Supplementary Material:Zero-Shot Vision Language Reasoning via Dual-layer Scene Graph Chain of Thoughts

Overview

These are the exact prompts used in our experiments for generating scene graphs and answering questions based on images. The prompts ensure relevance, accuracy, and structured output for various tasks, including scene graph generation, reasoning, and multiple-choice answering.

1 Only Object Prompt

USER: `<image>` For the provided image and its associated question, generate a object list in JSON format that includes the following:

1. Objects that are relevant to answering the question.
2. Object attributes that are relevant to answering the question.
3. Object relationships that are relevant to answering the question.

Question: "{question}"

Focus on accuracy and relevance to the question.

Object List (in JSON format only):

ASSISTANT:

2 Global Scene Graph Prompt

USER: `<image>` For the provided image, generate a comprehensive global scene graph in JSON format that includes the following:

1. Objects that are present in the image with their confidence scores.
2. Object attributes in the image.
3. Object relationships: Spatial and semantic relationships between objects with confidence scores.

Do not be repetitive. Focus on accuracy and avoid hallucination. Only include objects and relationships that are clearly visible in the image. Provide confidence scores between 0.0 and 1.0 for all objects and relationships.

Scene Graph (JSON format only):

ASSISTANT:

3 Query Scene Graph Prompt

USER: <image>
Analyze the image and generate a scene graph in valid JSON format for the question below.

Available Objects: {readable_objects}
Question: "{question}"
Generate a JSON response with this exact structure:

```
{
    "objects": [
        {"name": "object_name", "attributes": ["attr1", "attr2"]}
    ],
    "relationships": [
        {"subject": "obj1", "predicate": "relation", "object": "
           obj2"}
    ]
}
```

Focus only on objects and relationships that help answer the question. Ensure the JSON is valid and complete.

ASSISTANT:

4 MM Bench: Final Answer Generation Prompt

```
choice_text = f"""
A. {choices['choice_a']}
B. {choices['choice_b']}
C. {choices['choice_c']}
D. {choices['choice_d']}
"""
```

```
prompt = f"""USER: <image>
Global Scene Graph: {global_sg}
Query Specific Scene Graph: {query_sg}
Question: {question}

Choices:{choice_text}

Instructions: Analyze the image and both the scene graphs to
   answer the multiple choice question.The global scene graph
   gives the overall context of the image and the query specific
   scene graph gives context relevant for answering the question.
    Answer with only the letter (A, B, C, or D) of the correct
   choice which is the most appropriate.

A:"""
```

Dataset: MMBench
Number of samples used: 1,000

Original dataset size: Approximately 3,000 samples
Evaluation metric: Accuracy (MCQWS)

5 SEED Images: Final Answer Generation Prompt

```
choice_text = f"""
A. {choices['choice_a']}
B. {choices['choice_b']}
C. {choices['choice_c']}
D. {choices['choice_d']}
"""
```

```
prompt = f"""USER: <image>
Global Scene Graph: {global_sg}
Query Specific Scene Graph: {query_sg}
Question: {question}

Choices:{choice_text}

Instructions: Analyze the image and both the scene graphs to
   answer the multiple choice question.The global scene graph
   gives the overall context of the image and the query specific
   scene graph gives context relevant for answering the question.
    Answer with only the letter (A, B, C, or D) of the correct
   choice which is the most appropriate.

A:"""
```

Dataset: SEED
Number of samples used: 1,000
Original dataset size: Approximately 10,000 samples
Evaluation metric: Accuracy (MCQWS)

6 WHOOPS!: Final Answer Generation Prompt

```
extraction="Answer with ONLY one or two words.Do not provide
   explainations or full sentences"
```

```
prompt = f"""USER: <image>
Global Scene Graph: {global_sg}
Query Specific Scene Graph: {query_sg}
Question: {question}



Instructions:Use both the query and global scene graphs to answer
    the question about the image.
{extraction}
```

```
A:"""
```

Dataset: WHOOPS!
Number of samples used: Full dataset(500 images)
Evaluation metric: BEM Score

Additional Notes on Ablation Studies

For the final answer generation task, in addition to using both the Global Scene Graph and Query Scene Graph as inputs, we also conducted ablation experiments where:

1. Only the Query Scene Graph was used while omitting the Global Scene Graph.
2. Neither the Global Scene Graph nor the Query Scene Graph was used, testing the model's ability to reason from image content alone.

These variants are not included in the main prompts above but were part of our evaluation to analyze the contribution of structured scene information to reasoning performance.