DeepSeek-OCR: Contexts Optical Compression

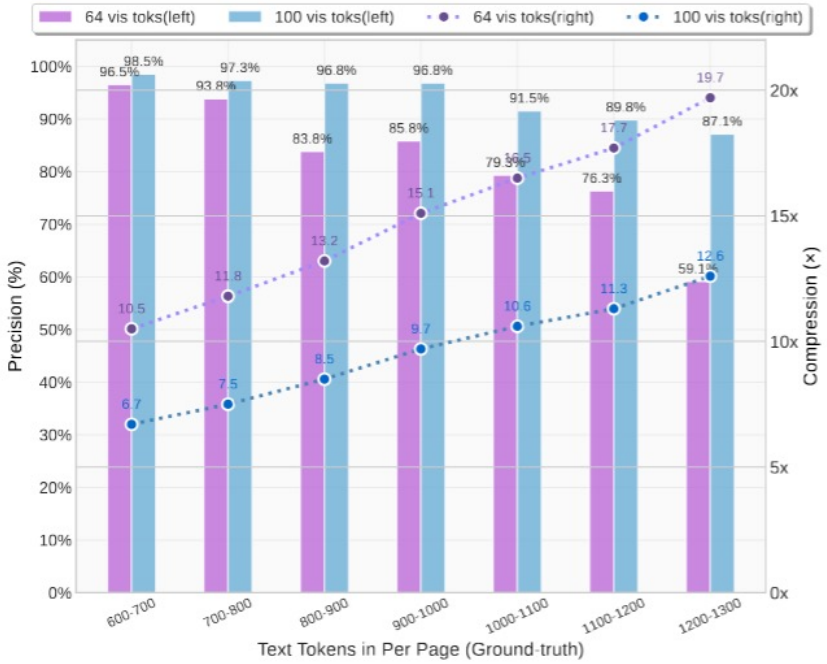
Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

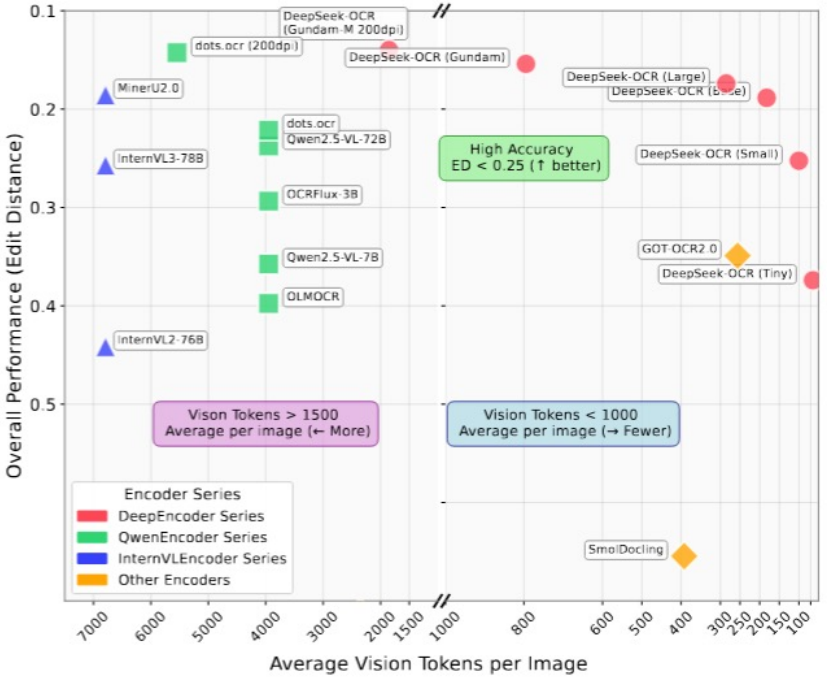
Presented by:

John Tan Chong Min

Huge compression with DeepSeek OCR



(a) Compression on Fox benchmark



(b) Performance on Omnidocbench

Can we use images to compress text?

Great accuracy to compression ratio

- Experiments show that when the number of text tokens is within 10 times that of vision tokens (i.e., a **compression ratio < 10×**), the model can achieve decoding (OCR) precision of **97%**

- Even at **a compression ratio of 20×**, the OCR accuracy still remains at about **60%**

Architecture

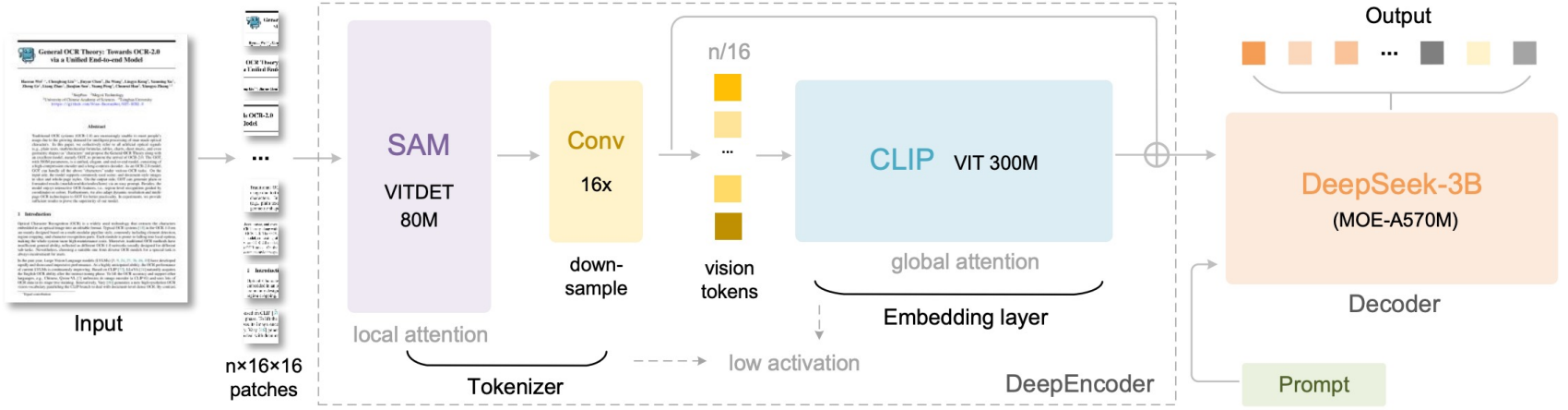
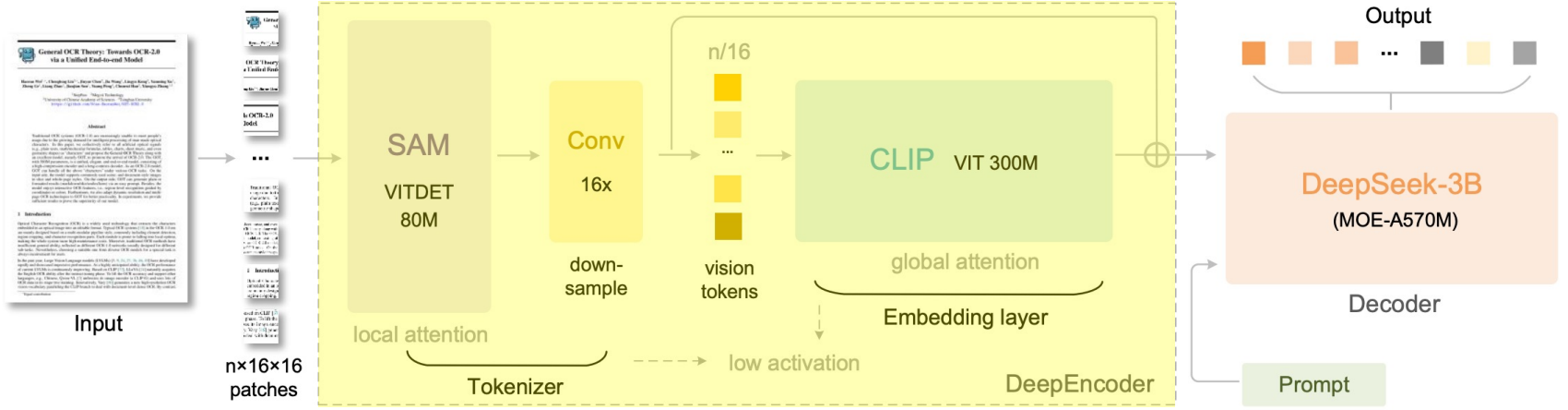


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three components: a SAM [17] for perception dominated by window attention, a CLIP [29] for knowledge with dense global attention, and a 16× token compressor that bridges between them.

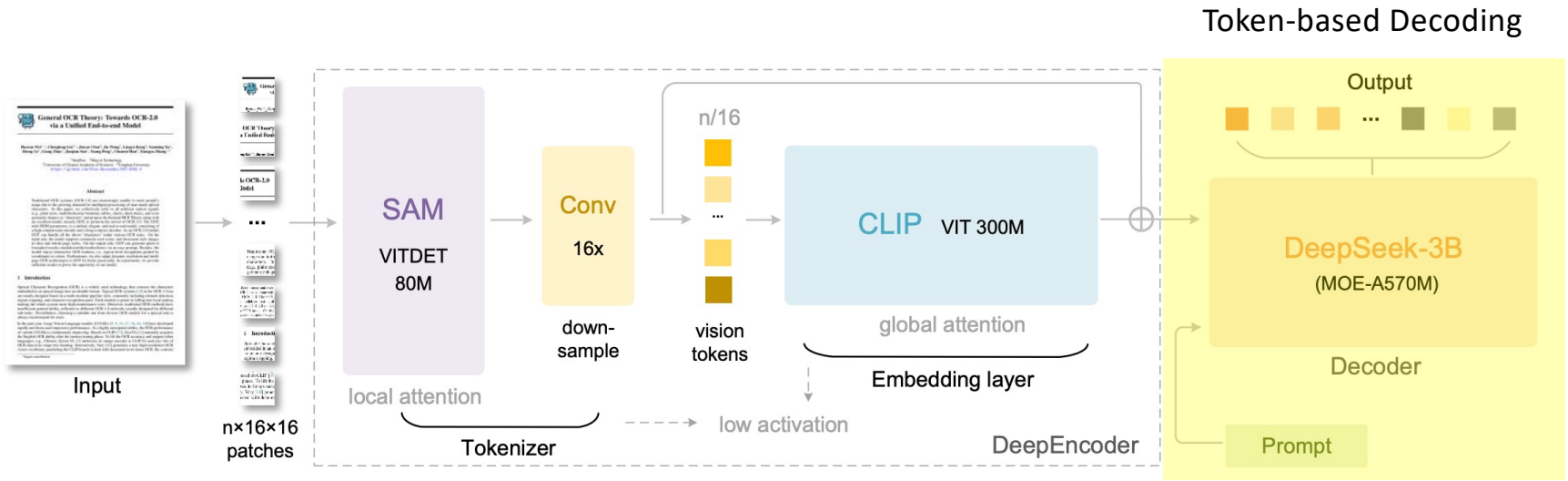
Encoder

Local focus (SAM) -> Compression (Conv) -> Text-based latents (CLIP)



- Extracts image features and tokenizes as well as compresses visual representations
- Uses Visual Transformers (ViT) in Segment Anything Model (SAM) and CLIP

Decoder



- Generates the required result based on image tokens and prompts
- Token-based generation of final text

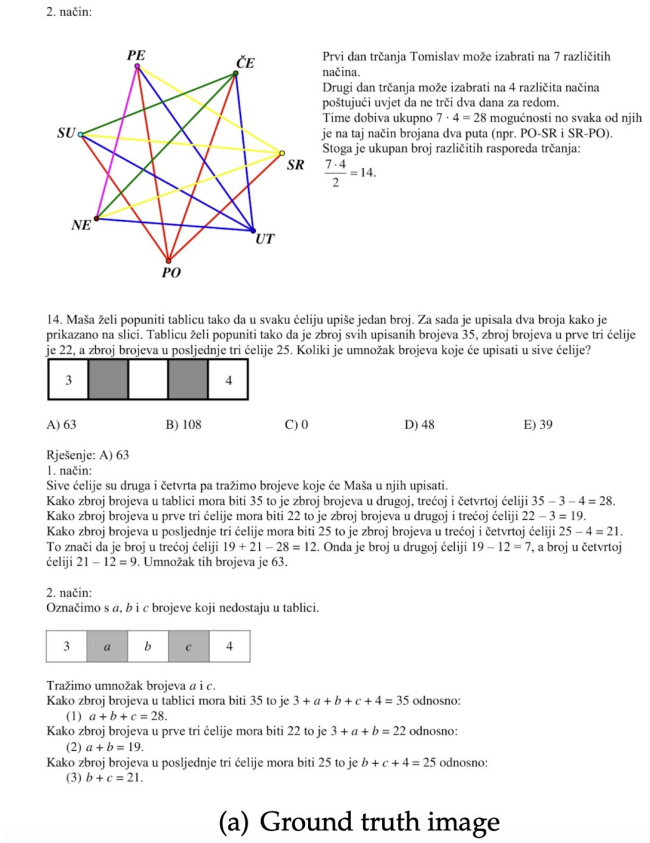
Training Data (Image Input -> Text Output)

- **OCR (70%):** Traditional OCR tasks such as scene image OCR and document OCR, Parsing tasks for complex artificial images, such as common charts, chemical formulas, and plane geometry parsing data

- **General vision data (20%):** Used to inject certain general image understanding capabilities into DeepSeek OCR and preserve the general vision interface

- **Text-only data (10%):** Converts image input of text to text output

OCR 1.0: Mapping document regions



2. način:

Prvi dan trčanja Tomislav može izabrati na 7 različitih načina.
Drugi dan trčanja može izabrati na 4 različita načina poštujući uvjet da ne trči dva dana za redom.
Time dobiva ukupno $7 \cdot 4 = 28$ mogućnosti no svaka od njih je na taj način brojana dva puta (npr. PO-SR i SR-PO).
Stoga je ukupan broj različitih rasporeda trčanja:
$\frac{7 \cdot 4}{2} = 14.$

14. Maša želi popuniti tablicu tako da u svaku ćeliju upiše jedan broj. Za sada je upisala dva broja kako je prikazano na slici. Tablicu želi popuniti tako da je zbroj svih upisanih brojeva 35, zbroj brojeva u prve tri ćelije je 22, a zbroj brojeva u posljednje tri ćelije 25. Koliki je umnožak brojeva koje će upisati u sve ćelije?

3				4

A) 63 B) 108 C) 0 D) 48 E) 39

Rješenje: A) 63
1. način:
Sive ćelije su druga i četvrta pa tražimo brojeve koje će Maša u njih upisati.
Kako zbroj brojeva u tablici mora biti 35 to je zbroj brojeva u drugoj, trećoj i četvrtoj ćeliji $35 - 3 - 4 = 28$.
Kako zbroj brojeva u prve tri ćelije mora biti 22 to je zbroj brojeva u drugoj i trećoj ćeliji $22 - 3 = 19$.
Kako zbroj brojeva u posljednje tri ćelije mora biti 25 to je zbroj brojeva u trećoj i četvrtoj ćeliji $25 - 4 = 21$.
To znači da je broj u trećoj ćeliji $19 + 21 - 28 = 12$. Onda je broj u drugoj ćeliji $19 - 12 = 7$, a broj u četvrtoj ćeliji $21 - 12 = 9$. Umnožak tih brojeva je 63.

2. način:
Označimo s a, b i c brojeve koji nedostaju u tablici.

3	a	b	c	4

Tražimo umnožak brojeva a i c.
Kako zbroj brojeva u tablici mora biti 35 to je $3 + a + b + c + 4 = 35$ odnosno:
(1) $a + b + c = 28$.
Kako zbroj brojeva u prve tri ćelije mora biti 22 to je $3 + a + b = 22$ odnosno:
(2) $a + b = 19$.
Kako zbroj brojeva u posljednje tri ćelije mora biti 25 to je $b + c + 4 = 25$ odnosno:
(3) $b + c = 21$.

(a) Ground truth image

```
<|ref|>text<|/ref|><|det|>[[55, 43, 130, 60]]<|/det|>
2. način:

<|ref|>image<|/ref|><|det|>[[70, 93, 450, 360]]<|/det|>


<|ref|>text<|/ref|><|det|>[[460, 95, 896, 132]]<|/det|>
Prvi dan trčanja Tomislav može izabrati na 7 različitih načina.

<|ref|>text<|/ref|><|det|>[[460, 131, 880, 168]]<|/det|>
Drugi dan trčanja može izabrati na 4 različita načina poštujući uvjet da ne trči
dva dana za redom.

<|ref|>text<|/ref|><|det|>[[460, 166, 941, 220]]<|/det|>
Time dobiva ukupno \(7 \cdot 4 = 28\) mogućnosti no svaka od njih je na taj način
brojana dva puta (npr. PO-SR i SR-PO). Stoga je ukupan broj različitih rasporeda
trčanja:

<|ref|>equation<|/ref|><|det|>[[460, 217, 550, 256]]<|/det|>
\[ \frac{7 \cdot 4}{2} = 14. \]

<|ref|>text<|/ref|><|det|>[[55, 397, 931, 452]]<|/det|>
14. Maša želi popuniti tablicu tako da u svaku ćeliju upiše jedan broj. Za sada
je upisala dva broja kako je prikazano na slici. Tablicu želi popuniti tako da je
zbroj svih upisanih brojeva 35, zbroj brojeva u prve tri ćelije je 22, a zbroj
brojeva u posljednje tri ćelije 25. Koliki je umnožak brojeva koje će upisati u
sve ćelije?

<|ref|>table<|/ref|><|det|>[[57, 450, 360, 500]]<|/det|>
<table><tr><td>3</td><td></td><td></td><td></td><td>4</td></tr></table>

<|ref|>text<|/ref|><|det|>[[55, 515, 110, 534]]<|/det|>
A) 63

<|ref|>text<|/ref|><|det|>[[230, 515, 293, 534]]<|/det|>
B) 108

<|ref|>text<|/ref|><|det|>[[405, 515, 450, 534]]<|/det|>
C) 0

<|ref|>text<|/ref|><|det|>[[581, 515, 636, 534]]<|/det|>
D) 48
```

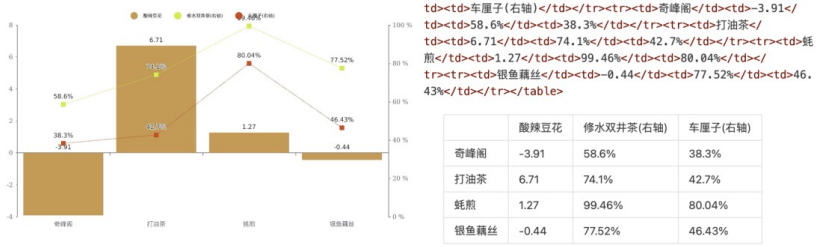
(b) Fine annotations with layouts

Use other OCR software to generate ground truth

Ground Truth contains:

- Bounding Boxes
- Label (text, image, table, equation etc.)
- Content

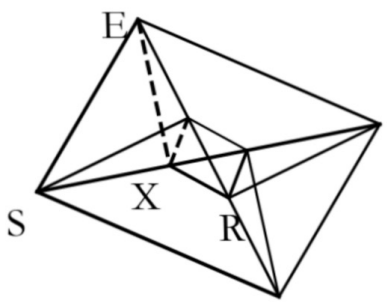
ORC 2.0: Advanced Figure Analysis



<table><tr><td></td><td>酸辣豆花</td><td>修水双井茶(右轴)</td><td>车厘子(右轴)</td></tr><tr><td>奇峰阁</td><td>-3.91</td><td>58.6%</td><td>38.3%</td></tr><tr><td>打油茶</td><td>6.71</td><td>74.1%</td><td>42.7%</td></tr><tr><td>蚝煎</td><td>1.27</td><td>99.46%</td><td>80.04%</td></tr><tr><td>银鱼藕丝</td><td>-0.44</td><td>77.52%</td><td>46.43%</td></tr></table>

	酸辣豆花	修水双井茶(右轴)	车厘子(右轴)
奇峰阁	-3.91	58.6%	38.3%
打油茶	6.71	74.1%	42.7%
蚝煎	1.27	99.46%	80.04%
银鱼藕丝	-0.44	77.52%	46.43%

(a) Image-text ground truth of chart




```
{
 "Line": {
  "line": [
   "(-8.0, -1.4) — (-6.0, 2.07) — (-3.99, 5.54) — (-3.19, 6.89)",
   "(-8.0, -1.4) — (-4.42, 0.39) — (-0.84, 2.18) — (-0.6, 2.3)",
   "(-8.0, -1.4) — (-4.36, -3.06) — (-0.72, -4.72) — (2.92, -6.38) — (3.19, -6.51)",
   "(-8.0, -1.4) — (-4.07, -0.65) — (-0.14, 0.1) — (3.79, 0.85) — (7.72, 1.6) — (8.0, 1.66)",
   "(-3.19, 6.89) — (-2.4, 2.97) — (-1.75, -0.2)",
   "(-3.19, 6.89) — (-1.5, 3.26) — (0.19, -0.37) — (1.88, -3.99) — (3.19, -6.51)",
   "(-3.19, 6.89) — (0.45, 5.23) — (4.09, 3.57) — (7.73, 1.91) — (8.0, 1.66)",
   "(-1.75, -0.2) — (-0.6, 2.3)",
   "(-1.75, -0.2) — (0.9, -2.3)",
   "(-0.6, 2.3) — (1.9, 0.2)",
   "(0.9, -2.3) — (1.9, 0.2)",
   "(0.9, -2.3) — (4.48, -0.51) — (8.0, 1.66)",
   "(1.9, 0.2) — (2.69, -3.72) — (3.19, -6.51)",
   "(3.19, -6.51) — (5.19, -3.04) — (7.2, 0.43) — (8.0, 1.66)"
  ],
  "line_type": [
   "-", "-", "-", "-", "—", "-", "-", "—", "-", "-", "-", "-", "-", "-"
  ],
  "line_endpoint": [
   "S: (-8.0, -1.4)",
   "E: (-3.19, 6.89)",
   "X: (-1.75, -0.2)",
   "R: (0.9, -2.3)"
  ]
 }
}
```


(b) Image-text ground truth of geometry

Figure 6 | For charts, we do not use OneChart's [7] dictionary format, but instead use HTML table format as labels, which can save a certain amount of tokens. For plane geometry, we convert the ground truth to dictionary format, where the dictionary contains keys such as line segments, endpoint coordinates, line segment types, etc., for better readability. Each line segment is encoded using the Slow Perception [39] manner.

General Vision and Text-only data

- **General Vision Data:** Generate relevant data for tasks such as caption, detection, and grounding. Makes use of CLIP's general visual knowledge

- **Text-only data:** Data processed to a length of 8192 tokens, which is also the sequence length for DeepSeek-OCR

Text Compression is possible using Vision

Table 2 | We test DeepSeek-OCR's vision-text compression ratio using all English documents with 600-1300 tokens from the Fox [21] benchmarks. Text tokens represent the number of tokens after tokenizing the ground truth text using DeepSeek-OCR's tokenizer. Vision Tokens=64 or 100 respectively represent the number of vision tokens output by DeepEncoder after resizing input images to 512×512 and 640×640.

Text Tokens	Vision Tokens =64		Vision Tokens=100		
	Precision	Compression	Precision	Compression	Pages
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

Text Compression as Memory Forgetting?

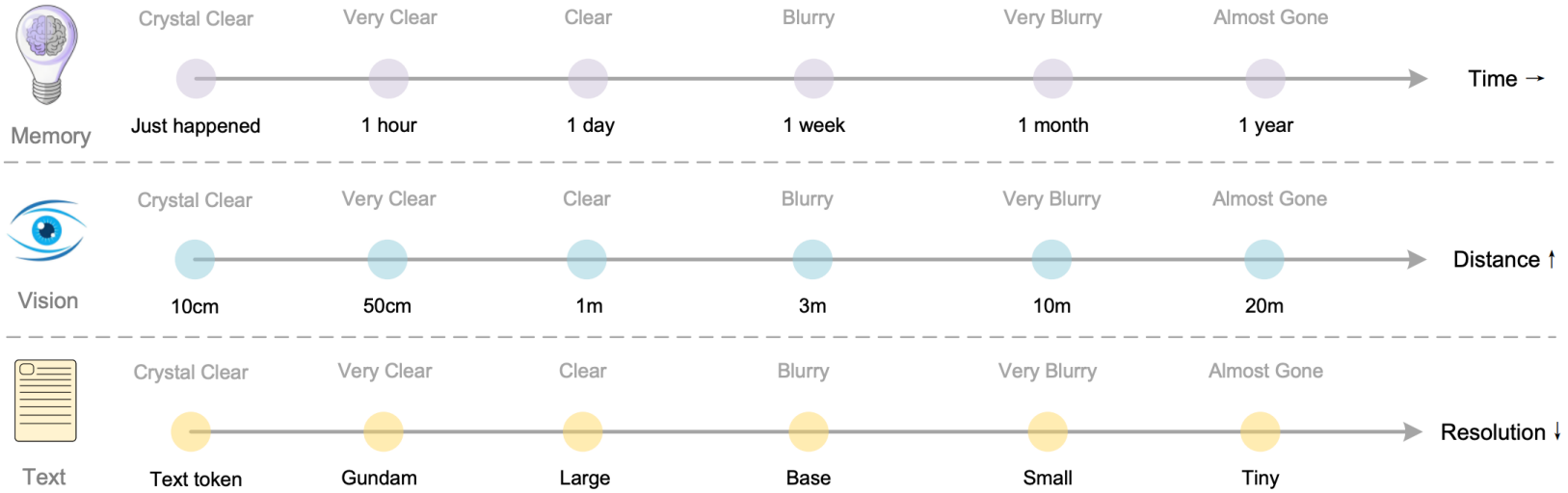


Figure 13 | Forgetting mechanisms constitute one of the most fundamental characteristics of human memory. The contexts optical compression approach can simulate this mechanism by rendering previous rounds of historical text onto images for initial compression, then progressively resizing older images to achieve multi-level compression, where token counts gradually decrease and text becomes increasingly blurred, thereby accomplishing textual forgetting.

Parallel Insights in VisionRAG paper

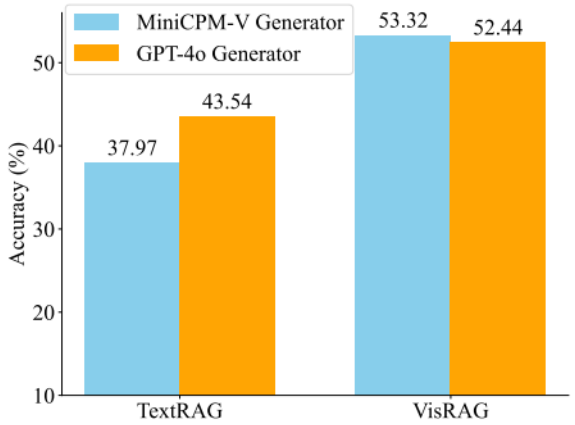
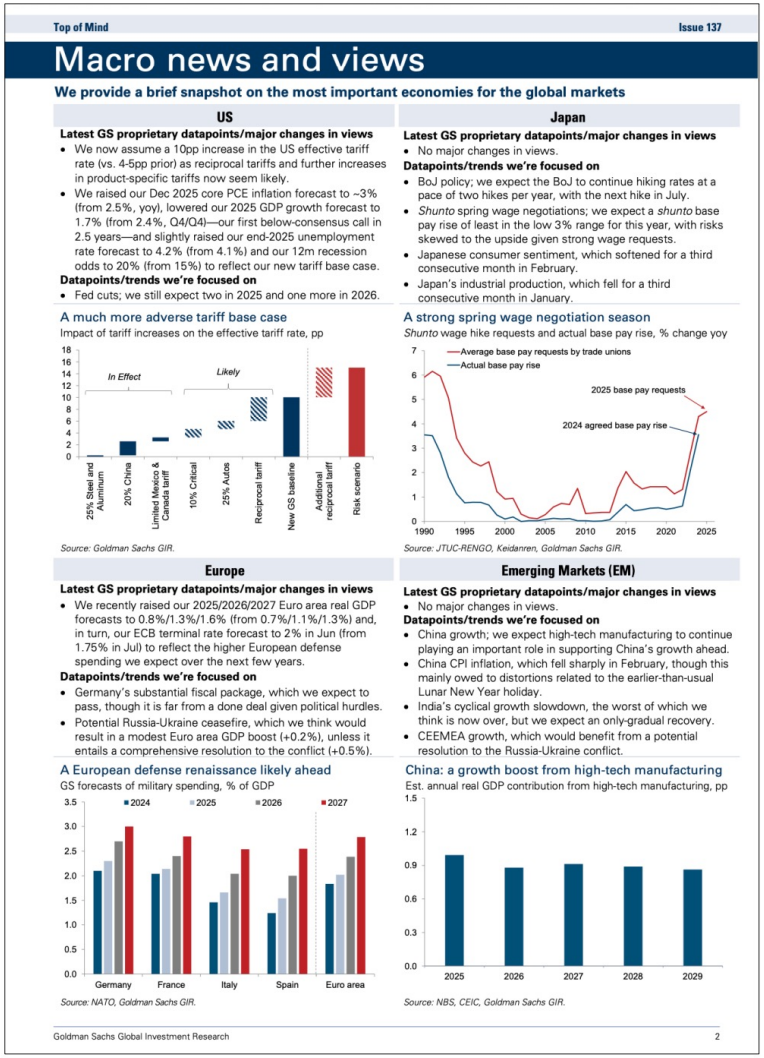


Figure 1: TextRAG vs. VisRAG on final generation accuracy. In TextRAG, parsed text serves as the basis for both retrieval and generation processes. In contrast, VisRAG leverages the original document image directly by using a VLM-based retriever and generator. Details can be found in Sec. 5.1.

- **Using images directly for RAG can improve retrieval compared to using text chunks for RAG**
- *My personal experience: Vision RAG can indeed search over larger amounts of data, but they are lossier in retrieving exact information*

VISRAG: Vision-based Retrieval-augmented Generation On Multi-modality Documents. Yu et. al. ICLR. 2025. https://arxiv.org/pdf/2410.10594

OCR Example Results

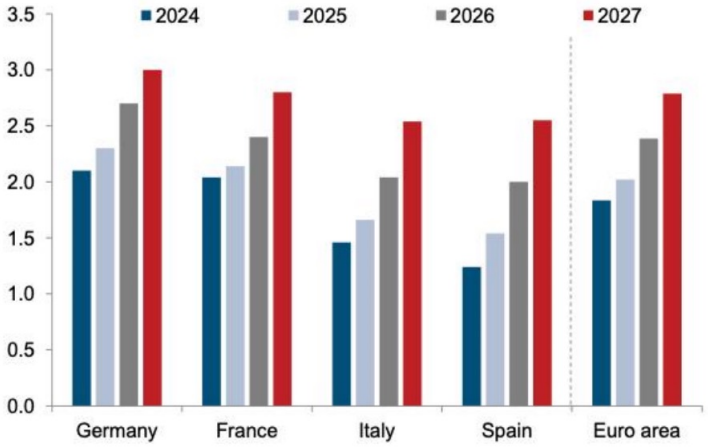


Input image

Result

Document Layout

<image>\nParse the figure.



	2024	2025	2026	2027
Germany	2.1	2.3	2.7	3.0
France	2.05	2.15	2.4	2.8
Italy	1.45	1.65	2.05	2.55
Spain	1.25	1.55	2.0	2.55
Euro area	1.85	2.05	2.4	2.8

Deep Parsing

Europe

Latest GS proprietary datapoints/major changes in views

We recently raised our 2025/2026/2027 Euro area real GDP forecasts to 0.8%/1.3%/1.6% (from 0.7%/1.1%/1.3%) and, in turn, our ECB terminal rate forecast to 2% in Jun (from 1.75% in Jul) to reflect the higher European defense spending we expect over the next few years.

Datapoints/trends we're focused on

Germany's substantial fiscal package, which we expect to pass, though it is far from a done deal given political hurdles. Potential Russia- Ukraine ceasefire, which we think would result in a modest Euro area GDP boost (+0.2%), unless it entails a comprehensive resolution to the conflict (+0.5%)

A European defense renaissance likely ahead

GS forecasts of military spending, % of GDP

Source: NATO, Goldman Sachs GIR.

Japan

Latest GS proprietary datapoints/major changes in views

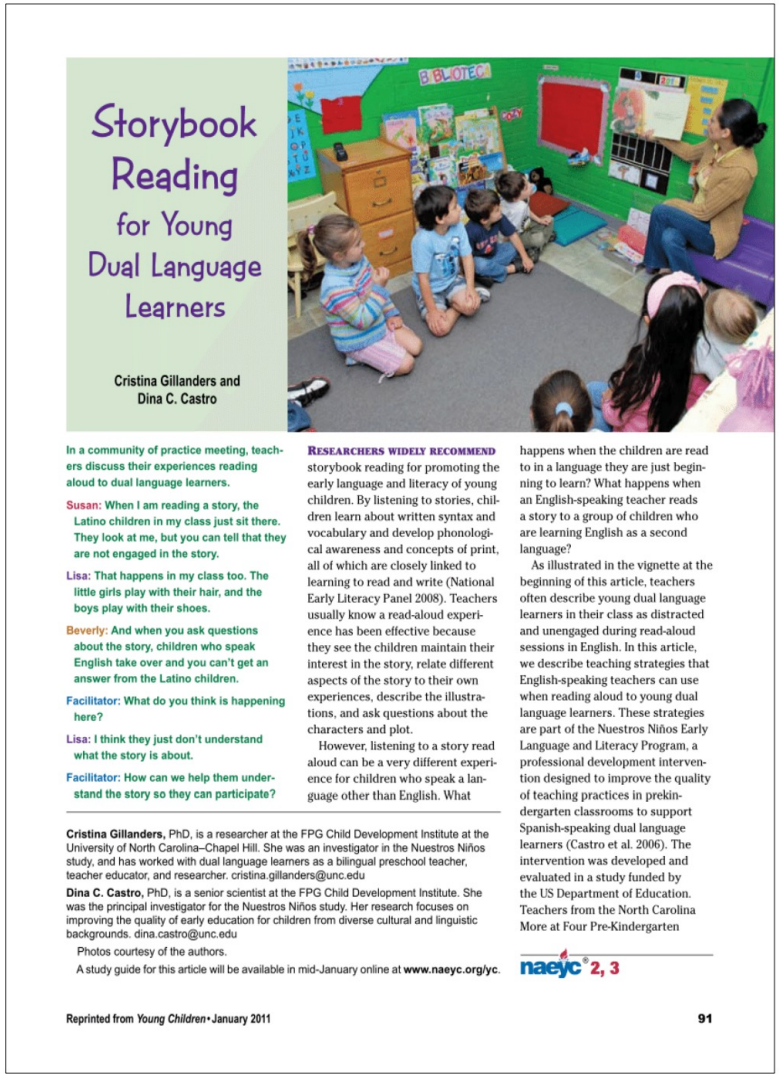
No major changes in views. Datapoints/trends we're focused on BoJ policy; we expect the BoJ to continue hiking rates at a pace of two hikes per year, with the next hike in July. Shunto spring wage negotiations; we expect a shunto base pay rise of least in the low 3% range for this year, with risks skewed to the upside given strong wage requests. Japanese consumer sentiment, which softened for a third consecutive month in February. Japan's industrial production, which fell for a third consecutive month in January.

A strong spring wage negotiation season

Shunto wage hike requests and actual base pay rise, % change yoy

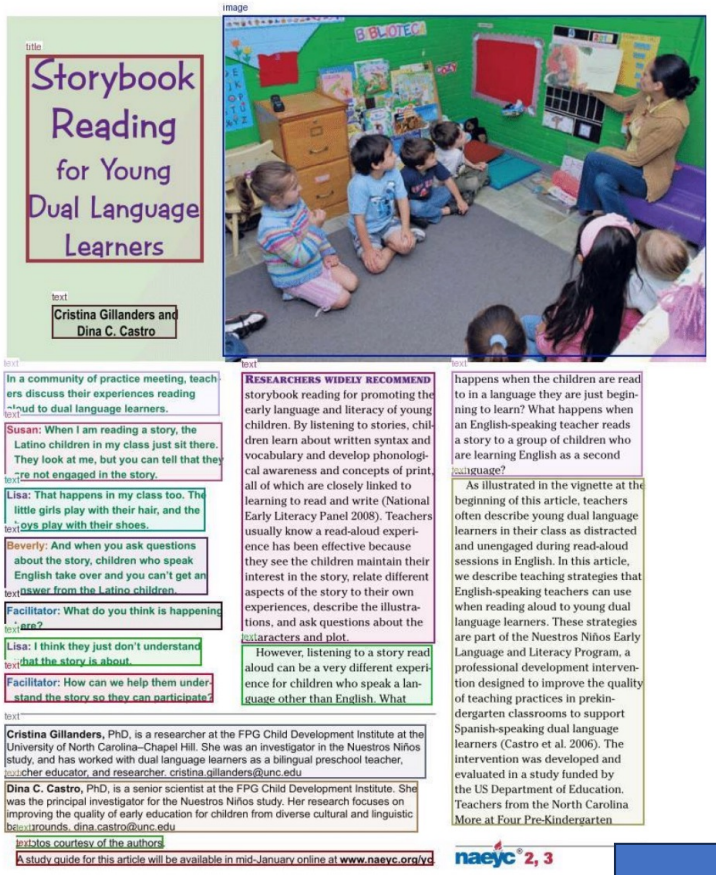
Rendering

Graph Reconstruction



Input image

<image>\n<|grounding|>Convert the document to markdown.



Result

Document Layout

<image>\nParse the figure.

The image depicts an indoor classroom setting with a group of children and an adult. The children are seated on the floor, facing a woman who is standing and appears to be reading or presenting to them. The woman is wearing a brown sweater and blue jeans. The children are dressed in various colors, with some wearing short pants and others in long pants.

The classroom has a green wall with educational posters and a bulletin board. The floor is covered with a gray carpet. To the left, there is a wooden dresser with a drawer partially open, and a chair is visible behind it. On the right side of the image, there is a purple bean bag chair.

The children are engaged with the woman, with some looking at her and others looking down or away. The room is well-lit, and the overall atmosphere seems to be one of attentiveness and learning.

The text "BIBLIOTECA" is visible on the wall, suggesting that the room may be part of a library or a section dedicated to books. The presence of educational materials and the organized layout of the room indicate that this is a space designed for learning and reading.

Deep Parsing

Storybook Reading for Young Dual Language Learners

Cristina Gillanders and Dina C. Castro

In a community of practice meeting, teachers discuss their experiences reading aloud to dual language learners.

Susan: When I am reading a story, the Latino children in my class just sit there. They look at me, but you can tell that they are not engaged in the story.

Lisa: That happens in my class too. The little girls play with their hair, and the boys play with their shoes.

Beverly: And when you ask questions about the story, children who speak English take over and you can't get an answer from the Latino children.

Facilitator: What do you think is happening here?

Lisa: I think they just don't understand what the story is about.

Facilitator: How can we help them understand the story so they can participate?

RESEARCHERS WIDELY RECOMMEND storybook reading for promoting the early language and literacy of young children. By listening to stories, children learn about written syntax and vocabulary and develop phonological awareness and concepts of print, all of which are closely linked to learning to read and write (National Early Literacy Panel 2008). Teachers usually know a read-aloud experience has been effective because they see the children maintain their interest in the story, relate different aspects of the story to their own experiences, describe the illustrations, and ask questions about the characters and plot.

However, listening to a story read aloud can be a very different experience for children who speak a language other than English. What

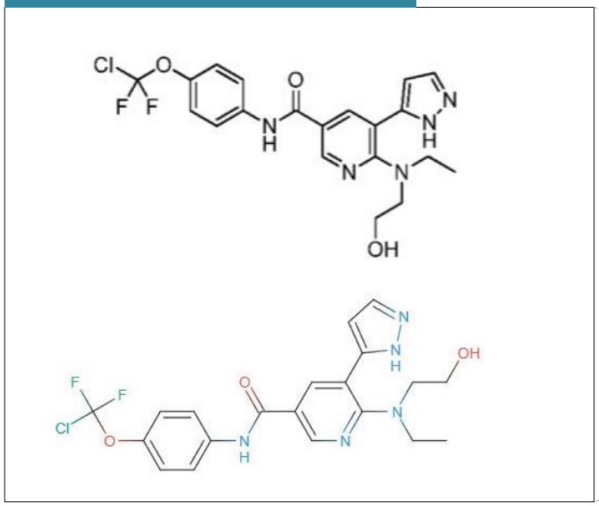
happens when the children are read to in a language they are just beginning to learn? What happens when an English- speaking teacher reads a story to a group of children who are learning English as a second language?

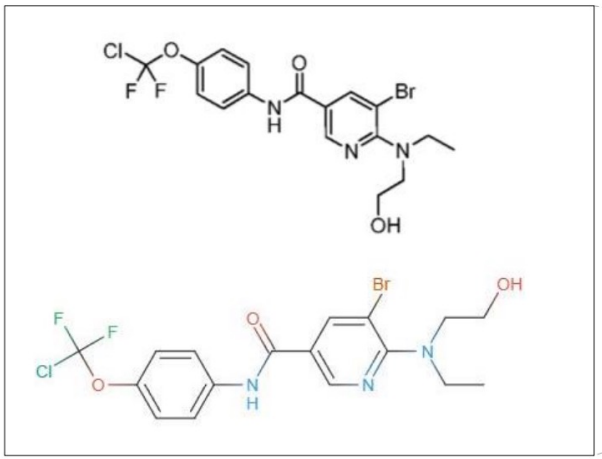
As illustrated in the vignette at the beginning of this article, teachers often describe young dual language
distracted and unengaged during read- aloud sessions in English. In this article, we describe teaching str
speaking teachers can use when reading aloud to young dual language learners. These strategies are par

Rendering

Image Understanding

<image>\nParse the figure.



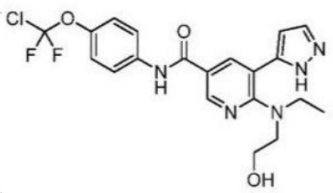


Deep Parsing

[00369] The title compound was prepared in an analogous fashion to that described in Stage 22.1 using 5- bromo- 6- chloro- N- (4- (chlorodifluoromethoxy)phenyl)nicotinamide (Stage 22.2) and 2- methylamino- ethanol to afford a white crystalline solid. HPLC (Condition 4) $t_R = 5.72$ min, UPLC- MS (Condition 3) $t_R = 1.14\text{min}, m/z = 452.2[M+H]^+$.

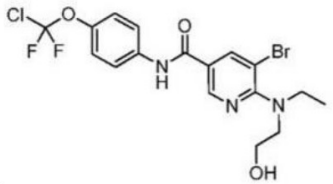
Example 24

N- (4- (Chlorodifluoromethoxy)phenvn- 6- (ethyl(2- hvdroxyethyl)amino)- 5- (IH- pyrazol- 5- vDnicotinamide



[00370] The title compound was prepared in an analogous fashion to that described in Example 26 using 5- bromo- N- (4- (chlorodifluoromethoxy)phenyl)- 6- (ethyl(2- hydroxyethyl)amino)nicotinamide (Stage 24.1) and 1- (tetrahydro- 2H- pyran- 2- yl)- 5- (4,4,5,5- tetramethyl- 1,3,2- dioxaborolan- 2- yl)- 1H- pyrazole to afford a yellow solid. UPLC- MS (Condition 3) $t_R = 1.02\text{min}, m/z = 452.2[M+H]^+, m/z = 450.1[M-H]^-$; ^1H-NMR (400 MHz, DMSO- d6) δ ppm 0.93 (t, $J = 7.09\text{Hz}, 3H$) 3.17 - 3.27 (m, 2 H) 3.35 - 3.43 (m, 2 H) 3.43 - 3.53 (m, 2 H) 4.59 (br. s, 1 H) 6.53 (d, $J = 1.96\text{Hz}, 1H$) 7.33 (d, $J = 9.05\text{Hz}, 2H$) 7.76 (br. s, 1 H) 7.82 - 7.95 (m, 2 H) 8.13 (d, $J = 2.45\text{Hz}, 1H$) 8.72 (d, $J = 2.45\text{Hz}, 1H$) 10.29 (s, 1 H) 12.98 (br. s, 1 H).

[00371] Stage 24.1 5- Bromo- N- (4- (chlorodifluoromethoxy)phenyl)- 6- (ethyl(2- hydroxyethyl)amino)nicotinamide

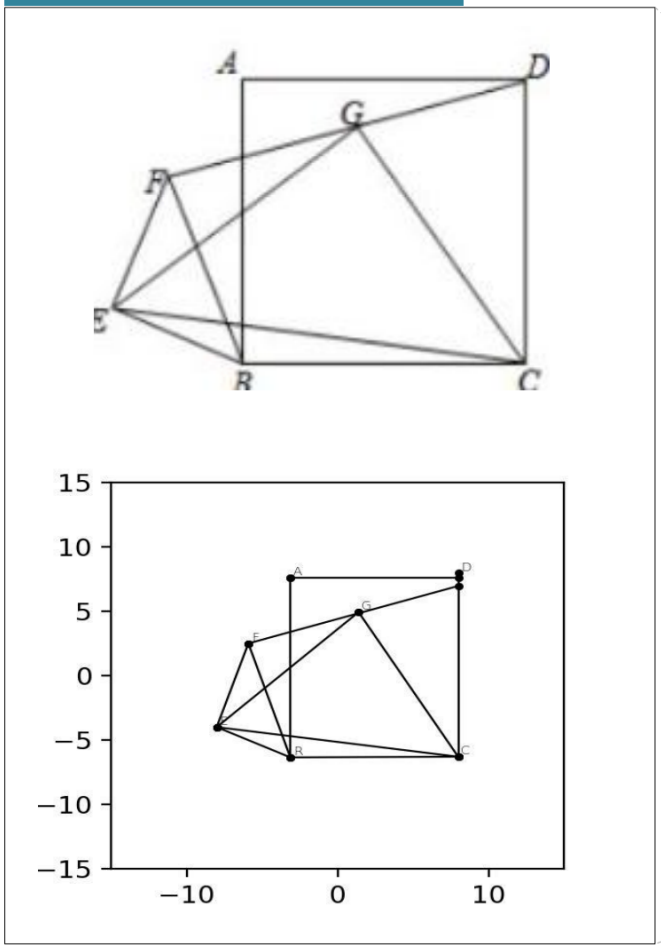


[00372] The title compound was prepared in an analogous fashion to that described in Stage 22.1 using 5- bromo- 6- chloro- N- (4- (chlorodifluoromethoxy)phenyl)nicotinamide (Stage

Rendering

Complex Image Reconstruction

<image>\nParse the figure.



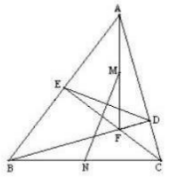
Deep Parsing

八年级数学下册几何证明题练习

1. 已知: $\triangle ABC$ 的两条高 BD, CE 交于点 F, 点 M, N, 分别是 AF, BC 的中点, 连接 ED, MN;

(1)证明: MN 垂直平分 ED;

(2)若 $\angle EBD = \angle DCE = 45°$, 判断以 M, E, N, D 为顶点的四边形的形状, 并证明你的结论;



2. 四边形 ABCD 是正方形, $\triangle BEF$ 是等腰直角三角形, $\angle BEF = 90°$, BE=EF, 连接 DF, G 为 DF 的中点, 连接 EG, CG, EC;

(1)如图1,若点E在CB边的延长线上,直接写出EG与GC的位置关系及 $\frac{EG}{GC}$ 的值;

(2)将图1中的△BEF绕点B顺时针旋转至图2所示位置,请问(1)中所得的结论是否仍然成立?若成立,请写出证明过程;若不成立,请说明理由;

(3)将图1中的△BEF绕点B顺时针旋转 $\alpha(0° < \alpha < 90°)$, 若 BE=1, AB= $\sqrt{2}$, 当 E, F, D 三点共线时, 求 DF 的长;

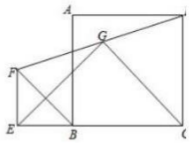


图1

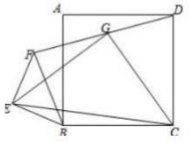
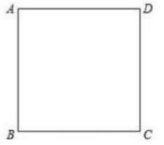


图2



备用图

Rendering

Complex Figure Reconstruction

<image>\nIdentify all objects in the image and output them in bounding boxes.

<image>\nLocate <|ref|>the teacher<|/ref|> in the image.

<image>\nLocate <|ref|>11-2=<|/ref|> in the image.

Bounding Box Identification

My Practical Experience using DeepSeek OCR

- Only good for text that is well structured in neat lines, with no background interference

- Misses out text within images

Overall Takeaway

- A model is only as good as its training data
- A model is only as good as the model architecture that enables it to preserve information (like positional information)
- Use the right model, trained on the right data, for the right tasks

Question to Ponder

- Should we use a Vision Transformer-based model to coordinate-sensitive positions like output bounding boxes?

- Should we compress text modality into visual images for processing? What are the advantages and drawbacks?

- Which is better, a Vision Language Model-based OCR, versus a traditional OCR software?