Data collection

① LLM inference for VQA model

From reflection: in solving the compositional question 'Is there a horse that is not white?', the VQA model in step3 incorrectly answer the subquestion 'What color of the horse?'

Prompt
1. Instruction
2. Desirable answer
3. Program of all steps
4. Intermediate result of all steps

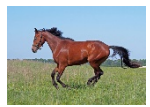
Correct answer is white.

② Collect data from open-vocabulary dataset for the LOC and SEG models

From reflection: the LOC/SEG model in step3 fails for the object 'horse'

③ Search on the Internet for the SELECT, REPLACE, and CLASSIFY models

From reflection: the SELECT/REPLACE/CLASSIFY models fails for the object 'horse'


'horse'


'horse'

Training

① Prompt tuning for LLMs

Correct Examples for Plan

Instruction: Is there a cow or a horse that is not white?
Plan:
Step1, Locate cow in the image.
Step2, Locate horse in the image.
Step3, Crop the image region of cow.
Step4, Crop the image region of horse.
Step5, Ask 'What the color of the cow?' for the crop image in Step3.
Step6, Ask 'What the color of the horse?' for the crop image in Step4.
Step7, Obtain the answer based on the color of cow and horse.
Step8, Output the answer.

Incorrect Examples for Plan

Instruction: Is there a cow or a horse that is not white?
Plan:
Step1, Locate cow
Step2, Locate horse
Step3, Count the number of cow
Step4, Count the number of horse
Step5, Obtain the answer based on the number of cow and horse
Step6, Output the answer
Reason: In Step3 and Step4, it should ask the color of the horse and cow, instead of counting the number

Correct Examples for Program

Instruction: Is there a cow or a horse that is not white?
Plan:
Step1, Locate cow in the image.
Step2, Locate horse in the image.
Step3, Crop the image region of cow.
Step4, Crop the image region of horse.
Step5, Ask 'What the color of the cow?' for the crop image in Step3.
Step6, Ask 'What the color of the horse?' for the crop image in Step4.
Step7, Obtain the answer based on the color of cow and horse.
Step8, Output the answer.
Program:
BOX0=LOC(image=IMAGE,object='cow')
BOX1=LOC(image=IMAGE,object='horse')
IMAGE0=CROP(image=IMAGE,box=BOX0)
IMAGE1=CROP(image=IMAGE,box=BOX1)
ANSWER0=VQA(image=IMAGE0, question='What the color of the cow')
ANSWER1=VQA(image=IMAGE1, question='What the color of the horse')
ANSWER2=EVAL(expr="'yes' if {ANSWER0}!='white' else 'no' ")
FINAL_RESULT=RESULT(var=ANSWER2)

Incorrect Examples for Program

Instruction: Is there a cow or a horse that is not white?
Plan:
Step1, Locate cow in the image.
Step2, Locate horse in the image.
Step3, Crop the image region of cow.
Step4, Crop the image region of horse.
Step5, Ask 'What the color of the cow?' for the crop image in Step3.
Step6, Ask 'What the color of the horse?' for the crop image in Step4.
Step7, Obtain the answer based on the color of cow and horse.
Step8, Output the answer.
Program:
BOX0=LOC(image=IMAGE,object='cow')
BOX1=LOC(image=IMAGE,object='horse')
IMAGE0=CROP(image=IMAGE,box=BOX0)
IMAGE1=CROP(image=IMAGE,box=BOX1)
ANSWER0=VQA(image=IMAGE0, question='What the color of the cow')
ANSWER1=VQA(image=IMAGE1, question='What the color of the horse')
ANSWER2=EVAL(expr="'yes' if {ANSWER0}!='white' else 'no' ")
FINAL_RESULT=RESULT(var=ANSWER2)
Reason: In Step7 of the Program, it does not consider the color of horse. It should be ANSWER2=EVAL(expr="'yes' if {ANSWER0}!='white' or {ANSWER1}!='white' else 'no' ").

Demosntration Pool

$\mathcal{D}$

② Prompt tuning for VFMs

Prompt Pool


prompt + CLIP → Loss → Horse: $\mathcal{P}$
BP


prompt + Maskerformer → Loss → Horse: $\mathcal{P}$
BP


prompt + OWL → Loss → Horse: $\mathcal{P}$
BP

Validation

① Prompt validation for LLMs

Demosntration pool

$\mathcal{D}$

prompt

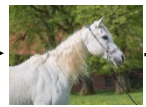
Is there a horse that is not white?

or
- Keep the prompt
- Remove the prompt

② Prompt validation for VFMs

Prompt Pool

$\mathcal{P}$


Validation data

Model

or
- Keep the prompt
- Remove the prompt