Pixel-Level Reasoning Segmentation via Multi-turn Conversations

Anonymous ACL submission

Abstract

Existing visual perception systems focus on region-level segmentation in single-turn dialogues, relying on complex and explicit query instructions. Such systems cannot reason at the pixel level and comprehend dynamic user intent that changes over interaction. Our work tackles this issue by introducing a novel task, Pixel-level Reasoning Segmentation (Pixel-level RS) based on multi-turn conversations, tracking evolving user intent via multi-turn interactions for fine-grained segmentation. To establish a benchmark for this novel task, we build a **P**ixel-level **R**eason**I**ng **S**egmentation Dataset Based on Multi-**T**urn Conversations (PRIST), comprising 24k utterances from 8.3k multi-turn conversational scenarios with segmentation targets. Building on PRIST, we further propose MIRAS, a **M**ulti-turn **I**nteractive **R**e**A**soning **S**egmentation framework, integrates pixel-level segmentation with robust multi-turn conversation understanding, generating pixel-grounded explanations aligned with user intent. The PRIST dataset and MIRSA framework fill the gap in pixel-level reasoning segmentation. Experimental results on the PRIST dataset demonstrate that our method outperforms current segmentation-specific baselines in terms of segmentation and LLM-based reasoning metrics. The code and data are available at: https://anonymous.4open.science/r/PixelRS/.

1 Introduction

Existing general multimodal large language models (MLLMs) (Bai et al., 2023; Zhu et al., 2023; Liu et al., 2024b) exhibit exceptional visual perception, enabling both image segmentation and textual reasoning, while they primarily rely on explicit human instructions for region-level grounding. Although some segmentation-specific works have explored grounded reasoning responses (Peng et al., 2023; You et al., 2023; Pi et al., 2023; Zhang et al., 2023b), they depend on user-provided regions to trigger reasoning. These perception systems still cannot actively comprehend user's nuanced intent in real-world scenarios. To alleviate this problem, Lai et al. (2023) proposes the reasoning segmentation task that aims to achieve segmentation based on a implicit reasoning query. Recent studies (Ren et al., 2024; Xia et al., 2024; Yuan et al., 2024) have extended this region-level task to encompass multi-object segmentation scenarios to advance development. However, these methods have two limitations: 1) They rely on single-turn ambiguous queries and cannot fully understand users' evolving intent. 2) They lack pixel-level segmentation and only achieve region-level segmentation through one-step explanations (e.g., segment roughly *all ingridients* in Figure 1(a)). In contrast, multi-turn interactions can progressively clarify vague and generalized instructions such as *"make a bread"*. As illustrated in Figure 1(b), the system through multi-turn interactions first guide to clarify the user's desired type of bread, providing targeted responses, and ultimately focuses on the user's specific needs, achieving pixel-level segmentation in final.

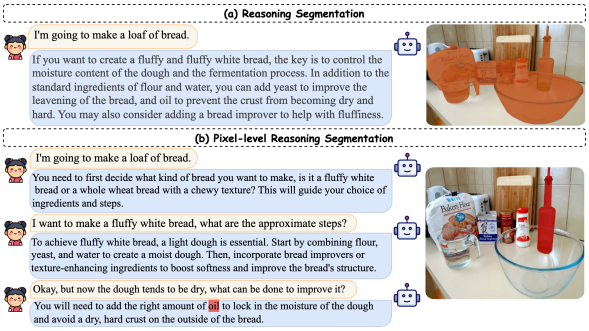


Figure 1: **RS vs. Pixel-level RS.** Pixel-level RS refines intent understanding and segmentation (e.g., *"oil bottle"*) through multi-turn interactions, while RS produces rough segmentation (e.g., *"all ingredients"*) and handles implicit single-turn queries poorly.

To address these challenges, we propose a novel task, Pixel-level Reasoning Segmentation (Pixel-level RS) based on multi-turn conversations, that refines both reasoning and segmentation through

multi-turn interactions, requiring the system to understand the evolving user intent and generating pixel-level explanations, and segmentation masks. Given the lack of benchmarks for pixel-level segmentation based on multi-turn reasoning, we build a **P**ixel-level **R**eason**I**ng **S**egmentation Dataset Based on Multi-**T**urn Conversation (**PRIST**), consisting of 24k utterances, 8.3k multi-turn conversational scenarios with specific segmentated targets, which provides a valuable resource for advancing Pixel-level RS research. PRIST focuses on pixel-level segmentation tasks while introducing new challenges in multi-turn reasoning and evolving intent comprehension. We design a progressive three-step dialogue automatic generation pipeline based on a reasoning tree to iteratively guide and generate dialogue content, inspired by Tree-of-Thought (ToT) (Yao et al., 2024). By integrating a multi-step reasoning chain with a tree structure, this approach facilitates deeper and broader reasoning in pixel-level segmentation training.

To further advance this novel task, we propose a **M**ulti-turn **I**nteractive **R**e**A**soning **S**egmentation framework, **MIRAS**, that enables pixel-level segmentation through progressive reasoning. MIRAS incorporates a dual-vision encoder that fuses multi-scale features to capture detailed visual information. To improve segmentation performance, we introduce a semantic region alignment strategy to inject semantic information into the mask decoder. Additionally, the framework supports multi-turn interactions to iteratively clarify user intent and ambiguous regions. Given the inherent subjectivity in reasoning tasks, manual assessments can be influenced by personal preferences. To ensure fairness, we develop comprehensive evaluation metrics leveraging Large Language Models (LLMs) to assess multi-turn reasoning segmentation across coherence, consistency, and accuracy dimensions. Our contributions can be summarized as follows:

- We propose a novel task, Pixel-level Reasoning Segmentation, that aims to achieve fine-grained segmentation through multi-turn conversations. Then, we build the PRIST dataset, including 8.3k high-quality multi-turn conversational scenarios and pixel-level segmentation targets.
- We develop a multi-turn reasoning segmentation framework, MIRAS, that facilitates pixel-level intentional understanding and segmentation through multi-turn interactions.
- Comprehensive experimental results of our proposed method on different metrics demonstrate both the utility of the PRIST dataset and the effectiveness of the model.

2 Related Work

2.1 Datasets for Reasoning Segmentation

Current reasoning segmentation datasets (Lai et al., 2023; Rasheed et al., 2024; Yuan et al., 2024) mainly focus on region-level segmentation and single-step reasoning, which fail to meet the comprehensive requirements of the pixel-level RS task. Traditional region-level segmentation datasets (Yu et al., 2016; Krishna et al., 2017) rely on simple and explicit instructions, lacking complex user intents understanding. To bridge this gap, ReasonSeg (Lai et al., 2023) is the first to propose a segmentation dataset based on complex queries, which is a small scale and does not support multi-turn interactions. While recent datasets (Yuan et al., 2024; Rasheed et al., 2024; Ren et al., 2024) have expanded in size and diversity, they remain focused on multi-object segmentation and provide limitations for multi-step reasoning and fine-grained grounding. In contrast, our PRIST dataset utilizes conversation to simulate a human-like multi-step reasoning process, which innovatively combines multi-turn reasoning with pixel-level segmentation. Detailed dataset comparisons are shown in Table 8 in Appendix B.3.

2.2 MLLMs for Region-level Segmentation

MLLMs have advanced vision-language perception tasks, with recent works (Peng et al., 2023; You et al., 2023; Zhang et al., 2023b) focusing on image-level visual dialogue. Some (Chen et al., 2023; Peng et al., 2023; Pi et al., 2023) achieve region-level understanding by incorporating positional information and boundary boxes, mainly relying on LLMs for region interpretation. Several models (Lai et al., 2023; Ren et al., 2024; Rasheed et al., 2024; Zhang et al., 2024) integrate segmentation-specific modules with LLMs for end-to-end training, enabling a more comprehensive understanding of regions. While these methods address pixel-level grounding, they still face limitations in complex reasoning. More comparison between models in Appendix A.2. Our proposed MIRAS overcomes these challenges by enhancing segmentation accuracy through interactive reasoning, progressively refining the boundaries.

3 PRIST

The pixel-level RS task takes an image $\mathcal{I}$ and a multi-turn dialogue $\mathcal{D}$ as input, then simultaneously generates a target segmentation mask $\mathcal{M}$ along

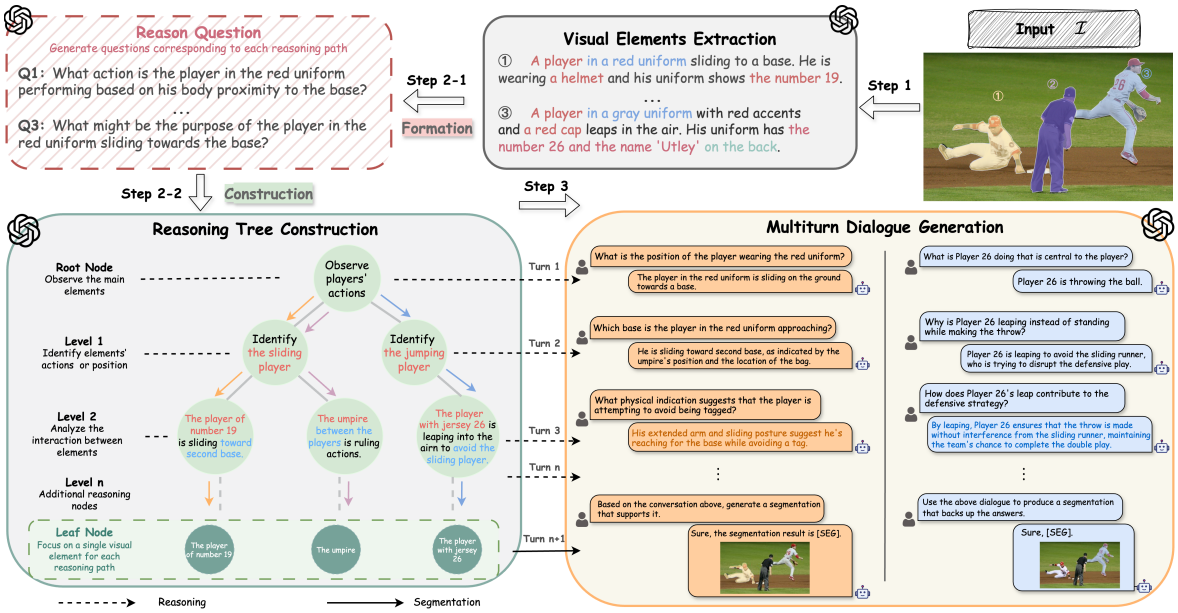


Figure 2: **The Generation Pipeline of PRIST Dataset.** i) *Step 1* extracts visible elements from images, establishing a semantic foundation for subsequent steps. ii) *Step 2-1* generates complex reasoning questions from these elements, while *Step 2-2* iteratively refines the questions through a reasoning tree, ensuring rigorous reasoning. iii) *Step 3* organizes the nodes in reasoning tree into a multi-turn dialogue format.

with a textual reasoning chain $\{a_1, a_2, \ldots, a_N\}$ that captures the complete dialogue history, where a_i denotes the system's response in the i-th turn. It is defined as follows:

$$(\mathcal{M}, \{a_1, a_2, \ldots, a_N\}) = \textbf{Model}(\mathcal{I}, \mathcal{D}). \quad (1)$$

Furthermore, we construct the pixel-level reasoning segmentation dataset based on the multi-turn conversation (PRIST) using a three-step progressive annotation pipeline, capturing fine-grained details through context-aware multi-turn dialogue.

3.1 Data Preparation

Given the focus on pixel-level segmentation, we select TextCaps[1] (Sidorov et al., 2020) as the image source due to its detailed visual information. To ensure a diverse range of scenes, we randomly select 280 images from each of the 10 major categories, resulting in a total of 2.8k images.

3.2 Generation Pipeline

We propose a three-step progressive automated annotation pipeline to create the PRIST Dataset, as illustrated in Figure 2. Appendix B.2 details the pipeline's prompts and output formats.

3.2.1 Visual Elements Extraction (Step-1)

We first extract N visible objects $\mathcal{O} = \{o_i\}_{i=1}^{N}$, from the input image $\mathcal{I}$. Each object o_i represents a distinct target for generating a dialogue. Specifically, we automatically identify visible elements with detailed attributes (e.g., color, position) to each object using GPT-4o (Achiam et al., 2023), along with corresponding textual descriptions (see Figure 7a). This step ensures that visual and semantic details are fully represented.

3.2.2 Reasoning Process Construction (Step-2)

Pixel-level RS addresses complex scenarios requiring multi-turn reasoning with implicit user instructions. To simulate such scenarios, we implement a "question-first, problem-solving" strategy as shown in Figure 2, where the reasoning process is refined after first forming the reasoning problem and then constructing the reasoning tree. To align with multi-turn reasoning, we propose a hierarchical reasoning tree that recursively decomposes complex questions into smaller subquestions, progressively focusing on segmentation targets. Each reasoning tree path connects related elements to build a logical chain for pixel-level segmentation.

Reasoning Question Formation (Step 2-1) This step expands a complex reasoning question $\mathcal{Q}_i$ for each target o_i, serving as the overall origin for next question decomposition and the theme for multi-turn dialogues in Step-3. To balance processing efficiency, we randomly select K objects ($2 \leq K \leq \min(N, 4)$) from the element set $\mathcal{O}$ in Step-1 as targets, as fewer may missing essential interactions while more escalate complexity.

$$\mathcal{Q}_i = \textbf{Formation}(\mathcal{I}, o_i), i = \{1, ..., K\}. \quad (2)$$

Reasoning Tree Construction (Step 2-2) The construction process (see Figure 7b) with GPT-

[1]TextCaps contains a total of 28k images.

4o initiates by establishing the elements o_i as leaf nodes, allowing their corresponding reasoning problems $\mathcal{Q}_i$ to develop a distinct path $\mathcal{P}_i$ within the reasoning tree $\mathcal{T}$. Through iterative decomposition, each $\mathcal{Q}_i$ evolves into a sequence of progressive QA pairs (q_n, a_n), with the tree's depth directly corresponding to the granularity of subquestions. This hierarchical expansion refines the problem-solving framework and progressively narrows pixel-level target localization. The resulting reasoning tree $\mathcal{T}$ explicitly captures the logical progression of complex questions, providing a structural foundation for the multi-turn dialogues in Step-3.

$$\begin{aligned} \mathcal{P}_i &= \textbf{Construction}(\mathcal{Q}_i, o_i), \\ &= \{(q_1, a_1), \cdots, (q_n, a_n)\}, \\ \mathcal{T} &= \{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_i\}_{i=1}^{K}, \end{aligned} \tag{3}$$

where n is the depth of the reasoning path. To manage computational resources and logical flexibility, we impose a constraint limiting each reasoning tree layer to a maximum of three child nodes.

3.2.3 Multi-turn Dialogue Generation (Step-3)

We further build multi-turn dialogue $\mathcal{D}_i$ based on the hierarchical reasoning tree from Step-2. Specifically, we integrate all nodes in each reasoning path $\mathcal{P}_i$, where each node represents a QA pair, to form the progressive multi-turn dialogue $\mathcal{D}_i$. Thus, each image can form K conversations, $\{\mathcal{D}_1, \cdots, \mathcal{D}_i\}_{i=1}^{K}$. To ensure responses fully integrate contextual information, prompts are designed to incorporate key elements and relationships from Step-1 (see Figure 7c), expanding understanding of landmarks, historical context, and scene interactions. Furthermore, pixel-level RS is designed to guide the model in performing fine-grained segmentation, with the final query in each dialogue being a segmentation-related instruction (e.g., *"Please segment the core objects according to the above dialogue"*).

	G1	G2	G3	G4	G5
IoU	0.82	0.88	0.91	0.86	0.81
Kappa	0.79	0.82	0.80	0.78	0.75

Table 1: Consistency Analysis Results for Mask Annotators. Gx denotes the x-th expert group.

3.3 Quality Assurance

We employ manual annotation to generate the true segmentation masks for each sample. Ten experts, selected through qualification tests, annotate masks and correct any commonsense errors in the dialogues. To ensure the quality of PRIST, we implement a double-check process. Specifically, experts are organized into five groups, with each group annotating the same set of samples. This set ensures that two annotators independently annotate every sample. More details are in Appendix B.1. The reliability of annotations across the five groups is assessed using IoU (Girshick et al., 2014), which measures the overlap between two annotators, and Cohen's Kappa (Cohen, 1960), which quantifies their consistency. As shown in Table 1, all groups achieve a result of IoU > 0.80 and Kappa > 0.75, demonstrating high annotation quality and strong consistency across the groups.

Statistic	**Number**
Total Images	2,800
Total Samples	8,320
Segmentation	
- Focus Classes	12
- Granularity (Coarse: Med.: Fine)	22%: 25%: 53%
Multi-turn Dialogue	
- Number of Utterances	24k
- Avg. / Max. Turns	4 / 7
- Avg. / Max. Dialogue Length	477.6 / 518

Table 2: **PRIST Statistics.** According to COCO's image standard (640 vs. 1024), mask granularity is categorized as "Fine" (< $(s \times 32)^2$ px), "Med." ($(s \times 32)^2$ to $(s \times 96)^2$ px), and "Coarse" (> $(s \times 96)^2$ px).

3.4 Dataset Analysis

As shown in Table 2, we provide detailed statistics of the PRIST dataset. PRIST contains 24k high-quality utterances, and 8.3k multi-turn conversational scenarios, with each scenario focusing on a single, fine-grained object. The dataset is divided into three subsets: train / validation / test splits, containing 7,312 / 500 / 508 samples, respectively.

To quantify pixel-level segmentation, we adopt the COCO mask size standard[2] for measuring target scales, aligning with existing datasets (Chen et al., 2015; Caesar et al., 2018) and providing a quantitative basis for fine-grained segment evaluation. Statistics in Table 2, fine-grained targets (the scaling factor s is 1.6) account for 53% of PRIST, surpassing the 41% of small targets in COCO (Chen et al., 2015). With a minimum mask area of 304px and a standardized image size of 1024 × 1024, PRIST meets the fine-grained requirements of pixel-level segmentation. We emphasizes exhibiting high diversity in categories and descriptions to enhance expressiveness. Illustrated the focus distribution of PRIST in Figure 3, the categories include four types: *Textual Content*, *Physical Objects*, *Visual Elements*, and *People*

[2] https://cocodataset.org/#detection-eval

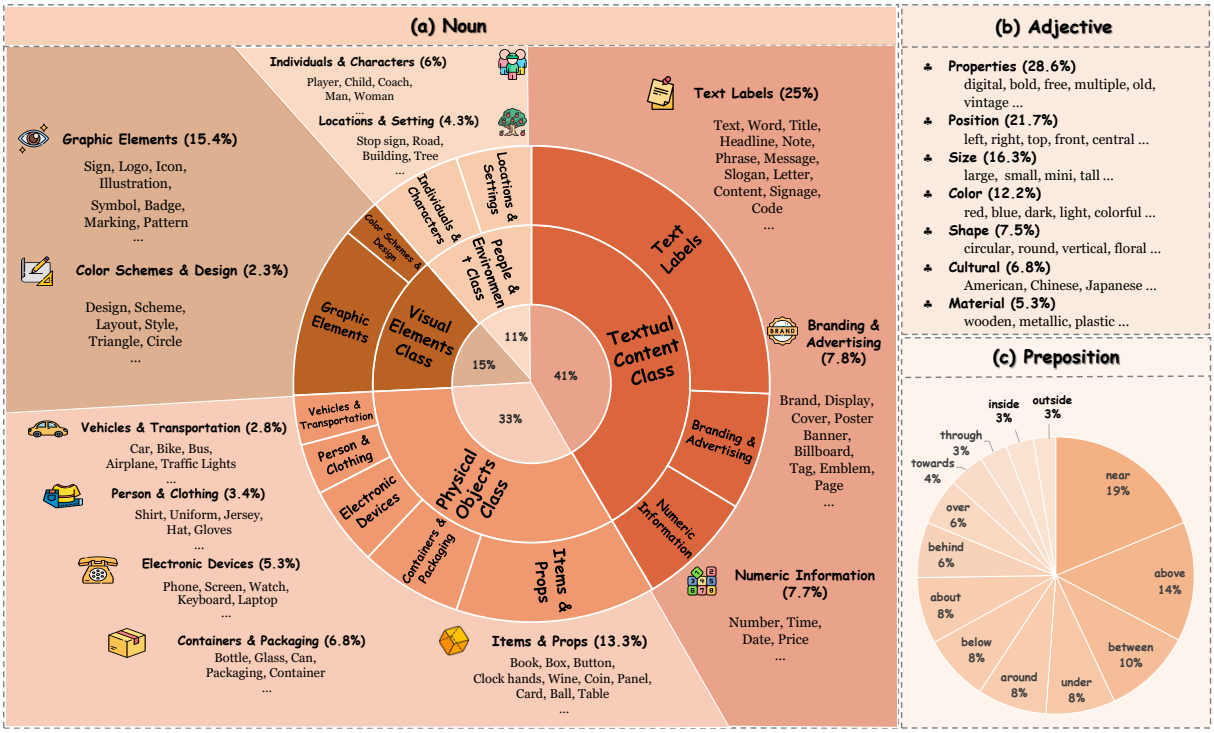


Figure 3: **The focus distribution of PRIST.** We analyze focus objects across 3 dimensions: noun, adjective and preposition, which capture fine granularity, diversity, and close spatial relationships between objects.

& Environment, which are further refined into 12 subcategories that cover objects from coarse- to fine-grained levels. At the descriptive level, a combination of *Noun*, *Adjective*, and *Preposition* is employed: nouns provide basic category information (e.g., "tree"), adjectives enrich focus details (e.g., "worn-out chair"), and prepositions describe spatial relationships (e.g., "a book under the table"). PRIST delivers rich semantic-spatial annotations, establishing a benchmark resource for Pixel-level RS advancement.

4 MIRAS

To further research the novel pixel-level RS task, we propose MIRAS, a framework that refines user intent through multi-turn interactions to achieve pixel-grounded explanations and segmentation.

4.1 Architecture

The architecture of MIRAS is illustrated in Figure 4, consisting of three core components: Visual Encoder, MLLM ($\mathcal{F}$, (Liu et al., 2024a)), and Mask Decoder ($\mathcal{D}_m$, (Kirillov et al., 2023)). To seamlessly connect reasoning with segmentation, we introduce a special token `[SEG]` as a placeholder for segmenting regions and enabling end-to-end processing. Two key modules are proposed for pixel-level RS. First, we integrate dual visual encoders (Li et al., 2023) to extract enriched visual features. Then, the semantic region alignment strategy is designed to further refine the model's focus by incorporating target semantic information.

Dual Visual Encoders We train the dual visual encoder using a high-resolution image ($\mathbf{X}_H$, 768 × 768 pixels) processed by ConvNext-L (Liu et al., 2022) paired with its low-resolution counterpart ($\mathbf{X}_L$, 336 × 336 pixels) processed by CLIP-L/14 (Radford et al., 2021), downsampled from $\mathbf{X}_H$. Then, different resolutions are fused by a cross-attention module (Lin et al., 2022) to enhance visual detail capturing. Note that $\mathbf{X}_H$ equals $\mathbf{X}_{\text{img}}$.

$$\begin{aligned}
\mathbf{X}'_H &= \textbf{ConvNext}(\mathbf{X}_H), \quad \mathbf{X}'_H \in \mathbb{R}^{H\times W\times 3}, \\
\mathbf{X}'_L &= \textbf{CLIP}(\mathbf{X}_L), \quad \mathbf{X}'_L \in \mathbb{R}^{H'\times W'\times 3}, \\
\mathbf{E}'_{\text{img}} &= \text{CrossAtten}(Q=\mathbf{X}'_L, K=\mathbf{X}'_H, V=\mathbf{X}'_H), \\
\mathbf{E}_{\text{img}} &= \text{MLP}(\mathbf{E}'_{\text{img}}) + \mathbf{E}'_{\text{img}},
\end{aligned} \tag{4}$$

Semantic Region Alignment We employ the SAM to obtain the pixel-level image features.

$$\mathbf{E}_{\text{seg}} = \mathcal{V}_{\text{pixel}}(\mathbf{X}_{\text{img}}), \quad \mathbf{E}_{\text{seg}} \in \mathbb{R}^{N\times 256}. \tag{5}$$

To provide clear segmentation intent, we design a novel segmentation prompt template `[OBJ]{CLASS}[SEG]`, where `{CLASS}` is the object description (e.g., `[OBJ]a bunch of grapes [SEG]` in Figure 4). We further utilize special tokens `[OBJ]` to extract relevant sub-sequence $\mathcal{H}_{\text{seg}}$ from $\mathcal{H}$ for segmentation and employ the cross-attention module to capture sufficient semantic information, which is crucial for efficient fine-

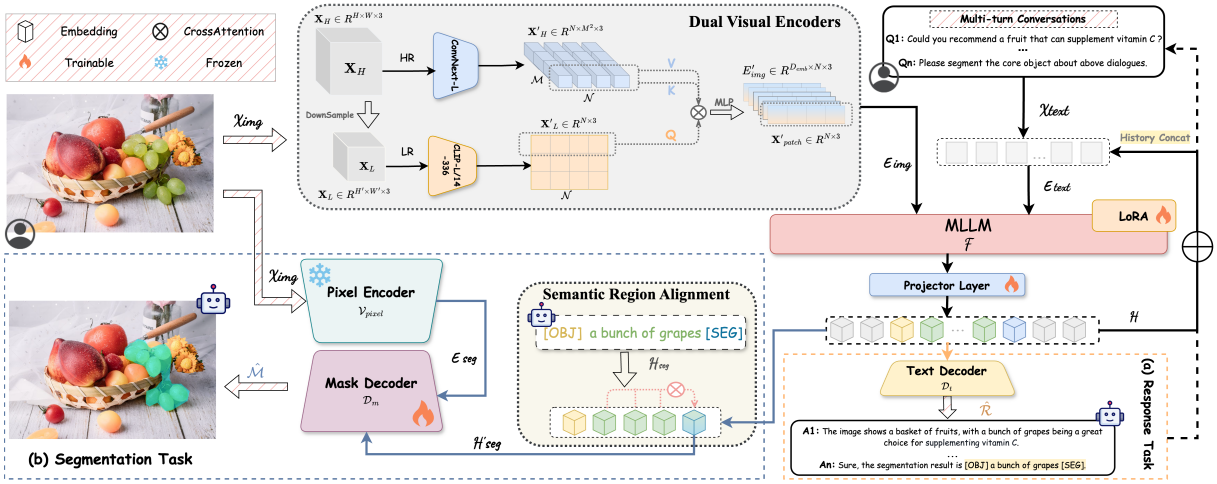


Figure 4: **Overview Architecture of MIRAS.** The model integrates MLLM and SAM modules by introducing a special token [SEG]. MIRAS can perform both (a) Multi-turn Response and (b) Segmentation tasks end-to-end.

grained segmentation. The method resolves potential mismatches in dimensions caused by varying lengths of {CLASS}, denoted as N^{sub}.

$$\mathcal{H}'_{\text{seg}} = \text{CrossAtten}(Q, K, V|\mathcal{H}_{\text{seg}}),$$
$$\mathcal{H}_{\text{seg}} \in \mathbb{R}^{N^{sub}\times 256}, \mathcal{H}'_{\text{seg}} \in \mathbb{R}^{1\times 256}, \tag{6}$$

The mask decoder $\mathcal{D}_m$ combines region features from the pixel encoder with the hidden features of [SEG] to produce the final mask.

$$\hat{\mathcal{M}} = \mathcal{D}_m(\mathbf{E}_{\text{seg}}, \mathcal{H}'_{seg}), \tag{7}$$

4.2 Training Process

We employ a two-stage training process to achieve efficient pixel-level reasoning segmentation. In Stage-1, mask-text alignment pretraining based on various datasets is conducted, followed by instruction fine-tuning using the PRIST dataset in Stage-2 (more details in Appendix C.2). The objectives remain consistent across both stages: the text generation loss $\mathcal{L}_{\text{t}}$ and a linear combination of per-pixel BCE loss $\mathcal{L}_{bce}$ and DICE loss $\mathcal{L}_{dice}$ for segmentation. Only the mask decoder and projection layer are trainable to balance efficiency and performance while keeping the image encoder and MLLM frozen. The training loss is formulated as:

$$\mathcal{L} = \lambda_t \mathcal{L}_t(\mathcal{R}, \hat{\mathcal{R}}) + \lambda_{bce} BCE(\mathcal{M}, \hat{\mathcal{M}}) + \lambda_{dice} DICE(\mathcal{M}, \hat{\mathcal{M}}) \tag{8}$$

where λ_{t}, λ_{bce} and λ_{dice} values 1.0, 2.0 and 0.5 separately, following LISA (Lai et al., 2023).

5 Experiment

Experiments evaluate our model's adaptability to the novel pixel-level RS task and its general performance in the classical referring expression segmentation (RES) task (Kazemzadeh et al., 2014; Yu et al., 2016). Appendix C shows backbone selection and more implementation details.

5.1 Evaluation Metrics

We establish a benchmark for Pixel-level RS across three evaluation aspects: pixel-level segmentation, conversation response, and reasoning quality. For segmentation, we employ the CIoU metric (Zheng et al., 2020) and propose pixel-wise mask precision, recall, and F1 metrics. The precision measures the accuracy of segmentation while recall evaluates the coverage. For response, the metrics including BLEU-4 (Papineni et al., 2002), ROUGE-L (Kingma, 2014), Dist-n (Li et al., 2015), and METEOR (Banerjee and Lavie, 2005). Considering the subjectivity of reasoning, we introduce LLM as a scoring tool using four metrics: Progressiveness **(PR)**, Logical Coherence **(LC)**, Content Consistency **(CC)**, and Target Relevance **(TR)**, with higher scores reflecting better performance across aspects. Meanwhile, we employ GPT-4o as a judge to assess dialogue reasoning quality. The model wins when its score surpasses that of the human response, as reflected by the **Win Rate (%)** metric. Evaluation criteria are detailed in Appendix C.4.

5.2 Baselines

We compare with three types of baselines: **1) General MLLMs.** We take advanced close- and open-source MLLMs to evaluate zero-shot on PRIST for pixel-level RS capability. **2) Segmentation-specific MLLMs.** LISA (Lai et al., 2023), PixelLM (Ren et al., 2024) and OMG-LLaVA (Zhang et al., 2024) are evaluated under zero-shot and fine-tuning settings to show PRIST's task-specific enhancement. **3) Segmentation-specific Models.** We compare three semantic segmentation models (e.g., LVIT (Yang et al., 2022)) on the RES task with MIRAS to verify its advantage of the architecture.

Model	CIoU	Pixel-wise			Response				
		Prec.	Recall	F1	BLEU-4	Dist-1/2	ROU_L.	MET.	BERTS.
Zero-shot									
InternVL2-8B* (Chen et al., 2024b)	8.26	7.59	11.55	9.16	1.20	7.5 / 41.0	18.44	23.98	84.78
Qwen2-VL-7B* (Wang et al., 2024a)	10.64	10.30	15.88	12.50	3.00	9.3 / 41.2	26.28	28.26	86.73
LLaVA-v1.5-7B* (Liu et al., 2024b)	11.25	11.35	25.90	15.68	3.21	5.6 / 27.5	16.02	18.41	78.81
LLaVA-v1.6-7B* (Liu et al., 2024a)	11.84	11.90	34.78	17.69	1.07	5.5 / 27.5	20.60	25.00	84.26
GPT-4o* (OpenAI, 2024)	14.13	17.35	35.01	23.18	4.30	9.1 / 42.9	26.35	28.55	87.62
LISA (Lai et al., 2023)	10.45	15.33	43.07	15.09	1.97	6.2 / 28.4	18.21	26.59	85.67
PixelLM (Ren et al., 2024)	9.87	17.21	35.36	14.68	1.34	6.7 / 30.1	14.93	21.26	85.13
OMG-LLaVA (Zhang et al., 2024)	9.67	16.67	77.80	27.46	8.70	11.2 / 42.0	23.47	27.90	87.30
MIRAS (*Stage-1*)	13.12	15.64	45.11	23.22	4.17	6.9 / 28.4	25.94	28.18	87.54
Fine-tuning									
LISA (Lai et al., 2023)	11.23	**26.23**	29.22	27.64	7.81	14.2 / 40.7	27.84	30.74	86.75
PixelLM (Ren et al., 2024)	10.32	20.95	18.84	11.71	9.97	11.6 / 38.0	30.63	32.99	87.80
OMG-LLaVA (Zhang et al., 2024)	13.84	21.54	49.31	29.98	**11.21**	12.3 / 35.3	30.59	39.18	**88.76**
MIRAS (*Stage-2*)	**14.72**	24.22	40.61	**30.34**	8.51	**15.7 / 49.2**	**30.82**	**40.06**	88.47

Table 3: **Results on Pixel-level RS.** MIRAS employs LLaVA v1.6 as its backbone due to its superior zero-shot performance. * denotes the general MLLMs, others are 7B segmentation-specific MLLMs.

5.3 Experimental Results and Analysis

As shown in Table 3, we comprehensively compare model performance across two critical dimensions: pixel-level segmentation and conversation response. Our proposed MIRAS demonstrates robust performance in pixel-level RS task, surpassing baselines and the closed-source GPT-4o in both segmentation and response metrics, establishing a new benchmark for the PRIST dataset. Notably, the performance improvement from PRIST fine-tuning exhibits universal applicability across different architectures. Table 4 further validates the enhanced reasoning capabilities of our method, with evaluation results approaching human expert proficiency.

5.3.1 Pixel-level Segmentation

As illustrated in the left part of Table 3, GPT-4o achieves CIoU 14.13 and precision 17.35, surpassing open-source models. While it surpasses stage-1 MIRAS, it remains below fine-tuned MIRAS (*stage-2*), indicating our framework can achieve closed-source model competency with efficient resource utilization. To enhance pixel-level performance, all segmentation-specific MLLMs are fine-tuned on the PRIST dataset under a consistent setting. Fine-tuning results in an increase in terms of CIoU and Precision, with OMG-LLaVA's CIoU ↑43%, Precision ↑29%, and LISA's Precision from 15.33 to 26.23 (↑71%), respectively. MIRAS (*stage-2*) establishes new benchmarks with precision 24.22, F1 30.34, and CIoU 14.72, demonstrating exceptional boundary delineation capabilities. Notably, this performance enhancement accompanies a precision-recall trade-off, i.e., recall decreases (LISA ↓32%, OMG-LLaVA ↓37%).

In zero-shot settings, all general MLLMs exhibit limited pixel-level segmentation, with GPT-4o slightly outperforming segmentation-specific MLLMs like PixelLM (Precision 17.35 vs. 17.21). This indicates that general MLLMs emphasize cross-task adaptability, while task-specific improvements rely on design-specific architecture (e.g., Mask Decoder). Additionally, the precision-recall trade-off observed in fine-tuned models reflects a strategic prioritization of segmentation specificity over generalizability in fine-grained tasks, which avoids the overgeneralization issues encountered in zero-shot settings, aligning with the objectives of pixel-level RS. An optimization choice validated by case studies in Appendix D.

5.3.2 Conversation Response

In the right of Table 3, GPT-4o demonstrates near-expert dialogue competence across metrics, approaching fine-tuned MLLMs, such as OMG-LLaVA, while outperforming open-source general MLLMs like Qwen2-VL-7B. MIRAS achieves the best performance in metrics such as Dist-1/2 (15.7/49.2), ROUGE_L (30.82), and METEOR (40.06), validating its ability to generate high-quality textual responses. GPT-4o's strong baseline performance underscores its inherent dialogic intelligence; however, domain adaptation remains essential for optimal performance. Most fine-tuned models show improvements in text metrics, highlighting PRIST's effectiveness in bridging visual-

textual semantic gaps. This demonstrates the framework's dual competence in simultaneously optimizing mask-text alignment and response coherence.

5.3.3 Reasoning Quality

Table 4 presents the reasoning quality evaluation based on LLMs (human scores detailed in Appendix C.4.3). Fine-tuning on the PRIST dataset led to improvements across all models, with an average Win Rate increase of approximately 10%. MIRAS achieved the SOTA with a Win Rate of 42% and the highest scores in all four reasoning metrics, closely approaching human expert levels. The overall improvement in four fine-tuned MLLMs' reasoning quality shows the substantial potential of PRIST dataset in enhancing reasoning capabilities, stemming from the extensive conceptual vocabulary it provides during fine-tuning.

Model	PR	LC	CC	TR	Win Rate(%)
Human	4.03	4.04	4.26	4.28	-
LISA(ft)	3.76	3.69	3.71	3.58	36 (↑11)
PixelLM(ft)	3.35	3.48	3.32	3.28	32 (↑13)
OMG-LLaVA(ft)	2.60	2.48	2.58	2.33	24 (↑8)
MIRAS(*Stage-2*)	**3.90**	**3.76**	**3.83**	**3.69**	**42 (↑11)**

Table 4: Comparison of the reasoning quality of domain-specific MLLMs fine-tuned on PRIST.

5.4 Generalization Segmentation

We compare segmentation-specific baselines on classical referring expression segmentation benchmarks to evaluate the generalizability of MIRAS. As detailed in Table 5, MIRAS' base configuration (last row) outperforms segmentation-specific models and demonstrates competitiveness against other MLLMs, even surpassing the latest OMG-LLaVA. Two findings emerge in results: (1) While fine-tuned MIRAS (*Stage-2*) shows a decline in general performance due to task-specific optimization, it retains advantages over Next-Chat and remains comparable to OMG-LLaVA. (2) The capacity of the base model determines the system's potential, evidenced by consistent improvements in MIRAS when evolving the foundational model from LLaVA-v1 (row 6) to LLaVA-v1.6 (row 9).

Model	refCOCO			refCOCO+			refCOCOg	
	Val	TestA	TestB	Val	TestA	TestB	Val	Test
GRIS (Wang et al., 2022)	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT (Yang et al., 2022)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES (Liu et al., 2023)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
LISA (v1) (Lai et al., 2023)	74.1	76.5	72.3	65.1	70.8	58.1	67.9	70.6
PixelLM (v1) (Ren et al., 2024)	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
MIRAS (*Stage-1*) (v1)	75.3	78.9	70.2	66.7	74.3	61.5	**73.2**	71.9
OMG-LLaVA (Zhang et al., 2024)	78.0	80.3	**74.1**	69.1	73.1	63.0	72.9	**72.9**
MIRAS (*Stage-2*) (v1.6)	76.9	79.8	72.8	68.8	74.4	62.5	71.8	70.8
MIRAS (*Stage-1*) (v1.6)	**78.4**	**80.5**	73.4	**72.1**	**74.8**	**63.4**	72.6	72.0

Table 5: **Results on the RES benchmark.** "v1/v1.6" indicates LLaVA version.

These impressive results are mainly attributed to MIRAS's convolutional backbone (i.e., ConvNeXt), which supports larger input images and enables semantic-assisted mask generation. This provides a solid foundation for achieving fine-grained segmentation in the next stage. However, this focus on task-specific patterns inherently introduces a trade-off, sacrificing some degree of generalization.

5.5 Ablation Study

To validate the effectiveness of the modules in MIRAS, we conduct the following ablation experiments on the general RES task, ensuring compatibility with the following framework by maintaining the base model as LLaVA-v1. **(1) Dual-visual Encoder** Table 6 illustrates that the dual-visual encoder improves performance (*Val* ↓ 2.6%, *Test* ↓ 0.9%) by supporting a higher resolution, which enhances the density of visual features and the ability to capture finer details. **(2) Semantic Region Alignment** The alignment strategy of injecting semantic information has achieved positive results, as shown in Table 6. When applied to the half dataset, it decreases *Val* by 1.4% and *Test* by 0.5%. Reducing the application to the full leads to further decline (*Val* ↓0.7%, *Test* ↓0.4%), highlighting its effectiveness in enhancing segmentation.

Architecture	refCOCOg	
	Val (U)	Test (U)
MIRAS (v1)	**73.2**	**71.9**
w/o Dual-visual Encoder	70.6	70.8
w/o Semantic Region Alignment (50%)	71.8	71.4
w/o Semantic Region Alignment	71.1 (↓)	71.0 (↓)

Table 6: **Ablation.** The metric is CIoU. 50% means half of the samples randomly added semantic information.

6 Conclusion

In this paper, we propose a novel task, Pixel-level Reasoning Segmentation, which focuses on fine-grained segmentation. To further advance, we construct a pixel-level reasoning segmentation dataset, PRIST, consisting of 24k utterances and 8.3k pixel-level segmentation targets, generated through a carefully designed three-stage progressive automatic annotation pipeline. Additionally, we present MIRAS, a framework designed for this task that combines segmentation with multi-turn interaction, along with LLM-based reasoning quality evaluation metrics. Comprehensive experiments on segmentation and reasoning demonstrate the effectiveness of the PRIST dataset and the superior performance of MIRAS, which advances research in pixel-level reasoning segmentation meaningfully.

Limitations

Although our research has achieved certain advancements in the pixel-level RS task, some limitations remain. PRIST is designed only for single-target segmentation, making it difficult to adapt to more complex scenarios, such as those with empty targets (i.e., no objects requiring segmentation) or multiple targets (i.e., simultaneously involving multiple distinct objects). Further exploring reasoning trees to model relationships among image elements and constructing datasets for multi-object, multi-level segmentation hold research potential. Additionally, we utilize the SAM model to efficiently assist MLLM in integrating text reasoning to achieve pixel-level segmentation. However, their integration of separate visual encoding modules creates structural redundancy, reducing efficiency. Developing a streamlined and efficient model architecture is an important direction for future work.

Ethics Statement

Pixel-level reasoning segmentation technology is a double-edged sword. On one hand, it demonstrates immense potential in fields (e.g., medical image analysis, autonomous driving, and intelligent surveillance), contributing to technological advancement and societal development. On the other hand, it is crucial to rigorously guard against risks related to privacy infringement and potential misuse. In our research, we meticulously selected key image features for constructing the training dataset to ensure the safety and representativeness of all samples. We employed an annotation process that aligns closely with ethical values for data collection, aiming to eliminate privacy breaches and the generation of harmful content. Before releasing the PRIST dataset, human experts will rigorously review all annotations and filter out inappropriate or risky data. Furthermore, users must agree to strict licensing terms to govern dataset usage. Importantly, although this technology excels in fine-grained visual understanding, it is not a substitute for human judgment. Its applications must always operate under human supervision, balancing innovation with ethical responsibility.

References

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. m^3 cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2023. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*.

Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. 2022. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE.

Chang Liu, Henghui Ding, and Xudong Jiang. 2023. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. *Preprint*, arXiv:2201.03545.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999.

OpenAI. 2024. Hello gpt-4o.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. 2023. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. *Preprint*, arXiv:2003.12462.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie

Zhou, Yu Qiao, et al. 2024b. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2024. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023a. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2023b. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321.

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

A Segmentation Task

A.1 Referring Expression Segmentation (RES)

The referring expression segmentation (RES) task (Kazemzadeh et al., 2014) involves receiving an image and a natural language expression referring to a specific object in the image (e.g., *"Please segment the apple in the image."*) as input, and then outputting the segmentation mask of that object. As a classic task in the field of semantic segmentation, it intuitively reflects the model's ability in visual localization. Refcoco, Refcoco+, and Refcocog provide mature evaluation benchmarks for this task. To ensure fairness in the comparison, we choose to compare segmentation-specific baseline models and conduct ablation studies of the model architecture on the RES task. This is because all segmentation-specific baselines support this task and are trained on the aforementioned datasets.

A.2 Comparisons of Segmentation-specific MLLMs

Table 7 provides a comprehensive comparison of recent segmentation-specific MLLMs in terms of

model architecture, pixel-level capabilities, and conversation capabilities. A few works (Zhang et al., 2023b; Lai et al., 2023; Ren et al., 2024; Rasheed et al., 2024) integrate specialized vision modules and LMMs, as indicated by the Region Enc. / Dec. The End-End Model (Pi et al., 2023; Chen et al., 2023; Peng et al., 2023) distinction separates models that leverage LMMs for region-level understanding from those employing external modules.

Method	Region Enc. / Dec.	Pixel-Level Seg. / Cap.	Conversation Multi-turn / Reason	End-End
VisionLLM (Wang et al., 2024b)	✗/✗	✗/✗	✗/✗	✗
DetGPT (Pi et al., 2023)	✗/✗	✗/✗	✓/✓	✓
Shikra (Chen et al., 2023)	✗/✗	✗/✓	✗/✗	✓
Kosmos-2 (Peng et al., 2023)	✗/✗	✗/✓	✗/✗	✓
GPT4RoI (Zhang et al., 2023b)	✓/✗	✗/✓	✓/✗	✓
LISA (Lai et al., 2023)	✗/✓	✓/✗	✗/✓	✗
PixelLM (Ren et al., 2024)	✗/✓	✓/✗	✗/✓	✓
GLaMM (Rasheed et al., 2024)	✓/✓	✓/✓	✗/✓	✓
OMG-LLaVA (Zhang et al., 2024)	✗/✓	✓/✓	✗/✓	✓
MIRAS (Ours)	✗/✓	✓/✓	✓/✓	✓

Table 7: Comparison of recent Segmentation-specific MLLMs.

B More Details about PRIST

B.1 Data Annotation

We recruit experts from the Computer Science department as annotators for our project, as they are familiar with the task requirements and objectives. Each annotator is compensated at a rate of 30 euros per hour for their work. Before starting annotation, all annotators underwent training and a small subset of data was pre-annotated, which was only considered a pass if the accuracy rate was at least 80%. The subsequent annotation process was carried out once the pre-annotation results met the required standards. Additionally, We used the open-source tool Labelme[3], which supports precisely outlining objects in arbitrary shapes and extracting masks, meeting the project's needs for accuracy.

Annotation process: To avoid hallucinations and errors in the model-generated dialogues, we enforced strict quality control on the texts generated by GPT-4o. A total of 10 annotators were recruited, with a two-layer pyramid structure employed to reduce subjective bias and enhance quality. First, six annotators were divided into three pairs, each independently annotating different data subsets. Next, an additional annotator per group resolved inconsistencies and checked the quality of consistent labels. Finally, an experienced annotator consolidated and reviewed all data to ensure high quality and consistency. Each annotator's main tasks included generating masks for target objects in images and correcting any commonsense errors in the dialogue content, such as mismatches in time or numbers between text and images.

[3] https://github.com/wkentaro/labelme

Figure 5: Wordcloud of the 200 popular focus-related words in PRIST.

B.2 Data Generation Process

We design a fully automated dataset annotation pipeline, leveraging multiple hierarchical levels in the visual domain to construct the PRIST dataset. The pipeline, entirely based on GPT-4o (*gpt-4o-2024-08-06*), incorporates CoT into a feedback loop to generate relevant multi-turn reasoning dialogues for various images. Each dialogue's final query is a segmentation-related instruction (e.g., *"Please segment the core objects according to the above dialogue"*), making PRIST suitable for pixel-level segmentation tasks and extending to general VQA (when ignoring the final instruction), providing a versatile foundation for multimodal dialogue research. The data is used to fine-tune MLLMs. We will release the PRIST dataset and the implementation of its automated annotation pipeline to support further research. For detailed information on the implementation of LLM prompts and output formats at three levels, refer to Figure 7.

B.3 Dataset Statistics and Analysis

As shown in Table 8, PRIST is compared with existing segmentation and multimodal reasoning benchmarks. PRIST combines the double advantages of pixel-level segmentation and multi-turn reasoning. In comparison with segmentation benchmarks (Yu et al., 2016; Krishna et al., 2017; Lai et al., 2023; Rasheed et al., 2024), the focus is on the granularity and richness of the segmentation targets. Figure 5 illustrates a visual word cloud of the 200 most popular focus-related terms extracted from PRIST. The prominence of each word in the cloud represents its frequency of occurrence in the dataset.

Benchmark	#Img.	#Reg.	#Samp.	Caption	Conversation Single. / Multi.	Segmentation Region. / Pixel.	Reasoning	#Step
Segmentation Benchmark								
RefCOCO (Yu et al., 2016)	20K	142K	142K	✓	✓/✗	✓/✗	✗	-
VG (Krishna et al., 2017)	82.4K	3.8M	3.8M	✓	✓/✗	✓/✗	✗	-
ReasonSeg (Lai et al., 2023)	1.2K	1.2K	1.2K	✓	✓/✗	✓/✗	✓	-
Osprey (Yuan et al., 2024)	100K	503K	724K	✓	✓/✗	✓/✓	✓	-
GranD (Rasheed et al., 2024)	11M	810M	7.5M	✓	✓/✓	✓/✓	✗	-
MUSE (Ren et al., 2024)	246K	910K	246K	✓	✓/✗	✓/✓	✓	-
Multimodal Reasoning Benchmark								
ScienceQA (Lu et al., 2022)	5.6K	-	5.6K	✓	✓/✓	✗/✗	✓	2.5
MMMU (Yue et al., 2024)	11.5K	-	11.5K	✓	✓/✗	✗/✗	✓	1.0
M^3COT (Chen et al., 2024a)	11K	-	11K	✓	✓/✗	✗/✗	✓	10.9
PRIST (Ours)	2.8K	8.3K	8.3K	✓	✓/✓	✓/✓	✓	4.0

Table 8: Comparison of existing segmentation and multimodal reasoning benchmarks. **#X**: the size of X, **Img.:** Image; **Reg.:** Segmentation Regions; **Samp.:** Conversational Samples; **Step:** Steps in the reasoning chain, averaged over all samples.

Compared to existing multimodal reasoning benchmarks (Lu et al., 2022; Zhang et al., 2023c; Chen et al., 2024a), PRIST enables a more in-depth and detailed reasoning process for each segmented object. The average text length per sample is 477 tokens, significantly higher than the 294 tokens in M^3CoT (Chen et al., 2024a). Each conversation contains an average of 4 turns, with a maximum of 7, surpassing ScienceQA's average of 2.5 (Lu et al., 2022). PRIST better simulates real-world interactions and human thought processes while ensuring more coherent and detailed reasoning chains for pixel-level reasoning segmentation, presenting a new challenge for the RS field.

C Experiment Setups

C.1 Implementation Details

Our experiments are all conducted on 2 NVIDIA A6000 (48G) GPUs. Aligned with region-level segmentation models (Rasheed et al., 2024; Zhang et al., 2024), we adopt LLaVAv1.6-7B [4] as the backbone $\mathcal{F}$ and ViT-H SAM [5] to instantiate pixel-level encoder $\mathcal{V}_{\text{pixel}}$ and mask decoder $\mathcal{D}_m$. The visual-language projection layer is implemented through 2-layer MLP and GELU activation following LLaVA-v1.6. The implementation is carried out in PyTorch, with Deepspeed Zero-2 optimization applied during two-stage training, and LoRA fine-tuning with $r = 8$ for the LLM. Different experimental setups are adopted at each training stage to facilitate model convergence. Detailed training configurations are provided in Table 9.

Config	Stage 1	Stage 2
Optimizer	AdamW	
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.95$	
Learning rate schedule	WarmipDecayLR	
Warmup iterations	100	
Weight decay	0	
Gradient accumulation steps	10	
Learning rate	3e-4	1e-5
Batch size	16	32
Training steps	50k	20k

Table 9: The training settings of MIRAS.

Backbone Selection In our framework, we adopt LLaVA v1.6 as the backbone due to its superior zero-shot performance compared to LLaVA v1.5 [6]. However, to ensure a fair comparison with other segmentation-specific MLLMs (Lai et al., 2023; Zhang et al., 2023a, 2024), we also conduct experiments using LLaVA v1.5 as the backbone. As shown in Table 10, the results indicate that the performance difference between the two versions is marginal, suggesting that both backbones are equally effective for our task. This consistency allows us to proceed with the latest LLaVA v1.6 while maintaining comparability with existing methods.

C.2 Two-Stage Training

We employ a two-stage training process to achieve efficient pixel-level reasoning segmentation. The

[4] https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf

[5] https://github.com/facebookresearch/segment-anything

[6] https://huggingface.co/liuhaotian/llava-v1.5-7b

Model	CIoU	Pixel-wise			Response				
		Prec.	Recall	F1	BLEU-4	Dist-1/2	ROU_L.	MET.	BERTS.
Stage-1									
MIRAS(v1.5)	12.35	15.42	49.93	22.53	3.98	7.0 / 28.6	26.84	27.98	87.32
MIRAS(v1.6)	13.12	15.64	45.11	23.22	4.17	6.9 / 28.4	25.94	28.18	87.54
Stage-2									
MIRAS(v1.5)	14.54	23.01	40.59	**31.86**	8.47	14.9 / 49.2	30.66	**41.15**	88.38
MIRAS(v1.6)	**14.72**	**24.22**	**40.61**	30.34	**8.51**	**15.7 / 49.2**	**30.82**	40.06	**88.47**

Table 10: The MLLM backbone selection of MIRAS.

specific details of each stage are as follows:

Stage 1: Mask-Text Alignment Pre-training The objective of this stage is to align mask-based region features with language embeddings. We collect mask-text pairs from various publicly available object-level datasets, including COCO (Chen et al., 2015), Ade20k (Zhou et al., 2019), and Mapillary (Neuhold et al., 2017), as well as region-level datasets like the RefCOCO series (Yu et al., 2016; Kazemzadeh et al., 2014) and PACO (Ramanathan et al., 2023). Additionally, we incorporate VQA and caption data, such as LLaVA-Instruct-80k (Liu et al., 2024a), COCO Caption(Chen et al., 2015). We mix these data in a 9:6:2:2 ratio and convert them into an instruction-following format for training, thereby enhancing its perceptual and conversational abilities.

Stage 2: Pixel-level Segmentation Fine-tuning At this stage, we maintain fixed model weights while fine-tuning our PRIST dataset to enhance fine-grained segmentation and reasoning capabilities. Subsequently, we integrate VQA data (Liu et al., 2024a) to enable MIRAS to follow user instructions (i.e., PRIST combined with VQA at a 4:1 ratio), enhancing its ability to handle complex pixel-level segmentation with precision.

Segmentation Prompt

Output Format:
<box>(x1, y1), (x2, y2)</box>

Please box out the position of focus and output the detection box in <box>(x1, y1), (x2, y2)</box> format, with coordinates representing the top-left and bottom-right corners of the detected area. If no focus is detected, return "no detected", and ensure a blank space follows all outputs.

C.3 General MLLMs Baselines

To evaluate the advanced general MLLMs on our task, we guide them with a unified segmentation prompt. Since these models do not directly generate masks, we first extract the bounding box coordinates and convert them into masks using a consistent function. The generated masks are then compared with ground truth masks. For fairness, we exclude the last round of dialogue in the final evaluation, as the models were not trained with identical segmentation instructions.

C.4 Evaluation

C.4.1 Pixel-level Segmentation Metrics

We adopt traditional methods for calculating precision and recall, applying them to the pixel-wise task. Specifically, each pixel is treated as a binary classification (i.e., 0,1) to quantify the model's ability in pixel-level segmentation.

Pixel-wise Precision measures the proportion of true positive samples among the pixels predicted as positive, reflecting the model's prediction accuracy. Its increase indicates the model can predict the pixels of the target region with higher confidence, thereby capturing the target more precisely.

$$Precision = \frac{TP}{TP + FP}$$

where TP represents pixels correctly predicted as the target, while FP represents non-target pixels incorrectly predicted as the target.

Pixel-wise Recall measures the proportion of actual positive samples that are correctly predicted as positive, reflecting the model's coverage of the target area. Its decrease indicates that the model is more strictly segmenting according to the target boundaries, thereby avoiding overgeneralization.

$$Recall = \frac{TP}{TP + FN}$$

where TP represents the pixels correctly predicted as the target, while FN represents the non-target pixels that are incorrectly predicted as the target.

C.4.2 LLM-based Metrics

Traditional metrics (e.g., BLEU-n, METEOR) often fail to capture deep semantics and logical coherence in dialogue. To address this, we employ GPT-4 as a scoring tool to evaluate the model's reasoning quality using four key metrics: ***1) Progressiveness (PR):*** Whether the current turn effectively sets up and guides the next. ***2) Logical Coherence (LC):*** Whether there is a smooth logical connection between the current and subsequent turns. ***3) Content Consistency (CC):*** Whether the dialogue revolves around the overall topic. ***4) Target Relevance (TR):*** Whether the dialogue stays focused on the target. We carefully designed prompts for each metric to ensure fairness and consistency. To ensure fairness and consistency, we carefully designed prompts for each metric. Ratings are assigned on a 5-point scale divided into three intervals (0-1, 2-3, 4-5), with clearly defined criteria for each range. The final score is calculated as the average across all four metrics, minimizing randomness and bias. Higher scores indicate better performance, reflecting overall improvement across evaluated aspects. GPT-4 strictly adheres to these guidelines, evaluating each dimension step by step. Notably, when using an LLM for scoring, to eliminate randomness, we repeat the scoring for each metric 5 times and calculate the average to determine the final score. Detailed evaluation prompts are shown in Figure 8.

C.4.3 Human Response

We randomly select 100 multi-turn conversations from the PRIST test set and invite three experts in computer science to respond masking the original answers. They all possess solid analytical skills and a deep understanding of model reasoning tasks. To ensure fairness in the evaluation, we also used the LLM-based evaluation criteria to score the three responses (as shown in Table 11). Finally, the average score from the three experts is used as the human evaluation value for each metric.

	PR	LC	CC	TR
Expert 1	4.06	4.10	4.38	4.29
Expert 2	3.98	4.02	4.21	4.31
Expert 3	4.05	4.00	4.19	4.24
Human	4.03	4.04	4.26	4.28

Table 11: The reasoning scores of three human experts based on LLM-based metrics.

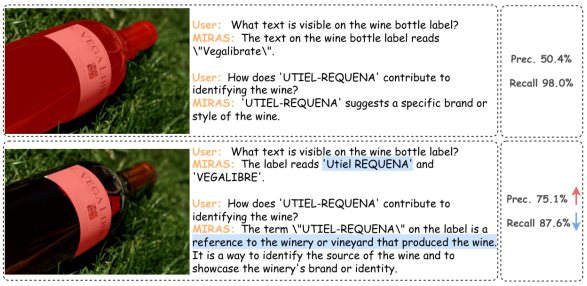


Figure 6: Qualitative results of the performance of MIRAS on Pixel-level RS. **Top:** Stage-1 segments *the entire bottle*. vs. **Bottom:** Stage-2 only segments *the label* and provides more accurate and comprehensive interpretations of *the "UTIEL-REQUENA" label*.

Additionally, we leverage GPT-4 as a judge to design an adversarial evaluation framework, using **Win Rate (%)** to assess reasoning quality in multi-turn dialogues. Human responses serve as the standard for model output comparison, the model is considered to achieve a win when its score surpasses the human response score.

D Case Study

To reflect the high quality of our constructed dataset, we present examples of PRIST test data in Figure 9, which effectively combine coherent and logically structured multi-turn conversations with fine-grained segmentation. We fine-tuned MIRAS on the PRIST dataset to enhance its understanding and localization of image details. As shown in Figure 6, the fine-tuned MIRAS (*Stage-2*) segments only *the label*, replacing *the entire bottle* in Stage-1, demonstrating stronger segmentation specificity (with a significant improvement in precision) while effectively mitigating the issue of overgeneralization (with a slight decrease in recall). This result also highlights the potential of PRIST in improving pixel-level reasoning capabilities.

To more intuitively demonstrate the advantages of our model in the pixel-level reasoning segmentation task, we also present several cases of user interactions with the chatbot. As shown in Figure 10, these cases illustrate MIRAS's coherence and consistency in multi-turn interactions and showcase its ability to achieve more precise fine-grained object segmentation.

 # Role: Visual Extractor
Requirements and Goals:
1. identify and list the elements visible in an image. \
2. Provide detailed descriptions of each element, with text, numbers, and other information.
3. List at least 3 elements visible in the image.
Output Format:

```
{"elements": [{
"name": "The name of the element1",
"text": "Text description of the element1",
"number": (optional)"Number of the element",
"caption": "Caption of the element"},
...]}
```

(a) Illustration of the visual elements extraction (step-1).

 # Role: Visual Reasoning Expert
Scoring Criteria:
1 Point (Avoid these): Abstract or vague questions, or too simple or too complex (not intuitive).\
2 Points: Direct questions that can be answered through basic reasoning based on visible elements. \
3 Points: Clear questions that require reasoning using details and commonsense, focusing on specific objects.
Requirements and Goals:
1. Design 3 reasoning questions based on the listed elements and details from the image and caption, with scores of 2-3 points.
2. Each question must involve reasoning and use common sense to analyze visible elements.
3. Display a reasoning tree structure: Present the reasoning tree structure showing the progression of reasoning for each question. Highlight any overlapping nodes or steps among the three questions to reflect common reasoning pathways.
4. Output the corresponding reasoning questions for each of the three paths.
Output Format:

```
{"root_node": "Root Node A",
 "levels": ["First Level: Level 1",
            "Second Level: Level 2"],
 "inference_nodes": [{
          "node_name": "Inference Node 1",
          "overlap": "Overlap of Path 1 and 3",    // "Overlap of Path x and y" or "Path x Independent"
          "reasoning_paths": [{
                    "path_name": "Path 1",
                    "path_description": "Reasoning Path 1 Description",
                    "reasoning_question": "Reasoning Question for Path 1" }],
                ...
                    },{
                   "node_name": "Inference Node n",
                   "overlap": "Path 2 Independent",
                   "reasoning_paths": [{
                             "path_name": "Path 2",
                             "path_description": "Reasoning Path 2 Description",
                             "reasoning_question": "Reasoning Question for Path 2" }]
                    }]
}
```

(b) Illustration of the reasoning tree construction (step-2).

 # Role: Multi-Turn Reasoning Dialogue Generator
Requirements and Goals:
1. Each dialogue must consist of 4 to 8 turns of questions and answers.
2. Each question should focus on a single, specific object visible in the image, with reasoning directly tied to the elements listed in the title and the image details.
3. Expand each reasoning path into a multi-round dialogue format according to the provided reasoning tree, demonstrating the logical progression of thought based on the identified paths.
Output Format:

```
{"Question1": {
      "Q1": "<First Question>",
      "A1": "<Answer to First Question>",
      "Q2": "<Second Question>",
      "A2": "<Answer to Second Question>",
      "Q3": "<Third Question>",
      "A3": "<Answer to Third Question>",
      ...},
 "Focus": "<Final Focus Object1>"},
...
}
```

(c) Illustration of the multi-turn dialogue generation (step-3).

Figure 7: The prompts and output formats of our dataset annotation pipeline.

 # Role: A Multiturn Dialogue Reasoning Evaluator

Requirements:

1. Please analyze the following multi-turn conversation and score it based on the following dimensions:
   - Progression: Does the conversation build logically from one turn to the next? Is the topic being developed effectively?
   - Logical Coherence: Is there a clear and natural logical connection between each turn in the conversation?
   - Content Consistency: Does each turn align with the overall topic and conversation goal?
   - Focal Goal Relevance: Does each turn focus on and contribute towards achieving the focal goal of the conversation?

   After providing scores for each dimension, include a summary that captures the strengths and weaknesses of the conversation.
2. Only output the reasoning scores in the above format.
3. Please evaluate the entire process to the content of the image.
4. Please focus on linking the dimensions between each round of conversation.

Evaluation Criteria:

1. Progression Score:
   4-5 points: Smoothly transitions to the next turn with sufficient setup for subsequent content.
   2-3 points: Partially advances the dialogue, but the connection feels somewhat awkward.
   0-1 points: Fails to advance the dialogue or is disconnected from the next turn.
2. Logical Coherence Score:
   4-5 points: Perfectly connected with rigorous logic.
   2-3 points: Generally clear logic, but with minor gaps or weak links.
   0-1 points: No logical connection or logically incoherent.
3. Content Consistency Score:
   4-5 points: Content is fully aligned and tightly focused on the goal.
   2-3 points: Generally consistent, but with a slight deviation.
   0-1 points: Deviates from the topic, losing the direction of the dialogue.
4. Focal Goal Relevance Score:
   4-5 points: Highly relevant and continuously advances the goal.
   2-3 points: Somewhat relevant, but with limited contribution.
   0-1 points: Deviates from the goal, failing to advance it.

Output in JSON Format:

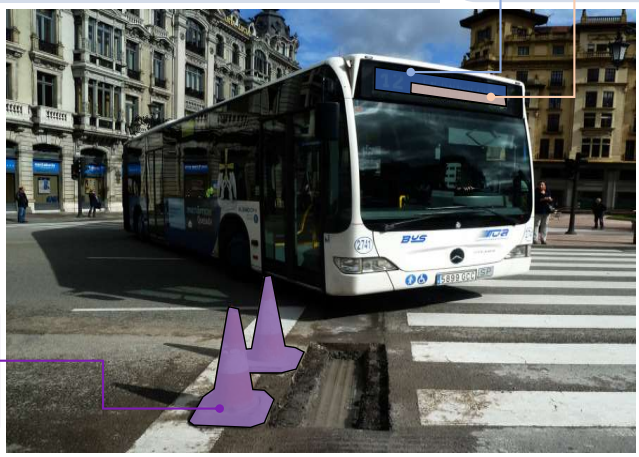
Your output should be a JSON object containing scores for each dimension, as well as the total score and a brief summary of the evaluation.

```
{
  "scores":  [
      {"name": "progression_score", "score": 4.5},
      {"name": "logical_coherence_score", "score": 4.0},
      {"name": "content_consistency_score", "score": 4.2},
      {"name": "focal_goal_relevance_score", "score": 4.8},
      {"name": "total_score", "score": 17.5},
    ],
  "summary": "The conversation demonstrates excellent progression and relevance to the focal goal. \
              However, there are some minor coherence issues and slight deviations from the main topic."
}
```

Figure 8: Illustration of the LLM-based evaluation of reasoning quality.

Q1: What information is displayed on the bus's front panel?
A1: The front of the bus displays the route number '12' and the destination 'CARISA-HOSPITAL'.
Q2: How can the route number and destination help identify the bus's route?
A2: The route number '12' and the destination 'CARISA-HOSPITAL' indicate the specific path the bus takes, which is directed towards the Carisa Hospital.
Q3: Using the conversation above, create a segmentation that validates the answers.
A3: The result is [OBJ]bus front display[SEG].

Name: Bus
Text: A white bus on route number 12 with 'PARQUE PRINCIPADO - CAU SA-HOSPITAL' displayed on the front panel. It has logos for ALSA and BYS on it and is crossing a zebra crossing.
Caption: White bus with route information crossing the street.
Number: 12



Name: Traffic cones
Text: Two orange-and-white traffic cones placed on the road near a road construction or manhole.
Caption: Orange-and-white traffic cones at a construction site.
Number: 2

Q1: What elements are present near the bus that suggest a road safety feature?
A1: Traffic cones are present near the bus, suggesting a road safety feature.
Q2: How do these traffic cones indicate an area of maintenance or hazard?
A2: Traffic cones are typically used to signal areas under maintenance or potential hazards, warning pedestrians and drivers to be cautious.
Q3: Please use the conversation above to outline a segmentation that aligns with the answers provided.
A3: The result is [OBJ]traffic cones[SEG].

Q1: What text is visible on the bus that indicates its destination?
A1: The text 'CARISA-HOSPITAL' is visible on the bus, which indicates its destination.
Q2: Why is this information important for passengers?
A2: This information helps passengers identify the correct bus that will take them to Carisa Hospital.
Q3: Segment the provided conversation to match the answers accordingly.
A3: The segmentation is [OBJ]text 'CARISA-HOSPITAL'[SEG].

Name: 10, 20 Euro Cent Coin
Text: Two gold colored Euro cent coins with an image of Europe and the values are 10 and 20.
Caption: Two Euro cent coins of 10 and 20.
Number: 2

Q1: What numbers are prominently displayed on each gold-colored coin?
A1: The gold-colored coins have the numbers 20 and 10 prominently displayed.
Q2: Which unique design element helps identify the value of these coins?
A2: The specific numbers and the Euro cent symbol help identify their value.
Q3: How does the unique design confirm the coin values?
A3: The design and numbers indicate the coins are 20 and 10 Euro cents.
Q4: Please segment the conversation above to align with the provided answers.
A4: Segmented as [OBJ]gold-colored coins with numbers 20 and 10[SEG].

Name: 2 Euro Cent Coin
Text: Two copper colored 2 Euro cent coins showing the number 2 and an image of Europe.
Caption: 2 Euro Cent Coin
Number: 2

Name: 1 Euro Cent Coin
Text: A copper colored 1 Euro cent coin showing the number 1 and a design of Europe.
Caption: 1 Euro Cent Coin
Number: 1

Q1: What is displayed on the larger silver-colored coin?
A1: The number 2 is prominently displayed on the largest silver-colored coin.
Q2: What does this number indicate regarding the coin's value?
A2: It indicates that the coin has a value of 2 Euros.
Q3: How does the prominent number assist in determining the value across all coins?
A3: The prominent number signifies the denomination directly, aiding in value identification.
Q4: Use the above conversation to create a segmentation that justifies the answers.
A4: The segmentation is [OBJ] 2 Euro silver-colored coin[SEG].

Q1: What color are the smallest coins visible on the table?
A1: The smallest coins are copper-colored.
Q2: How does the size of these copper coins relate to their value?
A2: Typically, smaller copper coins have lower values.
Q3: Which specific copper coin is the smallest in size?
A3: The 1-cent coin is the smallest copper-colored coin.
Q4: Please divide the conversation above to correspond with the given answers.
A4: Segmented as [OBJ] 1-cent copper coin[SEG].

Figure 9: **The samples of PRIST.** The figure displays two examples from the PRIST dataset, created through the automated annotation pipeline. It offers a range of semantic labels and attributes for the identified objects, as well as contextualized multi-turn conversations.

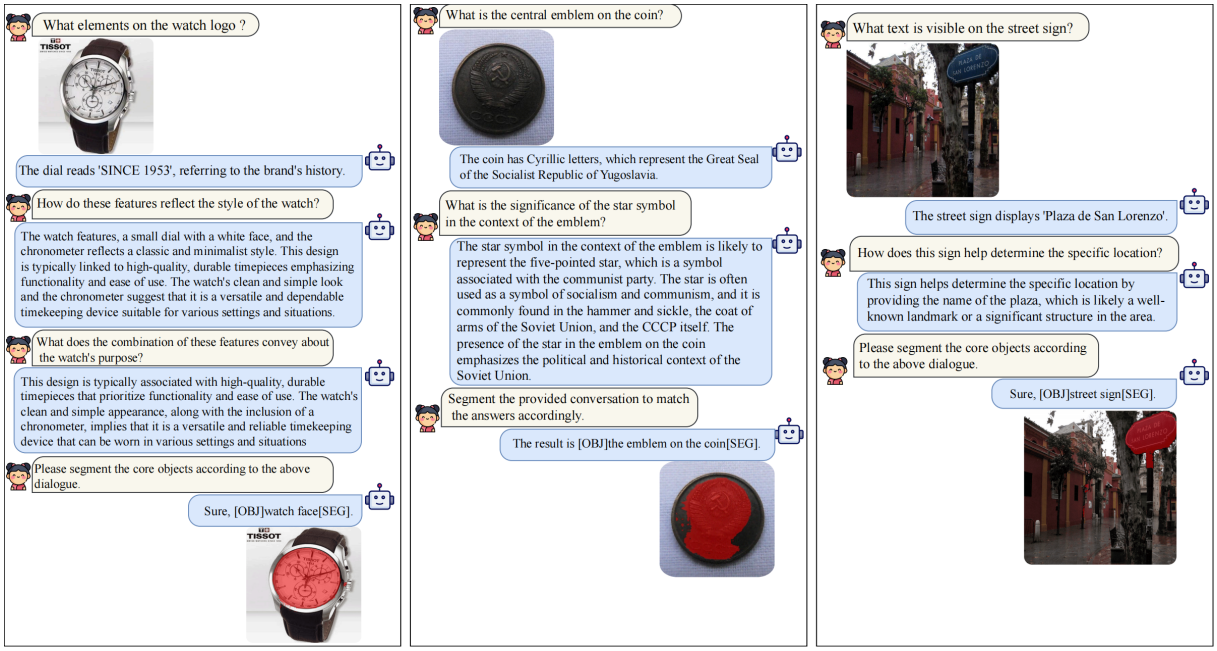


Figure 10: Pixel-level segmentation and multi-turn conversational interactions facilitated by MIRAS. **Left:** The model segments only the watch face instead of the entire watch, and identifies the three sub-dials and the "SINCE 1953" logo. **Center:** The model segments only the emblem on the coin instead of the whole coin, and infers that it is related to Soviet communism. **Right:** The model segments only the road sign, and accurately recognizes the text "Plaza de San Lorenzo".