Table 1. Examples of task instruction following formats.

Task	Two randomly chosen examples from many
ROC	What is the anatomical structure located at the coordinates <box>? What is the object located at the coordinates <box>?
RC	Can you describe the content within the coordinates <box>? Can you describe what is seen at the coordinates <box>?
VG	Where is the <object> located in the image? What are the coordinates for the given <object> in the image?
MIA	Based on the image, what can you infer about the overall medical condition? How would you interpret the overall medical condition shown in the image?

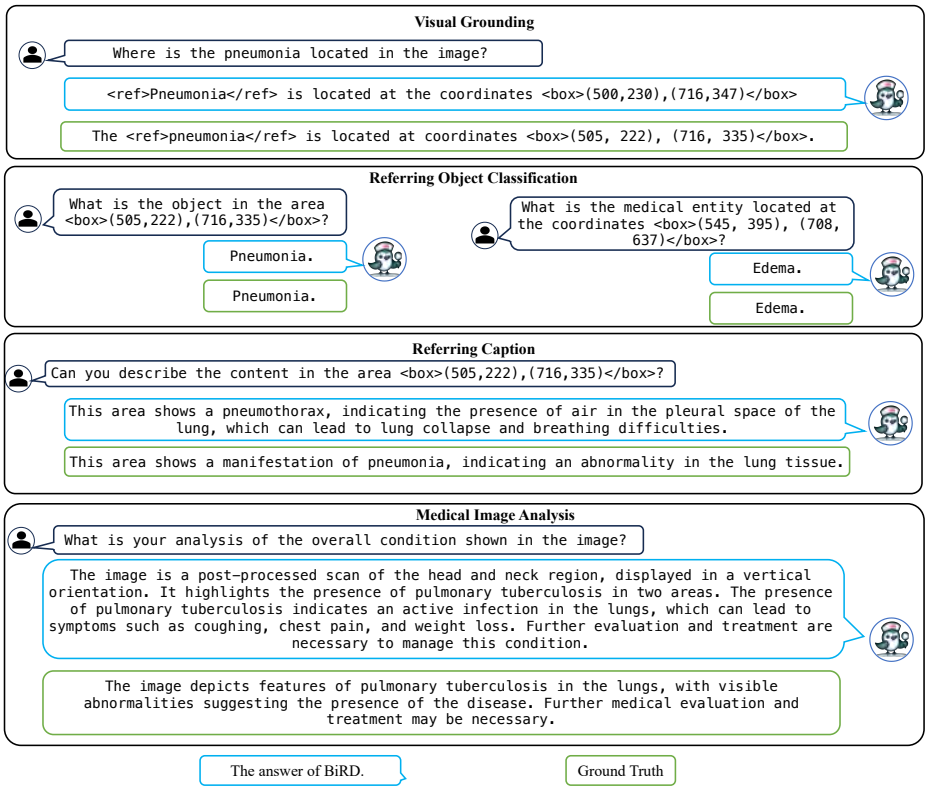

Fig. 1. More visualization examples of the BiRD model's performance across various tasks. Please note, due to space constraints, the actual images inputted during testing are not displayed in the figures.

```
messages = [ {"role":"system", "content": f"""You are an AI visual assistant capable of analyzing a
single medical image. Suppose you can see the given image, as you will receive 1 Global_caption,
describing the contents observed within the image. Additionally, Objects are provided, indicating the
specific entities' locations and their detailed coordinates within the image. Coordinates are
represented as bounding boxes (x1, y1, x2, y2) and are normalized to a 0 to 1 scale based on the
original image size. Each entity is indexed by the name of the entity.

Your task is to generate multiple dialogues between the person asking about the image (User) and you
(Assistant) answering their questions.
To answer such questions, you first need to understand the visual content in the medical image and
respond based on medical background knowledge or reasoning. Increase the challenge of the
questions by not including detailed information about the visual content in the questions.

Generate answers for the following four types of questions:
 1. Visual Grounding (VG): The user's question must ask for the location information of a certain
entity or type of entity. There are two types of answers: a. If the entity does not exist, directly answer
that it does not exist; b. If the entity exists, the entity's name in the answer must be formatted as
<ref>entity name</ref>, and the corresponding coordinates must be indicated.

  2. Referring-Object (RO): The user's question must use the given entity's coordinates to ask about
the category of this area, without mentioning any entity names. The answer must only contain the
word for the entity name of this area, without any other words.

  3. Grounded Captioning (GC): The user's question must use the existing entity's coordinates to ask
for a simple description of the content within these coordinates. The answer should provide a brief
description based on the entity's relevant attributes and visual features, without including any
coordinates.

  4.Medical Imaging Interpretation (MII): The question asks for a diagnosis of the entire image.
Answer: Provide a comprehensive diagnostic analysis of the entire image based on the
"Global_caption."

Please note all questions and answers should not mention the source of information, always respond
as if you are directly looking at the image, and return the results in English."""} ]

for sample in fewshot_samples:
    messages.append({"role":"user", "content":sample['context']})
    messages.append({"role":"assistant", "content":sample['response']})
messages.append({"role":"user", "content":query})
```

Fig. 2. In this example, we provide the prompt used to generate the refer-and-ground instruction tuning data.