CLOVA: A Closed-Loop Visual Assistant with Tool Usage and Update

Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, Qing Li

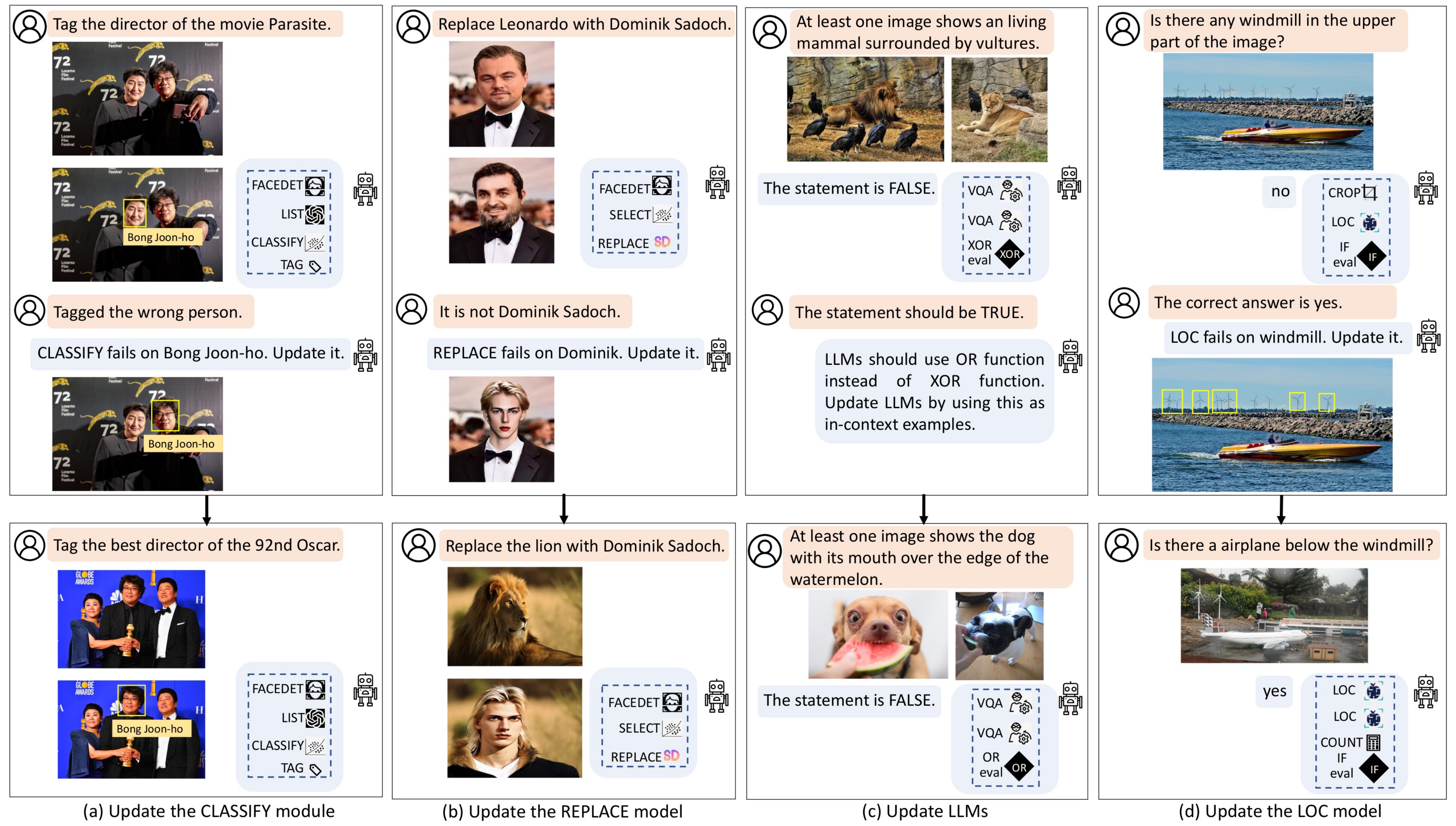
北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

Project

CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction

Motivation: Can VLM improve from user's feedback?


Tag the director of the movie Parasite.

Tagged the wrong person.
CLASSIFY fails on Bong Joon-ho. Update it.
Tag the best director of the 92nd Oscar.

(a) Update the CLASSIFY module

Replace Leonardo with Dominik Sadoch.

It is not Dominik Sadoch.
REPLACE fails on Dominik. Update it.
Replace the lion with Dominik Sadoch.

(b) Update the REPLACE model

At least one image shows an living mammal surrounded by vultures.
The statement is FALSE.

The statement should be TRUE.
LLMs should use OR function instead of XOR function. Update LLMs by using this as in-context examples.
At least one image shows the dog with its mouth over the edge of the watermelon.
The statement is FALSE.

(c) Update LLMs

Is there any windmill in the upper part of the image?
no

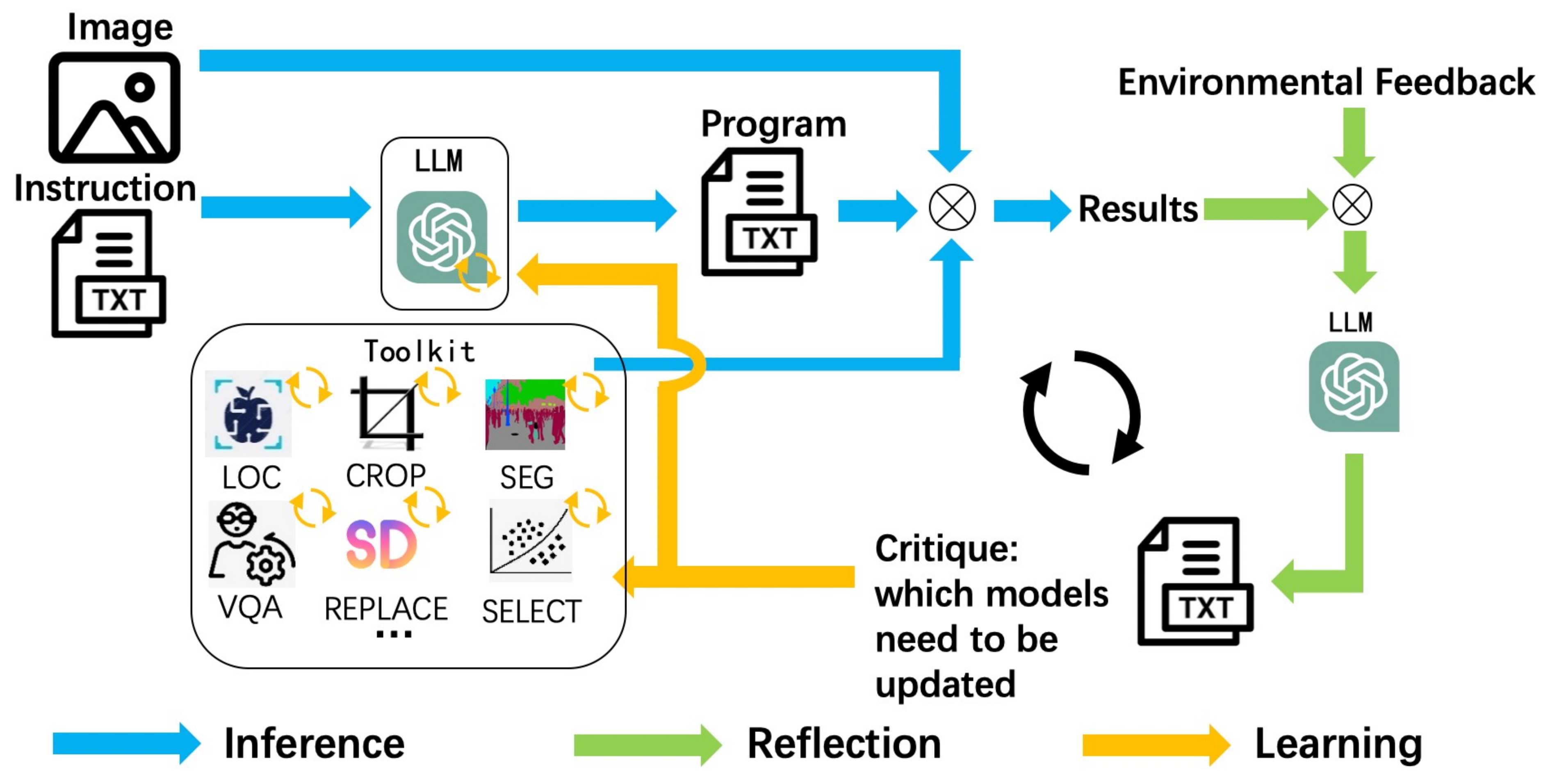
The correct answer is yes.
LOC fails on windmill. Update it.
Is there a airplane below the windmill?
yes

(d) Update the LOC model

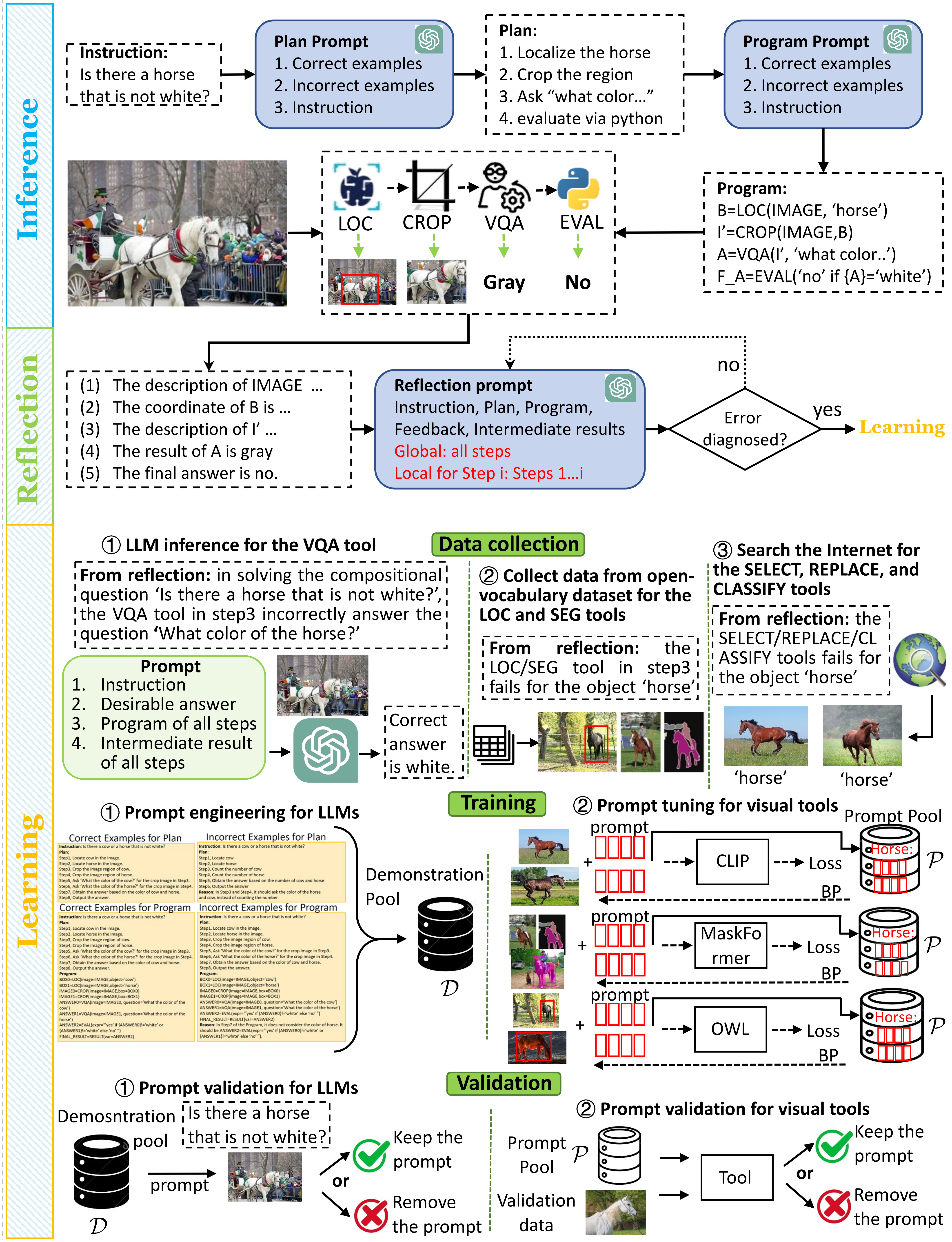
Challenges: (1) how to diagnose errors? (2) how to collect training data? (3) how to efficiently update the model?

Proposal: **CLOVA** = **Inference** + **Reflection** + **Learning**

❑ **improve from feedback** via a closed-loop learning framework



CLOVA = Inference + Reflection + Learning



Instruction: Is there a horse that is not white?

Plan Prompt
1. Correct examples
2. Incorrect examples
3. Instruction

Plan:
1. Localize the horse
2. Crop the region
3. Ask "what color…"
4. evaluate via python

Program Prompt
1. Correct examples
2. Incorrect examples
3. Instruction

Program:
```
B=LOC(IMAGE, 'horse')
I'=CROP(IMAGE,B)
A=VQA(I', 'what color..')
F_A=EVAL('no' if {A}='white')
```

(1) The description of IMAGE …
(2) The coordinate of B is …
(3) The description of I' …
(4) The result of A is gray
(5) The final answer is no.

Reflection prompt
Instruction, Plan, Program, Feedback, Intermediate results
Global: all steps
Local for Step i: Steps 1…i

Data collection

① LLM inference for the VQA tool

From reflection: in solving the compositional question 'Is there a horse that is not white?', the VQA tool in step3 incorrectly answer the question '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 tools

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

③ Search the Internet for the SELECT, REPLACE, and CLASSIFY tools

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

Training

① Prompt engineering 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' or {ANSWER1}!='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' ").

② Prompt tuning for visual tools

Validation

① Prompt validation for LLMs

② Prompt validation for visual tools

Experiments

Main results

	Method	GQA	NLVRv2	Editing	Tagging
E2E	Otter [27]	48.2	48.2	-	-
	MMICL [83]	64.4	62.2	-	-
Tool	GPT4TOOLs [75]	41.2	45.4	17.8	
	Visual ChatGPT [75]	43.2	51.6	21.7	-
	InternGPT [40]	44.8	39.4	-	-
	HuggingGPT [62]	46.0	44.0	-	-
	ViperGPT [67]	47.2	-	-	-
	VISPROG [11]	49.8	60.8	40.2	0.393
	CLOVA (Ours)	**54.6**	**65.6**	**65.4**	**0.502**

Ablation studies

	Method	GQA	NLVRv2
Reflection	w/o local reflection	52.0	65.2
	w/o global reflection	53.6	64.2
	w/o intermediate results	48.8	61.2
	w/o plan	50.0	62.6
	Ours	**54.6**	**65.6**
Prompt Engineering for LLMs	w/o incorrect cases	46.1	61.4
	w/o correct cases	48.2	63.2
	w/o validation	44.2	61.0
	Ours	**54.6**	**65.6**
Prompt Tuning for visual tools	w/o validation	42.8	62.8
	Ours	**54.6**	**65.6**

Different LLMs

Dataset	Method	LLaMA2-7B	GPT-3.5	GPT-4
GQA	Baseline	39.2	46.4	52.6
	+ Update LLMs	56.8	51.6	56.6
	+ Update visual tools	60.2	54.6	60.4
NLVRv2	Baseline	50.0	60.2	64.8
	+ Update LLMs	59.2	63.6	68.8
	+ Update visual tools	63.8	65.6	69.2

Method	GQA	NLVRv2	Editing	Tagging
LLama2-7B	39.2	50.0	31.2	0.308
LLama2-7B + Ours	**60.2**	**63.8**	**47.6**	**0.357**
Mistral-7B	20.4	34.6	29.0	0.205
Mistral-7B + Ours	**31.4**	**42.2**	**46.5**	**0.303**

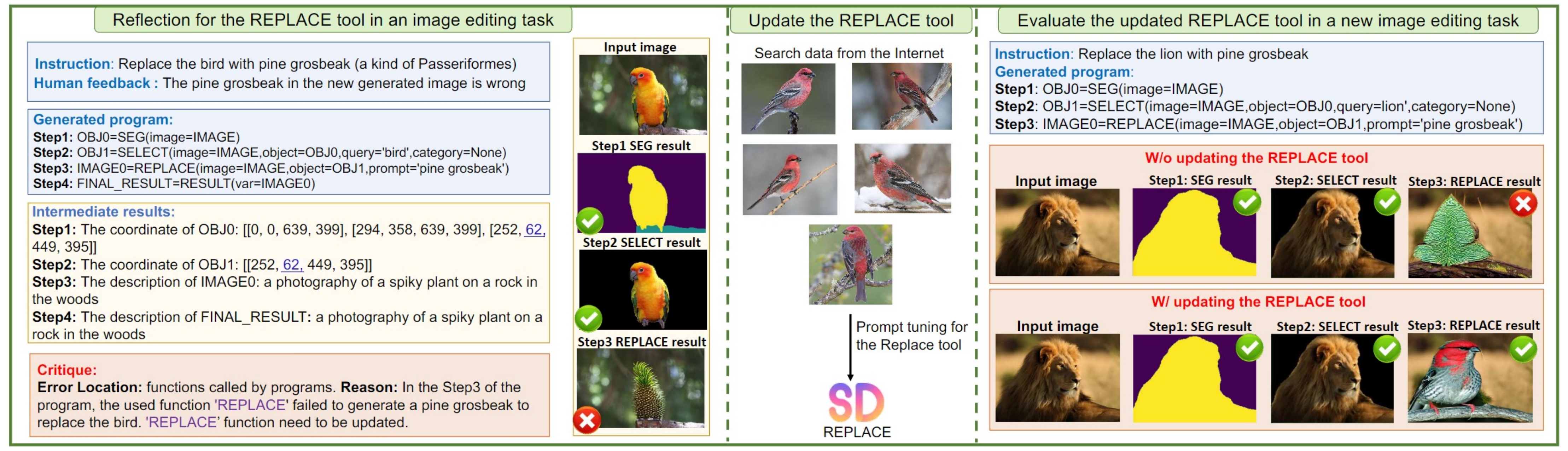
Quantitative observation:

❑ CLOVA achieves SOTA among tool-usage VLMs.

❑ CLOVA is robust to different LLMs, including open and closed ones.

❑ Update both LLM and visual tools bring significant improvements.

Qualitative example


Reflection for the REPLACE tool in an image editing task

Instruction: Replace the bird with pine grosbeak (a kind of Passeriformes)
Human feedback : The pine grosbeak in the new generated image is wrong

Generated program:
Step1: OBJ0=SEG(image=IMAGE)
Step2: OBJ1=SELECT(image=IMAGE,object=OBJ0,query='bird',category=None)
Step3: IMAGE0=REPLACE(image=IMAGE,object=OBJ1,prompt='pine grosbeak')
Step4: FINAL_RESULT=RESULT(var=IMAGE0)

Intermediate results:
Step1: The coordinate of OBJ0: [[0, 0, 639, 399], [294, 358, 639, 399], [252, 62, 449, 395]]
Step2: The coordinate of OBJ1: [[252, 62, 449, 395]]
Step3: The description of IMAGE0: a photography of a spiky plant on a rock in the woods
Step4: The description of FINAL_RESULT: a photography of a spiky plant on a rock in the woods

Critique:
Error Location: functions called by programs. **Reason:** In the Step3 of the program, the used function 'REPLACE' failed to generate a pine grosbeak to replace the bird. 'REPLACE' function need to be updated.

Update the REPLACE tool

Search data from the Internet
Prompt tuning for the Replace tool
REPLACE

Evaluate the updated REPLACE tool in a new image editing task

Instruction: Replace the lion with pine grosbeak
Generated program:
Step1: OBJ0=SEG(image=IMAGE)
Step2: OBJ1=SELECT(image=IMAGE,object=OBJ0,query=lion',category=None)
Step3: IMAGE0=REPLACE(image=IMAGE,object=OBJ1,prompt='pine grosbeak')

Takeaway Message

We build **CLOVA**, the first VLM that can **improve from feedback** via a closed-loop learning framework with **inference**, **reflection**, **learning** phases.

❑ Use both correct and incorrect examples for prompts to generate plans & programs.

❑ Propose a global-local reflection scheme to diagnose errors.

❑ Apply hard/soft prompt tuning to update tools with limited data.

Code & Examples: clova-tool.github.io