to HTML format. We then convert the HTML files to SVGs and save them as PNG images. We find that this rendering method is more than  $20 \times$  faster than directly using the  $\LaTeX$ .
- **Molecular formulas.** We first download the *ChEMBL\_25* file that contains 2M smile sources. Then we use the Mathpix-markdown-it tool and *rdkit.Chem* package to gather about 1M of molecular formula image-text pairs.
- **Table.** From the crawled *.tex* files, we extract about 0.3M table sources and render them into images. Instead of Mathpix-markdown-it, we directly utilize the  $\LaTeX$  as the rendering tool due to its better rendering effects for advanced tables.
- **Full page data.** Using the Nougat Blecher et al. (2023) method, we obtain about 0.5M English markdown PDF-text pairs. Besides, following Vary Wei et al. (2023; 2024), we gather another 0.5M Chinese markdown pairs. We transfer their contents to Mathpix format. Furthermore, we additionally add 0.2M in-house data, which is directly labeled using Mathpix, including books, papers, and financial reports.

**More general OCR data.** We hope GOT can deal with more general optical artificial “characters”. Accordingly, we collect three related challenging tasks and generate the corresponding data. They are sheet music, geometric shapes, and charts, respectively.

- **Sheet music.** Music is a precious part of the cultural heritage and optical music recognition Calvo-Zaragoza et al. (2020); Ríos-Vila et al. (2024) plays an important role in achieving automatic recognition and transcription of sheet music. We choose the GrandStaff Ríos-Vila et al. (2023) dataset as the source to render. The dataset of polyphonic music scores provides the *Humdrum\*\*kern* transcriptions from the excerpts of music. In addition to the existing approximately 100K image-text samples, we also extract some text samples to re-render via the Verovio Python Package. We mainly add new backgrounds from white to real paper styles and randomly add the title and

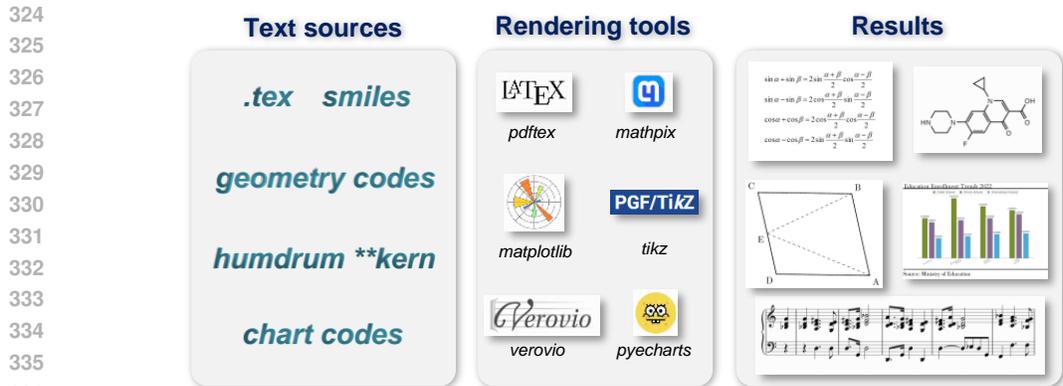


Figure 3: We use six rendering tools to run data engines to make the GOT work well on diverse OCR tasks. We utilize the L<sup>A</sup>T<sub>E</sub>X for tables, Mathpix-markdown-it for math/molecular formulas, Tikz for simple geometric shapes, Verovio for sheet music, and Matplotlib/Pyecharts for charts, respectively.

author information. Note that we only render single-system sheet music due to we don’t have professionals in the relevant field and we do not know how to assemble single-system sheets to a full page. After rendering, we collect about 0.5M samples.

- **Geometric shape.** Geometry is a key capability of LVLMs and is a necessary step towards AGI. GOT is expected to transform optical geometric elements into TikZ Mertz & Slough (2007) text format. TikZ contains some concise commands to produce basic geometric elements and they can be compiled using L<sup>A</sup>T<sub>E</sub>X. We employ TikZ-style points and lines and use the simplest point-line spatial relationship to construct simple basic geometric shapes (*e.g.*, circles, rectangles, triangles, and combined shapes) as well as simple function curves (*e.g.*, straight lines, parabolas, ellipses, hyperbolas, and so on). Through this method, we obtained approximately 1M geometric Tikz data. The geometric rendering is complicated, and our current work is only a preliminary attempt. GOT can only recognize basic geometry at present, yet we believe that with the development of synthetic data technology and OCR-2.0, future models will be able to identify complex geometric shapes.
- **Chart.** Charts are crucial in data visualization and data analysis of several research fields. The proposed GOT refers to the chart structural extraction sub-task as “Chart OCR”, which converts the visual knowledge (*e.g.*, title, source, x-title, y-title, and values) on the chart image into an editable output with a table/Python-dict format. Following OneChart Chen et al. (2024a), the chart image-text pairs are rendered using Matplotlib and Pyecharts tools. Because GOT is only an OCR model, we don’t need the elements of the chart synthesized to be semantically related. Thus, we just randomly extract entity texts (for the title, source, x-title, y-title, etc) from the open-access NLP corpus. The numerical values are random numbers under a controlled distribution. Through this method, we obtained 2M chart data, with half from Matplotlib and half from Pyecharts.

### 3.4 CUSTOMIZING NEW OCR FEATURES BY POST-TRAINING THE DECODER

After compressing the general visual information of the diverse OCR-2.0 optical signals via the above two steps, GOT is ready to perform image-level OCR tasks in various scenarios. Based on this perceptually savvy vision encoder, GOT can be easily tuned to meet the users’ needs for input and output. Here, we customize GOT to enable three new features, *i.e.*, fine-grained, multi-page, and dynamic resolution OCR, by only post-training the decoder part.

#### 3.4.1 FINE-GRAINED DATA ENGINE FOR INTERACTIVE OCR

As a high-interactivity feature, fine-grained OCR Liu et al. (2024a) is the region-level visual perception controlled by spatial coordinates or colors. The user can add box coordinates or color text in the question prompt to request recognition within the region of interest (RoI), avoiding the output of other irrelevant characters. For the natural fine-grained OCR, the source images are from open-source datasets, including RCTW Shi et al. (2017), ReCTS Liu et al. (2019a), and ShopSign Zhang et al. (2019), and COCO-Text Veit et al. (2016) dataset. The datasets mentioned above provide the text

378 bounding boxes, so we can use them to produce fine-grained (region/color prompt) OCR data directly.  
379 For the document-level fine-grained OCR, following Fox Liu et al. (2024a), we filter out those  
380 with the scanned format in the downloaded PDF files and parse the left part using Python packages  
381 (Fitz/PDFminer). We record the page-level images, bounding boxes of each line/paragraph, and the  
382 corresponding texts to produce the ground truth of the box-guided OCR sub-task. For such a task,  
383 each coordinate value is first normalized and then magnified 1000 times. For the color-guided task,  
384 we choose the most commonly used colors (red, green, and blue) as the frame colors and draw them  
385 via the corresponding bounding box on the original image. Overall, we gather about 600K samples.

### 3.4.2 MULTI-CROP DATA ENGINE FOR ULTRA-LARGE-IMAGE OCR

388 GOT supports  $1024 \times 1024$  input size, which is enough for commonly used OCR tasks, e.g., scene  
389 OCR or A4-page PDF OCR. However, dynamic resolution is required for some scenes with huge  
390 images, such as two-page PDF horizontal stitching (commonly occurring when reading papers).  
391 Thanks to our high compression rate encoder, the dynamic resolution of GOT is achieved under a  
392 large sliding window ( $1024 \times 1024$ ), ensuring that our model can complete extreme resolution OCR  
393 tasks with acceptable image tokens. We use the InternVL-1.5 Chen et al. (2024b) cropping method  
394 with tiles max to 12. The ultra-resolution images are synthesized using the single-page PDF data  
395 mentioned above, including horizontal and vertical stitching, leading to 500K image-text pairs.

### 3.4.3 MULTI-PAGE DATA ENGINE FOR BATCHED PDF-FILE OCR

398 For OCR tasks, it is reasonable to use a “for loop” for multi-page processing. We introduce the  
399 multi-page OCR (without “for loop”) feature for GOT due to some formatted PDF data making it  
400 difficult to break pages (to obtain text that is completely incompatible with each page) to further scale  
401 up, such as *.tex* in Arxiv. We hope that with GOT, researchers no longer have to worry about PDF  
402 ground truth page breaks (e.g., Nougat Blecher et al. (2023)), as they can train on multiple pages  
403 directly. To realize such a feature, we randomly sample 2-8 pages from our Mathpix formatted PDF  
404 data and join them together to form a single round OCR task. Each selected page contains text that is  
405 less than 650 tokens, to ensure that the overall length does not exceed 8K. In total, we generate about  
406 200K multi-page OCR data, most of which are interlaced between Chinese and English pages.

## 408 4 EXPERIMENTS

### 410 4.1 IMPLEMENT DETAILS

412 We use  $8 \times 8$  L40s GPUs to train GOT. In the pre-training stage, we optimize all parameters with a  
413 global batch size of 128 and train for 3 epochs. We utilize the AdamW Loshchilov & Hutter (2019)  
414 optimizer and a cosine annealing scheduler Loshchilov & Hutter (2016) with a start learning rate of  
415  $1e-4$ . The max token length in this stage is set to 4096. In the joint-training stage, we put the max  
416 token length to 6000 and train the model with the same optimizer settings as stage 1 for 1 epoch. In  
417 the last post-training stage, we expand the max token length to 8192 to allow the model to support  
418 multi-patch/page OCR features. In this stage, the beginning learning rate is  $2e-5$ , and the epoch is 1.  
419 During each train-data process, 80% of the data from the previous stage is sampled for the following  
420 stage to ensure that the basic ability does not degrade when adding new features.

### 422 4.2 MAIN RESULTS

423 In this section, we verify the performance of GOT on 5 different OCR tasks, including 1) plain  
424 document OCR; 2) scene text OCR; 3) fine-grained document OCR; 4) formatted (Mathpix mark-  
425 down) document OCR; 5) more general character OCR. Note that the test data for each benchmark  
426 undergoes strict text similarity filtering to ensure that it is not included in the training data. Sources  
427 of each test benchmark and model performance analysis are as follows.

#### 429 4.2.1 PLAIN DOCUMENT OCR PERFORMANCE

430 We use the open-source Fox Liu et al. (2024a) benchmark to test the performance of GOT and popular  
431 LVLMS on both Chinese and English PDF OCR. The metrics we used are those commonly in OCR

Table 1: Performance comparison of dense English (en) and Chinese (zh) OCR on document-level pages. The results of other models are from the previous work Liu et al. (2024a).

Method	Size	Edit Distance↓		F1-score↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
UReader Ye et al. (2023b)	7B	0.718	-	0.344	-	0.296	-	0.469	-	0.103	-	0.287	-
LLaVA-NeXT Liu et al. (2024c)	34B	0.430	-	0.647	-	0.573	-	0.881	-	0.478	-	0.582	-
InternVL-ChatV1.5Chen et al. (2024b)	26B	0.393	0.265	0.751	0.816	0.698	0.784	0.917	0.866	0.568	0.622	0.663	0.717
Nougat Blecher et al. (2023)	250M	0.255	-	0.745	-	0.720	-	0.809	-	0.665	-	0.761	-
TextMonkey Liu et al. (2024d)	7B	0.265	-	0.821	-	0.778	-	0.906	-	0.671	-	0.762	-
DocOwl1.5 Hu et al. (2024)	7B	0.258	-	0.862	-	0.835	-	0.962	-	0.788	-	0.858	-
Vary Wei et al. (2023)	7B	0.092	0.113	0.918	0.952	0.906	0.961	0.956	0.944	0.885	0.754	0.926	0.873
Vary-toy Wei et al. (2024)	1.8B	0.082	0.142	0.924	0.914	0.919	0.928	0.938	0.907	0.889	0.718	0.929	0.832
Qwen-VL-Plus Bai et al. (2023b)	-	0.096	0.121	0.931	0.895	0.921	0.903	0.950	0.890	0.893	0.684	0.936	0.828
Qwen-VL-Max Bai et al. (2023b)	>72B	0.057	0.091	0.964	0.931	0.955	0.917	<b>0.977</b>	0.946	0.942	0.756	<b>0.971</b>	0.885
Fox Liu et al. (2024a)	1.8B	0.046	0.061	0.952	0.954	0.957	0.964	0.948	0.946	0.930	0.842	0.954	0.908
<b>GOT</b>	580M	<b>0.035</b>	<b>0.038</b>	<b>0.972</b>	<b>0.980</b>	<b>0.971</b>	<b>0.982</b>	0.973	<b>0.978</b>	<b>0.947</b>	<b>0.878</b>	0.958	<b>0.939</b>

tasks, i.e., edit distance, F1-score, precision, recall, BLEU, and METEOR. Due to the lengthy text of the document, we use word-level segmentation to calculate each indicator. As shown in Table 1, with only 580M, GOT achieves advanced performance on pure text OCR in the document, proving the excellent PDF text perception and recognition ability.

Table 2: Performance of English (en) and Chinese (zh) OCR for scene texts. On these common image-level OCR tasks, GOT can achieve better results compared to other popular models.

Method	Size	Edit Distance↓		F1-score↑		Precision↑		Recall↑		BLEU↑		METEOR↑	
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
UReader Ye et al. (2023b)	7B	0.568	-	0.661	-	0.843	-	0.569	-	0.258	-	0.488	-
LLaVA-NeXT Liu et al. (2024c)	34B	0.499	-	0.558	-	0.637	-	0.538	-	0.379	-	0.678	-
TextMonkey Liu et al. (2024d)	7B	0.331	-	0.743	-	0.827	-	0.710	-	0.521	-	0.728	-
DocOwl1.5 Hu et al. (2024)	7B	0.334	-	0.788	-	0.887	-	0.751	-	0.525	-	0.708	-
InternVL-ChatV1.5Chen et al. (2024b)	26B	0.267	0.123	0.834	0.913	<b>0.942</b>	<b>0.934</b>	0.790	0.902	0.587	0.588	0.744	0.876
Qwen-VL-Max Bai et al. (2023b)	>72B	0.182	0.168	0.881	0.867	0.891	0.878	0.888	0.873	0.586	0.572	0.848	0.845
<b>GOT</b>	580M	<b>0.112</b>	<b>0.096</b>	<b>0.926</b>	<b>0.928</b>	0.934	0.914	<b>0.927</b>	<b>0.954</b>	<b>0.676</b>	<b>0.641</b>	<b>0.896</b>	<b>0.928</b>

#### 4.2.2 SCENE TEXT OCR PERFORMANCE

We collect 400 natural images, half in Chinese and half in English, as the scene text OCR benchmark. All the ground truth in this benchmark are manually corrected. Because the text in the scene image is relatively short, we use character-level segmentation to calculate various metrics. As shown in Table 2, we can see that GOT also works well on natural images, demonstrating the model’s excellent performance on most basic OCR tasks (both document and scene texts).

#### 4.2.3 FORMATTED DOCUMENT OCR PERFORMANCE

Converting the optical PDF image to a markdown-like format is an important feature of an OCR model. To verify this ability of GOT, we carefully prepare 90 pages of samples as a high-quality benchmark. The benchmark, containing both Chinese and English document pages, is first generating pseudo-labels via Mathpix, and then manually correcting for errors. In Table 3, we can see the single-scale (1024×1024) GOT can yield satisfactory results. When we use multi-crop inference, the performance of GOT is further lifted especially on formulas and tables with small texts. The results prove the effectiveness of GOT on documents with formatted outputs. Besides, the dynamic resolution scheme is a good choice when processing higher-resolution images.

#### 4.2.4 FINE-GRAINED OCR PERFORMANCE

We report the fine-grained OCR metrics of GOT. As shown in Table 4, the GOT is overall better than Fox Liu et al. (2024a) on both the bounding box-based and color-based referential OCR tasks, indicating that our model enjoys excellent interactive OCR capabilities.

Table 3: Performances of formatted document (Chinese/English) and more general OCR. Single means the input is the vanilla image and multi-crop represents the dynamic resolution strategy.

	Types	Edit Distance↓	F1-score↑	Precision↑	Recall↑	BLEU↑	METEOR↑
Markdown document	<b>single:</b>						
	All text	0.097	0.942	0.944	0.942	0.877	0.876
	Formula	0.269	0.749	0.771	0.751	0.512	0.716
	Table	0.254	0.867	0.857	0.897	0.756	0.760
	<b>multi-crop:</b>						
	All text	0.086	0.953	0.948	0.960	0.896	0.903
	Formula	0.159	0.865	0.858	0.882	0.628	0.828
Table	0.220	0.878	0.861	0.919	0.779	0.811	
General	Sheet music	0.046	0.939	0.963	0.939	0.900	0.923
	Geometry	0.061	0.884	0.882	0.888	0.766	0.882

Table 4: Comparison of fine-grained document OCR. Without the need to tune the vision encoder, GOT can easily achieve excellent capabilities of box-guided OCR and color-guided OCR.

Metrics	English					Chinese			
	box			color		box		color	
	DocOwl1.5	Fox	GOT	Fox	GOT	Fox	GOT	Fox	GOT
Edit Distance ↓	0.435	0.059	<b>0.041</b>	0.064	<b>0.034</b>	0.042	<b>0.033</b>	0.114	<b>0.040</b>
F1-score ↑	0.670	0.957	<b>0.970</b>	0.940	<b>0.966</b>	0.955	<b>0.965</b>	0.884	<b>0.957</b>
Precision ↑	0.886	0.962	<b>0.973</b>	0.942	<b>0.970</b>	0.966	<b>0.974</b>	0.902	<b>0.969</b>
Recall ↑	0.617	0.955	<b>0.969</b>	0.942	<b>0.964</b>	0.947	<b>0.958</b>	0.873	<b>0.948</b>
BLEU ↑	0.478	0.914	<b>0.926</b>	0.868	<b>0.910</b>	0.885	<b>0.898</b>	0.778	<b>0.884</b>
METEOR ↑	0.569	0.955	<b>0.966</b>	0.938	<b>0.961</b>	0.934	<b>0.942</b>	0.848	<b>0.931</b>

Table 5: Performance on number-centric chart OCR. With sufficient optimization of visual perception and dense information compression, GOT surpasses the popular models by a large margin.

	Metric	Deplot (1.3B)	UniChart (0.26B)	ChartVLM (7.3B)	GPT-4V (>100B)	Qwen-VL (>72B)	GOT (0.58B)
ChartQA-SE	AP@strict	0.614	0.423	0.718	0.504	0.586	<b>0.747</b>
	AP@slight	0.709	53.18	0.814	0.606	0.685	<b>0.845</b>
	AP@high	0.729	0.560	0.842	0.643	0.727	<b>0.867</b>
PlotQA-SE	AP@strict	0.031	0.105	0.038	0.073	0.005	<b>0.133</b>
	AP@slight	0.165	0.260	0.468	0.194	0.042	<b>0.596</b>
	AP@high	0.265	0.269	0.540	0.223	0.120	<b>0.640</b>

#### 4.2.5 MORE GENERAL OCR PERFORMANCE

We utilize the sheet music, geometry, and chart benchmarks to verify GOT’s more general OCR performance. For the first two tasks, we separately render 100 and 180 additional samples as benchmarks, and as can be seen in Table 3, GOT still performs well on these new OCR tasks. For chart OCR, we use structure-extraction version Chen et al. (2024a) ChartQA Masry et al. (2022) and PlotQA Methani et al. (2020) as benchmarks. In Table 5, the chart OCR ability of GOT is even much better than the chart-specific models Liu et al. (2023a); Masry et al. (2023); Xia et al. (2024) and popular LVLMS OpenAI (2023); Bai et al. (2023b). All results demonstrate the effectiveness of our model on more general OCR tasks.

## 5 CONCLUSION

This paper presents a primary OCR-2.0 model that is structurally simpler than OCR-1.0 systems, focuses more on pure OCR tasks than LVLMS, and enjoys superior performance. OCR-2.0 integrates various pan-OCR tasks into one model and is a valuable research direction in model design, data engineering, and application scenarios. We want the simple, elegant, effective, and promising GOT OCR-2.0 model to attract more attention to such a task.

## REFERENCES

- 540  
541  
542 Casia-hwdb2-line. [https://huggingface.co/datasets/Teklia/](https://huggingface.co/datasets/Teklia/CASIA-HWDB2-line)  
543 CASIA-HWDB2-line, 2024a.
- 544  
545 iam-line. <https://huggingface.co/datasets/Teklia/IAM-line>, 2024b.
- 546  
547 Norhand-v3-line. <https://huggingface.co/datasets/Teklia/NorHand-v3-line>,  
548 2024c.
- 549  
550 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
551 Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu,  
552 Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan,  
553 Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin  
554 Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng  
555 Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou,  
556 Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*,  
557 2023a.
- 558  
559 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,  
560 and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization,  
561 text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.
- 562  
563 Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical  
564 understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.
- 565  
566 Jorge Calvo-Zaragoza, Jan Hajič Jr, and Alexander Pacha. Understanding optical music recognition.  
567 *ACM Computing Surveys (CSUR)*, 53(4):1–35, 2020.
- 568  
569 Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun,  
570 Chunrui Han, and Xiangyu Zhang. Onechart: Purify the chart structural extraction via one  
571 auxiliary token. *arXiv preprint arXiv:2404.09987*, 2024a.
- 572  
573 Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi  
574 Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial  
575 multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.
- 576  
577 Yuning Du, Chenxia Li, Ruoyu Guo, Cheng Cui, Weiwei Liu, Jun Zhou, Bin Lu, Yehua Yang, Qiwen  
578 Liu, Xiaoguang Hu, et al. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. *arXiv preprint*  
579 *arXiv:2109.03144*, 2021.
- 580  
581 Jiayi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui  
582 Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal  
583 pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431,  
584 2022.
- 585  
586 Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei  
587 Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding.  
588 *arXiv preprint arXiv:2403.12895*, 2024.
- 589  
590 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-  
591 training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*,  
592 2023a.
- 593  
594 Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun  
595 Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models.  
596 In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102,  
597 2023b.
- 598  
599 Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer  
600 backbones for object detection. In *European conference on computer vision*, pp. 280–296. Springer,  
601 2022.

- 594 Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection  
595 with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis  
596 and machine intelligence*, 45(1):919–931, 2022.
- 597 Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian  
598 Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document  
599 understanding. *arXiv preprint arXiv:2405.14295*, 2024a.
- 600 Chenglong Liu, Haoran Wei, Jinze Yang, Jintao Liu, Wenxi Li, Yuchen Guo, and Lu Fang. Gigahu-  
601 mandet: Exploring full-body detection on gigapixel-level images. In *Proceedings of the AAAI  
602 Conference on Artificial Intelligence*, volume 38, pp. 10092–10100, 2024b.
- 603 Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton  
604 Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual  
605 language reasoning by plot-to-table translation. In *Findings of the 61st Annual Meeting of the  
606 Association for Computational Linguistics*, 2023a. URL [https://arxiv.org/abs/2212.  
607 10505](https://arxiv.org/abs/2212.10505).
- 608 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.
- 609 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee.  
610 Llava-next: Improved reasoning, ocr, and world knowledge, January 2024c. URL [https:  
611 //llava-vl.github.io/blog/2024-01-30-llava-next/](https://llava-vl.github.io/blog/2024-01-30-llava-next/).
- 612 Xi Liu, Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong  
613 Wang, Minghui Liao, et al. Icdar 2019 robust reading challenge on reading chinese text on  
614 signboard. *arXiv preprint arXiv:1912.09641*, 2019a.
- 615 Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection  
616 via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019b.
- 617 Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey:  
618 An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*,  
619 2024d.
- 620 Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv  
621 preprint arXiv:1608.03983*, 2016.
- 622 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- 623 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A bench-  
624 mark for question answering about charts with visual and logical reasoning. *arXiv preprint  
625 arXiv:2203.10244*, 2022.
- 626 Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A  
627 universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint  
628 arXiv:2305.14761*, 2023.
- 629 Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document  
630 images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*,  
631 pp. 2200–2209, 2021.
- 632 Andrew Mertz and William Slough. Graphics with tikz. *The PracTEX Journal*, 1:1–22, 2007.
- 633 Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over  
634 scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer  
635 Vision*, pp. 1527–1536, 2020.
- 636 OpenAI. Gpt-4 technical report, 2023.
- 637 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
638 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
639 models from natural language supervision. In *International conference on machine learning*, pp.  
640 8748–8763. PMLR, 2021.

- 648 Antonio Ríos-Vila, David Rizo, José M Iñesta, and Jorge Calvo-Zaragoza. End-to-end optical  
649 music recognition for pianoform sheet music. *International Journal on Document Analysis and*  
650 *Recognition (IJ DAR)*, 26(3):347–362, 2023.
- 651 Antonio Ríos-Vila, Jorge Calvo-Zaragoza, and Thierry Paquet. Sheet music transformer: End-to-end  
652 optical music recognition beyond monophonic transcription. *arXiv preprint arXiv:2402.07596*,  
653 2024.
- 654  
655 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi  
656 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An  
657 open large-scale dataset for training next generation image-text models. *Advances in Neural*  
658 *Information Processing Systems*, 35:25278–25294, 2022.
- 659  
660 Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian  
661 Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017*  
662 *14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pp.  
663 1429–1434. IEEE, 2017.
- 664 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and  
665 Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference*  
666 *on computer vision and pattern recognition*, pp. 8317–8326, 2019.
- 667  
668 Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and  
669 benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*,  
670 2016.
- 671  
672 Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui  
673 Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models.  
674 *arXiv preprint arXiv:2312.06109*, 2023.
- 675  
676 Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui  
677 Han, and Xiangyu Zhang. Small language model meets with reinforced vision vocabulary. *arXiv*  
*preprint arXiv:2401.12503*, 2024.
- 678  
679 Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou,  
680 Botian Shi, Junchi Yan, and Yu Qiao. Chartx & chartvlm: A versatile benchmark and foundation  
681 model for complicated chart reasoning, 2024.
- 682  
683 Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu,  
684 Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model  
for document understanding. *arXiv preprint arXiv:2307.02499*, 2023a.
- 685  
686 Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian,  
687 Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding  
688 with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023b.
- 689  
690 Chongsheng Zhang, Guowen Peng, Yuefeng Tao, Feifei Fu, Wei Jiang, George Almpantidis, and  
691 Ke Chen. Shoptsign: A diverse scene text dataset of chinese shop signs in street views. *arXiv*  
*preprint arXiv:1903.10412*, 2019.
- 692  
693 Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Adaptive boundary  
694 proposal network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF International*  
695 *Conference on Computer Vision*, pp. 1305–1314, 2021.
- 696  
697 Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher  
698 Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language  
models. *arXiv preprint arXiv:2205.01068*, 2022.
- 699  
700 Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document  
701 layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*,  
pp. 1015–1022. IEEE, 2019.

## A APPENDIX

In this section, we provide sufficient output results of GOT to show its outstanding OCR performance. We also demonstrate the format of the corresponding input prompt for different types of OCR tasks.

### Prompt: OCR with format:

```

6 1 Kinetic theory
Note that
|2e(e, u) - u|^2 = 1 and |e + u|^2 = 2(1 + (e, u)) = 2(e + u, u).
We prepare the proof by the following lemma.
Lemma 1.3. Let Φ be an appropriate test function and u ∈ S^2. Then (cf. (1.17))
∫_{S^2} Φ(e + u) de = 4 ∫_{S^2_+(u)} (e, u) Φ(2e(e, u)) de (1.20)
= 4 ∫_{S^2_+(u)} |(e, u)| Φ(2e(e, u)) de = 2 ∫_{S^2} |(e, u)| Φ(2e(e, u)) de.
Proof. Introducing spherical coordinates φ_1 ∈ [0, π], φ_2 ∈ [0, 2π] such that
e_1 = cos φ_1, e_2 = sin φ_1 cos φ_2, e_3 = sin φ_1 sin φ_2
and u = (1, 0, 0), one obtains
∫_{S^2} Φ(e + u) de = (1.21)
∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 Φ(1 + cos φ_1, sin φ_1 cos φ_2, sin φ_1 sin φ_2).
On the other hand, using the elementary properties
sin 2α = 2 sin α cos α, 1 + cos 2α = 2 cos^2 α,
one obtains
∫_{S^2_+(u)} (e, u) Φ(2e(e, u)) de = ∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 cos φ_1 ×
Φ(2 cos^2 φ_1, 2 cos φ_1 sin φ_1 cos φ_2, 2 cos φ_1 sin φ_1 sin φ_2)
= ∫_0^π dφ_1 ∫_0^{2π} dφ_2 1/2 sin 2φ_1 ×
Φ(1 + cos 2φ_1, sin 2φ_1 cos φ_2, sin 2φ_1 sin φ_2)
= 1/4 ∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 Φ(1 + cos φ_1, sin φ_1 cos φ_2, sin φ_1 sin φ_2).
Comparing (1.21) and (1.22) gives (1.20).
Proof of Theorem 1.2. Using Lemma 1.3 with
Φ(z) = B(v, w, z - w) [f(v + |v-w|/2 z) f(w - |v-w|/2 z) - f(v)f(w)]

```

### Output:

```

Note that
|2e(e, u) - u|^2 = 1 and |e + u|^2 = 2(1 + (e, u)) = 2(e + u, u).
We prepare the proof by the following lemma.
Lemma 1.3. Let Φ be an appropriate test function and u ∈ S^2. Then (cf. (1.17))
∫_{S^2} Φ(e + u) de = 4 ∫_{S^2_+(u)} (e, u) Φ(2e(e, u)) de
= 4 ∫_{S^2_+(u)} |(e, u)| Φ(2e(e, u)) de = 2 ∫_{S^2} |(e, u)| Φ(2e(e, u)) de
Proof. Introducing spherical coordinates φ_1 ∈ [0, π], φ_2 ∈ [0, 2π] such that
e_1 = cos φ_1, e_2 = sin φ_1 cos φ_2, e_3 = sin φ_1 sin φ_2
and u = (1, 0, 0), one obtains
∫_{S^2} Φ(e + u) de =
∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 Φ(1 + cos φ_1, sin φ_1 cos φ_2, sin φ_1 sin φ_2).
On the other hand, using the elementary properties
sin 2α = 2 sin α cos α, 1 + cos 2α = 2 cos^2 α,
one obtains
∫_{S^2_+(u)} (e, u) Φ(2e(e, u)) de = ∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 cos φ_1 ×
Φ(2 cos^2 φ_1, 2 cos φ_1 sin φ_1 cos φ_2, 2 cos φ_1 sin φ_1 sin φ_2)
= ∫_0^π dφ_1 ∫_0^{2π} dφ_2 1/2 sin 2φ_1 ×
Φ(1 + cos 2φ_1, sin 2φ_1 cos φ_2, sin 2φ_1 sin φ_2)
= 1/4 ∫_0^π dφ_1 ∫_0^{2π} dφ_2 sin φ_1 Φ(1 + cos φ_1, sin φ_1 cos φ_2, sin φ_1 sin φ_2).
Comparing (1.21) and (1.22) gives (1.20).
Proof of Theorem 1.2. Using Lemma 1.3 with
Φ(z) = B(v, w, z - w) [f(v + |v-w|/2 z) f(w - |v-w|/2 z) - f(v)f(w)]

```

Method (val set)	Backbone	Input size	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
<b>Center-guided (anchor-based):</b>								
Faster R-CNN [Ren et al., 2015] w/ FPN	ResNet-101	1000 × 600	36.2	59.1	39.0	18.2	39.0	48.2
RetinaNet w/ FPN [Lin et al., 2017b]	ResNet-101	1333 × 800	40.8	61.1	44.1	24.1	44.2	51.2
Cascade R-CNN [Cai and Vasconcelos, 2018]	ResNet-101	1333 × 800	42.8	62.1	46.3	23.7	45.5	55.2
YOLOv4 [Bochkovskiy et al., 2020]	CSPDarkNet-53	608 × 608	43.5	65.7	47.3	26.7	46.7	53.3
<b>Center-guided (anchor-free):</b>								
CenterNet [Zhou et al., 2019]	Hourglass-104	512 × 512	42.1	61.1	45.9	24.1	45.5	52.8
FCOS [Tian et al., 2019]	ResNeXt-101	1333 × 800	42.1	62.1	45.2	25.6	44.9	52.0
Reppoints [Yang et al., 2019]	ResNet-101	1333 × 800	45.0	66.1	49.0	26.6	48.6	57.5
<b>Corner-guided:</b>								
CenterNet [Duan et al., 2019]	Hourglass-104	511 × 511	44.9	62.4	48.1	25.6	47.4	57.4
CenterNet [Duan et al., 2019]	Hourglass-104	511 × 511	45.8	63.0	49.3	25.0	48.2	58.7
CPN [Koussas et al., 2020]	Hourglass-104	511 × 511	47.0	65.0	51.0	26.5	50.2	60.7
CornerNet baseline [Law and Deng, 2018]	Hourglass-104	511 × 511	40.5	56.5	43.1	19.4	42.7	53.9
w/ Corner Affinity [Improvement]	Hourglass-104	511 × 511	46.3	64.0	49.9	27.4	49.3	58.7
			+ 5.8	+ 7.5	+ 6.8	+ 8.0	+ 6.6	+ 4.8
<b>Method (val set)</b>								
DETR [Carion et al., 2020]	Backbone	Feature	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
DETR [Carion et al., 2020]	ResNet-50	Encoder	43.3	63.1	45.9	22.5	47.3	61.1
Sparse R-CNN [Sun et al., 2021]	ResNet-101	Encoder	44.9	64.7	47.7	23.7	49.5	62.3
Sparse R-CNN [Sun et al., 2021]	ResNet-101	FPN	42.8	61.2	45.7	20.7	44.6	57.6
w/ Corner Affinity	ResNet-101	FPN	44.1	62.1	47.2	26.1	46.3	59.7
	Hourglass-52	Single	43.2	60.8	46.1	25.1	46.8	57.7
	Hourglass-104	Single	45.1	62.9	48.3	26.7	48.5	59.8

Method (test-dev set)	Backbone	Input size	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
<b>Center-guided (anchor-based):</b>								
Faster R-CNN [Ren et al., 2015] w/ FPN	ResNet-101	1000 × 600	36.2	59.1	39.0	18.2	39.0	48.2
RetinaNet w/ FPN [Lin et al., 2017b]	ResNet-101	1333 × 800	40.8	61.1	44.1	24.1	44.2	51.2
Cascade R-CNN [Cai and Vasconcelos, 2018]	ResNet-101	1333 × 800	42.8	62.1	46.3	23.7	45.5	55.2
YOLOv4 [Bochkovskiy et al., 2020]	CSPDarkNet-53	608 × 608	43.5	65.7	47.3	26.7	46.7	53.3
<b>Center-guided (anchor-free):</b>								
CenterNet [Zhou et al., 2019]	Hourglass-104	512 × 512	42.1	61.1	45.9	24.1	45.5	52.8
FCOS [Tian et al., 2019]	ResNeXt-101	1333 × 800	42.1	62.1	45.2	25.6	44.9	52.0
Reppoints [Yang et al., 2019]	ResNet-101	1333 × 800	45.0	66.1	49.0	26.6	48.6	57.5
<b>Corner-guided:</b>								
CenterNet [Duan et al., 2019]	Hourglass-104	511 × 511	44.9	62.4	48.1	25.6	47.4	57.4
CenterNet [Duan et al., 2019]	Hourglass-104	511 × 511	45.8	63.0	49.3	25.0	48.2	58.7
CPN [Koussas et al., 2020]	Hourglass-104	511 × 511	47.0	65.0	51.0	26.5	50.2	60.7
CornerNet baseline [Law and Deng, 2018]	Hourglass-104	511 × 511	40.5	56.5	43.1	19.4	42.7	53.9
w/ Corner Affinity	Hourglass-104	511 × 511	46.3	64.0	49.9	27.4	49.3	58.7
Improvement			+ 5.8	+ 7.5	+ 6.8	+ 8.0	+ 6.6	+ 4.8
<b>Method (val set)</b>								
DETR [Carion et al., 2020]	Backbone	Feature	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
DETR [Carion et al., 2020]	ResNet-50	Encoder	43.3	63.1	45.9	22.5	47.3	61.1
DETR [Carion et al., 2020]	ResNet-101	Encoder	44.9	64.7	47.7	23.7	49.5	62.3
Sparse R-CNN [Sun et al., 2021]	ResNet-101	FPN	42.8	61.2	45.7	20.7	44.6	57.6
Sparse R-CNN [Sun et al., 2021]	ResNet-101	FPN	44.1	62.1	47.2	26.1	46.3	59.7
w/ Corner Affinity	Hourglass-52	Single	43.2	60.8	46.1	25.1	46.8	57.7
	Hourglass-104	Single	45.1	62.9	48.3	26.7	48.5	59.8

$$\begin{aligned}
 d_L(C_L, \{v'\}) &= |C_L| + |\{v'\}| + 2(d_T(C_L, v') - 1) \\
 &= |C_v| - 1 + |S_v^*| + 2(\text{rad } T - 1) \\
 &= |C_v| + |S_v^*| + 2(d_T(C_v, S_v^*) - 1) \\
 &\quad - 1 + 2(\text{rad } T - d_T(C_v, S_v^*)) \\
 &= d_L(C_v, S_v^*) + 1 + 2(\text{rad } T - 1 - d_T(C_v, S_v^*)) \\
 &= e_L(C_v) + 1 + 2(\text{rad } T - 1 - d_T(C_v, S_v^*)).
 \end{aligned}$$

$$\begin{aligned}
 d_L(C_L, \{v'\}) &= |C_L| + |\{v'\}| + 2(d_T(C_L, v') - 1) \\
 &= |C_v| - 1 + |S_v^*| + 2(\text{rad } T - 1) \\
 &= |C_v| + |S_v^*| + 2(d_T(C_v, S_v^*) - 1) \\
 &\quad - 1 + 2(\text{rad } T - d_T(C_v, S_v^*)) \\
 &= d_L(C_v, S_v^*) + 1 + 2(\text{rad } T - 1 - d_T(C_v, S_v^*)) \\
 &= e_L(C_v) + 1 + 2(\text{rad } T - 1 - d_T(C_v, S_v^*)).
 \end{aligned}$$

Figure 4: The formatted text OCR ability of GOT. GOT works well on full-page texts and table/formula slice texts. These input forms are the most commonly used in document OCR, which proves that GOT has great prospects in application.

756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

## Prompt: OCR:

[21], and GuidedBackpropagation [22]) to explain image captioning predictions with respect to the image content and the words of the sentence generated so far. These approaches provide high-resolution image explanations for CNN models [22], [23]. LRP also provides plausible explanations for LSTM architectures [24], [25]. Figure 1 shows an example of the explanation results of attention-guided image captioning models. Taking LRP as an example, both positive and negative evidence is shown in two aspects: 1) for image explanations, the contribution of the image input is visualized as heatmaps; 2) for linguistic explanations, the contribution of the previously generated words to the latest predicted word is shown.

The explanation results in Figure 1 exhibit intuitive correspondence of the explained word to the image content and the related sequential input. However, to our best knowledge, few works quantitatively analyze how accurate the image explanations are grounded to the relevant image content and whether the highlighted inputs are used as evidence by the model to make decisions. We study the two questions by quantifying the grounding property of attention and explanation methods and by designing an ablation experiment for both the image explanations and linguistic explanations. We will demonstrate that explanation methods can generate image explanations with accurate spatial grounding property, meanwhile, reveal more related inputs (pixels of the image input and words of the linguistic sequence input) that are used as evidence for the model decisions. Also, explanation methods can disentangle the contributions of the image and text inputs and provide more interpretable information than purely image-centered attention.

With explanation methods [26], we have a deeper understanding of image captioning models beyond visualizing the attention. We also observe that image captioning models sometimes hallucinate words from the learned sentence correlations without looking at the images and sometimes use irrelevant evidence to make predictions. The hallucination problem is also discussed in [27], where the authors state that it is possibly caused by language priors or visual mis-classification, which could be partially due to the biases present in the dataset. The image captioning models tend to generate those words and sentence patterns that appear more frequently during training. The language priors are helpful, though, in some cases, [28] incorporates the inductive bias of natural language with scene graphs to facilitate image captioning. However, language bias is not always correct, for example, not only men ride snowboards [29] and bananas are not always yellow [30], [31]. To this end, [29] and [31] attempted to generate more grounded captions by guiding the model to make the right decisions using the right reasons. They adopted additional annotations, such as the instance segmentation annotation and the human-annotated rank of the relevant image patches, to design new losses for training.

In this paper, we reduce object hallucination by a simple *LRP-inference fine-tuning* (LRP-IFT) strategy, without any additional annotations. We firstly show that the explanations, especially LRP, can weakly differentiate the grounded (true-positive) and hallucinated (false-positive) words. Secondly,

based on the findings that LRP reveals the related features of the explained words and that the sign of its relevance scores indicates supporting versus opposing evidence (as shown in Figure 1), we utilize LRP explanations to design a re-weighting mechanism for the context representation. During fine-tuning, we up-scale the supporting features and down-scale the opposing ones using a weight calculated from LRP relevance scores. Finally, we use the re-weighted context representation to predict the next word for fine-tuning.

LRP-IFT is different from standard fine-tuning which weights the gradients of parameters with small learning rates to gradually adapt the model parameters. Instead, it pinpoints the related features/evidence for a decision and guides the model to tune more on those related features. This fine-tuning strategy resembles how we correct our cognition bias. For example, when we see a green banana, we will update the color feature of bananas and keep the other features such as the shape.

We will demonstrate that LRP-IFT can help to de-bias image captioning models from frequently occurring object words. Though language bias is intrinsic, we can guide the model to be more precise when generating frequent object words rather than hallucinate them. We implement the LRP-IFT on top of pre-trained image captioning models trained with Flickr30K [32] and MSCOCO2017 [33] datasets and effectively improve the mean average precision (mAP) of predicted frequent object words evaluated across the test set. At the same time, the overall performance in terms of sentence-level evaluation metrics is maintained.

The contributions of this paper are as follows:

- We establish explanation methods that disentangle the contributions of the image and text inputs and explain image captioning models beyond visualizing attention.
- We quantitatively measure and compare the properties of explanation methods and attention mechanisms, including tasks of finding the related features/evidence for model decisions, grounding to image content, and the capability of debugging the models (in terms of providing possible reasons for object hallucination and differentiating hallucinated words).
- We propose an LRP-inference fine-tuning strategy that reduces object hallucination and guides the models to be more precise and grounded on image evidence when predicting frequent object words. Our proposed fine-tuning strategy requires no additional annotations and successfully improves the mean average precision of predicted frequent object words.

In the rest of this paper, Section II introduces recent image captioning models, the state-of-the-art explanation methods for neural networks, and other related works. In Section III, we will introduce the image captioning model structures applied in this paper. The adaptations of explanation methods to attention-guided image captioning models are summarized in Section IV. The analyses of attention and explanations and our proposed LRP-inference fine-tuning strategy are introduced in Section V.

## Output:

[21], and GuidedBackpropagation [22]) to explain image captioning predictions with respect to the image content and the words of the sentence generated so far. These approaches provide high-resolution image explanations for CNN models [22], [23]. LRP also provides plausible explanations for LSTM architectures [24], [25]. Figure 1 shows an example of the explanation results of attention-guided image captioning models. Taking LRP as an example, both positive and negative evidence is shown in two aspects: 1) for image explanations, the contribution of the image input is visualized as heatmaps; 2) for linguistic explanations, the contribution of the previously generated words to the latest predicted word is shown.

The explanation results in Figure 1 exhibit intuitive correspondence of the explained word to the image content and the related sequential input. However, to our best knowledge, few works quantitatively analyze how accurate the image explanations are grounded to the relevant image content and whether the highlighted inputs are used as evidence by the model to make decisions. We study the two questions by quantifying the grounding property of attention and explanation methods and by designing an ablation experiment for both the image explanations and linguistic explanations. We will demonstrate that explanation methods can generate image explanations with accurate spatial grounding property, meanwhile, reveal more related inputs (pixels of the image input and words of the linguistic sequence input) that are used as evidence for the model decisions. Also, explanation methods can disentangle the contributions of the image and text inputs and provide more interpretable information than purely image-centered attention.

With explanation methods [26], we have a deeper understanding of image captioning models beyond visualizing the attention. We also observe that image captioning models sometimes hallucinate words from the learned sentence correlations without looking at the images and sometimes use irrelevant evidence to make predictions. The hallucination problem is also discussed in [27], where the authors state that it is possibly caused by language priors or visual mis-classification, which could be partially due to the biases present in the dataset. The image captioning models tend to generate those words and sentence patterns that appear more frequently during training. The language priors are helpful, though, in some cases, [28] incorporates the inductive bias of natural language with scene graphs to facilitate image captioning. However, language bias is not always correct, for example, not only men ride snowboards [29] and bananas are not always yellow [30], [31]. To this end, [29] and [31] attempted to generate more grounded captions by guiding the model to make the right decisions using the right reasons. They adopted additional annotations, such as the instance segmentation annotation and the human-annotated rank of the relevant image patches, to design new losses for training.

In this paper, we reduce object hallucination by a simple *LRP-inference fine-tuning* (LRP-IFT) strategy, without any additional annotations. We firstly show that the explanations, especially LRP, can weakly differentiate the grounded (true-positive) and hallucinated (false-positive) words. Secondly, based on the findings that LRP reveals the related features of the explained words and that the sign of its relevance scores indicates supporting versus opposing evidence (as shown in Figure 1), we utilize LRP explanations to design a re-weighting mechanism for the context representation. During fine-tuning, we up-scale the supporting features and down-scale the opposing ones using a weight calculated from LRP relevance scores. Finally, we use the re-weighted context representation to predict the next word for fine-tuning.

LRP-IFT is different from standard fine-tuning which weights the gradients of parameters with small learning rates to gradually adapt the model parameters. Instead, it pinpoints the related features/evidence for a decision and guides the model to tune more on those related features. This fine-tuning strategy resembles how we correct our cognition bias. For example, when we see a green banana, we will update the color feature of bananas and keep the other features such as the shape. We will demonstrate that LRP-IFT can help to de-bias image captioning models from frequently occurring object words. Though language bias is intrinsic, we can guide the model to be more precise when generating frequent object words rather than hallucinate them. We implement the LRP-IFT on top of pre-trained image captioning models trained with Flickr30K [32] and MSCOCO2017 [33] datasets and effectively improve the mean average precision (mAP) of predicted frequent object words evaluated across the test set. At the same time, the overall performance in terms of sentence-level evaluation metrics is maintained.

The contributions of this paper are as follows:

- We establish explanation methods that disentangle the contributions of the image and text inputs and explain image captioning models beyond visualizing attention.
- We quantitatively measure and compare the properties of explanation methods and attention mechanisms, including tasks of finding the related features/evidence for model decisions, grounding to image content, and the capability of debugging the models (in terms of providing possible reasons for object hallucination and differentiating hallucinated words).
- We propose an LRP-inference fine-tuning strategy that reduces object hallucination and guides the models to be more precise and grounded on image evidence when predicting frequent object words. Our proposed fine-tuning strategy requires no additional annotations and successfully improves the mean average precision of predicted frequent object words.

In the rest of this paper, Section II introduces recent image captioning models, the state-of-the-art explanation methods for neural networks, and other related works. In Section III, we will introduce the image captioning model structures applied in this paper. The adaptations of explanation methods to attention-guided image captioning models are summarized in Section IV. The analyses of attention and explanations and our proposed LRP-inference fine-tuning strategy are introduced in Section V.

Figure 5: The plain text (document) OCR ability of GOT. For double-column documents with high text density, GOT can still handle them well, proving the excellent text perception ability.

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

**Prompt: OCR**



**Output:**

禾不锈钢 6米剪板折弯



Tiredness

The same source quotes yet another as stating, "You cannot be saved if you don't believe in the Trinity."

The same source quotes yet another as stating, "You cannot be saved if you don't believe in the Trinity."

**Prompt: OCR/ [green]OCR/ OCR with format:**

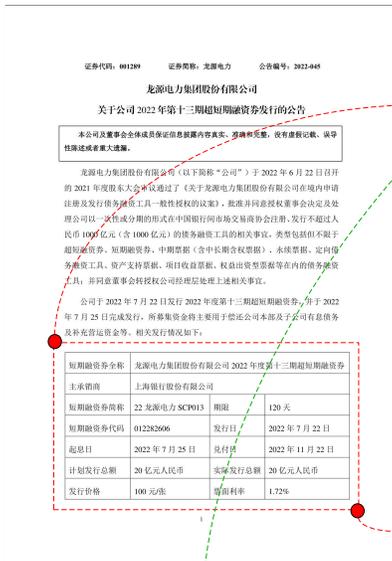


双十二 返场狂欢 12.13-12.15 爆款6折起

返场狂欢

双十二返场狂欢  
12.13-12.15  
爆款6折起

**Prompt: [x1,y1,x2,y2] OCR with format:**



[x1,y1]

短期融资券全称	龙源电力集团股份有限公司 2022 年度第十三期超短期融资券		
主承销商	上海银行股份有限公司		
短期融资券简称	22 龙源电力 SCP013	期限	120 天
短期融资券代码	012282606	发行日	2022 年 7 月 22 日
起息日	2022 年 7 月 25 日	兑付日	2022 年 11 月 22 日
计划发行总额	20 亿元人民币	实际发行总额	20 亿元人民币
发行价格	100 元/张	票面利率	1.72%

[x2,y2]



\text{{饅饅面}}

Figure 6: Scene OCR and fine-grained OCR results of GOT. We equip GOT with more interactive fine-grained OCR tasks, allowing it to output OCR results of regions of interest based on prompts.

### Prompt: OCR with format upon the patch reference:

The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-20)

#### GigaHumanDet: Exploring Full-Body Detection on Gigapixel-Level Images

Chenglong Liu<sup>1,2\*</sup>, Haoran Wei<sup>3,4\*</sup>, Jinze Yang<sup>1</sup>, Jintao Liu<sup>1</sup>,  
Wenci Li<sup>1</sup>, Yuchen Guo<sup>1,5</sup>, Lu Fang<sup>1,6</sup>

<sup>1</sup>MOECC Technology  
<sup>2</sup>BNRist, Tsinghua University  
<sup>3</sup>BNRist, Tsinghua University  
<sup>4</sup>Shanghai Jiao Tong University  
{luchenglong20, weihoran18, yangjinze20, lijintao201}@mails.ac.cn  
wenci.li@sjtu.edu.cn, yuchen.guo@sjtu.edu.cn, fanglu@sjtu.edu.cn

**Abstract**  
Performing person detection in super-high-resolution images has been a challenging task. For such a task, modern detectors, which usually encode a box using center and width/height, struggle with accuracy due to two factors: 1) Human characteristics: people come in various postures and the center with high freedom is difficult to capture robust visual pattern; 2) Image characteristics: due to vast scale diversity of input (gigapixel-level), distance regression (for width and height) is hard to pinpoint, especially for a person, with substantial scale, who is near the camera. To address these challenges, we propose GigaHumanDet, an innovative solution aimed at further enhancing detection accuracy for gigapixel-level images. GigaHumanDet employs the corner matching method to avoid the potential issues of a high degree of freedom in corner predicting. To better distinguish similar-looking persons and enforce instance consistency of corner pairs, an instance-guided learning approach is designed to capture discriminative individual semantics. Further, we devise reliable shape-aware bodies equipped with a multi-precision strategy as the human corner matching guidance to be appropriately adapted to the single-view large scene. Experimental results on PANDA and STCcrowd datasets show the superiority and strong adaptability of our design. Notably, our model achieves 82.4% in terms of AP<sub>50</sub>, outperforming current state-of-the-arts by more than 10%.

**Introduction**  
Person detection is a fundamental and critical task for human-centric visual analysis. Recently, the resolution has reached gigapixel level (e.g., 25k × 14k pixels) (Wang et al., 2020), posing a challenge for object detectors to cover the analysis at large-scale quickly with clear local details. Most modern detectors fall into the center-regression-guided type (Cai and Vasconcelos 2018; Tian et al., 2019; Heun et al., 2021; Zhang et al., 2022), which locates objects via centers and prefers clear four boundaries to regress width/height (see Figure 1 (a)). But for the human-centric task in gigapixel images, it seems that the requirements of

the above methods cannot be perfectly met. Numerous individuals exhibit various postures, leading to the instability of human center visual pattern (e.g., Figure 1 (a)); the center of the bending man lies on the ground while the center of the standing woman is occluded, further making it harder to determine four boundaries. Besides, because people are at different distances from the gigapixels (Yuan et al., 2017), there is vast variation in human size (e.g., Figure 1 (c)) from

10092

The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-20)

**Figure 2: Framework of GigaHumanDet.** By discriminative semantics and robust body embeddings, our model performs gigapixel-level full-body detection by matching corner pairs.

125 × 336 to 602 × 1465). Thus, long-range regression errors directly affect the quality of boxes produced by modern detectors that regress width/height, and the box may even be mistakenly considered a negative sample since [w] with the ground truth is under threshold (e.g., yields a semi-finished box that only covers half of the body (Figure 1(b)).

Instead of the above detectors, another is the corner-semantic-guided type (Law and Deng 2018), which estimates two corners of the box via boundary clues. We find this type is more suitable for gigapixel-level human detection. As the best knowledge, the degree of freedom of determining a corner is lower than that of a center (a perfect corner needs four bodies while a center only needs two), and manually generating a human annotation box via two corners is a strong prior knowledge for us. In Figure 1 (d), although the ideal center sample is occluded, each corner point can still be located by two clear boundary clues. Further, the object representation of corner detectors do not rely on distance regression, so they are robust to multi-scale changes of the object and enjoy greater potential to generate higher-quality boxes in such large-scale single-view scenes. However, existing corner-semantic-guided methods designed for common scenarios can not work well in the extremely large-scale full-body detection task, especially their heuristic corner-matching algorithms. For example, CornerNet (Law and Deng 2018) determines corner pairs by local response similarity, which can cause severe confusion in detecting humans with similar appearances in the large spatial scale scene. Cornerlet (Dui et al., 2019) predicts one corner point to filter out the false positives (FP), but it still can not perform well when applied to a gigapixel image, because the central region of regression FP often covers the center point of a third person. Therefore, we may ask: Can we devise a robust corner matching algorithm for gigapixel-level full-body detection?

To answer the question and expand the applicability of corner-semantic-guided detectors for full-body detection on gigapixel-level images, we propose GigaHumanDet. GigaHumanDet predicts two developed corners of the bounding box, each of which is expressed explicitly and requires only two bodies to be determined, improving the robustness in gigapixel crowds with various postures. To acquire discriminative corner semantics for similar-looking persons, we

devise an instance-guided learning strategy as shown in Figure 2. For reliable matching corners, we devise shape-aware bodies which encodes the body shape embeddings at the corresponding corner location. To further purify the body shape embedding and make it more tolerant to the drastic scale variation, a Gaussian-inspired multi-precision regression strategy is devised to alleviate the difficulty and inaccuracy of once long-range regression. Note that the distance regression is decoupled from the corner prediction, so the corner pair can decide an accurate object box without being affected by the distance regression error. Equipped with robust bodies, GigaHumanDet can reap promising performance by pairing corners with consistent body shapes. Experimental results on gigapixel-level PANDA (Wang et al., 2020) benchmark show GigaHumanDet yields a new SOTA accuracy in terms of 82.4% on AP<sub>50</sub>, boosting 60.8% than CornerNet baseline and surpassing other methods by more than 10%. Further, the competitive results on STCcrowd verify the applicability of our method to general pedestrian detection. Our contributions can be summarized as follows:

- This paper unleashes the power of corner modeling approach on gigapixel-level full-body detection, and we demonstrate that it is more suitable and robust than center-regression-guided methods for this task.
- We design instance-guided learning and multi-precision strategy to acquire discriminative corner semantics. We propose shape-aware bodies to provide reliable corner-matching guidance for large-scale dense scenes.
- GigaHumanDet achieves the SOTA accuracy on the gigapixel-level detection task and outperforms other advanced methods by 10%.

**Related Work**

**Object Detection on High-Resolution Images**  
Object detection on large-scale HR images has become a challenging task. A gigapixel-level human-centric PANDA dataset (Wang et al., 2020) is published and its resolution has reached 25,000 × 14,000. Due to wide FoV and high resolution, pedestrians have various postures as well as occlusions, and drastic scale changes exist, which degrades the accuracy of modern detectors developed on COCO (Lin et al., 2014).

**Center-Guided and Corner-Guided Methods**  
The center-guided detectors utilize centers and width/height to encode boxes. Most of them (He et al., 2017; Cai and Vasconcelos 2018; Ce et al., 2021) take the center as a reference point and regress object size. Different from the above methods, the corner-guided detector (Law and Deng 2018; Dui et al., 2019) is proposed to estimate corners and match them to compose the final box. But for gigapixel images, humans with similar appearances cause great confusion when matching corners. Our GigaHumanDet employs the corner modeling method and tackles the matching problem by setting robust body-shape embedding for each corner.

**Pedestrian Detection**  
The pedestrian detection task has been widely studied (Cai et al., 2016). Some full-body detectors try to lift the accu-

10093

### Output:

The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-20)

#### GigaHumanDet: Exploring Full-Body Detection on Gigapixel-Level Images

Chenglong Liu<sup>1,2\*</sup>, Haoran Wei<sup>3,4\*</sup>, Jinze Yang<sup>1</sup>, Jintao Liu<sup>1</sup>,  
Wenci Li<sup>1</sup>, Yuchen Guo<sup>1,5</sup>, Lu Fang<sup>1,6</sup>

<sup>1</sup>MOECC Technology  
<sup>2</sup>BNRist, Tsinghua University  
<sup>3</sup>BNRist, Tsinghua University  
<sup>4</sup>Shanghai Jiao Tong University  
{luchenglong20, weihoran18, yangjinze20, lijintao201}@mails.ac.cn  
wenci.li@sjtu.edu.cn, yuchen.guo@sjtu.edu.cn, fanglu@sjtu.edu.cn

**Abstract**  
Performing person detection in super-high-resolution images has been a challenging task. For such a task, modern detectors, which usually encode a box using center and width/height, struggle with accuracy due to two factors: 1) Human characteristics: people come in various postures and the center with high freedom is difficult to capture robust visual pattern; 2) Image characteristics: due to vast scale diversity of input (gigapixel-level), distance regression (for width and height) is hard to pinpoint, especially for a person, with substantial scale, who is near the camera. To address these challenges, we propose GigaHumanDet, an innovative solution aimed at further enhancing detection accuracy for gigapixel-level images. GigaHumanDet employs the corner matching method to avoid the potential issues of a high degree of freedom in corner predicting. To collect distinguish similar-looking persons and enforce instance consistency of corner pairs, an instance-guided learning approach is designed to capture discriminative individual semantics. Further, we devise reliable shape-aware bodies equipped with a multi-precision strategy as the human corner matching guidance to be appropriately adapted to the single-view large scene. Experimental results on PANDA and STCcrowd datasets show the superiority and strong adaptability of our design. Notably, our model achieves 82.4% in terms of AP<sub>50</sub>, outperforming current state-of-the-arts by more than 10%.

**Introduction**  
Person detection is a fundamental and critical task for human-centric visual analysis. Recently, the resolution has reached gigapixel level (e.g., 25k × 14k pixels) (Wang et al., 2020), posing a challenge for object detectors to cover the analysis at large-scale spatial range with clear local details. Most modern detectors fall into the center-regression-guided type (Cai and Vasconcelos 2018; Tian et al., 2019; Heun et al., 2021; Zhang et al., 2022), which locates objects via centers and prefers clear four boundaries to regress width/height (see Figure 1 (a)). But for the human-centric task in gigapixel images, it seems that the requirements of

the above methods cannot be perfectly met. Numerous individuals exhibit various postures, leading to the instability of human center visual pattern (e.g., Figure 1 (a)); the center of the bending man lies on the ground while the center of the standing woman is occluded, further making it harder to determine four boundaries. Besides, because people are at different distances from the gigapixels (Yuan et al., 2017), there is vast variation in human size (e.g., Figure 1 (c)) from

10092

The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-20)

**Figure 2: Framework of GigaHumanDet.** By discriminative semantics and robust body embeddings, our model performs gigapixel-level full-body detection by matching corner pairs.

125 × 336 to 602 × 1465). Thus, long-range regression errors directly affect the quality of boxes produced by modern detectors that regress width/height, and the box may even be mistakenly considered a negative sample since [w] with the ground truth is under threshold (e.g., yields a semi-finished box that only covers half of the body (Figure 1(b)).

Instead of the above detectors, another is the corner-semantic-guided type (Law and Deng 2018), which estimates two corners of the box via boundary clues. We find this type is more suitable for gigapixel-level human detection. As the best knowledge, the degree of freedom of determining a corner is lower than that of a center (a perfect corner needs four bodies while a center only needs two), and manually generating a human annotation box via two corners is a strong prior knowledge for us. In Figure 1 (d), although the ideal center sample is occluded, each corner point can still be located by two clear boundary clues. Further, the object representation of corner detectors do not rely on distance regression, so they are robust to multi-scale changes of the object and enjoy greater potential to generate higher-quality boxes in such large-scale single-view scenes. However, existing corner-semantic-guided methods designed for common scenarios can not work well in the extremely large-scale full-body detection task, especially their heuristic corner-matching algorithms. For example, CornerNet (Law and Deng 2018) determines corner pairs by local response similarity, which can cause severe confusion in detecting humans with similar appearances in the large spatial scale scene. Cornerlet (Dui et al., 2019) predicts one corner point to filter out the false positives (FP), but it still can not perform well when applied to a gigapixel image, because the central region of regression FP often covers the center point of a third person. Therefore, we may ask: Can we devise a robust corner matching algorithm for gigapixel-level full-body detection?

To answer the question and expand the applicability of corner-semantic-guided detectors for full-body detection on gigapixel-level images, we propose GigaHumanDet. GigaHumanDet predicts two developed corners of the bounding box, each of which is expressed explicitly and requires only two bodies to be determined, improving the robustness in gigapixel crowds with various postures. To acquire discriminative corner semantics for similar-looking persons, we

devise an instance-guided learning strategy as shown in Figure 2. For reliable matching corners, we devise shape-aware bodies which encodes the body shape embeddings at the corresponding corner location. To further purify the body shape embedding and make it more tolerant to the drastic scale variation, a Gaussian-inspired multi-precision regression strategy is devised to alleviate the difficulty and inaccuracy of once long-range regression. Note that the distance regression is decoupled from the corner prediction, so the corner pair can decide an accurate object box without being affected by the distance regression error. Equipped with robust bodies, GigaHumanDet can reap promising performance by pairing corners with consistent body shapes. Experimental results on gigapixel-level PANDA (Wang et al., 2020) benchmark show GigaHumanDet yields a new SOTA accuracy in terms of 82.4% on AP<sub>50</sub>, boosting 60.8% than CornerNet baseline and surpassing other methods by more than 10%. Further, the competitive results on STCcrowd verify the applicability of our method to general pedestrian detection. Our contributions can be summarized as follows:

- This paper unleashes the power of corner modeling approach on gigapixel-level full-body detection, and we demonstrate that it is more suitable and robust than center-regression-guided methods for this task.
- We design instance-guided learning and multi-precision strategy to acquire discriminative corner semantics. We propose shape-aware bodies to provide reliable corner-matching guidance for large-scale dense scenes.
- GigaHumanDet achieves the SOTA accuracy on the gigapixel-level detection task and outperforms other advanced methods by 10%.

**Related Work**

**Object Detection on High-Resolution Images**  
Object detection on large-scale HR images has become a challenging task. A gigapixel-level human-centric PANDA dataset (Wang et al., 2020) is published and its resolution has reached 25,000 × 14,000. Due to wide FoV and high resolution, pedestrians have various postures as well as occlusions, and drastic scale changes exist, which degrades the accuracy of modern detectors developed on COCO (Lin et al., 2014).

**Center-Guided and Corner-Guided Methods**  
The center-guided detectors utilize centers and width/height to encode boxes. Most of them (He et al., 2017; Cai and Vasconcelos 2018; Ce et al., 2021) take the center as a reference point and regress object size. Different from the above methods, the corner-guided detector (Law and Deng 2018; Dui et al., 2019) is proposed to estimate corners and match them to compose the final box. But for gigapixel images, humans with similar appearances cause great confusion when matching corners. Our GigaHumanDet employs the corner modeling method and tackles the matching problem by setting robust body-shape embedding for each corner.

**Pedestrian Detection**  
The pedestrian detection task has been widely studied (Cai et al., 2016). Some full-body detectors try to lift the accu-

10093

Figure 7: Dynamic resolution of GOT for high-resolution images. In the dual-page paper reading mode shown in the figure (data is from Liu et al. (2024b)), the input resolution of the original GOT is not sufficient to handle it. Therefore, we adapt dynamic resolution technology to make the model no longer limited to the size of the image.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

Prompt: OCR with format across multi pages:



Output:



Figure 8: Multi-page (document) OCR ability of GOT. With this feature, researchers can continue to train the GOT with multi-page PDF-text pairs, such as Arxiv paper with .tex file.

Prompt: OCR with format:

Output:

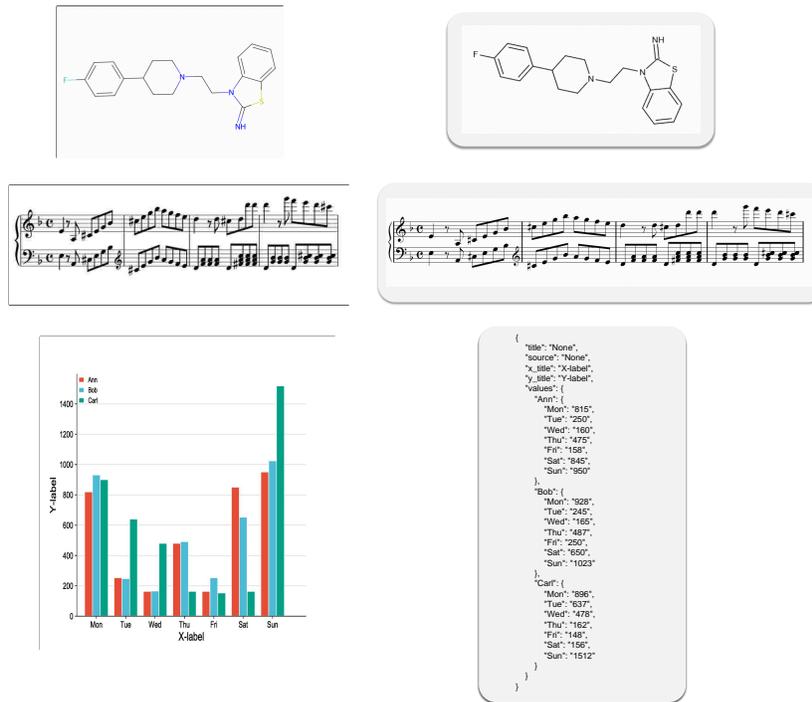


Figure 9: More general OCR results. GOT can process molecular formulas, sheet music, and charts.

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

**Prompt: OCR with format:**

**AEON NEWS RELEASE** 

2019年3月22日  
株式会社ジーフット

**ブレストージュシューズ ショッピングショップオープン!!**

**トレーディングポスト ホワイトレーベル Trading Post WHITE LABEL 二子玉川店オープンのお知らせ**

株式会社ジーフット（東京都中央区新川、代表取締役社長：堀江泰文）の子会社となる株式会社ブレストージュシューズ（東京都中央区新川、代表取締役社長：武井秀善）は、Trading Post のセカンドライン Trading Post WHITE LABEL のショッピングショップ「Trading Post WHITE LABEL 二子玉川店」を麻布テラーラースクエア二子玉川店内に2019年3月29日（金）にオープンいたします。



**Trading Post WHITE LABEL 二子玉川店**

これまでパーソナルオーダーのスーツやシャツなどで人気のビジネスウェアのセレクトショップ 麻布テラーにて期間限定ポップアップイベントを行ってまいりましたが、この度トレーディングポスト初となるショッピングショップを麻布テラーラースクエア二子玉川店にオープンすることとなりました。

こだわりを持ち、本物を求めるお客さまへTrading Post オリジナルアイテムを中心に上質で本物のこだわりアイテムを国内外からセレクト。麻布テラー、Trading Post WHITE LABEL の多彩な商品展開やフィッティングなどの接客サービスを通じ、お客さまへ、より満足いただけるトータルコーディネートをご提案致します。

（主な取り扱いブランド）  
Trading Post, Soffice&Solid, CARMINA, Tricker's, Allen Edmonds, etc

**Output:**

**NEWS RELEASE**

2019年3月22日  
株式会社ジーフット

**ブレストージュシューズ ショッピングショップオープン!!**

トレーディングポスト ホワイトレーベル  
Trading Post WHITE LABEL 二子玉川店オープンのお知らせ

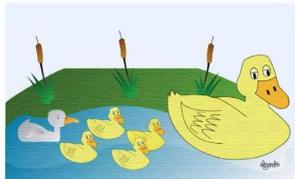
株式会社ジーフット（東京都中央区新川、代表取締役社長：堀江泰文）の子会社となる株式会社ブレストージュシューズ（東京都中央区新川、代表取締役社長：武井秀善）は、Trading Post のセカンドライン Trading Post WHITE LABEL のショッピングショップ「Trading Post WHITE LABEL 二子玉川店」を麻布テラーラースクエア二子玉川店内に2019年3月29日（金）にオープンいたします。

**Trading Post WHITE LABEL 二子玉川店**

これまでパーソナルオーダーのスーツやシャツなどで人気のビジネスウェアのセレクトショップ 麻布テラーにて期間限定ポップアップイベントを行ってまいりましたが、この度トレーディングポスト初となるショッピングショップを麻布テラーラースクエア二子玉川店にオープンすることとなりました。

こだわりを持ち、本物を求めるお客さまへTrading Post オリジナルアイテムを中心に上質で本物のこだわりアイテムを国内外からセレクト。麻布テラー、Trading Post WHITE LABEL の多彩な商品展開やフィッティングなどの接客サービスを通じ、お客さまへ、より満足いただけるトータルコーディネートをご提案致します。

（主な取り扱いブランド）  
Trading Post, Soffice&Solid, CARMINA, Tricker's, Allen Edmonds, etc



—No es un pavo, por cierto —dijo la pata—.  
Fijense en la elegancia con que nada, y en lo derecho que se mantiene. Sin duda que es uno de mis pequeñitos. Y si uno lo mira bien, se da cuenta pronto de que es realmente muy guapo. ¡Cuac, cuac! Vamos, vengan conmigo y déjenme enseñarles el mundo y presentarlos al corral entero. Pero no se separen mucho de mí, no sea que los pisoteen. Y anden con los ojos muy abiertos, por si viene el gato.

Y con esto se encaminaron al corral. Había allí un escándalo espantoso, pues dos familias se estaban

- No es un pavo, por cierto - dijo la pata-.
- Fijense en la elegancia con que nada, y en lo derecho que se mantiene. Sin duda que es uno de mis pequeñitos. Y si uno lo mira bien, se da cuenta pronto de que es realmente muy guapo. ¡Cuac, cuac! Vamos, vengan conmigo y déjenme enseñarles el mundo y presentarlos al corral entero. Pero no se separen mucho de mí, no sea que los pisoteen. Y anden con los ojos muy abiertos, por si viene el gato.
- Y con esto se encaminaron al corral. Había allí un escándalo espantoso, pues dos familias se estaban

Figure 10: We do not specifically introduce additional OCR capabilities for GOT other than Chinese and English. Yet the PDF data we crawled may contain a small amount of text in other languages, leading to the GOT seeming to have the ability to recognize other languages. However, we cannot guarantee the OCR quality of other languages. Therefore, we recommend fine-tuning the model with corresponding data if this feature is needed.