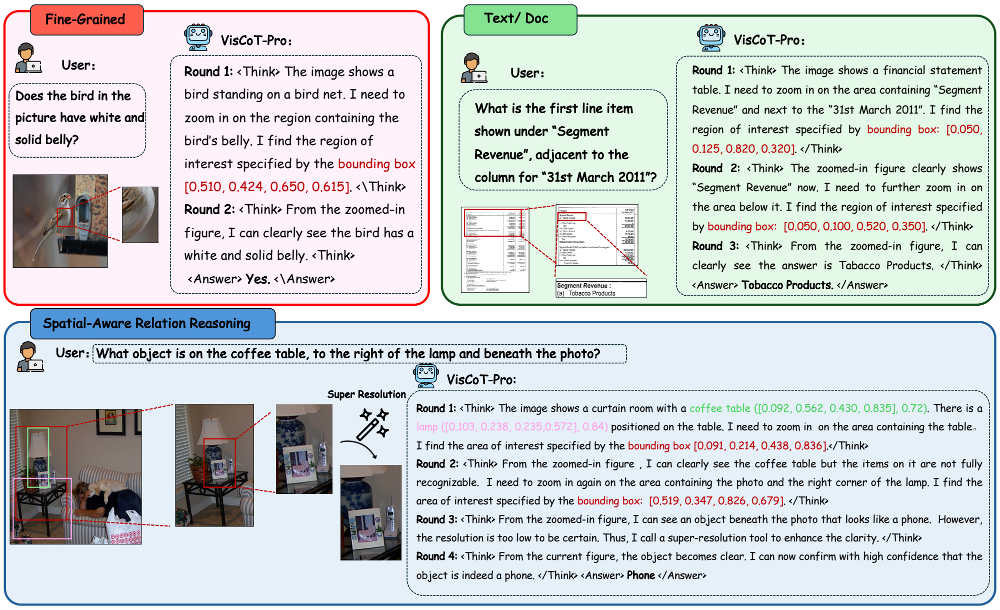
VisCoT-Pro: Advancing Visual Reasoning with a Large-scale Comprehensive Visual Chain-of-Thought Dataset

Anonymous authors
Paper under double-blind review



User: Does the bird in the picture have white and solid belly?

VisCoT-Pro:

Round 1: <Think> The image shows a bird standing on a bird net. I need to zoom in on the region containing the bird's belly. I find the region of interest specified by the bounding box [0.510, 0.424, 0.650, 0.615]. <\Think>

Round 2: <Think> From the zoomed-in figure, I can clearly see the bird has a white and solid belly. <Think>

<Answer> **Yes.** <\Answer>

User: What is the first line item shown under "Segment Revenue", adjacent to the column for "31st March 2011"?

VisCoT-Pro:

Round 1: <Think> The image shows a financial statement table. I need to zoom in on the area containing "Segment Revenue" and next to the "31st March 2011". I find the region of interest specified by bounding box: [0.050, 0.125, 0.820, 0.320]. </Think>

Round 2: <Think> The zoomed-in figure clearly shows "Segment Revenue" now. I need to further zoom in on the area below it. I find the region of interest specified by bounding box: [0.050, 0.100, 0.520, 0.350]. </Think>

Round 3: <Think> From the zoomed-in figure, I can clearly see the answer is Tabacco Products. </Think>

<Answer> **Tobacco Products.** </Answer>

User: What object is on the coffee table, to the right of the lamp and beneath the photo?

VisCoT-Pro:

Round 1: <Think> The image shows a curtain room with a coffee table ([0.092, 0.562, 0.430, 0.835], 0.72). There is a lamp ([0.103, 0.238, 0.235,0.572], 0.84) positioned on the table. I need to zoom in on the area containing the table。 I find the area of interest specified by the bounding box [0.091, 0.214, 0.438, 0.836].</Think>

Round 2: <Think> From the zoomed-in figure , I can clearly see the coffee table but the items on it are not fully recognizable. I need to zoom in again on the area containing the photo and the right corner of the lamp. I find the area of interest specified by the bounding box: [0.519, 0.347, 0.826, 0.679]. </Think>

Round 3: <Think> From the zoomed-in figure, I can see an object beneath the photo that looks like a phone. However, the resolution is too low to be certain. Thus, I call a super-resolution tool to enhance the clarity. </Think>

Round 4: <Think> From the current figure, the object becomes clear. I can now confirm with high confidence that the object is indeed a phone. </Think> <Answer> **Phone** </Answer>

Figure 1: An MLLM trained on our **VisCoT-Pro** benchmark emulates a human-like visual reasoning process to solve a complex query. Instead of naively processing the full image, the model learns a dynamic global-to-local workflow: it first assesses the entire scene, then iteratively identifies and zooms in on relevant regions to gather fine-grained evidence. This multi-step, spatially-aware visual Chain-of-Thought enables the model to ground its reasoning in specific details and solve complex spatial problems that challenge conventional models.

Abstract

Chain-of-Thought (CoT) prompting has emerged as a powerful technique for eliciting complex reasoning in Large Language Models (LLMs). However, its potential within multimodal large language models (MLLMs) remains largely unrealized. A primary bottleneck is the lack of suitable datasets and benchmarks: existing visual-CoT resources are often limited in scale and diversity, or fail to capture the human-like, spatially-aware reasoning required for genuine visual understanding. To address these limitations, we introduce **VisCoT-Pro**, a large-scale and comprehensive benchmark designed to advance visual CoT reasoning. Our benchmark comprises two key components: (1) the main **VisCoT-Pro** dataset with 506k examples covering four domains, featuring multi-round, human-like step-by-step supervision that is substantially larger and more detailed than prior resources, and (2) **VisCoT-Pro-Max**, a 165k subset with richer step rationales and 3D grounding via depth-informed annotations, produced with stronger GPT-4.1-series guidance. We conduct extensive experiments on the state-of-the-art Qwen2.5-VL model. Training on VisCoT-Pro not only yields substantial improvements in the model's intrinsic step-by-step visual reasoning capabilities but also demonstrates remarkable generalization, significantly boosting performance on existing academic benchmarks. This highlights our dataset's ability to equip VLMs with robust, transferable reasoning skills, enabling them to better understand and think about the visual world. We release VisCoT-Pro as a foundational resource, providing the community with both a high-quality training corpus and a reliable benchmark to catalyze future research in visual CoT.

1 INTRODUCTION

Recent Multimodal Large Language Models (MLLMs) have achieved remarkable progress in practical visual understanding, largely by pairing high-capacity language models with powerful visual encoders through sophisticated alignment pipelines (OpenAI, 2023; Zhu et al., 2023; Yin et al., 2023; Bai et al., 2023). Foundational models such as LLaVA (Liu et al., 2023b) and its successors, including InternVL (Chen et al., 2024a), Qwen2.5-VL (Yang et al., 2024), and MiniCPM-V (Yao et al., 2024), demonstrate state-of-the-art performance across a diverse spectrum of tasks. They excel at visual question answering (Li et al., 2024a), fine-grained visual grounding (Peng et al., 2024), and optical character recognition (Zhang et al., 2023), establishing them as powerful and versatile tools for real-world knowledge access, assistance, and creative work.

However, while MLLM architectures have advanced, their underlying reasoning paradigm often remains rudimentary. In the unimodal text domain, complex reasoning has been revolutionized by techniques like Chain-of-Thought (CoT) prompting, which trains models to articulate step-by-step rationales before arriving at an answer (Wei et al., 2022). This deliberative process has unlocked dramatic gains in arithmetic, commonsense, and symbolic reasoning by making the model's inferential pathway an explicit object of supervision. In stark contrast, the multimodal domain has yet to experience a similar paradigm shift (Liu et al., 2024b; Shao et al., 2024; Man et al., 2025). Most MLLMs are still optimized via a direct input-to-answer format, which provides no supervision on intermediate cognitive steps. This approach encourages models to learn superficial shortcuts, fosters an over-reliance on spurious linguistic priors, and ultimately limits their potential, leaving them prone to hallucination when faced with complex, multi-step visual queries (Ke et al., 2025).

We posit that this stagnation is rooted in a fundamental misalignment between the dominant training paradigms and the process of human cognition. Humans approach detail-oriented visual problems by first scanning the global scene to form a coarse hypothesis, then iteratively narrowing their attention to candidate regions for closer inspection. This global-to-local process often involves targeted *manipulations*—such as cropping or zooming—to inspect fine details and inter-object spatial relations until the query is resolved (Qi et al., 2024). While most VLMs acquire the necessary intrinsic skills for this process (e.g., object grounding, OCR) during pre-training, teaching them to execute this human-like cognitive workflow is stymied by a foundational bottleneck: the inadequacy of existing visual reasoning benchmarks. These resources suffer from three critical shortcomings: **(1)** they offer insufficient scale and domain coverage, limiting the ability to learn generalizable reasoning patterns (Zhang et al., 2025b; Man et al., 2025; Sarch et al., 2025); **(2)** they provide inadequate supervision for multi-round reasoning, often reducing the complex process to a single static crop without the necessary stepwise rationales (Li et al., 2024b; Ye et al., 2024; Qi et al., 2024); and **(3)** their grounding and QA annotations remain overwhelmingly two-dimensional, neglecting the depth-aware spatial reasoning essential for understanding the physical world (Shao et al., 2024; Sarch et al., 2025; Wu et al., 2025). This scarcity of data that faithfully represents the human cognitive workflow has become a significant barrier, impeding the development of MLLMs that can move beyond simple perception towards genuine visual cognition.

To address these fundamental challenges, we introduce **VisCoT-Pro**, a large-scale, comprehensive benchmark designed to instill human-like, spatially-aware reasoning in MLLMs. Our benchmark is composed of a broad 506k-example primary dataset and a high-quality 165k-example subset, **VisCoT-Pro-Max**, which features richer annotations and more complex reasoning. Our unified data construction process begins by enriching each image with pseudo-3D signals, combining monocular depth estimates and semantic segmentation with existing 2D ground-truth boxes. Using these 3D-aware scene representations, we then prompt powerful MLLMs to generate a multi-round visual CoT that emulates the human reasoning workflow. For the subset, we leverage stronger GPT-4.1-series guidance to produce more detailed rationales. Across the entire benchmark, each step provides a brief scene description, identifies a region of interest for closer inspection, and concludes with a justifying rationale. This fine-grained supervision is designed to discourage shortcut learning, promote a global-to-local "zoom-and-verify" behavior, and enhance model generalization. Crucially, we introduce explicit 3D grounding—a feature most prominent in our subset, by training the model to exploit ordinal depth and spatial relations directly from a single image, thereby strengthening its depth-aware reasoning capabilities.

In summary, this paper makes the following key contributions:

- We construct and release **VisCoT-Pro**, a new large-scale dataset of 506k examples across four diverse domains. Each example features multi-round, human-like step-by-step supervision, making it substantially larger and more detailed than prior resources.
- We curate **VisCoT-Pro-Max**, a 165k high-quality subset produced with stronger GPT-4.1-series guidance. This subset features richer step rationales and explicit 3D grounding via depth-informed annotations, specifically targeting advanced spatial reasoning.
- We design a novel training and inference pipeline that effectively leverages our annotations to instill sophisticated spatial CoT reasoning in MLLMs. Using this pipeline, our models achieve state-of-the-art results across a range of metrics, demonstrating improved fidelity and robustness in multi-domain visual reasoning.

2 Related Works

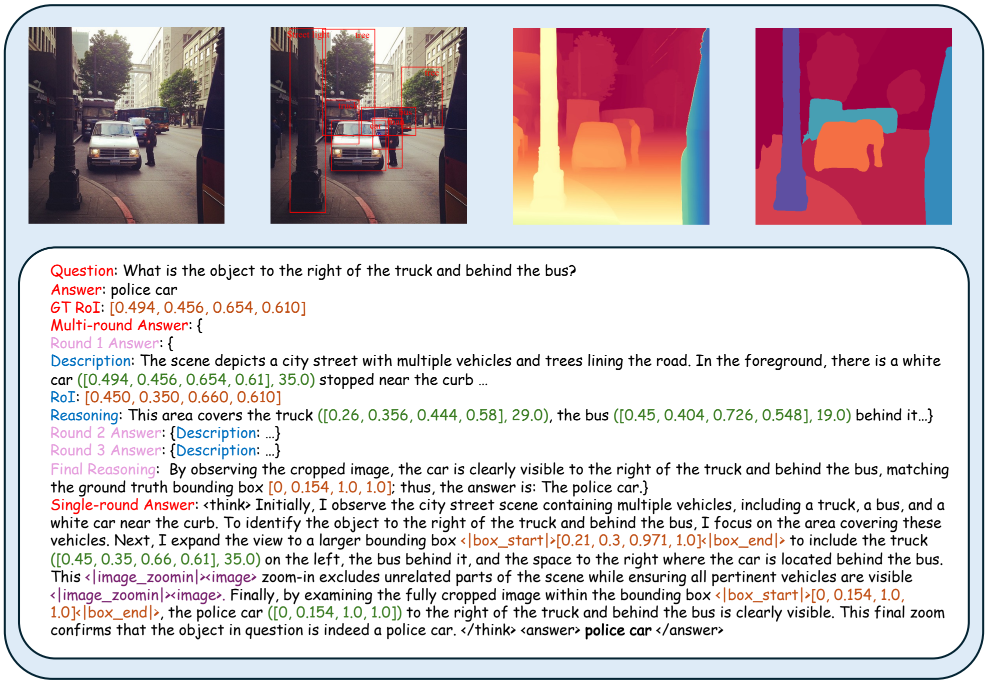
2.1 Multimodal Large Language Models

The development of powerful Multimodal Large Language Models (MLLMs) is a central pursuit in multimodal learning (Liu et al., 2024a). Modern systems typically pair a visual encoder (e.g., a ViT (Dosovitskiy et al., 2020)) with a projection module (e.g., an MLP or Q-Former (Li et al., 2023a)) to map image features into the language embedding space of an LLM for autoregressive decoding (Yang et al., 2024). Leading model families—including LLaVA-OneVision (Li et al., 2024a), InternVL (Chen et al., 2024a), Qwen-VL (Bai et al., 2023), LLaVA-UHD (Xu et al., 2024), InternVL-3 (Zhu et al., 2025), Qwen-2.5-VL (Bai et al., 2025), and Gemini-2.5-Pro (Comanici & Team, 2025)—demonstrate increasingly sophisticated multimodal reasoning and high-resolution perception. Nevertheless, a fundamental challenge persists: when targets are diminutive or spatial relationships are intricate, these models can still falter. The prevailing paradigm of processing an entire high-resolution image naively makes it difficult to allocate computation effectively, encouraging models to expend resources on irrelevant regions or rely on spurious textual priors rather than performing deliberate spatial verification. This reveals a critical gap between a model's raw perceptual capacity and its ability to apply that capacity strategically, motivating a shift away from monolithic image processing towards more efficient and human-like reasoning workflows, which our work aims to facilitate.

2.2 Multimodal Reasoning

Inspired by the success of Chain-of-Thought (CoT) in improving the interpretability and reliability of LLMs (Wei et al., 2022; Zhang et al., 2022; Gao et al., 2025), several lines of research have attempted to instill similar deliberative reasoning in Multimodal Large Language Models (MLLMs). Model-centric approaches aim to elicit this behavior through methods like in-context learning (Zhang et al., 2024; Mitra et al., 2024; Gupta & Kembhavi, 2023; Gao et al., 2024; Chen et al., 2024b) or targeted supervised fine-tuning and reinforcement learning (Liu et al., 2025b; Yang et al., 2025; Zhou et al., 2025; Zhang et al., 2025a; Liu et al., 2025a). A more fundamental, data-centric line of work seeks to supervise the reasoning process explicitly, but existing efforts have notable limitations. A significant body of work simplifies the complex human workflow into a single intermediate localization step. Datasets like $V^{\star}$, CogCom, VPD, Visual CoT, DualFocus, and Chain-of-Spot fall into this category, but this under-specification risks encouraging models to learn shortcuts and spurious correlations (Wu & Xie, 2024; Qi et al., 2024; Hu et al., 2024; Shao et al., 2024; Cao et al., 2024; Liu et al., 2024b). While other datasets do generate multi-round chains, which is a step in the right direction (Zhang et al., 2025b; Dong et al., 2025; Man et al., 2025), they are often constrained by limited scale and narrow domain coverage. Furthermore, nearly all prior data-centric efforts focus exclusively on 2D grounding, neglecting the explicit depth reasoning essential for understanding physical spatial relations (Shao et al., 2024; Sarch et al., 2025; Wu et al., 2025). Our work directly addresses these gaps by providing a large-scale, multi-domain benchmark with explicit supervision for multi-round, 3D-aware spatial reasoning.

3 VISCOT-PRO



Question: What is the object to the right of the truck and behind the bus?
Answer: police car
GT RoI: [0.494, 0.456, 0.654, 0.610]
Multi-round Answer: {
Round 1 Answer: {
Description: The scene depicts a city street with multiple vehicles and trees lining the road. In the foreground, there is a white car ([0.494, 0.456, 0.654, 0.61], 35.0) stopped near the curb …
RoI: [0.450, 0.350, 0.660, 0.610]
Reasoning: This area covers the truck ([0.26, 0.356, 0.444, 0.58], 29.0), the bus ([0.45, 0.404, 0.726, 0.548], 19.0) behind it…}
Round 2 Answer: {Description: …}
Round 3 Answer: {Description: …}
Final Reasoning: By observing the cropped image, the car is clearly visible to the right of the truck and behind the bus, matching the ground truth bounding box [0, 0.154, 1.0, 1.0]; thus, the answer is: The police car.}
Single-round Answer: <think> Initially, I observe the city street scene containing multiple vehicles, including a truck, a bus, and a white car near the curb. To identify the object to the right of the truck and behind the bus, I focus on the area covering these vehicles. Next, I expand the view to a larger bounding box <|box_start|>[0.21, 0.3, 0.971, 1.0]<|box_end|> to include the truck ([0.45, 0.35, 0.66, 0.61], 35.0) on the left, the bus behind it, and the space to the right where the car is located behind the bus. This <|image_zoomin|><image> zoom-in excludes unrelated parts of the scene while ensuring all pertinent vehicles are visible <|image_zoomin|><image>. Finally, by examining the fully cropped image within the bounding box <|box_start|>[0, 0.154, 1.0, 1.0]<|box_end|>, the police car ([0, 0.154, 1.0, 1.0]) to the right of the truck and behind the bus is clearly visible. This final zoom confirms that the object in question is indeed a police car. </think> <answer> **police car** </answer>

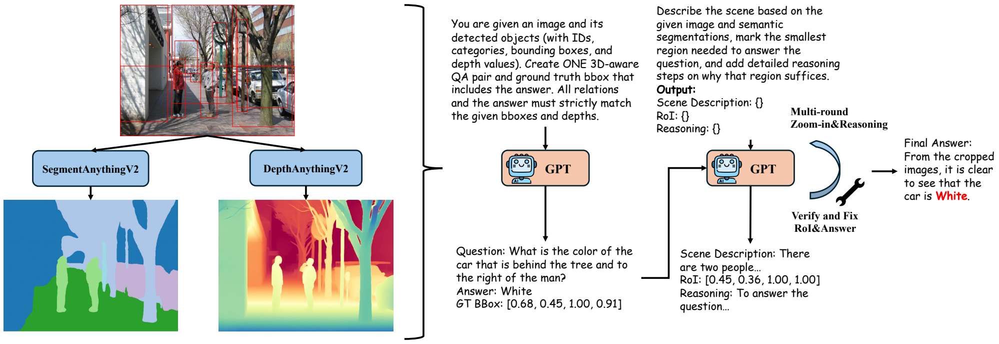
Figure 2: Overview of one data sample. For each image–question pair, we provide a gold region of interest (bounding box) and a compact multi-round visual chain-of-thought: each round offers a scene sketch, an optional zoom to a predicted RoI, and a brief rationale. When available, depth cues indicate ordinal ordering. The annotations are concise and process-oriented, enabling spatially grounded reasoning on fine details and complex relations.

As detailed in section 1, existing visual reasoning datasets suffer from three persistent limitations—insufficient scale and domain coverage, lack of multi-round stepwise supervision, and minimal depth-aware grounding—necessitating a resource that trains MLLMs to follow human-like visual reasoning. We address these gaps by curating **VisCoT-Pro**—a large, spatially aware visual chain-of-thought (CoT) corpus that explicitly supervises the *process* of visual reasoning rather than only final answers. As illustrated in fig. 2, each sample consists of a question, an answer, and a *multi-round* CoT that mirrors human global-to-local problem solving: every round provides (i) a brief *scene description*, (ii) a predicted *region of interest* (RoI, via bounding box) when zoom is warranted, and (iii) a short *rationale* explaining why that RoI suffices. Beyond 2D cues, we attach pseudo-3D signals—monocular depth and semantic segmentation—so that the chain can reference ordinal depth and part/region evidence when needed. This unified annotation format encourages models to localize, zoom, and verify iteratively, reducing shortcut learning and promoting depth-aware spatial reasoning.

To ensure broad coverage while keeping the focus on process supervision, **VisCoT-Pro** spans **four domains**—text/doc understanding, fine-grained recognition, general VQA, and spatial-aware relational reasoning—continuing and extending prior category choices (table 1). In total, the primary set contains **506k** examples, and we further release a **165k** high-fidelity subset, **VisCoT-Pro-Max**, with richer rationales and stronger depth-informed grounding. Together, these resources offer detailed, stepwise supervision (see fig. 2) designed to cultivate global-to-local "zoom–and–verify" behaviors and robust reasoning over small objects and complex 2D/3D relations.

3.1 DATASET GENERATION

VisCoT-Pro



You are given an image and its detected objects (with IDs, categories, bounding boxes, and depth values). Create ONE 3D-aware QA pair and ground truth bbox that includes the answer. All relations and the answer must strictly match the given bboxes and depths.

Question: What is the color of the car that is behind the tree and to the right of the man?
Answer: White
GT BBox: [0.68, 0.45, 1.00, 0.91]

Describe the scene based on the given image and semantic segmentations, mark the smallest region needed to answer the question, and add detailed reasoning steps on why that region suffices.
Output:
Scene Description: {}
RoI: {}
Reasoning: {}

Scene Description: There are two people...
RoI: [0.45, 0.36, 1.00, 1.00]
Reasoning: To answer the question...

Final Answer: From the cropped images, it is clear to see that the car is White.

Figure 3: Pipeline for **VisCoT-Pro** and **VisCoT-Pro-Max** data generation and supervision. Given an input image, we derive semantic segments and monocular depth to form an object list with categories, bounding boxes, and ordinal depth; a generator then produces a 3D-aware QA pair and target box. A second stage emits a compact, multi-round visual CoT—scene sketch, predicted RoI, and rationale—while iteratively zooming and verifying (with RoI/answer fix) until the final answer and finalized annotations are obtained.

Building on the Visual-CoT seed, we expand each image–question–answer triple with *process-level* supervision. For every example, the model (GPT-4.1-Nano(Achiam et al., 2023)) is prompted to produce a concise scene description, a normalized region of interest (RoI; $[x_1, y_1, x_2, y_2] \in [0, 1]^4$, and a brief rationale. We enforce coverage by adjusting the RoI to tightly contain the ground-truth box and iteratively refining via global-to-local zoom. The refinement terminates when the RoI area is no more than twice the GT area or when a small round budget (≤ 3) is reached. This yields multi-round chains that are compact yet faithful, providing stepwise evidence aligned with the final answer while discouraging shortcut learning.

VisCoT-Pro-Max

As illustrated in fig. 3, **VisCoT-Pro-Max** augments the above pipeline with explicit spatial priors to elicit depth-aware reasoning. We first derive pseudo-3D cues per image—monocular depth and semantic segmentation (object IDs, categories, pixel boxes, and ordinal depths)—and feed these structured signals, together with the image, to a stronger generator (GPT-4.1-Mini (Achiam et al., 2023)). The model is instructed to create 3D-aware questions whose relations jointly involve 2D layout (*left of*, *above*) and depth (*in front of*, *behind*), outputting a consistent GT box for the target. We then apply the same verify-and-fix routine with multi-round zoom (round$\leq$ 4) to obtain concise descriptions, RoIs, and rationales at each step. In addition to these *multi-round* traces, we provide a *single-round* distilled variant that compacts the multi-step chain into one rationale and a final RoI. As shown in *Single-round Answer* in fig. 2, our **VisCoT-Pro-Max** also enables single-pass answering while preserving explicit process supervision. The result is a depth-informed visual CoT corpus tailored for small-object queries and complex 2D/3D relations. More details of the prompt design and algorithms are provided in appendix B.

3.2 Dataset Analysis

We visualize corpus statistics in fig. 4 and summarize coverage in table 1. RoIs skew heavily toward small regions—most notably in text/doc sources—showing that answer-critical evidence occupies only a tiny fraction of the image; on average, the annotated box is compact (≈ 247.8 pixels, 13.2% of the image), reinforcing the need to localize, zoom, and verify rather than process full frames blindly. Most examples resolve in 2–3 rounds, with harder spatial/3D cases extending to 4; meanwhile, the *average per-round response length* remains concise and grows with task difficulty, which provides much more detailed reasoning steps than existing datasets (Shao et al., 2024; Zhang et al., 2025b; Qi et al., 2024), indicating efficient yet expressive supervision. Compared with prior datasets, **VisCoT-Pro** offers markedly larger scale (506k), richer and explicit multi-round reasoning, broader domain diversity, and a substantial depth-aware subset (**VisCoT-Pro-Max**, 165k) that equips models for spatially grounded 2D/3D reasoning.

Table 1: **Overview of the VisCoT-Pro dataset.** It spans four distinct domains and aggregates diverse source datasets, providing broad coverage of visual data styles.

Domain	Source Dataset	Train/Val Size		GPT Model	Dataset Description
Text/Doc	TextVQA (Singh et al., 2019)	16k	526	4.1-nano	Images with text
	TextCaps (Sidorov et al., 2020)	32k	846	4.1-nano	Images with text
	DocVQA (Mathew et al., 2021)	50k	846	4.1-nano	Doc Images
	DUDE (Van Landeghem et al., 2023)	11k	559	4.1-nano	Doc Images
	SROIE (Huang et al., 2019)	2k	685	4.1-nano	Invoice Images
Fine-Grained Understanding	Birds-200-2011 (Wah et al., 2011)	10k	491	4.1-nano	Images of birds
General VQA	Flickr30k (Plummer et al., 2015)	126k	1455	4.1-nano	Images
	Visual7W (Zhu et al., 2016)	30k	994	4.1-nano	Images
Spatial Relatio Reasoning	VSR (Liu et al., 2023a)	3k	404	4.1-nano	Images
	GQA (Hudson & Manning, 2019)	165k	978	4.1-mini	Images **(with spatial-aware detailed reasoning steps)**
	Open images (Kuznetsova et al., 2020)	43k	944	4.1-nano	Images

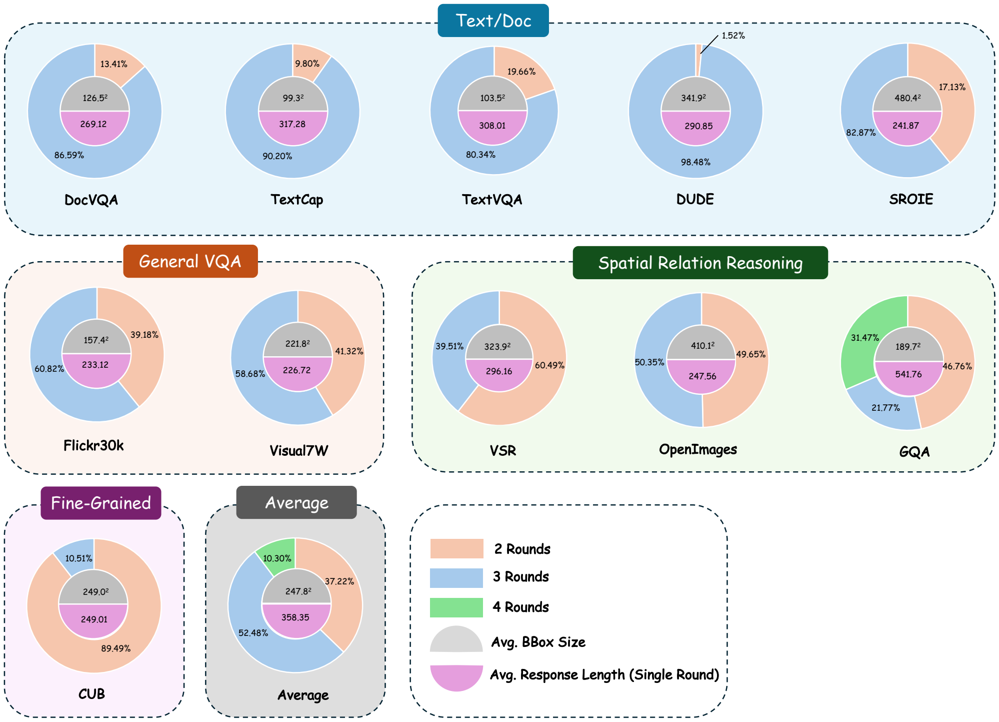


Figure 4: Statistics of the proposed **VisCoT-Pro** dataset. We visualize the CoT bbox distribution, average bbox size, and average length of response in each round for each source dataset.

4 Enhancing MLLMs with CoT Reasoning Capabilities

Formulation and Training

Given an image I and textual query Q, our model generates a multi-step reasoning process $Y = (a_0, a_1, \dots, a_T)$ to derive the final answer (fig. 5). At step t, the action $a_t = (r_t, b_{t+1})$ consists of a textual reasoning snippet r_t and a bounding box b_{t+1} for the next region of interest. Generation is conditioned on prior actions and their visual inputs: the visual context at step t is obtained by cropping I with b_t from the previous step, and we denote features by $\mathcal{V}(\text{crop}(I, b_t))$. The process is initialized with b_0 as the full image. The model auto-regressively outputs the tokens of each a_t—both the rationale and the serialized box coordinates—based on the initial query and the full history of preceding visual and textual data:

$$a_t \sim P_\theta(\cdot|Q, a_0, \dots, a_{t-1}, \mathcal{V}(\text{crop}(I, b_0)), \dots, \mathcal{V}(\text{crop}(I, b_t))). \tag{1}$$

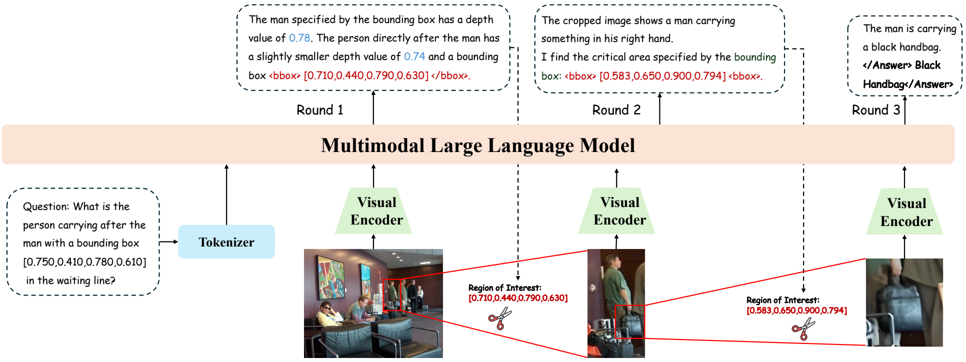

The man specified by the bounding box has a depth value of 0.78. The person directly after the man has a slightly smaller depth value of 0.74 and a bounding box <bbox> [0.710,0.440,0.790,0.630] </bbox>.
The cropped image shows a man carrying something in his right hand.
I find the critical area specified by the bounding box: <bbox> [0.583,0.650,0.900,0.794] <bbox>.
The man is carrying a black handbag. </Answer> Black Handbag</Answer>
Round 1
Round 2
Round 3
Multimodal Large Language Model
Question: What is the person carrying after the man with a bounding box [0.750,0.410,0.780,0.610] in the waiting line?
Tokenizer
Visual Encoder
Visual Encoder
Visual Encoder
Region of Interest: [0.710,0.440,0.790,0.630]
Region of Interest: [0.583,0.650,0.900,0.794]

Figure 5: **Overview of VisCoT-Pro paradigm**. The model iteratively processes the query by first generating a textual rationale and a bounding box for the next region of interest. It then crops the original image to this region, extracts new visual features, and appends them to the context to inform the next reasoning step, creating a zoom-and-verify sequence.

We fine-tune this model on **VisCoT-Pro** via Supervised Fine-Tuning (SFT) using Qwen2.5-VL-7B (Bai et al., 2025) as the base. During fine-tuning, we apply LoRA (Hu et al., 2022) for efficient adaptation. The objective maximizes the likelihood of the ground-truth sequence Y given (I, Q) in a standard autoregressive manner, predicting the next token at each step. Concretely, we minimize the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{(I,Q)\in\mathcal{D}} \sum_{t=1}^{|Y|} \log P_\theta(Y_t | I, Q, Y_{<t}), \tag{2}$$

where $\mathcal{D}$ is the training set, θ the trainable parameters, and Y_t the t-th token of Y. The sequence Y is formed by serializing the multi-step CoT output, converting each b_i into discrete tokens, so the model is trained end-to-end to produce both textual reasoning steps and precise coordinates for focusing visual attention.

5 EXPERIMENT

Training Details. We apply **VisCoT-Pro** and **VisCoT-Pro-Max** to the Qwen2.5-VL-7B (Bai et al., 2025) model. We train 2 epochs for the baseline model using **VisCoT-Pro** with the **VisCoT-Pro-Max** excluded and one additional epoch with all dataset. The learning rate is $2e^{-5}$ for the LLM backbone and projector, $2e^{-6}$ for ViT. The batch size is 12 per device. More details can be found in appendix A.

Benckmarks. We follow the Visual-CoT protocol (Shao et al., 2024) and use its 11 source test sets (fig. 4). As motivated in section 1, we group tasks into four domains—text/doc understanding, fine-grained recognition, general VQA, and spatial relational reasoning—to stress both perception and multi-step spatial inference. To assess zero-shot generalization beyond our domains, we further report results on MME (Fu et al., 2023), a comprehensive LVLM evaluation suite, and on V*Bench (Wu et al., 2024), which targets high-resolution detail following and visual search. For automatic scoring, we follow prior MLLM work (Li et al., 2023b; Luo et al., 2023; Shao et al., 2024) and use an LLM-based judge to assign a scalar score in [0,1] per example; higher is better. The judging prompt and calibration details are provided in the appendix.

5.1 COMPARISON WITH STATE-OF-THE-ART MLLMS

Comparison on Visual-CoT-Benchmark. Against strong open-source baselines, our models deliver the best overall performance as shown in table 2. **VisCoT-Pro-Max**–7B attains the highest average (0.781), edging Qwen-VL-2.5-7B (0.770) and clearly surpassing InternVL-2.5-8B (0.738), LLaVA-NeXT-8B (0.705), and VisCoT-7B (0.580). Relative to the original Visual-CoT system—which supervises only a single RoI step—our multi-round, spatially grounded CoT yields large gains across most datasets (*e.g.*, +0.23 average vs VisCoT-7B), supporting the view that stepwise localization and verification reduce shortcutting. By domain, we are strongest on fine-grained recognition (Birds-200-2011: 0.809) and general VQA (Flickr30k: 0.779; Visual7W: 0.711), and competitive on spatial

Table 2: Comparison with state-of-the-art MLLMs on Visual CoT benchmark.

MLLM	Doc/Text					Fine-grained	General VQA		Spatial Relation Reasoning			Average
	DocVQA	TextCaps	TextVQA	DUDE	SROIE	Birds-200-2011	Flickr30k	Visual7W	GQA	Open images	VSR	
VisCoT-7B (Shao et al., 2024)	0.476	0.675	0.775	0.386	0.470	0.559	0.668	0.558	0.631	**0.822**	0.614	0.580
LLaVA-NeXT-8B (Chen & Xing, 2024)	0.728	0.775	0.850	0.581	0.666	0.715	0.755	0.703	**0.736**	0.559	0.647	0.705
InternVL-2.5-8B (Zhu et al., 2025)	0.846	0.829	0.907	0.716	0.907	0.747	0.713	0.681	0.689	0.502	**0.737**	0.738
CoF-SFT-7B (Zhang et al., 2025b)	0.955	0.867	0.934	0.813	0.979	0.641	0.606	0.686	0.674	0.503	0.657	0.748
Qwen-VL-2.5-7B (Bai et al., 2025)	**0.964**	**0.871**	**0.952**	**0.817**	**0.987**	0.681	0.772	0.690	0.651	0.498	0.705	0.770
VisCoT-Pro-7B	0.889	0.847	0.905	0.745	0.914	0.798	0.765	**0.711**	0.642	0.723	0.645	0.779
VisCoT-Pro-Max-7B	0.887	0.815	0.901	0.733	0.924	**0.809**	**0.779**	0.707	0.669	0.732	0.653	**0.781**

relation reasoning (notably OpenImages: 0.732). On doc/text OCR, Qwen-VL-2.5-7B remains at the top, while our scores are slightly lower, which is consistent with extremely small RoIs where multi-round zoom may incur minor drift. Nevertheless, our cross-domain balance lifts the overall average beyond all baselines, indicating stronger comprehensive capability. Although a few single datasets favor specialized systems (e.g., Qwen on OCR, InternVL on VSR), our models exhibit no pronounced weaknesses and deliver the best overall averages, validating that large-scale, multi-round, and depth-aware visual CoT supervision improves both in-domain effectiveness and zero-shot robustness.

Zero-shot generalization. On external suites, **VisCoT-Pro**–7B achieves the best V* score (0.603) among compared methods and competitive MME (0.695) as shown in table 3. **VisCoT-Pro-Max**–7B further improves MME to 0.751 (still below InternVL-2.5-8B at 0.848) while remaining comparable on V* (0.590). These results—obtained without exposure to those benchmarks—suggest that scaling diverse, spatially grounded CoT (especially with depth-aware traces) enhances transfer to fine-detail following and broad multimodal skills.

5.2 Ablation Study

Table 5 disentangles the effects of our two datasets and an inference-time super-resolution (SR) heuristic. Relative to the vanilla Qwen baseline (Avg 0.770, strong on Doc/Text 0.920), training on **VisCoT-Pro** yields clear gains on General VQA (+0.005) and especially Relation (+0.080) and Fine-grained (+0.117), lifting the average to 0.779. Adding **VisCoT-Pro-Max** further improves General VQA (0.750), Relation (0.693), and Fine-grained (0.809), reaching the best overall average (0.781). Applying SR with Real-ESRGAN (Wang et al., 2021) to very small RoIs (128×128) offers negligible net benefit (Avg 0.778) and slightly hurts Doc/Text. We attribute this to: (i) high redundancy in text pixels—SR adds little beyond legible character shapes; (ii) no SR seen during training, limiting exploitation of added details; and (iii) SR triggers rarely because most RoIs exceed the threshold (cf. dataset statistics). Overall, the ablation confirms that our large-scale, multi-round, depth-aware supervision—not post-hoc SR—is the primary driver of the model's strong, balanced improvements in complex visual reasoning.

5.3 User Study

In a blinded study with 30 raters on 20 randomly sampled items per setting, we scored four criteria—Answer Accuracy (AA), Grounded Faithfulness (GF), Stepwise Clarity & Sufficiency (SCS), and Efficiency & Brevity (EB). The results in table 4 show that training on **VisCoT-Pro** markedly improves GF and SCS (tighter RoIs, clearer global to local chains), yielding a corresponding rise

Table 3: Comparison with state-of-the-art MLLMs on additional benchmarks.

Method	Visual CoT	MME	V*
VisCoT-7B (Shao et al., 2024)	0.580	0.701	0.445
LLaVA-NeXT-8B (Chen & Xing, 2024)	0.705	0.666	0.597
InternVL-2.5-8B (Zhu et al., 2025)	0.738	**0.848**	0.597
VisCoT-Pro-7B	0.798	0.695	**0.603**
VisCoT-Pro-Max-7B	**0.809**	0.751	0.590

Table 4: **Human evaluation (1–5).** AA = Answer Accuracy; GF = Grounded Faithfulness; SCS = Stepwise Clarity & Sufficiency; EB = Efficiency & Brevity. Mean is an unweighted average.

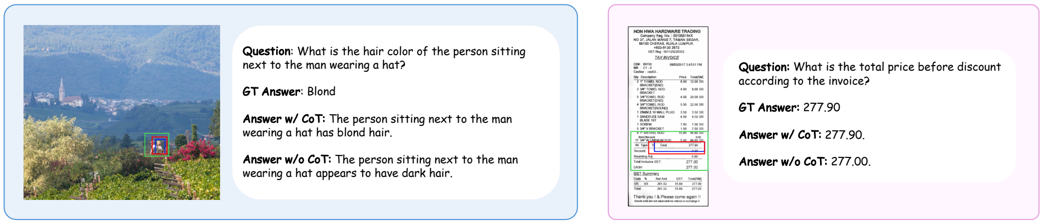
Method	AA	GF	SCS	EB	Mean
VisCoT-7B (Shao et al., 2024)	3.24	2.78	2.82	3.66	3.18
LLaVA-NeXT-8B (Chen & Xing, 2024)	3.6	3.0	3.0	**3.83**	3.41
InternVL-2.5-8B (Zhu et al., 2025)	3.84	3.32	3.26	3.73	3.52
VisCoT-Pro-7B	4.03	4.20	4.13	3.67	4.00
VisCoT-Pro-Max-7B	**4.12**	**4.45**	**4.34**	3.77	**4.11**

Table 5: **Ablation study** on dataset selection and Super Resolution.

VisCoT-Pro	VisCoT-Pro-Max	Super Res	Doc/ Text	General VQA	Relation Reasoning	Fine-grained	Average
			0.920	0.739	0.598	0.681	0.770
✓			0.864	0.744	0.678	0.798	0.779
✓	✓		0.856	**0.750**	0.693	0.809	**0.781**
✓	✓	✓	0.838	0.748	**0.694**	**0.811**	0.778

in AA, with EB roughly unchanged. Adding **VisCoT-Pro-Max** brings the most significant gains in GF/SCS—raters highlighted better depth use and fewer reasoning leaps—translating to the highest AA while maintaining competitive EB. Overall, the human study confirms that our multi-round, depth-aware supervision is the primary driver of quality improvements.

5.4 Visualization



Question: What is the hair color of the person sitting next to the man wearing a hat?

GT Answer: Blond

Answer w/ CoT: The person sitting next to the man wearing a hat has blond hair.

Answer w/o CoT: The person sitting next to the man wearing a hat appears to have dark hair.

Question: What is the total price before discount according to the invoice?

GT Answer: 277.90

Answer w/ CoT: 277.90.

Answer w/o CoT: 277.00.

Figure 6: Visualization results of **VisCoT-Pro** to illustrate the difference between various inference modes. Model-generated bounding boxes are shown in green (first-round) and red (second-round), while ground truth (GT) bounding boxes are in blue. Best viewed in color and zoomed in.

This section presents qualitative results in fig. 6, showcasing our model's visual CoT ability: the model first localizes evidence via predicted RoIs, then iteratively zooms and fuses fine-grained crops with the global view to produce the final answer. We compare three settings—our full CoT pipeline, **VisCoT-Pro** with ground-truth RoIs (GT BBox), and **VisCoT-Pro** without CoT (w/o CoT). The full and GT-BBox variants consistently focus on the correct regions and recover small, detail-sensitive cues (e.g., attributes, prices), while the w/o CoT baseline often mis-localizes or relies on superficial priors. These examples illustrate that accurate region selection and depth-aware, stepwise reasoning directly translate into higher answer fidelity and fewer hallucinations.

6 Conclusion

In summary, we close three gaps in visual CoT—limited scale/coverage, missing multi-round process supervision, and weak depth awareness—by releasing **VisCoT-Pro** (506k) and **VisCoT-Pro-Max** (165k) with detailed annotations: compact global to local stepwise rationales, RoI boxes, and in Pro-Max, pseudo-3D cues (monocular depth, semantic segmentation, ordinal relations). These comprehensive resources span text/doc, fine-grained, general VQA, and spatial relational reasoning, providing rich signals for faithful, spatially grounded inference. Models trained on **VisCoT-Pro** series achieve higher accuracy, grounded faithfulness, and generalization across Visual-CoT, MME, and V*Bench, with corroborating human studies. We hope **VisCoT-Pro** series serves as a strong baseline and widely applicable foundation for future spatially aware visual reasoning.

Ethics statement. We confirm that this research adheres to the ICLR Code of Ethics. Our work is built upon publicly available datasets and pretrained models, and we foresee no direct negative societal impacts or ethical concerns arising from our methodology.

Reproducibility statement. To empower the community to verify our results and build upon our work, we are committed to releasing the complete source code, including all experimental scripts, concurrent with the paper's publication.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

S. Bai, A. Yang, and Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yuhang Cao, Pan Zhang, Xiaoyi Dong, Dahua Lin, and Jiaqi Wang. Dualfocus: Integrating macro and micro perspectives in multi-modal large language models. *arXiv preprint arXiv:2402.14767*, 2024.

Lin Chen and Long Xing. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. `https://github.com/xiaoachen98/Open-LLaVA-NeXT`, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024a.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1254–1262, 2024b.

G. Comanici and Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *CVPR*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Chaoyou Fu, Xinyue Yang, Xingyu Chen, Mouxing Sun, Linjiang Qiu, Yi Huang, Yixiao Li, Tianyu Cheng, Jinghao Shi, Zhenyu Xiao, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. Clova: A closed-loop visual assistant with tool usage and update. In *CVPR*, pp. 13258–13268, 2024.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In *ICLR*, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, pp. 14953–14962, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *CVPR*, pp. 9590–9601, 2024.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, Ranjay Krishna, Jiajun Wu, and Hamid Rezatofighi. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*, 2025.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024b.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, volume 36, 2024a.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023b.

Zhiyuan Liu, Yuting Zhang, Feng Liu, Changwang Zhang, Ying Sun, and Jun Wang. Othink-mr1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning. *arXiv preprint arXiv:2503.16081*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024b.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14268–14280, 2025.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pp. 2200–2209, 2021.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pp. 14420–14431, 2024.

OpenAI. Gpt-4v(ision) system card. 2023. URL `https://api.semanticscholar.org/CorpusID:263218031`.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2403.16999*, 2024.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pp. 742–758. Springer, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, volume 35, pp. 24824–24837, 2022.

Pan Wu, Zekai Li, Wenhai Wang, Yuhui Li, Feng Lin, Tong Zhang, Zhaowei Zeng, Kai Zhang, Le Lu, Yu Qiao, and Jifeng Dai. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Penghao Wu and Saining Xie. $v^\star$: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, pp. 13084–13094, 2024.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. Voco-llama: Towards vision compression with large language models. *arXiv preprint arXiv:2406.12275*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025b.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint*, 2023.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. In *CVPR*, pp. 13215–13224, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

J. Zhu, Z. Chen, W. Wang, et al. Internvl3: Exploring advanced training and test-time strategies for open multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pp. 4995–5004, 2016.

A Implementation Details

We train both VisCoT-Pro-7B and VisCoT-Pro-Max-7B on four H200 GPUs (140 GB each). The implementation is carried out in PyTorch with DeepSpeed ZeRO-2 optimization. Detailed training configurations are provided in table 6 and table 7. All evaluations are performed on four A800 GPUs (80 GB each).

Configuration	**VisCoT-Pro**
Batch size	96
Learning rate	1.0×10^{-5}
Epochs	2
Optimizer	AdamW
LoRA rank	32
Image max pixels	262,144 (512 × 512)
Image min pixels	12,544 (112 × 112)
Cutoff length	8192

Table 6: Training configurations for **VisCoT-Pro**.

Configuration	**VisCoT-Pro-Max**
Batch size	96
Learning rate	1.0×10^{-5}
Epochs	1
Optimizer	AdamW
LoRA rank	32
Image max pixels	262,144 (512 × 512)
Image min pixels	12,544 (112 × 112)
Cutoff length	8192

Table 7: Training configurations for **VisCoT-Pro-Max**.

B Data Generation Details

For completeness, the full data–generation procedures are summarized in Appendix as algorithms 1 and 2. algorithm 1 details the **VisCoT-Pro** pipeline: given an image–QA pair and a GT box, a generator produces per-round *(description, RoI, rationale)* triplets; the RoI is mapped to pixel space and corrected via fix_{RoI} to cover the GT, then iteratively refined by cropping and re-prompting until either the area ratio falls below the threshold (Area(RoI) $\leq N \cdot$ Area($B^\star$)) or the round budget ($R_{\max}$=3) is reached, followed by a brief final justification. algorithm 2 extends this to **VisCoT-Pro-Max** by deriving pseudo-3D cues (monocular depth and semantic segmentation), forming an object list with ordinal depths, optionally generating a 3D-aware QA, and then running the same multi-round crop–refine loop (`local2global` + fix_{RoI}) with a larger budget ($R_{\max}$=4). Together, these algorithms make the dataset's *process-level* supervision explicit while cleanly separating the 2D-only and depth-aware variants.

Algorithm 1: VisCoT-Pro: Multi-round spatial CoT generation

```
Require: Image $I$, question $q$, answer $a$, GT box $B^\star$; max rounds $R_{\max}$=3; area ratio threshold $N$=2; generator $\mathcal{G}$ (GPT-4.1-Nano)
Ensure: Multi-round chain $\{(\text{desc}_t, \text{AoI}_t, \text{reason}_t)\}_{t=1}^{T}$ and final crop justification
$t \leftarrow 1$; $(W, H) \leftarrow \text{size}(I)$
(Round 1) Prompt $\mathcal{G}$ with $(I, q, a)$ to obtain $\text{desc}_1$, $\hat{A}_1 \in [0,1]^4$, $\text{reason}_1$
$A_1 \leftarrow \text{ratio2xyxy}(\hat{A}_1; W, H)$
$A_1 \leftarrow$ AdjustAoI$(A_1, B^\star)$    ▷ ensure coverage and in-bounds
while $t < R_{\max}$ and $\text{Area}(A_t) > N \cdot \text{Area}(B^\star)$ do
    $I_{t+1} \leftarrow \text{crop}(I, A_t)$
    Prompt $\mathcal{G}$ with $(I_{t+1}, q, a)$ to get $\text{desc}_{t+1}$, $\hat{A}_{t+1}$, $\text{reason}_{t+1}$
    Map $\hat{A}_{t+1}$ to global coords: $A_{t+1} \leftarrow \text{local2global}(\hat{A}_{t+1}; A_t)$
    $A_{t+1} \leftarrow$ AdjustAoI$(A_{t+1}, B^\star)$
    $t \leftarrow t+1$
end while
$T \leftarrow t$; $A_{\text{final}} \leftarrow A_T$; $I_{\text{final}} \leftarrow \text{crop}(I, A_{\text{final}})$
Query $\mathcal{G}$ once more with $(I_{\text{final}}, q, a, B^\star)$ for a one-sentence final justification
return $\{(\text{desc}_t, \text{AoI}_t, \text{reason}_t)\}_{t=1}^{T}$ and final justification
function ADJUSTAOI$(A, B^\star)$
    $A \leftarrow A \cup B^\star$    ▷ expand to include GT box
    return clipToImage$(A)$
end function
```

Algorithm 2: VisCoT-Pro-Max: Spatial-aware spatial CoT generation

```
Require: Image $I$, optional seed QA $(q_0, a_0, B_0^\star)$; max rounds $R_{\max}$=4; area ratio threshold $N$=2; generators $\mathcal{G}_{\text{Q}}$, $\mathcal{G}_{\text{CoT}}$ (GPT-4.1-Mini)
Ensure: 3D-aware QA $(q, a, B^\star)$ and multi-round chain $\{(\text{desc}_t, \text{AoI}_t, \text{reason}_t)\}_{t=1}^{T}$
Compute pseudo-3D cues: depth map $D \leftarrow \text{MonocularDepth}(I)$, semantic segmentation $S \leftarrow \text{Seg}(I)$
Build structured object list $\mathcal{O}$ with categories, $[x_1, y_1, x_2, y_2]$, and ordinal depths from $(S, D)$
if no seed QA then
    $(q, a, B^\star) \leftarrow \mathcal{G}_{\text{Q}}(I, \mathcal{O})$    ▷ joint 2D (*left-of*, *above*) + depth (*in front of*/*behind*)
else
    $(q, a, B^\star) \leftarrow (q_0, a_0, B_0^\star)$
end if
$t \leftarrow 1$; (Round 1) Prompt $\mathcal{G}_{\text{CoT}}$ with $(I, q, a, \mathcal{O})$ to get $\text{desc}_1$, $\hat{A}_1$, $\text{reason}_1$
$A_1 \leftarrow \text{ratio2xyxy}(\hat{A}_1)$; $A_1 \leftarrow$ AdjustAoI$(A_1, B^\star)$
while $t < R_{\max}$ and $\text{Area}(A_t) > N \cdot \text{Area}(B^\star)$ do
    $I_{t+1} \leftarrow \text{crop}(I, A_t)$
    Prompt $\mathcal{G}_{\text{CoT}}$ with $(I_{t+1}, q, a)$ to get $\text{desc}_{t+1}$, $\hat{A}_{t+1}$, $\text{reason}_{t+1}$
    $A_{t+1} \leftarrow \text{local2global}(\hat{A}_{t+1}; A_t)$; $A_{t+1} \leftarrow$ AdjustAoI$(A_{t+1}, B^\star)$
    $t \leftarrow t+1$
end while
return $(q, a, B^\star)$ and $\{(\text{desc}_t, \text{AoI}_t, \text{reason}_t)\}_{t=1}^{t}$
```

C Prompts Design

We provide the prompts used in dataset generation and GPT-evaluation here.

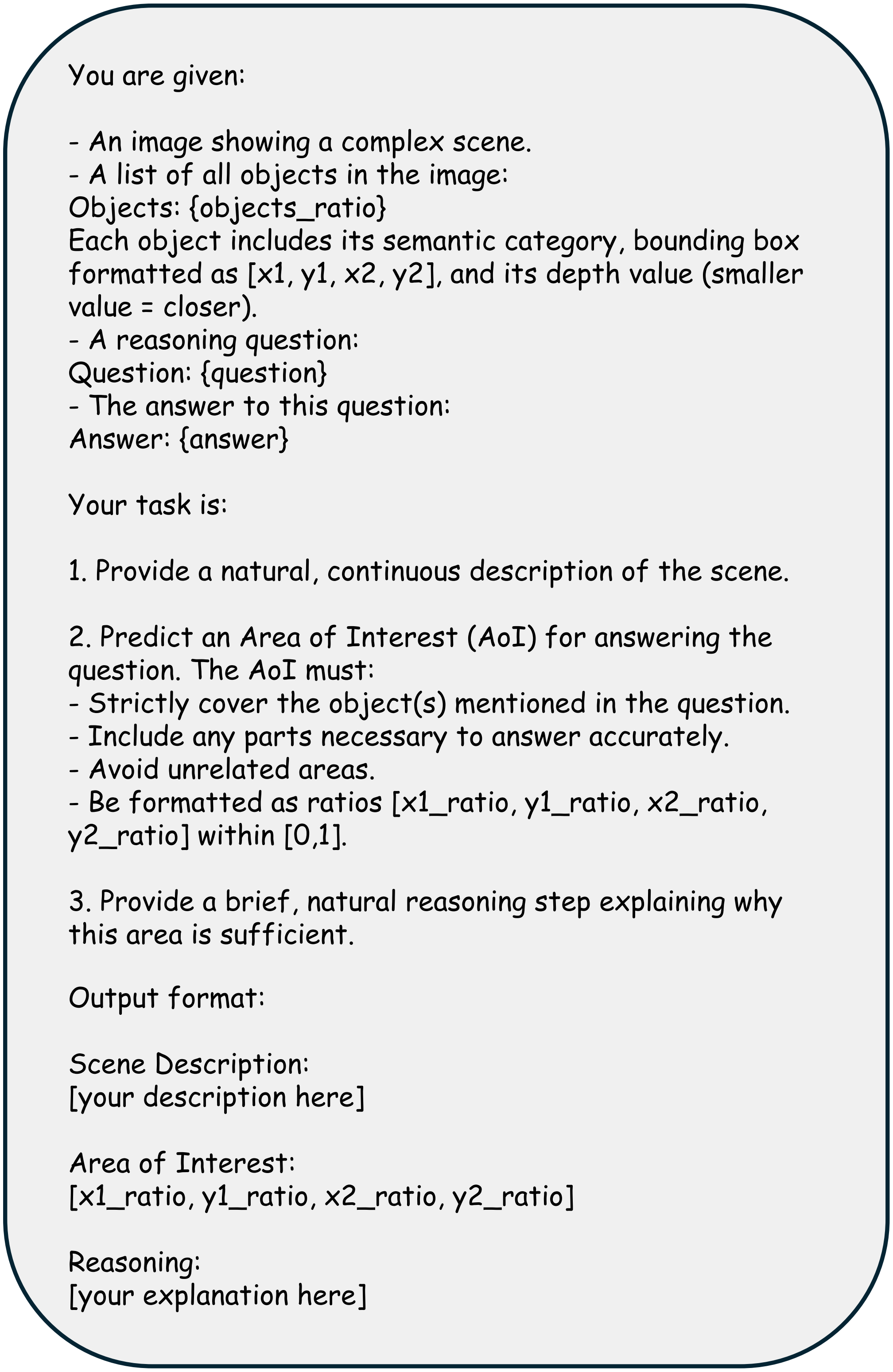
D More Data Examples

We show more examples in our dataset here.

E More inference Examples

F The Use of Large Language Models

LLMs are only used for for editorial support. Its role was strictly limited to improving grammar, phrasing, and overall readability. The LLMs did not contribute to any core scientific aspects of the work, including the research methodology, data analysis, or the formulation of results and conclusions. The intellectual contributions presented herein are entirely those of the authors.

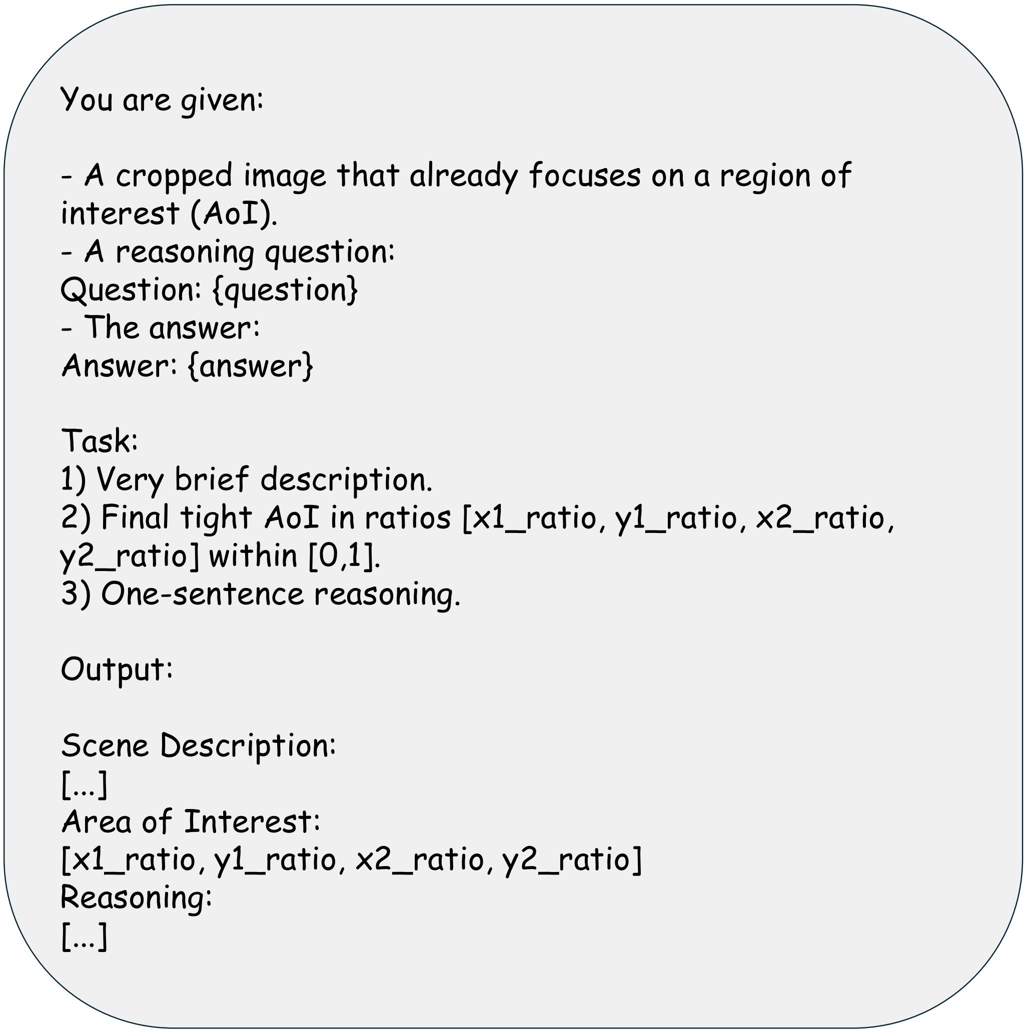


```
You are given:

- An image showing a complex scene.
- A list of all objects in the image:
Objects: {objects_ratio}
Each object includes its semantic category, bounding box
formatted as [x1, y1, x2, y2], and its depth value (smaller
value = closer).
- A reasoning question:
Question: {question}
- The answer to this question:
Answer: {answer}

Your task is:

1. Provide a natural, continuous description of the scene.

2. Predict an Area of Interest (AoI) for answering the
question. The AoI must:
- Strictly cover the object(s) mentioned in the question.
- Include any parts necessary to answer accurately.
- Avoid unrelated areas.
- Be formatted as ratios [x1_ratio, y1_ratio, x2_ratio,
y2_ratio] within [0,1].

3. Provide a brief, natural reasoning step explaining why
this area is sufficient.

Output format:

Scene Description:
[your description here]

Area of Interest:
[x1_ratio, y1_ratio, x2_ratio, y2_ratio]

Reasoning:
[your explanation here]
```

Figure 7: Propmpts used for **First-Round AoI on Full Image** Generates a scene description, the initial normalized Area of Interest, and brief reasoning on the original image (optionally with semantic mask)

```
You are given:

- A cropped image that already focuses on a region of
interest (AoI).
- A reasoning question about this cropped scene:
Question: {question}
- The answer to this question:
Answer: {answer}

Task:
1) Short description of the cropped scene.
2) Refine a new AoI in ratio [x1_ratio, y1_ratio, x2_ratio,
y2_ratio] within [0,1], tight but sufficient.
3) Brief reasoning for why this area suffices.

Output:

Scene Description:
[...]
Area of Interest:
[x1_ratio, y1_ratio, x2_ratio, y2_ratio]
Reasoning:
[...]
```

Figure 8: Propmpts used for **Second-Round AoI Refinement (Cropped Image)** Tightens the AoI on the first crop with a short description and justification when the initial region is too large.



```
You are given:

- A cropped image that already focuses on a region of
interest (AoI).
- A reasoning question:
Question: {question}
- The answer:
Answer: {answer}

Task:
1) Very brief description.
2) Final tight AoI in ratios [x1_ratio, y1_ratio, x2_ratio,
y2_ratio] within [0,1].
3) One-sentence reasoning.

Output:

Scene Description:
[...]
Area of Interest:
[x1_ratio, y1_ratio, x2_ratio, y2_ratio]
Reasoning:
[...]
```

Figure 9: Propmpts used for **Third-Round Final AoI (Further Crop)** Produces the final tight AoI with a minimal description and one-sentence rationale on the second crop.

You are given:

- A final cropped region already focused on the answer.
- The reasoning question:
Question: {question}
- The final answer:
Answer: {answer}
- The ground truth bounding box for the answer in this cropped region as ratios [x1, y1, x2, y2]:
Ground Truth BBox: {gt_bbox}

Task: Write ONE short sentence directly explaining how, by looking at this crop, you recognize the answer, and include the final answer and the grounding bbox.

Output:
Reasoning:
[your one-sentence explanation]

Figure 10: Propmpts used for **One-Sentence Visual Evidence on Final Crop (with GT Box)** Explains in one sentence how the final crop reveals the answer, explicitly including the final answer and the normalized GT bounding box.

```
You are given multiple rounds of reasoning and zoom-in areas of interest
(AoIs), leading to the final answer.

Question: {question}

Round 1:
- Description: {r1_desc}
- Area of Interest: {r1_area}
- Reasoning: {r1_reason}

Round 2 (if any):
- Description: {r2_desc}
- Area of Interest: {r2_area}
- Reasoning: {r2_reason}

Round 3 (if any):
- Description: {r3_desc}
- Area of Interest: {r3_area}
- Reasoning: {r3_reason}

Final Reasoning Step:
{final_reason}
Final Answer:
{final_answer}

Task:
- Summarize the reasoning across rounds into one coherent chain inside
<think>...</think>.
- Each round must mention:
1) what is observed,
2) why it is relevant,
3) the bounding box as <|box_start|>[x1, y1, x2, y2]<|box_end|>,
4) then insert <|image_zoomin|><image> to represent zooming in.
- Use natural, continuous sentences to describe the zooming process.
- Conclude with a short statement showing how the final zoom reveals the
answer.
- After </think>, output only the final answer in <answer>...</answer>.
Example output:
<think> To determine the jersey number of the player taking the shot, I
need to locate the player near the free-throw line where the action is likely
happening.
However, the relevant details are not clearly visible. To improve visibility, I
need to explore step by step. I start by zooming in on the region around the
free-throw line within the bounding box <|box_start|>[810, 555, 1598,
1080]<|box_end|>.
After zooming in on this area, I obtain a refined visual embedding
<|image_zoomin|><image>, which helps me locate the player taking the shot.
Next, I focus on the identified player within the bounding box
<|box_start|>[990, 750, 1103, 1020]<|box_end|>.
I zoom in again and extract a new visual embedding
<|image_zoomin|><image>, which clearly shows that the player is wearing the
number 8 jersey. </think> <answer> 8 </answer>

Now produce your own output for the given rounds and final answer.
Output (strict):
<think> [summary reasoning chain across rounds with boxes + zooms]
</think> <answer> {final_answer} </answer>
```

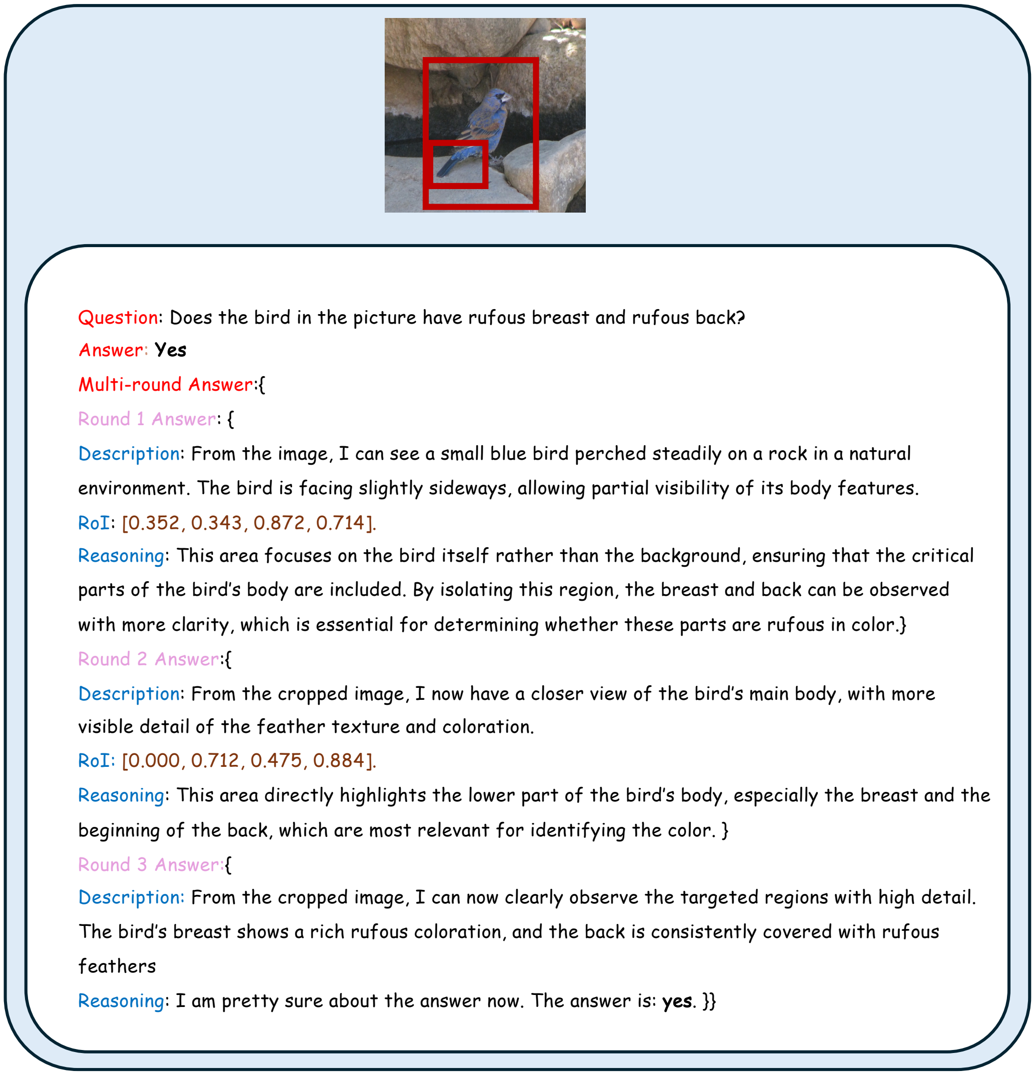
Figure 11: Propmpts used for **Multi-Round Chain Summarization (<think> + <answer>)**

```
You are given an image and its detected objects (with IDs, categories, bounding boxes, and depth values).

Objects:
{objects}

The objects use [x1, y1, x2, y2] pixel bounding boxes (top-left to bottom-right). Depth is in [0,1], where 1 = closest (in front) and 0 = farthest.

Create ONE 3D-aware QA pair.

Requirements:
- The question must include at least one 2D spatial relation (left of, right of, above, below, on, under) AND one depth relation (in front of, behind), No other condition is needed.
- The answer is one word: the category of the unique target object.
- Full answer is a short sentence with the answer in context.
- Output the **ground truth bounding box** of the answer object, copied exactly from the objects.
- All relations and the answer must strictly match the given bboxes and depths.

Output format (strict):
Question: <your question here>
Answer: <single word>
Full Answer: <short sentence>
Ground Truth BBox: [x1, y1, x2, y2]
```

Figure 12: Propmpts used for **3D-Aware QA Generation (Spatial + Depth Relations)**Creates a single QA pair that combines 2D spatial relations and depth relations, and outputs the target object's ground-truth bounding box.

```
GRADER_SYSTEM_PROMPT = (
    "You are an automatic grader. Given Question /
Standard answer / Model's answer, "
    "output a similarity score in [0,1] and nothing else,
formatted exactly as: "
    "score: <score> (up to two decimals).\n"
    "Rules: normalize case/whitespace; unify
number/time formats; exact or formatting-only →
1.00; "
    "synonyms or mutual containment without
contradiction → ~0.95–1.00; minor deviations →
~0.85–0.94; "
    "partial overlap → ~0.75–0.84;
unrelated/contradictory → ~0.00. "
    "For numbers use smaller relative error ⇒ higher
score; for intervals use larger overlap ⇒ higher score.
"
    "Grade only the core answer."
)
```

Figure 13: Propmpts used for Evaluation for the visual CoT benchmark using the ChatGPT



Question: Does the bird in the picture have rufous breast and rufous back?

Answer: **Yes**

Multi-round Answer:{

Round 1 Answer: {

Description: From the image, I can see a small blue bird perched steadily on a rock in a natural environment. The bird is facing slightly sideways, allowing partial visibility of its body features.

RoI: [0.352, 0.343, 0.872, 0.714].

Reasoning: This area focuses on the bird itself rather than the background, ensuring that the critical parts of the bird's body are included. By isolating this region, the breast and back can be observed with more clarity, which is essential for determining whether these parts are rufous in color.}

Round 2 Answer:{

Description: From the cropped image, I now have a closer view of the bird's main body, with more visible detail of the feather texture and coloration.

RoI: [0.000, 0.712, 0.475, 0.884].

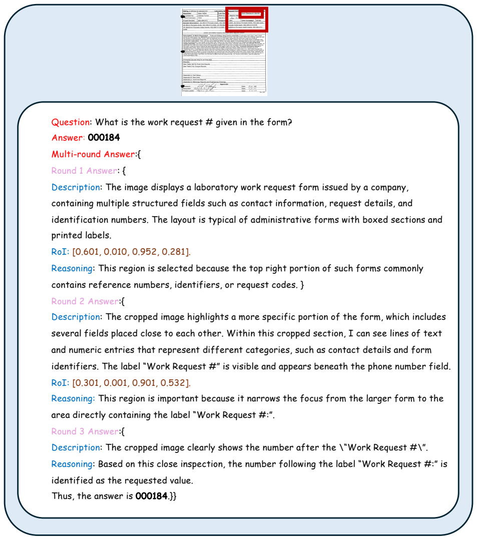
Reasoning: This area directly highlights the lower part of the bird's body, especially the breast and the beginning of the back, which are most relevant for identifying the color. }

Round 3 Answer:{

Description: From the cropped image, I can now clearly observe the targeted regions with high detail. The bird's breast shows a rich rufous coloration, and the back is consistently covered with rufous feathers

Reasoning: I am pretty sure about the answer now. The answer is: **yes**. }}

Figure 14: More data examples in **VCT-Pro**.



Question: What is the work request # given in the form?

Answer: **000184**

Multi-round Answer:{

Round 1 Answer: {

Description: The image displays a laboratory work request form issued by a company, containing multiple structured fields such as contact information, request details, and identification numbers. The layout is typical of administrative forms with boxed sections and printed labels.

RoI: [0.601, 0.010, 0.952, 0.281].

Reasoning: This region is selected because the top right portion of such forms commonly contains reference numbers, identifiers, or request codes. }

Round 2 Answer:{

Description: The cropped image highlights a more specific portion of the form, which includes several fields placed close to each other. Within this cropped section, I can see lines of text and numeric entries that represent different categories, such as contact details and form identifiers. The label "Work Request #" is visible and appears beneath the phone number field.

RoI: [0.301, 0.001, 0.901, 0.532].

Reasoning: This region is important because it narrows the focus from the larger form to the area directly containing the label "Work Request #:".

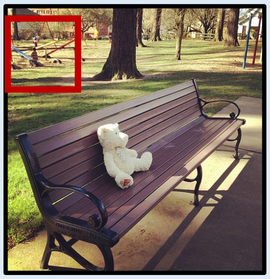
Round 3 Answer:{

Description: The cropped image clearly shows the number after the \"Work Request #\".

Reasoning: Based on this close inspection, the number following the label "Work Request #:" is identified as the requested value.

Thus, the answer is **000184**.}}

Figure 15: More data examples in **VCT-Pro**.



Question: What is the person doing?

Answer: **Sitting on a seesaw**

Multi-round Answer:{

Round 1 Answer: {

Description: The image shows what at first appears to be a teddy bear placed on a wooden bench, but upon closer inspection, it seems to be part of a larger outdoor scene. The background resembles a park or playground setting, with natural light and surrounding greenery.

RoI: [0.000, 0.119, 0.312, 0.378].

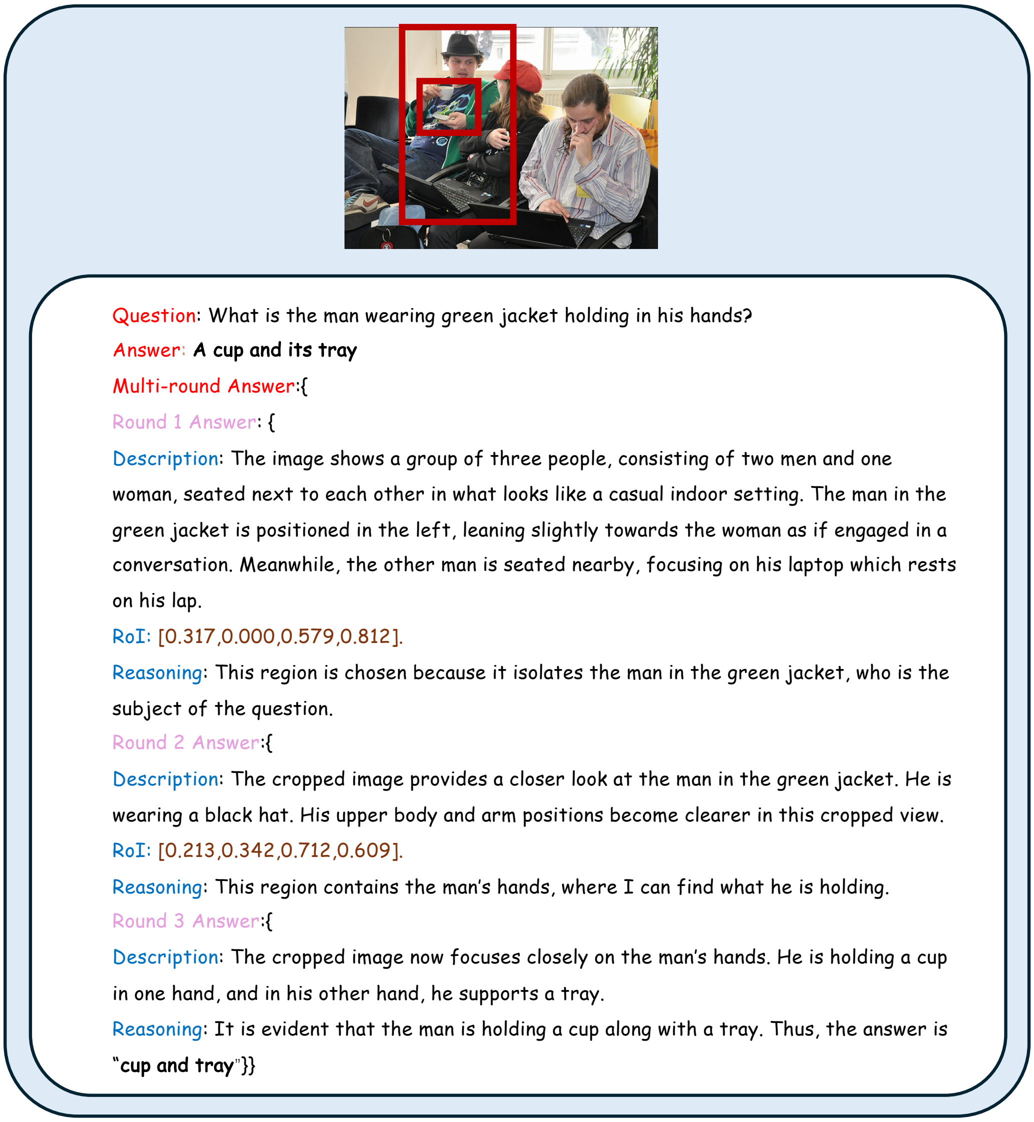
Reasoning: This area is chosen because it contains shapes and textures that resemble a person's outline rather than just background elements. Although the initial impression is ambiguous, it is the most likely region to include the subject of interest.

Round 2 Answer:{

Description: The cropped image reveals a person dressed in red clothing. The figure is no longer confused with surrounding objects, and the structure of the playground equipment becomes visible in this view.

Reasoning: I can clearly recognize the person's activity. The body posture indicates that the individual is sitting on a seesaw. Thus, the answer is: **the person is sitting on a seesaw**. }}

Figure 16: More data examples in **VCT-Pro**.



Question: What is the man wearing green jacket holding in his hands?

Answer: **A cup and its tray**

Multi-round Answer:{

Round 1 Answer: {

Description: The image shows a group of three people, consisting of two men and one woman, seated next to each other in what looks like a casual indoor setting. The man in the green jacket is positioned in the left, leaning slightly towards the woman as if engaged in a conversation. Meanwhile, the other man is seated nearby, focusing on his laptop which rests on his lap.

RoI: [0.317,0.000,0.579,0.812].

Reasoning: This region is chosen because it isolates the man in the green jacket, who is the subject of the question.

Round 2 Answer:{

Description: The cropped image provides a closer look at the man in the green jacket. He is wearing a black hat. His upper body and arm positions become clearer in this cropped view.

RoI: [0.213,0.342,0.712,0.609].

Reasoning: This region contains the man's hands, where I can find what he is holding.

Round 3 Answer:{

Description: The cropped image now focuses closely on the man's hands. He is holding a cup in one hand, and in his other hand, he supports a tray.

Reasoning: It is evident that the man is holding a cup along with a tray. Thus, the answer is "**cup and tray**"}}

Figure 17: More data examples in **VCT-Pro**.

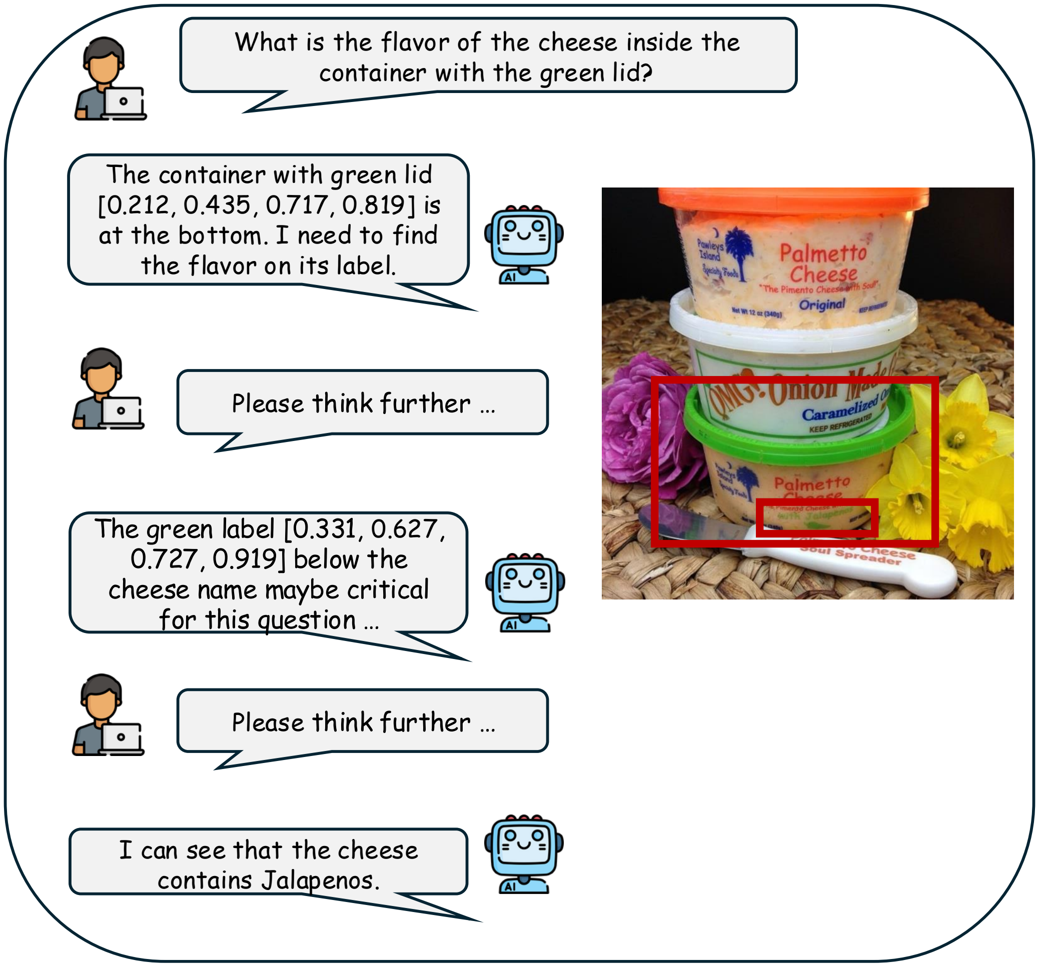
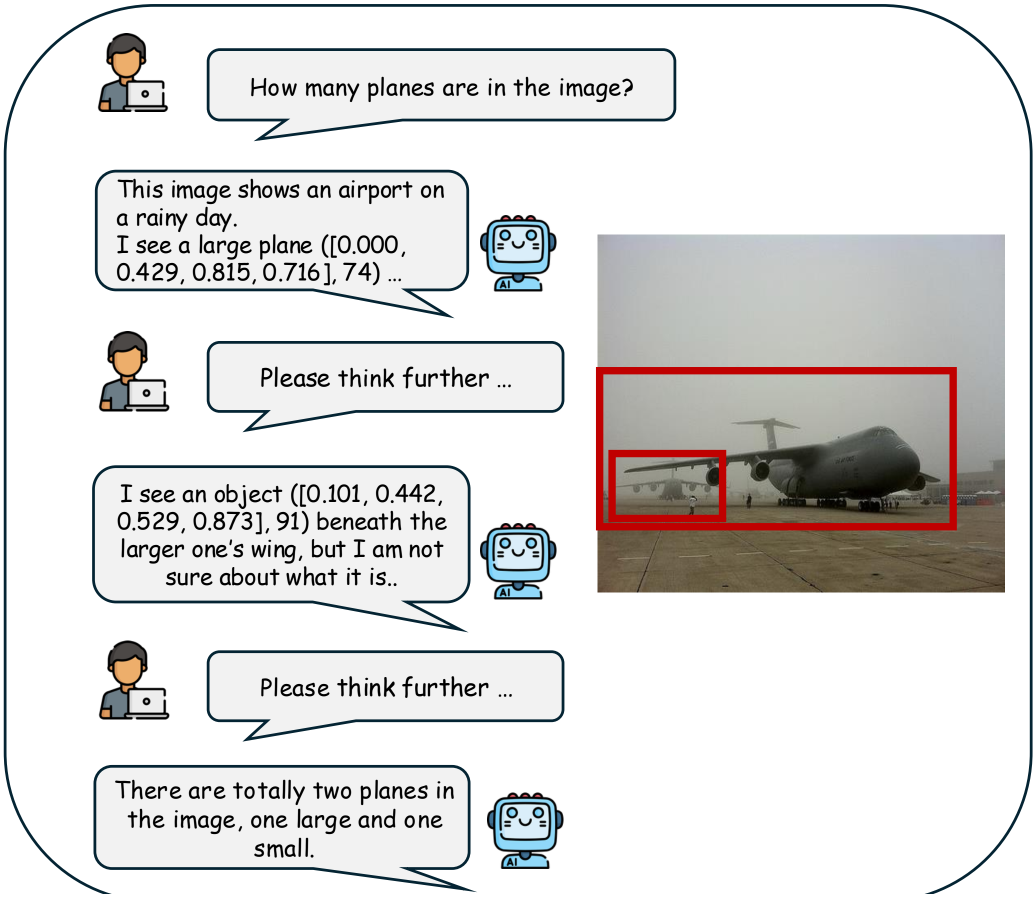

What is the flavor of the cheese inside the container with the green lid?
The container with green lid [0.212, 0.435, 0.717, 0.819] is at the bottom. I need to find the flavor on its label.
Please think further ...
The green label [0.331, 0.627, 0.727, 0.919] below the cheese name maybe critical for this question ...
Please think further ...
I can see that the cheese contains Jalapenos.

Figure 18: More inference examples in **VCT-Pro**.



How many planes are in the image?

This image shows an airport on a rainy day.
I see a large plane ([0.000, 0.429, 0.815, 0.716], 74) ...

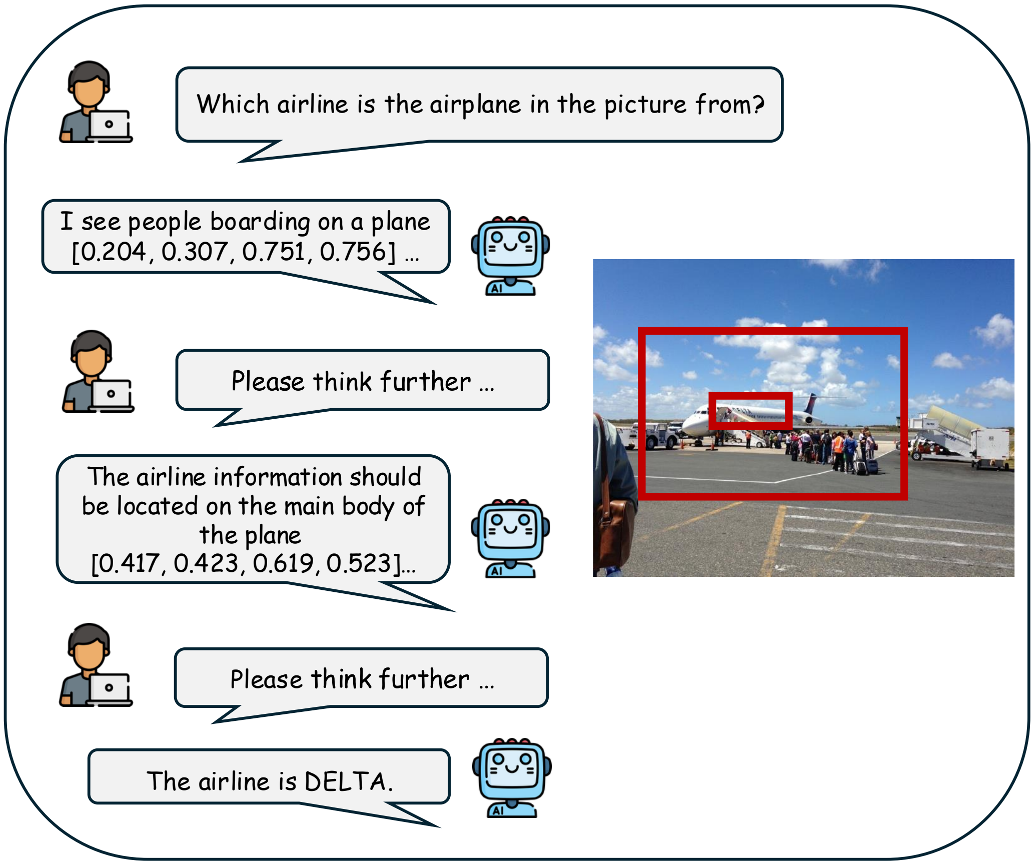
Please think further ...

I see an object ([0.101, 0.442, 0.529, 0.873], 91) beneath the larger one's wing, but I am not sure about what it is..

Please think further ...

There are totally two planes in the image, one large and one small.

Figure 19: More inference examples in **VCT-Pro**.



Which airline is the airplane in the picture from?

I see people boarding on a plane [0.204, 0.307, 0.751, 0.756] ...

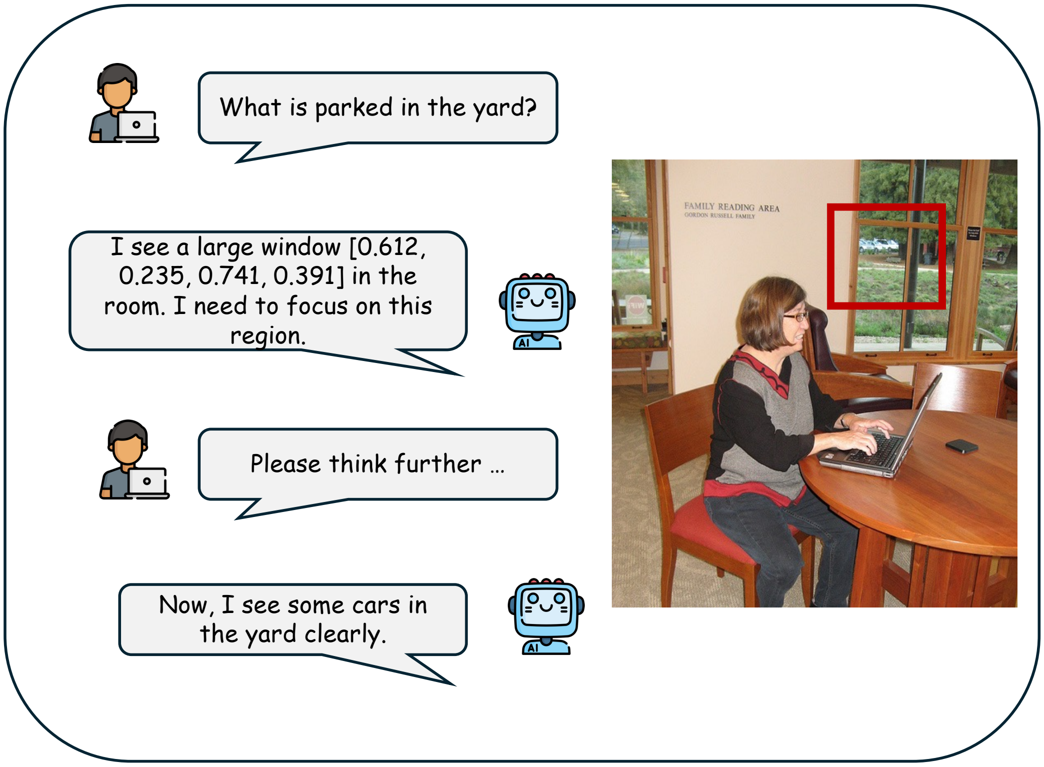
Please think further ...

The airline information should be located on the main body of the plane [0.417, 0.423, 0.619, 0.523]...

Please think further ...

The airline is DELTA.

Figure 20: More inference examples in **VCT-Pro**.



What is parked in the yard?

I see a large window [0.612, 0.235, 0.741, 0.391] in the room. I need to focus on this region.

Please think further ...

Now, I see some cars in the yard clearly.

Figure 21: More inference examples in **VCT-Pro**.