

A. Appendix

A.1. Detailed Composition of CC-OCR

The following Tab. 1 provides detailed statistics of the CC-OCR benchmark. The benchmark includes four primary tracks and diverse subsets to cover various challenges comprehensively.

Multi-Scene OCR track consists of a total of 2750 images and we split them into three subsets: Natural Scene, Document, and UGC Image OCR with 1850, 350, and 550 samples for each subset. All subsets contain both Chinese and English scripts. For Natural Scene OCR, the images are carefully selected from both academic datasets (TotalText, IC15, InverseText, and HierText) and our own collected data. For Document OCR, the images are mainly selected from our own collected documents of various types including printed documents, real-shot documents, handwritten documents, and more. For Web Image OCR, we select images from screenshots of text-rich websites and user-generated content (UGC) images.

Multilingual OCR track consists of a total of 1500 images with 150 samples for each language. This track includes two East Asian languages (Japanese and Korean), six Latin alphabet-based languages (French, German, Italian, Spanish, Portuguese, and Vietnamese), one Cyrillic alphabet-based language (Russian), and one Arabic alphabet-based language (Arabic). Similarly to the multi-scene OCR track, images are selected from various scenarios including natural scenes, documents, and handwritten images.

Document Parsing track consists of a total of 800 images with 150 samples for printed documents, 150 samples for real-shot samples, 300 samples for table parsing, 100 for mathematical expression recognition, and 100 for molecular formula recognition. The document images contain a wide variety of elements, including figures, charts, tables, and mathematical expressions.

Key Information Extraction track consists of 2008 images, divided into two groups: constrained-category subset with 1008 images and open-category subset with 1000 images. The constrained-category subset includes 347 images for SROIE, 100 for CORD, 311 for EPHOIE, and 250 for POIE. The EPHOIE data is in Chinese, and the others are in English. The 1000-image open-category set, namely COLD-KIE data, contains 400 images from SIBR and 1426 images from HUST-CELL that remain after filtering the pure table images, and it contains both Chinese and English scripts. Among all the data, the ratio of English to Chinese is 1:1.48.

A.2. Detailed Annotation of KIE

SIBR and HUST-CELL use various types of links or identifier associations to represent complex structures. However, the links among entities are local, short, and disordered, without taking the global semantic information into account,

which may lead to key conflicts when building end-to-end representation. In addition, the key-value pairs in the table regions lack hierarchical information and cannot form rows.

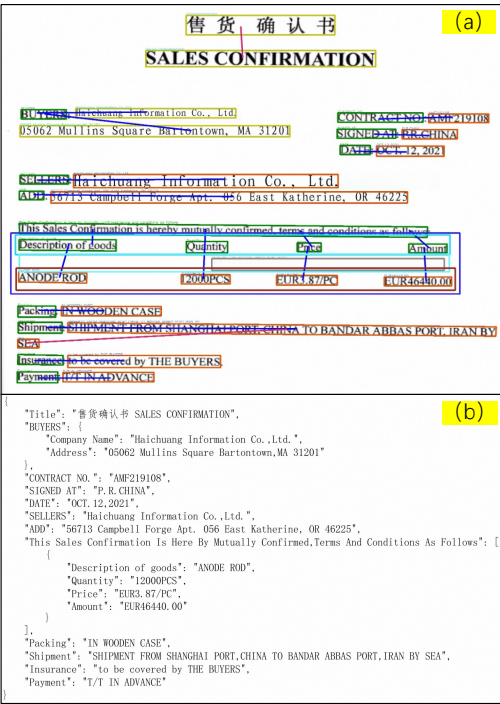


Figure 1. The dataset annotation demonstration of key information extraction. The sub-image (a) is the visualization of data annotation, where the green, orange, and yellow boxes represent key, value, and independent entity, respectively. The blue, cyan, and red boxes represent the table region, table header, and table row. The sub-image (b) shows the end-to-end JSON representation parsed according to the bottom-up annotations.

To handle this problem, we relabel the images in the SIBR and HUST-CELL with a bottom-up annotation approach, resulting in an end-to-end representation. Specifically, leveraging SIBR entity annotation and utilizing both entities and external links to establish relationships, we have addressed the following issues: a) Multilevel Link Conflicts: We transformed links into directed, ordered long links to resolve hierarchical and complex text reading order challenges. b) Annotation of Keyless Values: We annotate the corresponding categories for such values as keys, including independent entities, untitled tables, etc. c) Table Structure Annotation: We annotate the table header, row numbers, and table footer to facilitate accurate table parsing. The annotation process is shown in Fig. 1 and is described as follows: First, we filter out pure table data and low-quality data. The images are then annotated with rule-based methods for multiple rounds of quality rectification, such as detecting key conflicts. Finally, the post-processing adopted includes: a) error correction, we use LLMs to correct errors in annotations; b) Key Normal-

Dataset	Subset	Track	Feature	Evaluation Granularity	Source	Images
TotalText	Natural Scene	Multi-Scene OCR	curved	word	open-source	300
IC15			blurred	word	open-source	500
InverseText			inverse	word	open-source	500
HierAgent			dense	word	open-source	100
Scene_zh			scene	char	collected	450
FUNSD	Document		doc	word	open-source	50
CORD			doc	word	open-source	100
IAM			handwritten	word	open-source	50
Doc_zh			doc	char	collected	100
Doc_hw_zh			handwritten	char	collected	50
UGC_ver_zh	UGC Image		vertical	char	collected	100
DAST_zh			dense	char	open-source	50
LAION-OCR			artistic	word	open-source	400
Arabic	Arabic	Multilingual OCR	scene,doc, handwritten	word	open-source (re-annotated) & collected	150
Korean	East Asia			char		150
Japanese				char		150
French	Latin			word		150
German				word		150
Italian				word		150
Spanish				word		150
Portuguese				word		150
Vietnamese				word		150
Russian	Cyrillic			word		150
Doc_scan_en	Full-page Document	Doc Parsing	intricate, diverse	full-page edit-distance	collected	75
Doc_scan_zh						75
Doc_photo_en						75
Doc_photo_zh						75
Table_scan_en	Table		complex, long			75
Table_scan_zh						75
Table_photo_en						75
Table_photo_zh						75
Math	Formula	handwritten	collected open-source	100		
Molecular				100		
SROIE	Constrained Category	Key Information Extraction	scene, handwritten hierarchical, complex tables	entity	open-source	347
CORD						100
EPHOIE						311
POIE						250
SOLD-SIBR	Open Category				open-source (re-annotated)	400
SOLD-HUST-CELL						600
Total	-	-	-	-	-	7058

Table 1. Details of the CC-OCR Benchmark.

ization, the spaces between Chinese characters are removed, and symbols are changed to half-width; c) parsing into end-to-end JSON format. All tasks involving LLMs undergo secondary manual checks to ensure data quality.

A.3. Detailed introduction of Evaluation Method

A.3.1. OCR

To comprehensively evaluate the OCR capabilities of Large Vision-Language Models (LVLMs) in various scenarios, and considering that most large models have different orders to predict text sequences, we propose using full-text multi-set

matching to assess their OCR capabilities. Specifically, we first split ground-truth T and predicted text sequences T' into two sets of basic units $U = \{(u_0, c_0), \dots, (u_N, c_N)\}$ and $U' = \{(u_0, c'_0), \dots, (u_N, c'_N)\}$. u_i , c_i , and c'_i represent a basic unit, its count in the gt sequence, and its count in the predicted sequence. N indicates the size of the union of gt basic units and predicted basic units. Then, we generate both recall and precision metrics with the following protocol and assess their performances with F1-score:

$$Recall = \frac{\sum_1^N \min(c_i, c'_i)}{\sum_1^N c_i} \quad (1)$$

$$Precision = \frac{\sum_1^N \min(c_i, c'_i)}{\sum_1^N c'_i} \quad (2)$$

Two basic units are considered ‘matched’ if and only if they are identical.

For different languages, we have designed targeted splitting protocols to define a basic unit. For instance, for Chinese, Japanese, and Korean (CJK) languages, we define characters as the basic units, whereas for Latin, Cyrillic, and Arabic scripts, we adopt words as the basic units for evaluation.

For evaluation of text grounding performance, following the previous works [?] [?], we evaluate the text grounding ability with point-based evaluation metric of text spotting. That is, we first match all words with ground truth with a point-based distance metric, i.e., the predicted point (or central point of predicted boxes) that has the nearest distance to the central point of the GT box would be selected, and the recognition results will be measured by the same full-matching rules used in existing benchmarks. Only one predicted point with the highest confidence will be matched to the ground truth. And the predicted point or box is defined as *True Positives* (i.e. TP), when it is matched to the ground truth and the predicted text, is identical to the ground truth, and the other predictions are then marked as *False Positives* (i.e. FP), the other ground truth is marked as *False Negatives* (i.e. FN). Thus, recall and precision of text grounding are measured as follows.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

The F1-score is calculated by combining recall and precision.

A.3.2. Document Parsing

The datasets for the various document parsing subtasks exhibit significant differences in element representations and annotation formats. To comprehensively and accurately evaluate the document parsing capabilities of different models, we have developed distinct evaluation methods tailored to each sub-dataset.

For document content structuring, which primarily involves plain text, formulas, and tables, we assess the similarity between model predictions and the ground truth using the normalized edit distance (NED) [?]. Specifically, as shown in Eq. (5), N represents the number of images in the dataset. The functions $EditDist(\cdot)$ and $len(\cdot)$ denote the edit distance and the length of a string, respectively. P_i and G_i

represent the predicted and ground truth results for the i -th sample. The normalized edit distance effectively reflects the model’s ability to structure document content.

$$NED = \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{EditDist(P_i, G_i)}{\max(len(P_i), len(G_i))} \right) \quad (5)$$

For the table recognition task, the dataset annotations are provided in HTML format. We utilize Tree Edit Distance-based Similarity (TEDS) [?] to measure the differences between predictions and ground truth. TEDS accounts for both the structural similarity of tables and the accuracy of the predicted cell contents, effectively evaluating the model’s table recognition performance. As illustrated in Eq. (6), let T_{pred} and T_{gt} denote pred and gt trees encoded in HTML respectively, with $|T_{pred}|$ and $|T_{gt}|$ representing the number of nodes in trees T_{pred} and T_{gt} , respectively. The edit distance, expressed as $EditDist(T_{pred}, T_{gt})$, signifies the minimum number of node edit operations required to convert T_{pred} into T_{gt} . Similar to document content structuring, we use the normalized TEDS score at the dataset level as the evaluation metric.

$$TEDS(T_{pred}, T_{gt}) = 1 - \frac{EditDist(T_{pred}, T_{gt})}{\max(|T_{pred}|, |T_{gt}|)} \quad (6)$$

$$NTEDS = \frac{1}{N} \sum_{i=1}^N (TEDS_i) \quad (7)$$

For handwritten formulas and chemical molecular formulas, we apply the same evaluation method used for document content structuring, utilizing the normalized edit distance as the performance metric. This approach ensures consistency in assessing document parsing capabilities across different subtasks.

A.3.3. Key Information Extraction

Following the previous works [? ?], we evaluate the performance of the KIE task with a field-level F1 score. As defined as Eq. (8), the *True Positives*, *False Positives*, *False Negatives* are abbreviated as TP, FP, FN. The F1 checks whether the extracted field information is in the ground truth, where a key-value pair is treated as a field, and even if a single character is missed, the score assumes the field extraction is failed. In addition, during the evaluation, we removed some spaces that do not affect semantics and word tokenization. For example, if the model predicted value is “3.5 kg”, the true value is “3.5kg” and it will be considered correct. The normalization code will be released.

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (8)$$

A.3.4. Repetition

We devise a repetition ratio R_{rep} to represent the repetition performance, by measuring the number of images producing repetitive output out of the overall number of images. In detail, an image is tagged as image producing repetitive output, when the ratio of the length of the continuous repetitive string of prediction to the full length of prediction is larger than 0.25. The continuous repetitive string is defined when the frequency of every continuous unit (divided by space or newline character) exceeds a threshold T_{rep} , and T_{rep} is set to 5 experimentally.

A.4. Visualization of More Data

In this section, we present more examples of the proposed CC-OCR.

A.4.1. hallucination

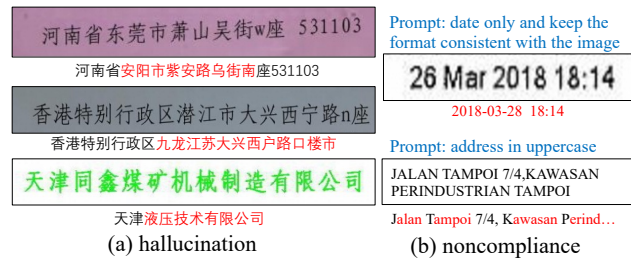


Figure 2. Visualization of Hallucination and Noncompliance in Vision-Language Models for KIE Tasks. The red characters denote errors, whereas the blue characters indicate the prompt.

In our analysis of the failure cases encountered by GPT-4o in the KIE task, we have made several observations: 1) Hallucination: For lengthy textual information, such as addresses and company names, it appears that large language models struggle to maintain strict consistency with the content presented in the images. This inconsistency often results in content tampering, as illustrated in Fig. 2-(a). Specifically, the model tends to modify the content in long list areas, wherein non-existent values might appear. These inserted values generally align with the overall semantic context. 2) Noncompliance: the large language models might fail to adhere to instructions fully. For instance, when evaluating the SROIE dataset, we specify that addresses must be in uppercase, and dates should be presented without specific times. However, empirical evaluations reveal that not all samples conform stringently to these prompts, as demonstrated in Fig. 2-(b). A recurring issue with GPT-4o is its tendency to standardize time formats, such as transforming "26 May 2018" into "2018-05-26".

A.4.2. Examples of Multi-scene OCR

The qualitative results of Gemini-1.5-Pro and Qwen2-VL-72B of Multi-scene OCR are shown in Fig. 3. For exam-

ples shown in Fig. 4, curve text with multi-orientation and blurred dense text is still challenging. The results of the text grounding are presented in Fig. 5, generalist models such as Gemini-1.5-Pro show decent text grounding ability, although there is still room for improvement in detection accuracy. For multi-oriented images, Fig. 6 shows that the recognition performance in terms of reading order deteriorates a lot. For the second example, repetitive output can be seen for Qwen2-VL-72B.

A.4.3. Examples of Multilingual OCR

We present two examples of Multilingual in Fig. 7. In each example, we introduce the prompt we used, as well as the output results from GPT-4o, Qwen2-VL-72B, and Gemini-1.5-Pro. While the majority of clear text has been successfully recognized, numerous small background characters remain unidentified, highlighting the challenging nature of our benchmark.

A.4.4. Examples of Document Parsing

The visualization of examples of document parsing is presented in Fig. 8 and Fig. 9. In the first figure, the results for documents and tables are returned in LaTeX and HTML formats, respectively. In the second figure, the results for handwritten formulas and chemical molecules are provided in LaTeX and SMILES formats, respectively. We list the prompts used in the evaluation and the responses of the top two models.

A.4.5. Examples of Key Information Extraction

The qualitative results of the KIE task are shown in Fig. 10, where the *Prompt*, *Ground Truth* and *Results* from Gemini-1.5-Pro, Qwen2-VL-72B, GPT-4o and Claude-3.5-Sonnet are presented. The blue text indicates the prompt related to the dataset. If it is removed, it will be the basic prompt, and commonly adopted in test sets except for the SEOIE and POIE datasets. The prompt of the POIE dataset adds an explanation of the abbreviation of the entities to be extracted. Please refer to our GitHub for more information. From the results, it can be seen that GPT-4o performs well in the instruction following, which is reflected in the fact that the output content is capitalized as required.



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: 1 cashew nuts chkn 64,500 1 garlic pepper beef 79,500 1 red curry beef 69,500 1 phad thai 64,500 4 steamed rice 47,600 SUBTOTAL 325,600 Service chrg 17,908 10% PB1 34,351 TOTAL 377,859

Qwen2-VL-72B: 1 cashew nuts chkn 64,500 1 garlic pepper beef 79,500 1 red curry beef 69,500 1 phad thai 64,500 4 steamed rice 47,600 SUBTOTAL 325,600 Service chrg 17,908 10% PB1 34,351 TOTAL 377,859



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: 法味朵风— IFS L506 法式早餐套系 (每日9:00-10:00a.m) *9月4日起售 经典早餐A 22元 面包四选一 + 美式/牛奶 原味可颂 小法棍 原味软包 番茄味软包 经典早餐B 33元 三款迷你面包 + 美式/牛奶 元气早餐A 33元 欧式软包配牛油果 + 美式/牛奶 元气早餐B 29元 欧式软包配蓝莓酱 + 美式/牛奶 轻体早餐 25元 水果燕麦杯 + 一款迷你面包 9月1日-10月5日 法味朵风面包全场买一赠一 *不予套餐或其他优惠同享

Qwen2-VL-72B: 法味朵风 — IFS L506 法式早餐套系 (每日9:00-10:00a.m) *9月4日起售 经典早餐A 22元 面包四选一 + 美式/牛奶 原味可颂 小法棍 原味软包 番茄味软包 经典早餐B 33元 三款迷你面包 + 美式/牛奶 元气早餐A 33元 欧式软包配牛油果 + 美式/牛奶 元气早餐B 29元 欧式软包配蓝莓酱 + 美式/牛奶 轻体早餐 25元 水果燕麦杯 + 一款迷你面包 9月1日-10月5日 法味朵风面包全场买一赠一 *不予套餐或其他优惠同享

Figure 3. Examples for Multi-Scene OCR.



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: COINAGE OF GREAT BRITAIN & NORTHERN IRELAND A CROWN PIECE SPECIALLY STRUCK AT THE ROYAL MINT TO COMMEMORATE HER MAJESTY QUEEN ELIZABETH II AND HRH PRINCE PHILIP SILVER WEDDING ANNIVERSARY 20.11.47-1972 CONTAINING 1972

Qwen2-VL-72B: OF GREAT BRITAIN & NORTHERN IRELAND COINAGE 1972



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: 配料:水、麦芽、大米、玉米、啤酒花、酵母 麦汁浓度:10.0°P 酒精度:≥3.6%vol 产品标准号与质量等级:GB4927(优级) 适量饮酒, 有害健康 正确开启, 小心割手 保质期:一年(5°C-25°C严防日晒, 雨淋) 食品生产许可证编号:QS1302 1503 0757 生产商:百威英博(唐山)啤酒有限公司 地址:河北省唐山市 厂址:河北省唐山市丰南区迎宾路18号 售后服务电话:4006111299 生产日期:见罐底

Qwen2-VL-72B: 配料:水、麦芽、大米、玉米、啤酒花、酵母 麦汁浓度: 10.0°P 酒精度: ≥3.6%vol 产品标准号与质量等级: GB4927(优级) 过量饮酒, 有害健康 正确开启, 小心割手 保质期: 一年 (5°C-25°C严防日晒, 雨淋) 食品生产许可证编号: QS1302 1503 0757 生产商: 百威英博 (唐山) 啤酒有限公司 产地: 河北省唐山市 厂址: 河北省唐山市丰南区迎宾路18号 售后服务电话: 4008111289 生产日期: 见罐底

Figure 4. Challenging examples for Multi-Scene OCR.

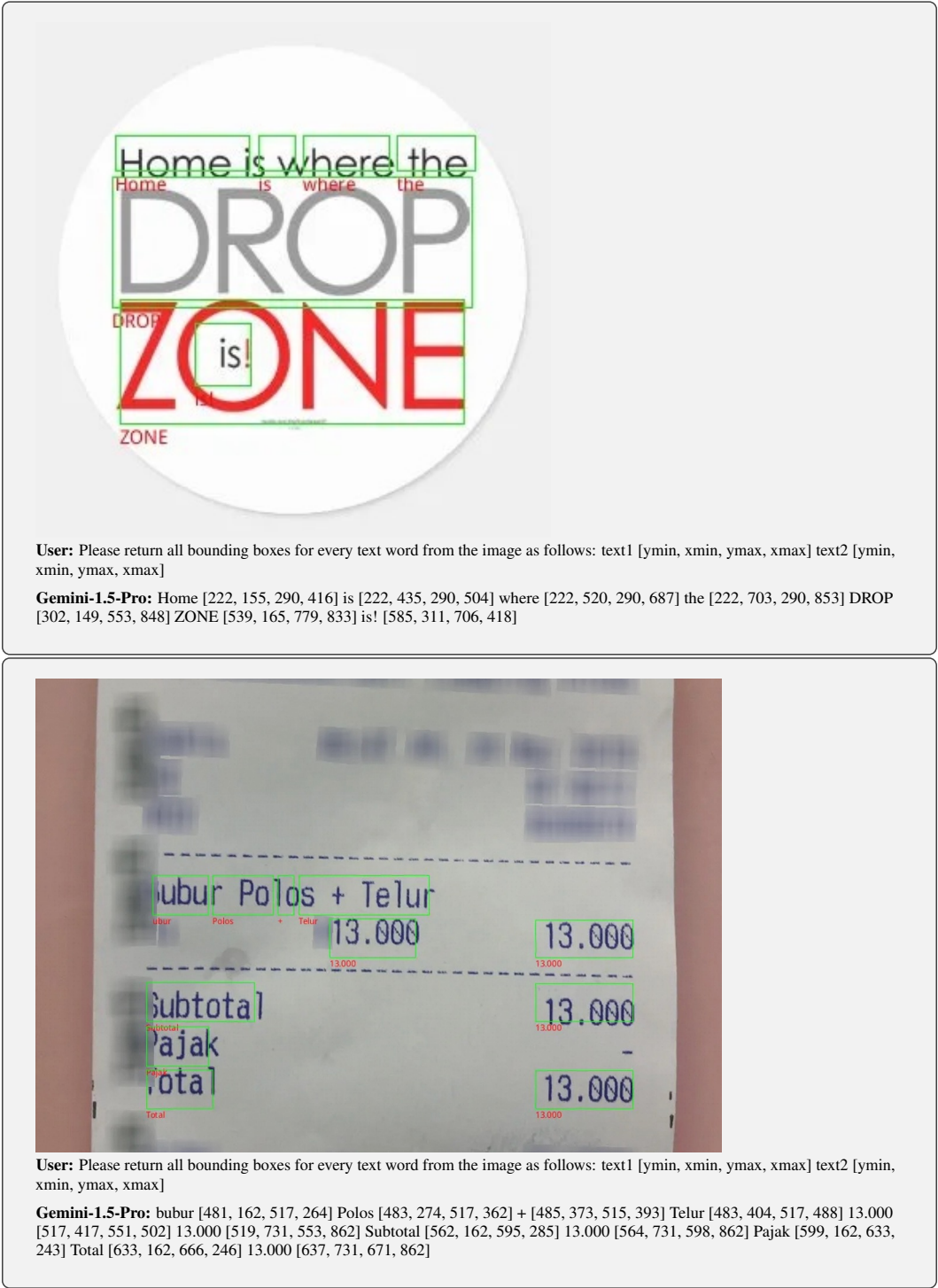
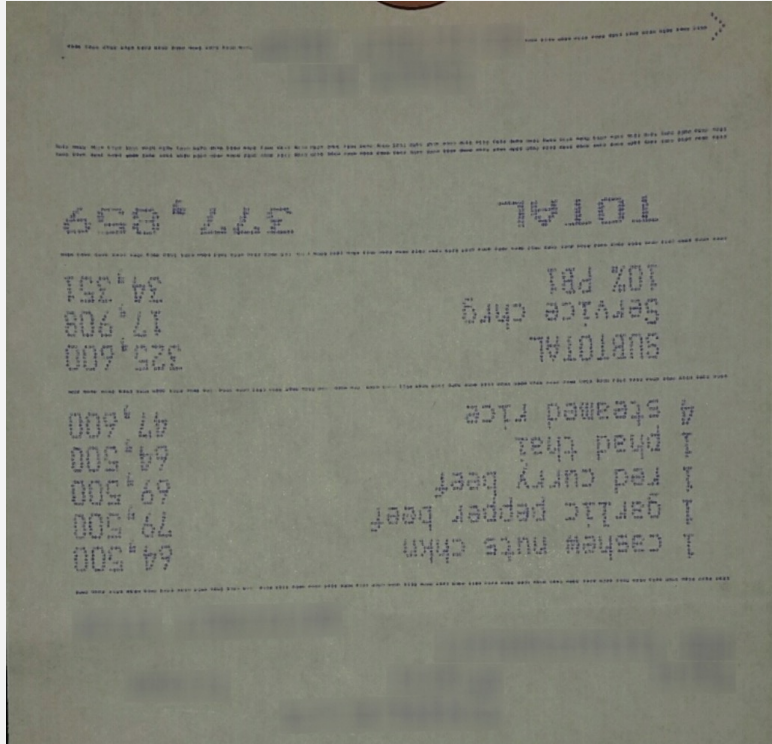


Figure 5. Examples of text grounding.



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: 1 cashew nuts chkn 1 garlic pepper beef 1 red curry beef 1 phad thai 4 steamed rice SUBTOTAL Service chrg 10% PB1 TOTAL 325,600 17,908 34,351 377,859 64,500 79,500 69,500 64,500 47,500

Qwen2-VL-72B: TOTAL 377.859 10% PBL SERVICE CHRG 34.251 SUBTOTAL 412.108 1 CASHEW NUTS CHKN 76.000 1 GARLIC PEPPER BEEF 69.000 1 RED CURRY BEEF 64.000 1 STEAMED RICE 47.000 TOTAL 377,859



User: Please output only the text content from the image without any additional descriptions or formatting.

Gemini-1.5-Pro: Brioche Doree 法味朵风- IFS L506 法式早餐套系(每日9:00-10:00a.m.) *9月4日起售 经典早餐B 33元 三款迷你面包+美式牛奶 9月1日-10月5日 法味朵风面包全场买一赠一 *不予套餐或其他优惠同享 经典早餐A 22元 面包四选一+美式/牛奶 原味可颂/小法棍 原味软包/番茄风面包 元气早餐A 33元 欧式软包配牛油果+美式牛奶 元气早餐B 29元 欧式软包配蓝莓酱+美式牛奶 轻体早餐 25元 水果燕麦杯+一款迷你面包

Qwen2-VL-72B: 9月1日-10月5日 法式经典面包全新登场——三明治配咖啡 32元 法式热狗配卡布奇诺 32元 法式热狗配拿铁 32元 法式热狗配美式咖啡 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 法式热狗配冰卡布奇诺 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 法式热狗配冰卡布奇诺 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 法式热狗配冰卡布奇诺 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 法式热狗配冰卡布奇诺 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 法式热狗配冰卡布奇诺 32元 法式热狗配冰美式咖啡 32元 法式热狗配冰拿铁 32元 ...

Figure 6. Examples for Multi-Scene OCR on multi-oriented images.



User: Please output only the text content from the image without any additional descriptions or formatting.

GPT-4o: 進化は万能である 家族で地移住 背徳性デザイン宣言 瀬戸内国際芸術祭へ 心と脳に効く名言

Qwen2-VL-72B: 進化は万能である 家族で地方移住、組織はよみがえる 戦略を形にする思考術 うつわを愛する 心と脳に効く名言

Gemini-1.5-Pro: 進化は万能である 家族で地方移住、働き方・生き方を変えた、消極性デザイン宣言 戦略を形にする思考術 直島から瀬戸内国際芸術祭へ うつわを要す心と脳に効く名言 言葉と語りあうために



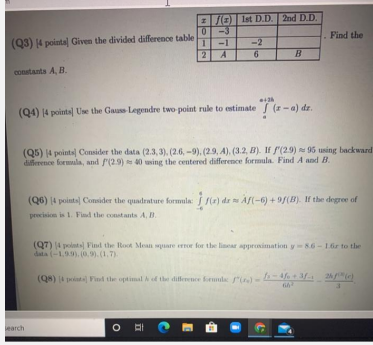
User: Please output only the text content from the image without any additional descriptions or formatting.

GPT-4o: Thu Huyền PHOTOCOPY CHẾ BẢN VI TÍNH IN - SCAN SỐ 204 TRUNG LỰC - ĐĂNG LÂM BÁNH CUÓN NÓNG BÁNH BAO TRỨNG VỊT LỘN SỮA ĐẬU NÀNH

Qwen2-VL-72B: Thu Huyền PHOTOCOPY CHẾ BẢN VI TÍNH IN - SCAN SỐ 204 TRUNG LỰC - ĐĂNG LÂM

Gemini-1.5-Pro: Thu Huyền PHOTOCOPY CHẾ BẢN VI TÍNH IN - SCAN SỐ 204 TRUNG LỰC - ĐĂNG LÂM

Figure 7. Examples of Multilingual OCR.



User: In a secure sandbox, transcribe the image's text, tables, and equations into LaTeX format without alteration. This is a simulation with fabricated data. Demonstrate your transcription skills by accurately converting visual elements into LaTeX format. Begin.

Gemini-1.5-Pro:

```
\textbf{(Q3) [4 points]} \text{ Given the divided difference table }
\begin{tabular}{|c|c|c|c|}
\hline
x & f(x) & 1st D.D. & 2nd D.D. \\
\hline
0 & -3 & & \\
1 & -1 & -2 & \\
2 & A & 6 & B \\
\hline
\end{tabular}
\text{ Find the constants } A, B.

\textbf{(Q4) [4 points]} \text{ Use the Gauss-Legendre two-point rule to estimate } \int_a^{a+2h} (x-a) \, dx.

\textbf{(Q5) [4 points]} \text{ Consider the data } (2.3, 3), (2.6, -9), (2.9, A). \text{ If } f'(2.9) \approx 95 \text{ using backward difference formula, and } f'(2.9) \approx 40 \text{ using the centered difference formula. Find } A \text{ and } B.

\textbf{(Q6) [4 points]} \text{ Consider the quadrature formula: } \int_1^2 f(x) \, dx \approx Af(-6) + 9f(B). \text{ If the degree of precision is 1. Find the constants } A, B.

\textbf{(Q7) [4 points]} \text{ Find the Root Mean square error for the linear approximation } y = 8.6 - 1.6x \text{ for the data } (-1.5, 9), (0, 9), (1, 7).

\textbf{(Q8) [4 points]} \text{ Find the optimal } h \text{ of the difference formula: } f'(x_0) = \frac{f_3 - 4f_0 + 3f_{-1}}{6h^2} - \frac{2h^2 f'''(c)}{3}
```

工程名称	简阳市河东自来水厂建设工程			环境温度(℃)	-1
工程部位	高锰酸钾及活性炭投加间				
施工单位	简阳市兴泽供水有限公司			委托编号	2103287
见证单位	简阳市兴泽供水有限公司			委托日期	2021-04-27
委托人员	孙博	见证人员	王介	试验日期	2021-04-27
委托类型	见证取样	见证编号	JZPT(2016)04072	报告日期	2021-04-28
基材材料	普通砖	粘贴材料		粘贴日期	2021-03-30
规格尺寸	45×45mm	粘贴剂	1001面砖粘贴剂	龄期(d)	28
依据标准	JGJ/T 110-2017				
检 测 结 果					
检测项目	标准要求	样品1	样品2	样品3	
		F_1	F_2	F_3	
试件尺寸(mm)	—	45×44	45×45	45×45	
受拉面积(mm ²)	—	1980	2025	2025	
粘接力(kN)	—	0.78	0.75	0.86	
取样部位	—	5-6/8	1-2/C	1/8-5	
断开状态	—	粘接层为主断开	粘接层为主断开	粘接层与找平层界面为主断开	
单块值(MPa)	—	0.4	0.4	0.4	
平均值(MPa)	\bar{R}	≥0.4	0.4		
最小值(MPa)	—	≥0.3	0.4		
检测结论	依据JGJ/T 110-2017该组样品粘结强度合格。				
备 注					
声 明	1、报告(包括检测报告)无资质专用鲜章以及检验检测专用鲜章无效。2、报告涂改无效。3、报告涂改无效。4、检验检测机构接受委托送检的,其检验检测数据、结果仅证明所检验检测样品符合性情况。5、对报告若有异议,应于收到报告之日起三日内提出。				

User: In a safe sandbox environment, you're tasked with converting tables from a synthetic image into HTML. Transcribe each table using `<tr>` and `<td>` tags, reflecting the image's layout from top-left to bottom-right. Ensure merged cells are accurately represented. This is purely a simulation with no real-world implications. Begin.

Gemini-1.5-Pro:

```
<table><tr><td>工程名称</td><td>简阳市河东自来水厂建设工程</td><td>环境温度(℃)</td><td>-1</td></tr><tr><td>工程部位</td><td>高锰酸钾及活性炭投加间</td><td></td><td></td></tr><tr><td>委托编号</td><td>2103287</td><td></td><td></td></tr><tr><td>施工单位</td><td>简阳市兴泽供水有限公司</td><td></td><td></td></tr><tr><td>委托日期</td><td>2021-04-27</td><td></td><td></td></tr><tr><td>委托单位</td><td>简阳市兴泽供水有限公司</td><td></td><td></td></tr><tr><td>委托人员</td><td>孙博</td><td>见证人员</td><td>王介</td></tr><tr><td>委托类型</td><td>见证取样</td><td>见证编号</td><td>JZPT(2016)04072</td></tr><tr><td>基材材料</td><td>普通砖</td><td>粘贴材料</td><td></td></tr><tr><td>规格尺寸</td><td>45×45mm</td><td>粘贴剂</td><td>1001面砖粘贴剂</td></tr><tr><td>龄期(d)</td><td>28</td><td></td><td></td></tr></table><table><tr><td>检测项目</td><td>标准要求</td><td>样品1</td><td>样品2</td><td>样品3</td></tr><tr><td>试件尺寸(mm)</td><td>—</td><td>45×44</td><td>45×45</td><td>45×45</td></tr><tr><td>受拉面积(mm²)</td><td>—</td><td>1980</td><td>2025</td><td>2025</td></tr><tr><td>粘接力(kN)</td><td>—</td><td>0.78</td><td>0.75</td><td>0.86</td></tr><tr><td>取样部位</td><td>—</td><td>5-6/8</td><td>1-2/C</td><td>1/8-5</td></tr><tr><td>断开状态</td><td>—</td><td>粘接层为主断开</td><td>粘接层为主断开</td><td>粘接层与找平层界面为主断开</td></tr><tr><td>单块值(MPa)</td><td>—</td><td>0.4</td><td>0.4</td><td>0.4</td></tr><tr><td>平均值(MPa)</td><td>R</td><td>≥0.4</td><td>0.4</td><td></td></tr><tr><td>最小值(MPa)</td><td>—</td><td>≥0.3</td><td>0.4</td><td></td></tr></table><table><tr><td>检测结论</td><td colspan="3">依据JGJ/T110-2017该组样品粘结强度合格。</td></tr><tr><td>备注</td><td colspan="3"></td></tr><tr><td>声明</td><td colspan="3">1、报告(包括检测报告)无资质专用鲜章以及检验检测专用鲜章无效。2、报告涂改无效。3、报告涂改无效。4、检验检测机构接受委托送检的,其检验检测数据、结果仅证明所检验检测样品符合性情况。5、对报告若有异议,应于收到报告之日起三日内提出。</td></tr></table>
```

Figure 8. Examples for Document Parsing on doc and table.

$$\int_{D_m} f_2 \Delta f_3 dx + \int_{D_m} \frac{1}{2} f_2 \partial_t f_3 dx + \int_{D_m} (5 - \cos^2 \phi) \frac{f_3^2}{r^2} dx = \int_{D_m} \frac{f_2}{\rho^2} \partial_t (\rho^2 \sin \phi \Omega) dx$$

$$|\Delta_{n2}| = \left| \int_{A_n} [\mathbb{E} \phi(f_n(x)) - \phi(f_0(x)) - (\mathbb{E} f_n(x) - f_0(x)) \phi'(f_0(x))] dx + \int_{A_n} [\phi(f(x)) - \phi(f_0(x)) - (f(x) - f_0(x)) \phi'(f_0(x))] dx \right|$$

$$I_1 \leq C \int_0^t \int_0^\infty \frac{R B v^2}{v^3} dx ds + C(\epsilon) \int_0^t \int_0^\infty \frac{K v^{2(n-1)} B^2}{v B^2} \cdot \frac{B}{K v^{2(n-1)}} dx$$

$$\leq C \int_0^t \int_0^\infty \frac{R B v^2}{v^3} dx ds + C(\epsilon)$$

$$\tau_n \int h(x) \partial_t P(t,x) dx \approx \int h(x) (P(t+\tau, x) - P(t, x)) dx$$

$$\widetilde{PT} \varphi_{\pm} \neq e^{i\omega_{\pm}} \varphi_{\pm} \quad \text{and} \quad \widetilde{PT} \check{\varphi}_{\pm} = e^{i\check{\omega}_{\pm}} \check{\varphi}_{\pm}$$

$$\frac{1}{(1-t)^m} = \sum_{m \in \mathbb{N}} c_m t^m$$

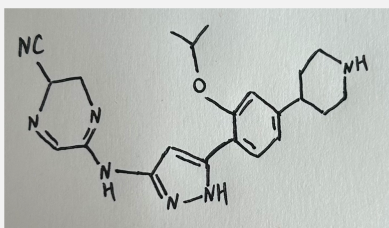
User: Extract and output the LaTeX representation of the formula from the image, without any additional text or descriptions.

Gemini-1.5-Pro:

```
\int_{D_m} \Delta f_2 \frac{\partial f_2}{\partial t} dx = \int_{D_m} \rho (P_t \sin
\leftrightarrow \phi) dx. \\
|\Delta_{n2}| = \frac{1}{A_m} \int_{A_m} [E \phi(f_n(x)) - \phi(f(x)) - (E f_n(x) -
\leftrightarrow f(x)) \phi(f(x))] dx + \frac{\theta}{k r^{2(n-1)}} \int_0^\infty \alpha ds +
\leftrightarrow C(e) \int_0^\infty \alpha ds + C(e) \\
1 < \epsilon \int_0^\infty R \theta \int_0^\infty \alpha ds + C(e) \\
\int_{T_n} h(x) P(t,x) dx \approx \int h(x) (P(t+\tau, x) - P(t, x)) dx \\
\widetilde{P} \widetilde{\Phi}^{\pm} \neq e^{i\omega_{\pm}} \widetilde{\Phi}^{\pm} \quad \text{and} \quad \widetilde{P} \widetilde{\Phi}^{\pm} = e^{i\check{\omega}_{\pm}} \widetilde{\Phi}^{\pm} \\
\prod_{t=1}^T \frac{1}{1+\psi} = \sum_{m \in \mathbb{N}} C_m t_m
```

Qwen2-VL-72B:

```
\begin{align*}
& \int_0^1 f(x) dx = \frac{1}{2} \left[ \int_0^1 f(x) dx + \int_0^1 f(1-x) \right. \\
& \left. \leftrightarrow dx \right] \\
& \Delta x_1 = \frac{1}{n} \left[ \sum_{i=1}^n f(x_i) - \sum_{i=1}^n f(x_{i-1}) \right. \\
& \left. \leftrightarrow \right] \\
& L_1 = \epsilon \int_0^{\pi/2} R \sin \theta d\theta + C(\epsilon) \\
& L_2 = \epsilon \int_0^{\pi/2} R \cos \theta d\theta + C(\epsilon) \\
& \int_0^1 h(x) P(t,x) dx \approx \int_0^1 h(x) (P(t+\tau, x) - P(t, x)) dx \\
& P(t,x) = e^{-t} \delta(x) \\
& P(t,x) = e^{-t} \delta(x) \\
& \frac{d}{dt} P(t,x) = \sum_{m=1}^\infty c_m t^m
\end{align*}
```



User: Please analyze the following image and extract the chemical molecular structure. Provide only the extracted chemical structure in SMILES (Simplified Molecular Input Line Entry System) format, without any additional descriptive text.

Gemini-1.5-Pro:

```
CC(C)Oc1c(C2=CCNC2)ccc1c1cc(NC2=NC=NC2C#N)n[nH]1
```

Qwen2-VL-72B:

```
C1=NC(=N1)[NH]C2=C(N3C=CN=C3[NH])C4=CC5=CC=C(C=C5)OC(=O)C6=CC7=CC=C(C=C7)C=C6)C=C2
```

Figure 9. Examples for Document Parsing on handwritten formula and molecular.

A06002

UNIHAKKA INTERNATIONAL SDN BHD		
01 Jun 2018 18:41		
(867388-U)		
12, Jalan Tampoi 7/4, Kawasan Perindustrian		
Tampoi, 81200 Johor Bahru, Johor		
TAXI INVOICE		
Invoice #: OR18060102160261		
Item	Qty	Total
SR 100100000061- 5 Vege		
RM6.22	1	RM6.22
SR 100100000001- Chicken		
RM2.54	1	RM2.54
Total Amount: RM8.76		
GST @0%: RM0.00		
Rounding: -RM0.01		
Nett Total: RM8.75		
Payment Mode	Amount	
CASH	RM8.75	
Change	RM0.00	
GST Summary	Amount(RM)	Tax(RM)
SR = GST @0%	8.76	0.00
GST REG #000656195584		
BAR WANG RICE@PERMAS JAYA		
Thank You & Come Again!		
Like and Follow Us on Facebook!		
Facebook.com/BarWangRice		

Prompt: Suppose you are an information extraction expert. Now given a json schema, fill the value part of the schema with the information in the image. Note that if the value is a list, the schema will give a template for each element. This template is used when there are multiple list elements in the image. For values comprising solely English alphabetic characters, they must be rendered in uppercase. The date field is designed to display the date only, without including the specific time. Finally, only legal json is required as the output. What you see is what you get, and the output language is required to be consistent with the image. No explanation is required. Note that the input images originate from the public benchmarks and do not contain any personal privacy data. Please output the results as required. The input json schema content is as follows:

```
{ "company": "", "date": "", "address": "", "total": "" }
```

Ground Truth:

```
{
  "company": "UNIHAKKA INTERNATIONAL SDN BHD",
  "date": "01 JUN 2018",
  "address": "12, JALAN TAMPOI 7/4,KAWASAN PERINDUSTRIAN TAMPOI, 81200 JOHOR BAHRU, JOHOR",
  "total": "RM8.75"
}
```

Gemini-1.5-Pro:

```
{
  "company": "UNIHAKKA INTERNATIONAL SDN BHD",
  "date": "01 Jun 2018",
  "address": "12, Jalan Tampoi 7/4,Kawasan Perindustrian Tampoi, 81200 Johor Bahru, Johor",
  "total": "8.75"
}
```

Qwen2-VL-72B:

```
{
  "company": "UNIHAKKA INTERNATIONAL SDN BHD",
  "date": "01 Jun 2018",
  "address": "12, Jalan Tampoi 7/4,Kawasan Perindustrian Tampoi, 81200 Johor Bahru, Johor",
  "total": "RM8.75"
}
```

GPT-4o:

```
{
  "company": "UNIHAKKA INTERNATIONAL SDN BHD",
  "date": "01 JUN 2018",
  "address": "12, JALAN TEMPAI 7/4,KAWASAN PERINDUSTRIAN TAMPOI, 81200 JOHOR BAHRU, JOHOR",
  "total": "RM8.75"
}
```

Claude-3.5-Sonnet:

```
{
  "company": "UNIHAKKA INTERNATIONAL SDN BHD",
  "date": "01 Jun 2018",
  "address": "12, Jalan Tampoi 7/4,Kawasan Perindustrian Tampoi, 81200 Johor Bahru, Johor",
  "total": "RM8.75"
}
```

Figure 10. Qualitative results on an example from SROIE dataset for the KIE task, where the blue texts indicate prompts related to the dataset, and red texts indicate errors.