# A Overview

Our supplementary includes the following sections:

- **Section B: Framework details.** Details for model design, implementation and training data.
- **Section C: Detection performance of the visual CoT bboxes.** Details for detection performance for the intermediate visual CoT bounding boxes.
- **Section D: More experiment results.** Additional performance evaluation and performance analysis.
- **Section E: Prompt design.** Prompt for generating the visual CoT dataset and evaluating the performance.
- **Section F: Limitations.** Discussion of limitations of our work.
- **Section G: Potential negative societal impacts.** Discussion of potential negative societal impacts of our work.
- **Section F: More visualization.** More Visualization of our dataset and demos.
- **Section I: Disclaimer.** Disclaimer for the visual CoT dataset and the related model.

Following NeurIPS Dataset and Benchmark track guidelines, we have shared the following artifacts:

| **Artifcat** | **Link** | **License** |
|---|---|---|
| Code Repository | `https://github.com/deepcs233/Visual-CoT` | Apache-2.0 license |
| Data | `https://huggingface.co/datasets/deepcs233/Visual-CoT` | CC BY 4.0 |
| Model Weights | `https://huggingface.co/collections/deepcs233/viscot-65fe883e2a0cdd3c59fc5d63` | Apache-2.0 license |

The authors are committed to ensuring its regular upkeep and updates.

# B Framework details

## B.1 Model details

We choose the pre-trained ViT-L/14 of CLIP [57] as the vision encoder and Vicuna-7/13B [13] as our LLM, which has better instruction following capabilities in language tasks compared to LLaMA [64]. Consider an input original image, we take the vision encoder to obtain the visual feature. Similar to LLaVA [40, 39], we use a simple linear layer to project the image features into the word embedding space to obtain the visual tokens $H_0$ which share the same dimensionality of the LLM.

## B.2 Implementation details

Following the setup described by Vicuna [13], our model undergoes a two-stage training process. In the first stage, we pre-train the model for 1 epoch using a learning rate of 2e-3 and a batch size of 128. For the second stage, we fine-tune the model for 1 epoch on our visual CoT dataset, employing a learning rate of 2e-5 and a batch size of 128. The Adam optimizer with zero weight decay and a cosine learning rate scheduler are utilized. To conserve GPU memory during fine-tuning, we employ FSDP (Full Shard Data Parallel) with ZeRO3-style. All models are trained using 32 × A100s. In the case of training the setting with a 7B LLM and a resolution of 224, the first/second pre-training stage completes within 1/16 hours.

## B.3 Training data details

We train the model on a reorganized Vision-Language dataset. The training data is a composite of three sources: the second stage data from LLaVA, data from Shikra's [6] second stage, and our visual CoT data. The inclusion of data from Shikra, which features various datasets with positional annotations, such as RefCOCO [24] for REC, visual gemone [27] for grounding caption. These datasets can enhance VisCoT's ability to accurately identify and understand locations within images. This enhancement is crucial for tasks requiring precise spatial awareness. We listed all training data in Table 6. We removed the images from the training set that are the same as those in the testing or validation set to prevent potential data leakage. Our training data includes three parts, and they are from LLaVA-1.5, a subset of Shikra, and our proposed visual CoT dataset separately.

Table 6: The overview of our training dataset.

| Dataset | Size | Source Datasets |
|---|---|---|
| LLaVA-1.5 | 665K | LLaVA, ShareGPT, VQAv2, GQA, OKVQA<br>OCRVQA, A-OKVQA, TextCaps, RefCOCO, VG |
| Shikra | 1.4M | RefCOCO(+/g), VG, PointQA-Local/Twice<br>Visual-7W, Flickr30K |
| Visual CoT dataset | 376K | TextVQA, TextCaps, DocVQA, Birds-200-2011<br>Flickr30K, InfographicsVQA, VSR, GQA, Open images |

# C Detection performance of the visual CoT bboxes

In Table 7, we present the detection performance based on the predicted CoT bounding boxes. A higher performance indicates that our VisCoT identifies the key regions with greater accuracy.

# D More experiment results

## D.1 Performance evaluation

In Tab. 8 and Tab. 9, we showcase the baseline performance of our model, where it directly answers questions without employing the visual CoT process.

**Multi-modal Large Language Models Benchmarks.** In Tab. 8, we evaluate our model on recently proposed MLLM benchmarks such as MME [16], POPE [36], MMbench [42], ScienceQA [43],

Table 7: Detection performance (Top-1 Accuracy@0.5) on the visual CoT benchmark. The ground truth bounding boxes used for computing the metric are the intermediate CoT bounding boxes annotated in our CoT benchmark.

| | | **Doc/Text** | | | | | **Chart** |
|---|---|---|---|---|---|---|---|
| MLLM | Res. | DocVQA | TextCaps | TextVQA | DUDE | SROIE | InfographicsVQA |
| VisCoT-7B | $224^2$ | 13.6 | 41.3 | 46.8 | 5.0 | 15.7 | 7.2 |
| VisCoT-7B | $336^2$ | 20.4 | 46.3 | 57.6 | 9.6 | 18.5 | 10.0 |

| | | **General VQA** | | **Relation Reasoning** | | | **Fine-grained** | **Average** |
|---|---|---|---|---|---|---|---|---|
| MLLM | Res. | Flickr30k | Visual7W | GQA | Open images | VSR | Birds-200-2011 | |
| VisCoT-7B | $224^2$ | 49.6 | 31.1 | 42.0 | 57.6 | 69.6 | 67.0 | 37.2 |
| VisCoT-7B | $336^2$ | 51.3 | 29.4 | 49.5 | 59.3 | 54.0 | 47.1 | 37.6 |

Table 8: Comparison with SoTA methods on 8 benchmarks. VisCoT achieves the best performance on the most of benchmarks, and ranks second on the other. For a fair comparison, VisCoT generates responses directly, without the visual CoT process. SQA [43]; $\text{VQA}^{\text{T}}$: TextVQA [61]; $\text{MME}^{\text{P}}$: MME-Preception [16]; $\text{MME}^{\text{C}}$: MME-Cognition [16]; POPE [36]; MMB: MMBench [42]; $\text{MMB}^{\text{CN}}$: MMBench-Chinese [42]. † uses 50M in-house instruction-finetuning data. * uses multiple vision encoders.

| Method | LLM | Res. | SQA | GQA | $\text{VQA}^{\text{T}}$ | POPE | $\text{MME}^{\text{P}}$ | $\text{MME}^{\text{C}}$ | MMB | $\text{MMB}^{\text{CN}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| BLIP-2 [31] | Vicuna-13B | $224^2$ | – | 41.0 | 42.5 | 85.3 | 1293.8 | – | – | – |
| InstructBLIP [14] | Vicuna-7B | $224^2$ | – | 49.2 | 50.1 | – | – | – | 36.0 | 23.7 |
| InstructBLIP [14] | Vicuna-13B | $224^2$ | – | 49.5 | 50.7 | 78.9 | 1212.8 | – | – | – |
| Shikra [6] | Vicuna-13B | $224^2$ | – | – | – | – | – | – | 58.8 | – |
| IDEFICS-9B [30] | LLaMA-7B | $224^2$ | 44.2 | 38.4 | 25.9 | – | – | – | 48.2 | 25.2 |
| IDEFICS-80B [30] | LLaMA-65B | $224^2$ | 68.9 | 45.2 | 30.9 | – | – | – | 54.5 | 38.1 |
| Qwen-VL† [3] | Qwen-7B | $448^2$ | 67.1 | 59.3 | **63.8** | – | – | – | 38.2 | 7.4 |
| Qwen-VL-Chat† [3] | Qwen-7B | $448^2$ | 68.2 | 57.5 | 61.5 | – | 1487.5 | **360.7** | 60.6 | 56.7 |
| LLaVA1.5 [40] | Vicuna-7B | $336^2$ | 66.8 | 62.0 | 58.2 | 85.9 | 1510.7 | – | 64.3 | 58.3 |
| LLaVA1.5 [40] | Vicuna-13B | $336^2$ | 71.6 | 63.3 | 61.3 | 85.9 | 1531.3 | 295.4 | 67.7 | **63.6** |
| SPHINX* [3] | LLaMA-13B | $224^2$ | 69.3 | 62.6 | 51.6 | 80.7 | 1476.1 | 310.0 | 66.9 | 56.2 |
| VisCoT | Vicuna-7B | $224^2$ | 68.2 | 63.1 | 55.4 | 86.0 | 1453.6 | 308.3 | **67.9** | 59.7 |
| VisCoT | Vicuna-13B | $224^2$ | 71.6 | **64.2** | 57.8 | 85.6 | 1480.0 | 255.4 | 66.9 | 60.5 |
| VisCoT | Vicuna-7B | $336^2$ | 68.3 | 62.0 | 61.0 | **86.5** | 1514.4 | 275.0 | 67.3 | 60.1 |
| VisCoT | Vicuna-13B | $336^2$ | **73.6** | 63.3 | 62.3 | 83.3 | **1535.7** | 331.8 | 67.4 | 61.6 |

TextVQA [61], GQA [21]. Our model still achieves comparative results across all benchmarks. This performance indicates that the visual CoT data we proposed not only enhances visual comprehension in CoT-specific scenarios but also boosts the model's overall visual understanding in standard inference setups. As demonstrated in Tab. 10, the implementation of visual CoT enables our model to achieve superior performance even with a lower resolution and a reduced number of visual tokens. This finding highlights the efficiency and effectiveness of the visual CoT approach in enhancing model accuracy.

**Visual grounding.** Furthermore, we evaluate VisCoT on REC benchmarks with RefCOCO [24], RefCOCO+ [48], and RefCOCOg [48] datasets. Our model outperforms the previous state-of-the-art models, including the specialist models such as G-DINO-L [41] and UNINEXT [78]. Notably, even with a minimal setup (7B LLM & 224 resolution), our approach outperforms methods that utilize higher resolutions or larger LLM models. This demonstrates that our dataset, enhanced with intermediate bounding boxes, significantly improves the model's precision in locating and understanding referred objects or regions. "Top-1 Accuracy@0.5" refers to the accuracy of a model in predicting the correct bounding box as the top prediction when the Intersection over Union (IoU) between the predicted and ground truth bounding boxes meets or exceeds 50%.

### D.2 Performance analysis

Tab. 4 shows that our baseline with visual CoT performs better than the model without CoT. We further investigate whether different bounding box sizes affect performance improvement. In Fig. 6,

Table 9: Performance (Top-1 Accuracy@0.5) on Referring Expression Comprehension (REC) tasks. For a fair comparison, VisCoT generates responses directly, without the visual CoT process.

| Method | Res. | RefCOCO+ | | | RefCOCO | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|---|---|
| | | val | test-A | test-B | val | test-A | test-B | val-u | test-u |
| *Specialist models* | | | | | | | | | |
| UNINEXT [78] | $640^2$ | 85.24 | 89.63 | 79.79 | 92.64 | 94.33 | 91.46 | 88.73 | 89.37 |
| G-DINO-L [41] | $384^2$ | 82.75 | 88.95 | 75.92 | 90.56 | 93.19 | 88.24 | 86.13 | 87.02 |
| *Generalist models* | | | | | | | | | |
| VisionLLM-H [69] | - | - | - | - | - | 86.70 | - | - | - |
| OFA-L [67] | $480^2$ | 68.29 | 76.00 | 61.75 | 79.96 | 83.67 | 76.39 | 67.57 | 67.58 |
| Shikra 7B [6] | $224^2$ | 81.60 | 87.36 | 72.12 | 87.01 | 90.61 | 80.24 | 82.27 | 82.19 |
| Shikra 13B [6] | $224^2$ | 82.89 | 87.79 | 74.41 | 87.83 | 91.11 | 81.81 | 82.64 | 83.16 |
| MiniGPT-v2-7B [5] | $448^2$ | 79.97 | 85.12 | 74.45 | 88.69 | 91.65 | 85.33 | 84.44 | 84.66 |
| MiniGPT-v2-7B-Chat [5] | $448^2$ | 79.58 | 85.52 | 73.32 | 88.06 | 91.29 | 84.30 | 84.19 | 84.31 |
| Qwen-VL-7B [3] | $448^2$ | 83.12 | 88.25 | 77.21 | 89.36 | 92.26 | 85.34 | 85.58 | 85.48 |
| Qwen-VL-7B-Chat [3] | $448^2$ | 82.82 | 88.59 | 76.79 | 88.55 | 92.27 | 84.51 | 85.96 | 86.32 |
| Ferret-7B [82] | $336^2$ | 80.78 | 87.38 | 73.14 | 87.49 | 91.35 | 82.45 | 83.93 | 84.76 |
| u-LLaVA-7B [77] | $224^2$ | 72.21 | 76.61 | 66.79 | 80.41 | 82.73 | 77.82 | 74.77 | 75.63 |
| SPHINX-13B [37] | $224^2$ | 82.77 | 87.29 | 76.85 | 89.15 | 91.37 | 85.13 | 84.87 | 83.65 |
| VisCoT-7B | $224^2$ | 85.68 | 91.34 | 80.20 | 90.60 | 93.49 | 86.65 | 85.29 | 86.04 |
| VisCoT-7B | $336^2$ | **87.46** | **92.05** | **81.18** | **91.77** | **94.25** | **87.46** | **88.38** | **88.34** |
| VisCoT-13B | $224^2$ | 86.26 | 91.20 | 80.57 | 91.40 | 93.53 | 87.26 | 86.62 | 86.79 |

Table 10: Performance on VQA benchmarks.

| Model | LLaVA-1.5-7B | VisCoT-7B (w/o COT) | VisCoT-7B | VisCoT-7B (w/o COT) | VisCoT-7B |
|---|---|---|---|---|---|
| Res. | $336^2$ | $224^2$ | $224^2$ | $336^2$ | $336^2$ |
| DocVQA | 21.6 | 14.4 | 39.0 | 29.4 | **49.3** |
| TextVQA | 58.2 | 55.5 | 62.9 | 60.2 | **66.9** |
| ChartQA | 17.7 | 14.2 | 19.2 | 17.5 | **22.8** |

we divide each evaluation dataset into five equal parts based on their relative bounding box sizes. We observe that the visual CoT usually achieve greater improvement when the corresponding bounding box is relative smaller.

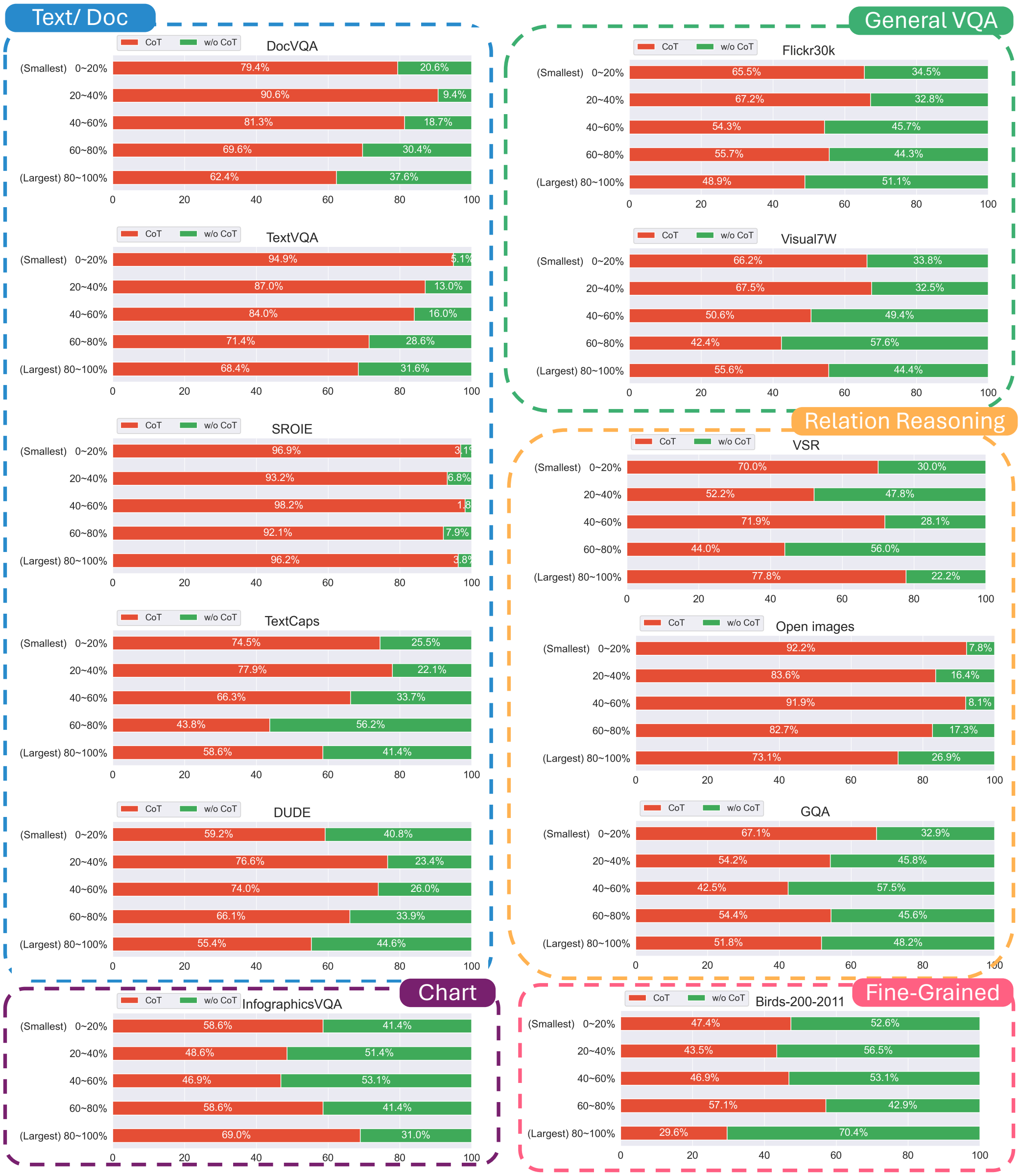


Figure 6: Visualization of performance improvement across different bounding box relative sizes for different source datasets. We find that visual CoT shows a larger improvement in cases where the queried object is relatively small. Red bars represent evaluation data samples where the model with CoT outperforms the model without CoT. Green bars indicate the opposite. The y-axis represents different ranges of relative sizes of bboxes $R$. For example, the 20-40% range indicates that the bboxes in this range occupy the relatively small 20-40% quantile within the entire dataset. For clarity, samples where both models achieve the same scores are omitted.

# E Prompt design

## E.1 Generating the dataset for TextCaps

You are an AI visual assistant, and you are seeing a single image. What you see is provided with several sentences and Ocr_tokens, describing the same image you are looking at. Ocr_tokens indicates the text in the image. Answer all questions as you are seeing the image. Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask THREE diverse questions and give corresponding answers. Again, do not ask about uncertain details. Do not just makeup questions and answers based on Ocr tokens. Your response should include questions asking about the textual information of the image, the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Please only ask questions that have definite answers:

- One can see the content in the image that the question asks about and can answer confidently;
- One can determine confidently from the image that it is not in the image. Do not ask any questions that cannot be answered confidently.
- One can not see the Ocr_tokens, so the question must not mention 'Ocr'

Craft Questions Around Ocr_tokens: Create questions that directly pertain to these identified words or phrases. Ensure that the question is structured in a way that the answer MUST be a word or phrase directly from the Ocr_tokens. Your answer cannot contain words outside of Ocr_tokens. The answers must be within three words.

Please follow the provided format:
Question: [question]
Answer: [answer]

Here is the context you need to process:
Image description: { }
Ocr_tokens: { }

### E.2 Generating the dataset for Flickr30k

You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Each sentence includes specific objects mentioned and their corresponding locations within the image (*e.g.*, [a peach] is located at [area: 95162] ) Answer all questions as you are seeing the image. Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.
The generated questions need closer examination of specific regions in the image to gather detailed information for answering. The generated answers must be based on the corresponding area.
When creating your questions, keep the following considerations in mind:

- Direct Alignment: Ensure the "Focus Area" specified in each question directly corresponds to the content of the question. For instance, if the question refers to "two women", the focus area should align with the portion described as "[Two women]" in the image description.
- Image-Only Basis: Respondents will only have access to the image itself and will NOT see the provided descriptions or area details. Ensure your questions can be answered by viewing the image alone.
- Avoid Repetition: Each question should be distinctive without overlapping content.
- Clarity and Precision: The answers to your questions should be both lucid and exact. Evade vagueness.
- Restricted Question Formats: Refrain from phrasing questions like "What's in region xx?" or "What happens in description 1?". The terms "description" and "region" should not appear in your questions & answers.
- MUST: The "Focus Area" you provide can answer the question you provide.

Please follow the provided format, area_id is a number:
Question: [question]
Focus Area: [area: area_id]
Answer: [answer]

Here is the data you need to process:
Describe 1: With a barn in the background a child puts her head through a hole in a cow cutout and smiles for the camera.
[a barn] is located at [area: 62407]
[a child] is located at [area: 62402]
[a hole] is located at [area: 62405]
. . .

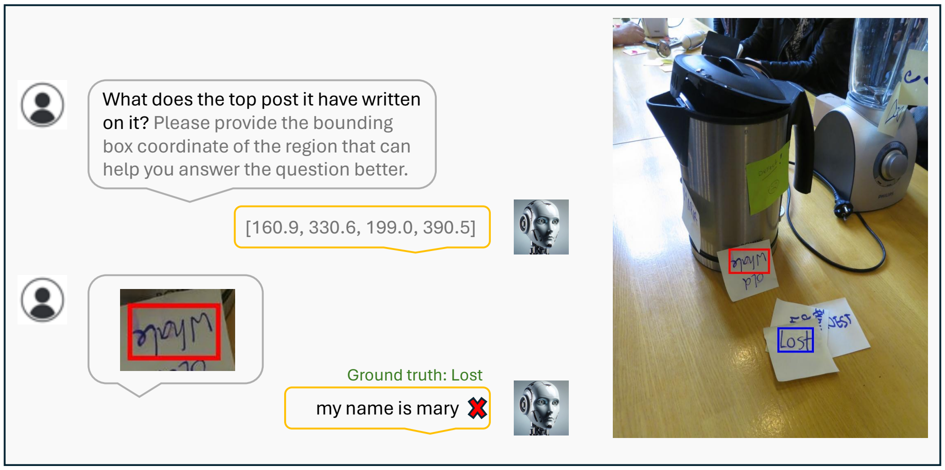
### E.3 Generating the dataset with detailed reasoning steps for GQA

You are an AI visual assistant, and you are seeing a single image. I will provide a question-answer pair along with the corresponding reasoning steps. The question and answer are based on an image. You need to generate the pure reasoning text in a step-by-step format, with each step clearly numbered (1. 2. 3. ... etc). The reasoning text should help solve the question and reach the final answer without including or hinting at the answer itself. The reasoning text must not include any ID numbers.

Here is the data you need to process:

Question: What appliance is to the right of the cabinet?
Answer: The appliance is a microwave.
Reasoning steps: [{"operation": "select", "dependencies": [], "argument": "cabinet (3588933)"},
{"operation": "relate", "dependencies": [0], "argument": "appliance,to the right of,s (1564001)"},
{"operation": "query", "dependencies": [1], "argument": "name"}]

. . .



What does the top post it have written on it? Please provide the bounding box coordinate of the region that can help you answer the question better.

[160.9, 330.6, 199.0, 390.5]

Ground truth: Lost

my name is mary ✗

Figure 7: Visualization results of the VisCoT. Model-generated bounding boxes are shown in red, while ground truth (GT) bounding boxes are in blue. In this case, our model incorrectly predicts the CoT region, leading to a wrong answer.

### E.4 Evaluation for the visual CoT benchmark using the ChatGPT

You are responsible for proofreading the answers, you need to give a score to the model's answer by referring to the standard answer, based on the given question. The full score is 1 point and the minimum score is 0 points. Please output the score in the form "score: <score>". The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score.

Question: { }
Standard answer: { }
Model's answer: { }

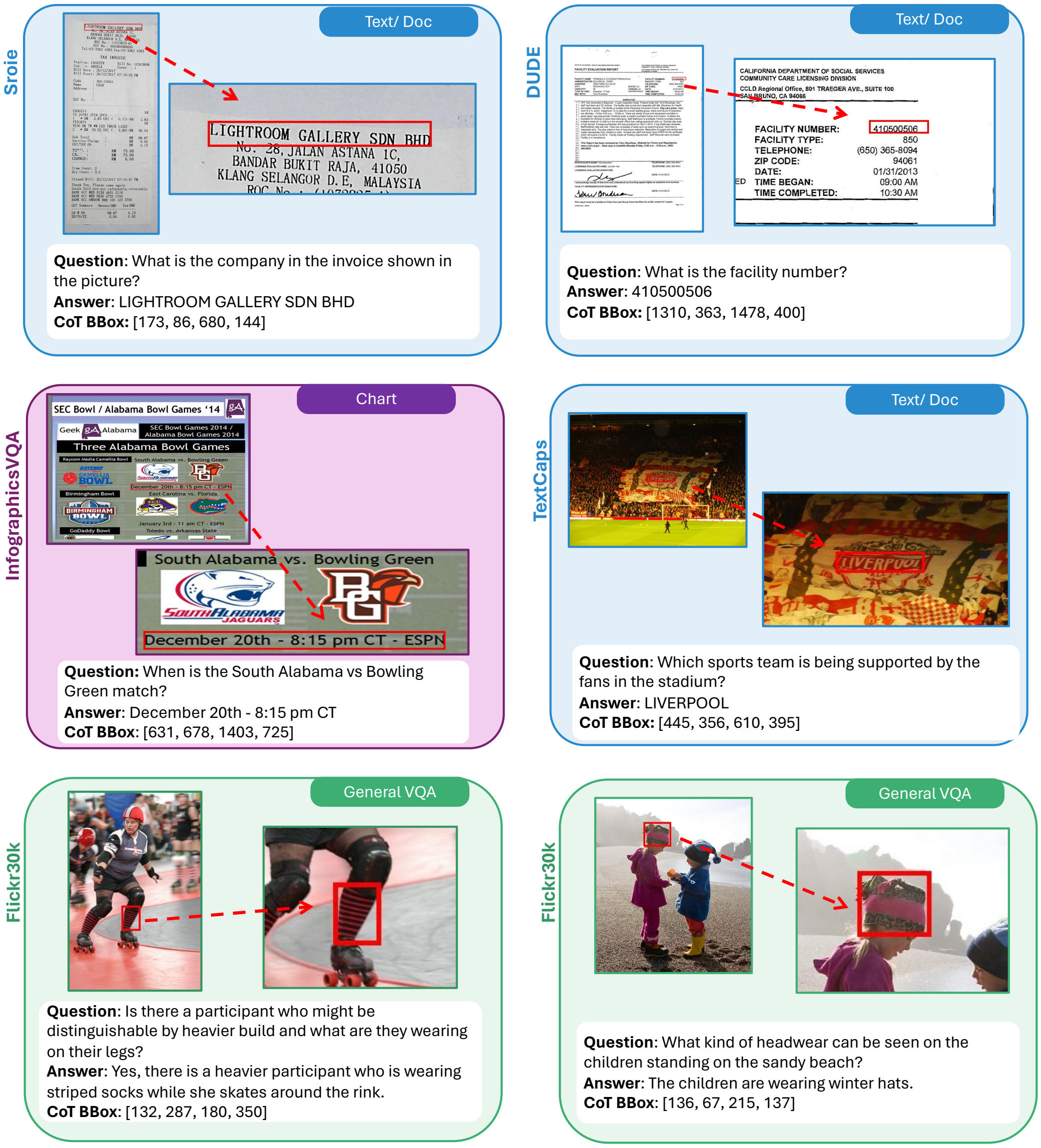
## F Limitations

In scenarios where the input image contains extensive information or the question is particularly complex, VisCoT may struggle to identify the most relevant region for answering the question. As shown in Figure 7, this challenge can sometimes result in the model being misled and producing incorrect responses.

Our data pipeline inherits the limitations of utilizing GPT-4 API. (1) Accuracy and Misinformation: Generated content may not always be accurate, which could lead to the spread of misinformation. To mitigate this, we have designed a comprehensive filtering script as a post-process to improve content quality. (2) Bias and Fairness: Since we do not have access to the training data of GPT-4, the generated instructional data might reflect inherent biases, potentially reinforcing social or cultural inequalities present in the base model training. In terms of data usage, we explicitly state that OpenAI's terms must be adhered to, and the data can only be used for research purposes.

## G Potential negative societal impacts

The potential negative societal impacts of our work are similar to other MLLMs and LLMs. The development of Visual CoT and MLLMs, while advancing AI, poses societal risks like increased privacy invasion, the perpetuation of biases, the potential for misinformation, job displacement, and ethical concerns regarding accountability and consent.



**Question**: What is the company in the invoice shown in the picture?
**Answer**: LIGHTROOM GALLERY SDN BHD
**CoT BBox:** [173, 86, 680, 144]

**Question**: What is the facility number?
**Answer:** 410500506
**CoT BBox:** [1310, 363, 1478, 400]

**Question**: When is the South Alabama vs Bowling Green match?
**Answer**: December 20th - 8:15 pm CT
**CoT BBox**: [631, 678, 1403, 725]

**Question**: Which sports team is being supported by the fans in the stadium?
**Answer**: LIVERPOOL
**CoT BBox:** [445, 356, 610, 395]

**Question**: Is there a participant who might be distinguishable by heavier build and what are they wearing on their legs?
**Answer**: Yes, there is a heavier participant who is wearing striped socks while she skates around the rink.
**CoT BBox**: [132, 287, 180, 350]

**Question**: What kind of headwear can be seen on the children standing on the sandy beach?
**Answer**: The children are wearing winter hats.
**CoT BBox**: [136, 67, 215, 137]

Figure 8: Examples in the visual CoT dataset, with corresponding question-answer annotations and visual CoT bboxes. The red bounding boxes in the images highlight the critical image regions that provide necessary and related information for answering the questions.

## H More visualization

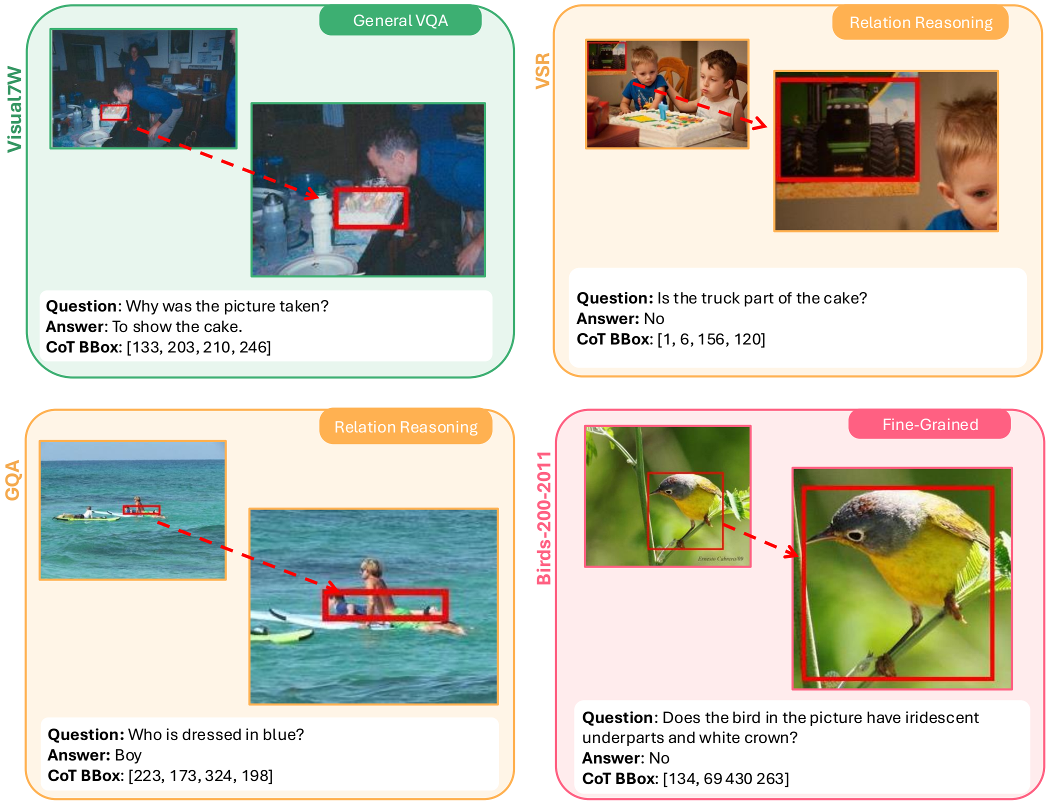
We provide more visualization results of our proposed visual CoT dataset in Fig. 8, Fig. 9.

We provide more visualization results of our VisCoT baseline in Fig. 10, Fig. 11, Fig. 12, Fig. 13.

## I Disclaimer

This dataset was collected and released solely for research purposes, with the goal of making the MLLMs dynamically focus on visual inputs and provide intermediate interpretable thoughts. The authors are strongly against any potential harmful use of the data or technology to any party.

**Intended Use.** The data, code, and model checkpoints are intended to be used solely for (I) future research on visual-language processing and (II) reproducibility of the experimental results reported



**Question**: Why was the picture taken?
**Answer**: To show the cake.
**CoT BBox**: [133, 203, 210, 246]

**Question:** Is the truck part of the cake?
**Answer:** No
**CoT BBox**: [1, 6, 156, 120]

**Question:** Who is dressed in blue?
**Answer:** Boy
**CoT BBox**: [223, 173, 324, 198]

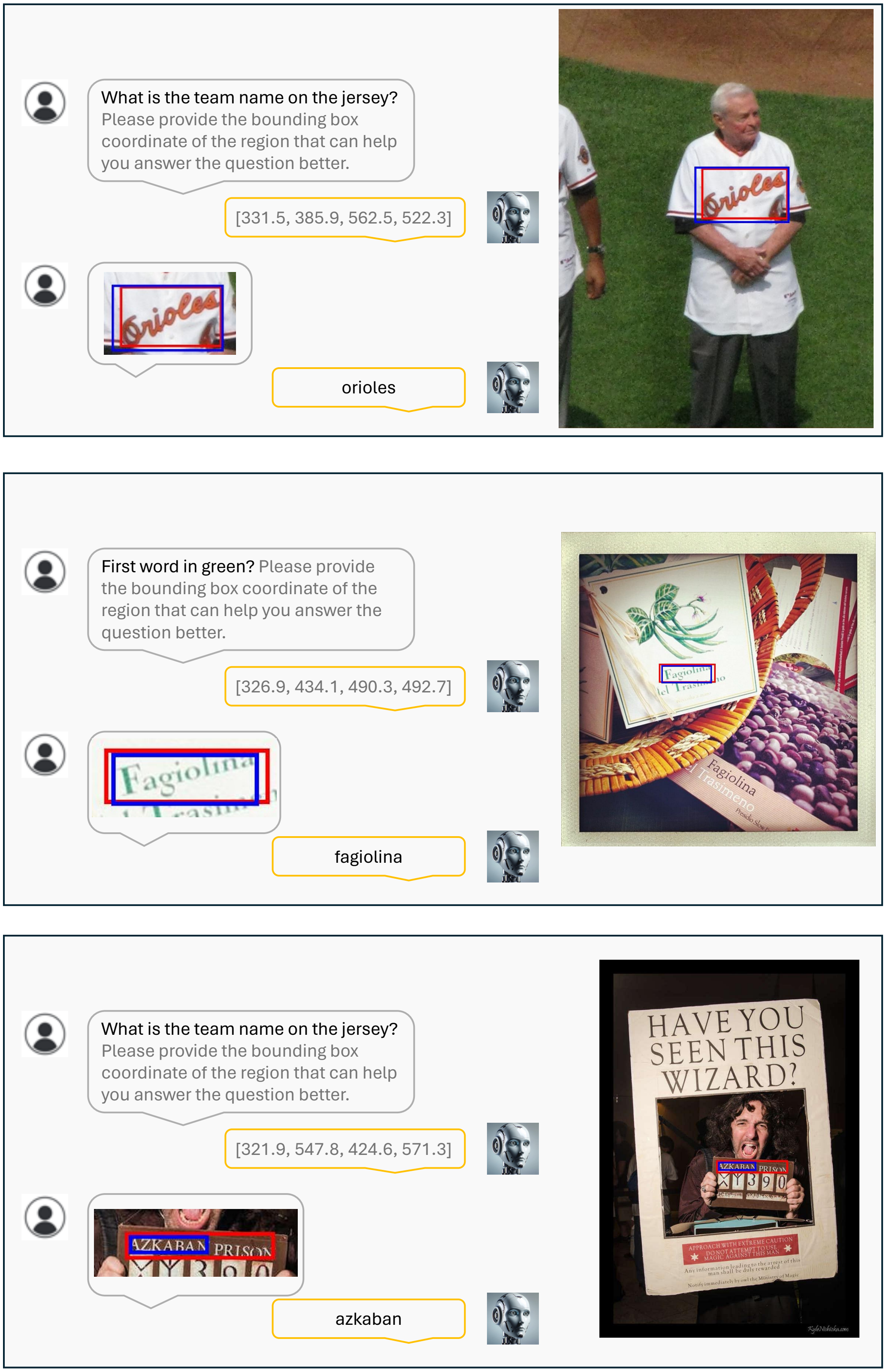
**Question**: Does the bird in the picture have iridescent underparts and white crown?
**Answer**: No
**CoT BBox**: [134, 69 430 263]

Figure 9: Examples in the visual CoT dataset, with corresponding question-answer annotations and visual CoT bboxes. The red bounding boxes in the images highlight the critical image regions that provide necessary and related information for answering the questions.

in the reference paper. The data, code, and model checkpoints are not intended to be used in clinical care or for any clinical decision making purposes.

**Primary Intended Use.** The primary intended use is to support AI researchers reproducing and building on top of this work. VisCoT and its associated models should be helpful for exploring various vision question answering (VQA) research questions.

**Out-of-Scope Use.** Any deployed use case of the model — commercial or otherwise — is out of scope. Although we evaluated the models using a broad set of publicly-available research benchmarks, the models and evaluations are intended for research use only and not intended for deployed use cases.



What is the team name on the jersey? Please provide the bounding box coordinate of the region that can help you answer the question better.

[331.5, 385.9, 562.5, 522.3]

orioles

First word in green? Please provide the bounding box coordinate of the region that can help you answer the question better.

[326.9, 434.1, 490.3, 492.7]

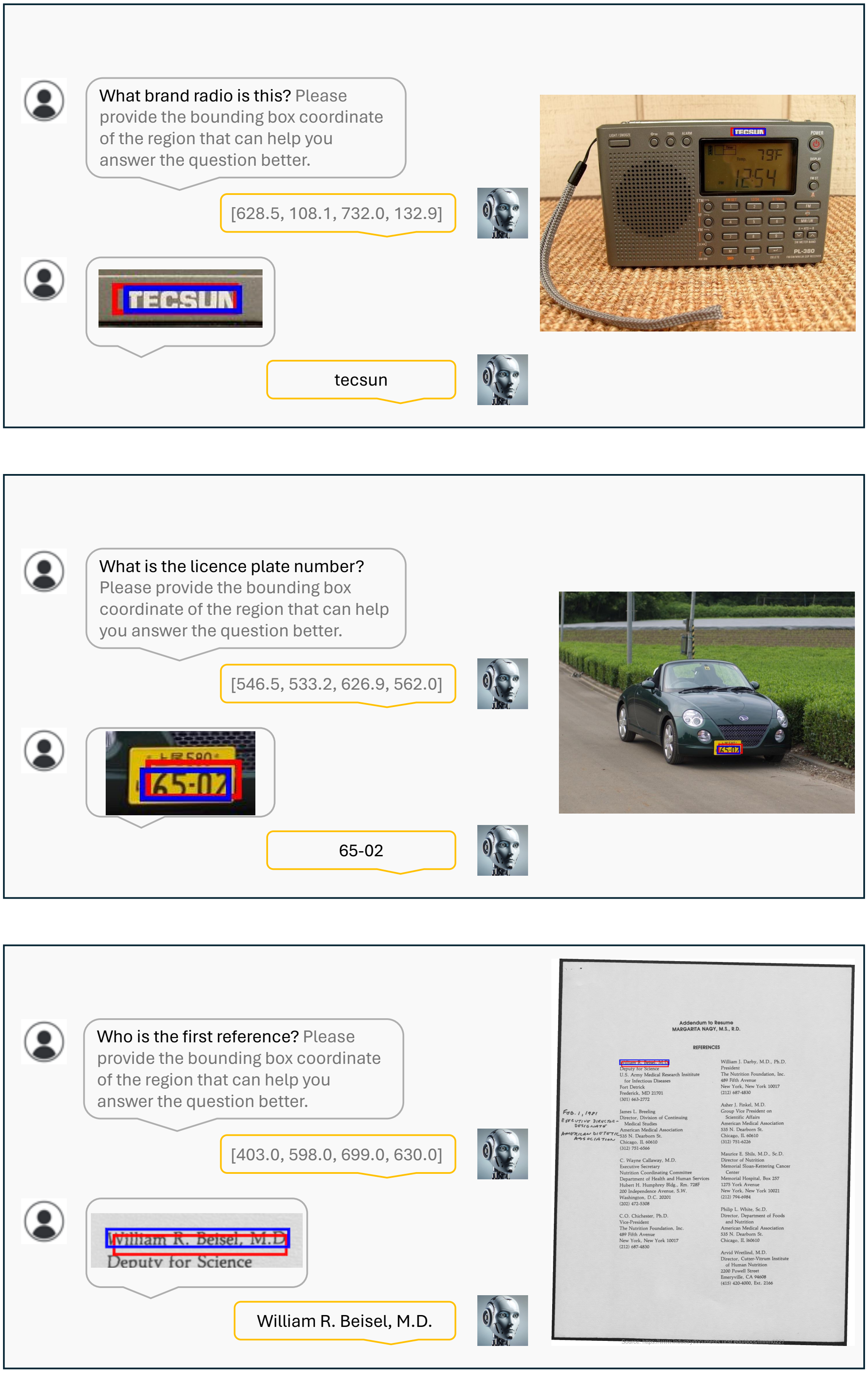
fagiolina

What is the team name on the jersey? Please provide the bounding box coordinate of the region that can help you answer the question better.

[321.9, 547.8, 424.6, 571.3]

azkaban

Figure 10: Visualization results of the VisCoT. Model-generated bounding boxes are shown in red, while ground truth (GT) bounding boxes are in blue.



What brand radio is this? Please provide the bounding box coordinate of the region that can help you answer the question better.

[628.5, 108.1, 732.0, 132.9]

tecsun

What is the licence plate number? Please provide the bounding box coordinate of the region that can help you answer the question better.

[546.5, 533.2, 626.9, 562.0]

65-02

Who is the first reference? Please provide the bounding box coordinate of the region that can help you answer the question better.

[403.0, 598.0, 699.0, 630.0]

William R. Beisel, M.D.

Addendum to Resume
MARGARITA NAGY, M.S., R.D.

REFERENCES

William R. Beisel, M.D.
Deputy for Science
U.S. Army Medical Research Insititute
for Infectious Diseases
Fort Detrick
Frederick, MD 21701
(301) 663-2772

Feb. 1, 1981
Executive Director-Designate
American Dietetic Association

James L. Breeling
Director, Division of Continuing
Medical Studies
American Medical Association
535 N. Dearborn St.
Chicago, IL 60610
(312) 751-6566

C. Wayne Callaway, M.D.
Executive Secretary
Nutrition Coordinating Committee
Department of Health and Human Services
Hubert H. Humphrey Bldg., Rm. 728F
200 Independence Avenue, S.W.
Washington, D.C. 20201
(202) 472-5308

C.O. Chichester, Ph.D.
Vice-President
The Nutrition Foundation, Inc.
489 Fifth Avenue
New York, New York 10017
(212) 687-4830

William J. Darby, M.D., Ph.D.
President
The Nutrition Foundation, Inc.
489 Fifth Avenue
New York, New York 10017
(212) 687-4830

Asher J. Finkel, M.D.
Group Vice President on
Scientific Affairs
American Medical Association
535 N. Dearborn St.
Chicago, IL 60610
(312) 751-6226

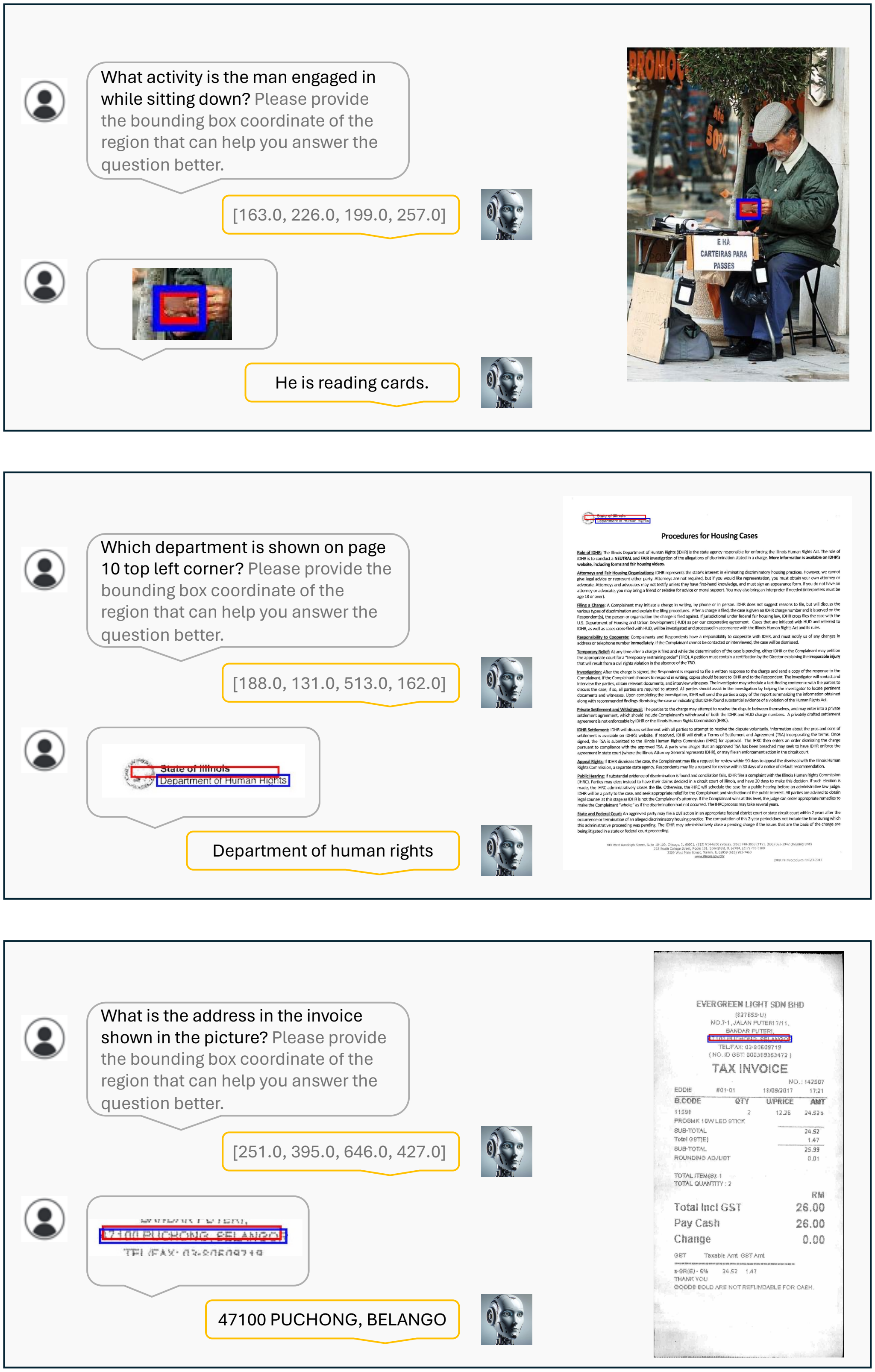
Maurice E. Shils, M.D., Sc.D.
Director of Nutrition
Memorial Sloan-Kettering Cancer
Center
Memorial Hospital, Box 257
1275 York Avenue
New York, New York 10021
(212) 794-6984

Philip L. White, Sc.D.
Director, Department of Foods
and Nutrition
American Medical Association
535 N. Dearborn St.
Chicago, IL l60610

Arvid Wretlind, M.D.
Director, Cutter-Vitrum Institute
of Human Nutrition
2200 Powell Street
Emeryville, CA 94608
(415) 420-4000, Ext. 2166

Source: https://www.industrydocuments.ucsf.edu/docs/fmnm0227

Figure 11: Visualization results of the VisCoT. Model-generated bounding boxes are shown in red, while ground truth (GT) bounding boxes are in blue.



What activity is the man engaged in while sitting down? Please provide the bounding box coordinate of the region that can help you answer the question better.

[163.0, 226.0, 199.0, 257.0]

He is reading cards.

Which department is shown on page 10 top left corner? Please provide the bounding box coordinate of the region that can help you answer the question better.

[188.0, 131.0, 513.0, 162.0]

Department of human rights

What is the address in the invoice shown in the picture? Please provide the bounding box coordinate of the region that can help you answer the question better.

[251.0, 395.0, 646.0, 427.0]

47100 PUCHONG, BELANGO

EVERGREEN LIGHT SDN BHD
(827859-U)
NO.7-1, JALAN PUTERI 7/11,
BANDAR PUTERI,
47100 PUCHONG, SELANGOR
TEL/FAX: 03-80609719
( NO. ID GST: 000389353472 )

TAX INVOICE

NO.: 142507
EDDIE #01-01 18/09/2017 17:21

| B.CODE | QTY | U/PRICE | AMT |
|---|---|---|---|
| 11598 | 2 | 12.26 | 24.52 s |
| PROBMK 10W LED STICK | | | |
| SUB-TOTAL | | | 24.52 |
| Total GST(E) | | | 1.47 |
| SUB-TOTAL | | | 25.99 |
| ROUNDING ADJUST | | | 0.01 |

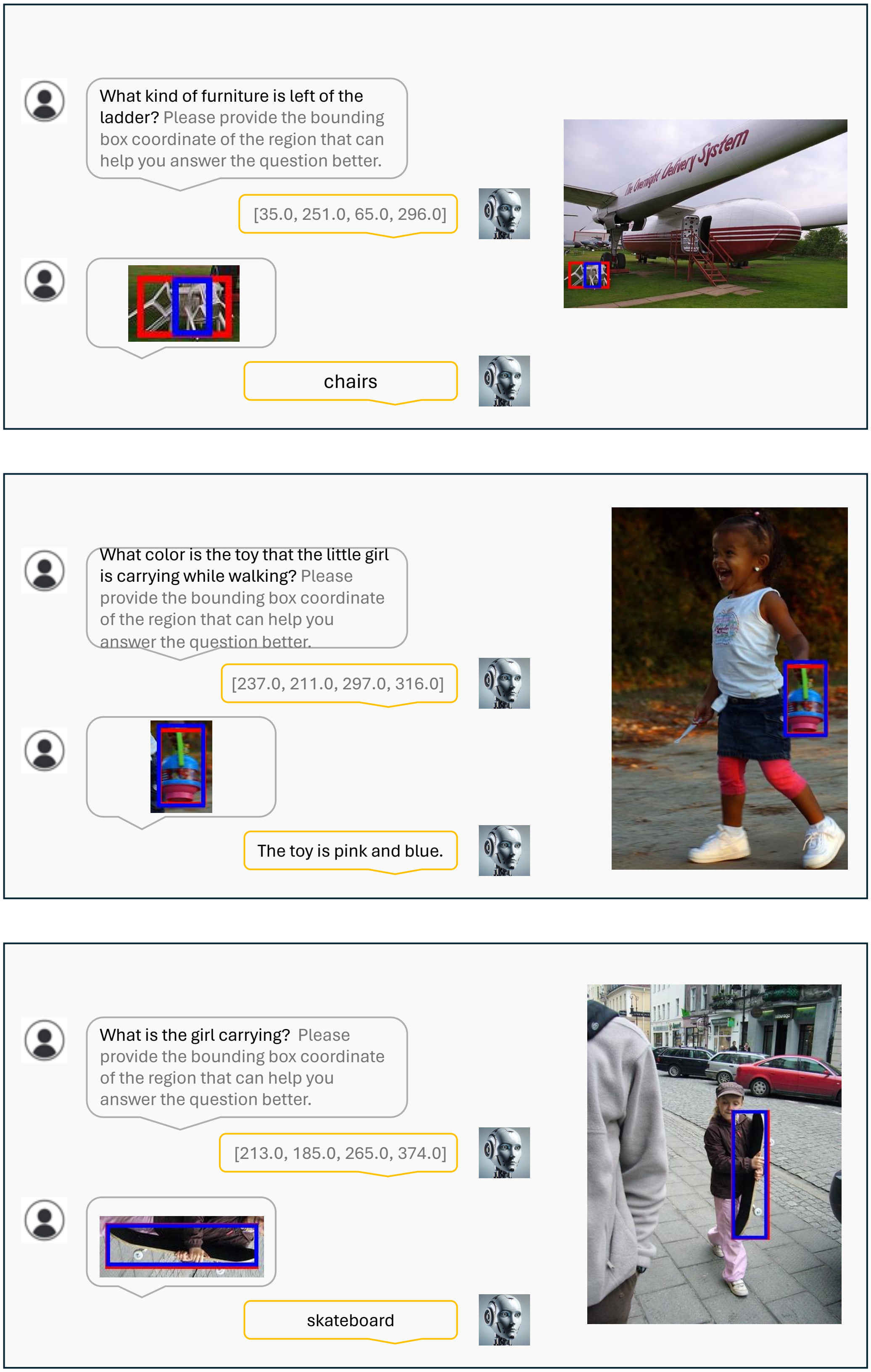
TOTAL ITEM(S): 1
TOTAL QUANTITY : 2

| | RM |
|---|---|
| Total Incl GST | 26.00 |
| Pay Cash | 26.00 |
| Change | 0.00 |

| GST | Taxable Amt | GST Amt |
|---|---|---|
| s-SR(E) - 6% | 24.52 | 1.47 |

THANK YOU
GOODS SOLD ARE NOT REFUNDABLE FOR CASH.

Figure 12: Visualization results of the VisCoT. Model-generated bounding boxes are shown in red, while ground truth (GT) bounding boxes are in blue.



What kind of furniture is left of the ladder? Please provide the bounding box coordinate of the region that can help you answer the question better.

[35.0, 251.0, 65.0, 296.0]

chairs

What color is the toy that the little girl is carrying while walking? Please provide the bounding box coordinate of the region that can help you answer the question better.

[237.0, 211.0, 297.0, 316.0]

The toy is pink and blue.

What is the girl carrying? Please provide the bounding box coordinate of the region that can help you answer the question better.

[213.0, 185.0, 265.0, 374.0]

skateboard

Figure 13: Visualization results of the VisCoT. Model-generated bounding boxes are shown in red, while ground truth (GT) bounding boxes are in blue.