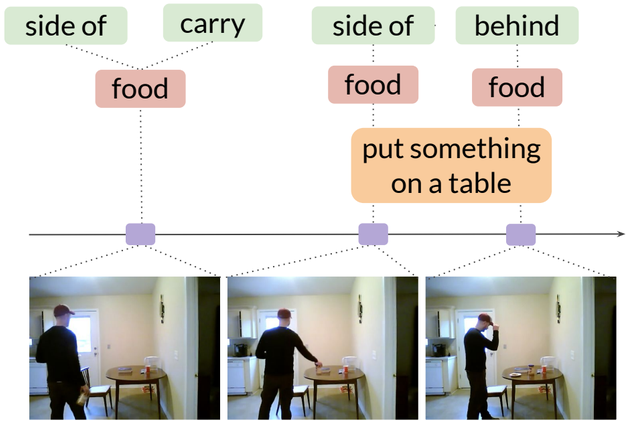
Input

Spatio-temporal scene graph



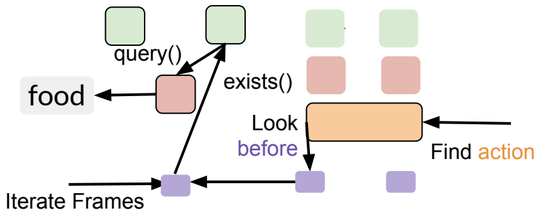
Video

Our benchmark generation process

Template1: What did they <relation><time><action>?

Program 1:

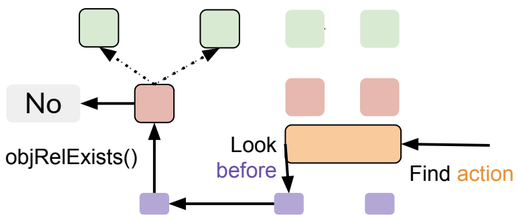
```
for frame before put something on a table:
  if exists(carry):
    return query(object)
```



Template2: Did they <relation> <object> <time><action>?

Program 2:

```
for frame before put something on a table:
  if objRelExists(food, behind):
    then "Yes"
else: "No"
```



Output

Question-Answer 1:
What did they carry before putting something on a table? - food

Alternate action reference:
What did they carry before the shortest action? - food

Question-Answer 2:
Did they go behind some food before putting something on a table? - No

Alternate object reference:
Did they go behind the object they carried before putting something on a table? - No

Legend: objects relationships actions time