



Multimodal Generalist Agents: Reasoning, Reflecting, and Learning like Humans

Machine Learning Lab @ BIGAI



Qing Li (李庆)

<https://liqing.io>

liqing@bigai.ai



Xiaojian Ma (马晓健)

<https://jeasinema.github.io>

maxiaojian@bigai.ai



Zhi Gao (高志)

<https://zhigao2017.github.io>

gaozhi@bigai.ai



Welcoming Diverse Approaches Keeps Machine Learning Strong

What technology counts as an “agent”? Instead of arguing, let's consider a spectrum along which various technologies are “agentic.”

Letters

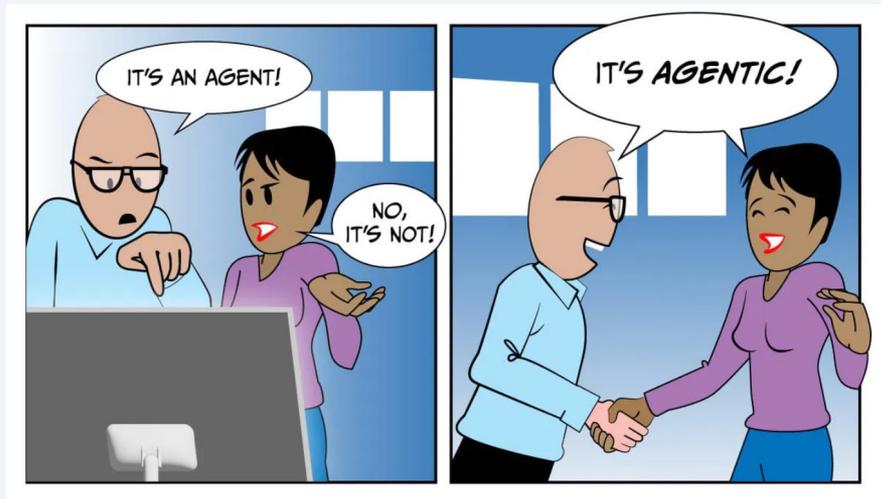
Technical Insights

📅 Published

Jun 13, 2024

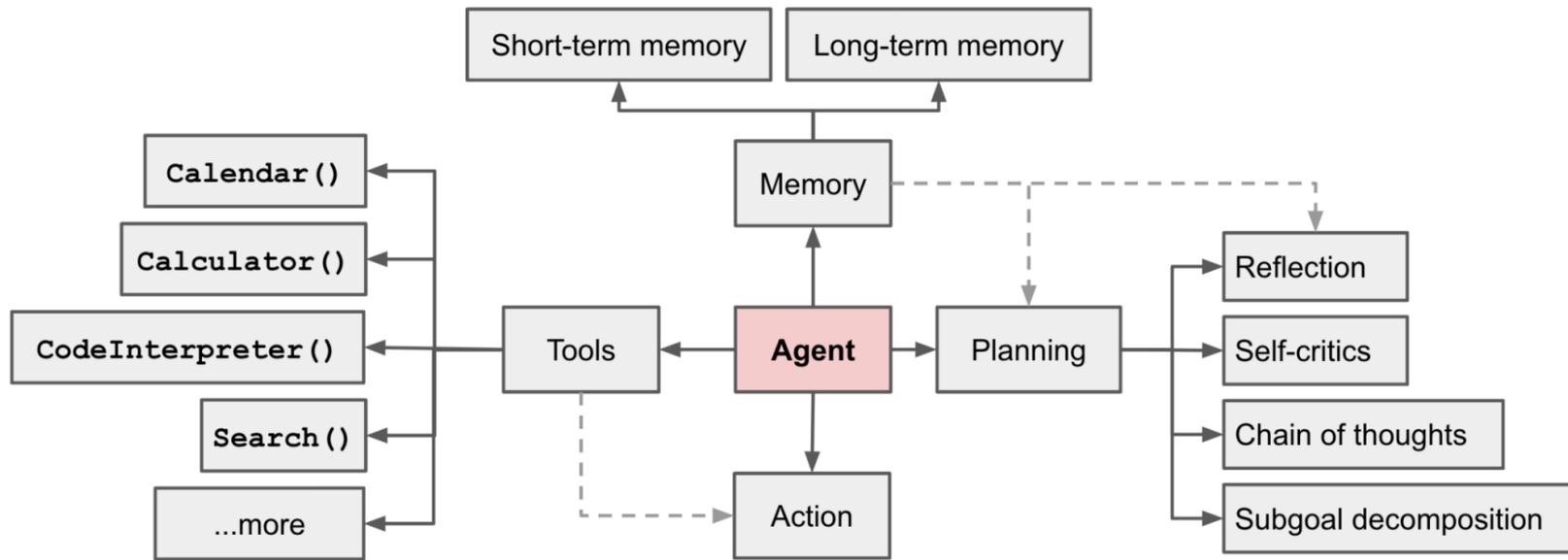
🕒 Reading time

2 min read



Prompts
In-context learning

Vector DB and SQL DB
Retrieval-Augmented Generation (RAG)



<https://lilianweng.github.io/posts/2023-06-23-agent/>

Basic I: Tool-Use and Planning

Web search tool

 You
What is the best coffee maker according to reviewers?

 Copilot
Searching for best coffee maker according to reviewers

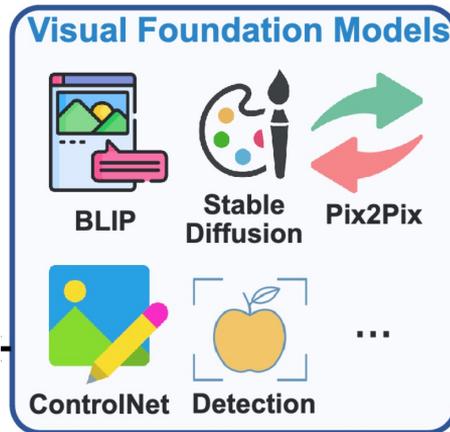
Example from Bing CoPilot

Code execution tool

 You
If I invest \$100 at compound 7% interest for 12 years, what do I have at the end?

```
principal = 100
interest_rate = 0.07
years = 12
value = principal*(1 + interest_rate)**years
```

Example from ChatGPT



Basic I: Tool-Use and Planning

Natural Language Visual Reasoning

LEFT:



RIGHT:



Statement: The left and right image contains a total of six people and two boats.

Program:

```
ANSWER0=Vqa(image=LEFT, question='How many people are in the image?')
ANSWER1=Vqa(image=RIGHT, question='How many people are in the image?')
ANSWER2=Vqa(image=LEFT, question='How many boats are in the image?')
ANSWER3=Vqa(image=RIGHT, question='How many boats are in the image?')
ANSWER4=Eval('{ANSWER0} + {ANSWER1} == 6 and {ANSWER2} + {ANSWER3} == 2')
RESULT=ANSWER4
```

Prediction: False

Basic II: Learning from Feedback



Please write code for {task}

```
def do_task(x): ...
```

```
def do_task_v2(x):
```

```
def do_task_v3(x):
```



Coder Agent
(LLM)

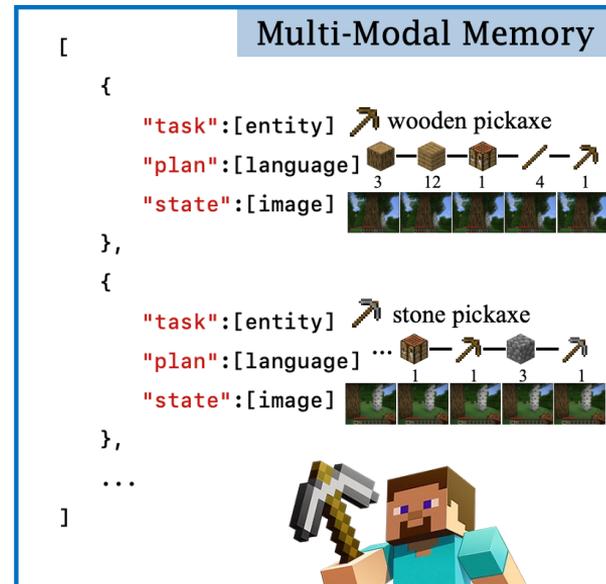
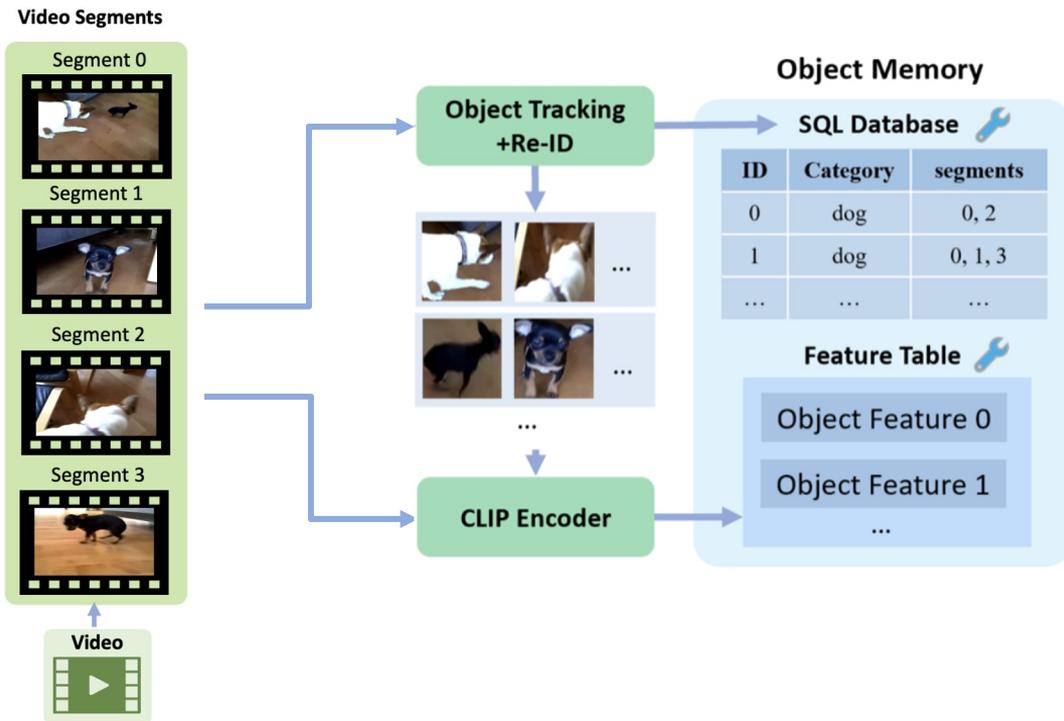
There's a bug on line 5. Fix it by ...

It failed Unit Test 3. Try changing ...

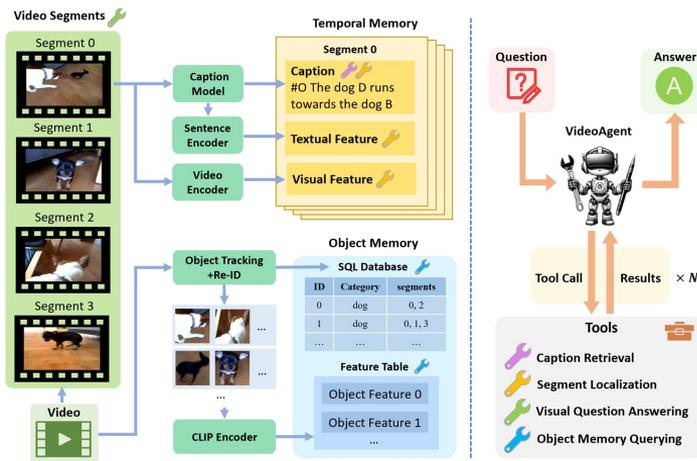


Critic Agent
(LLM)

Basic III: Memory

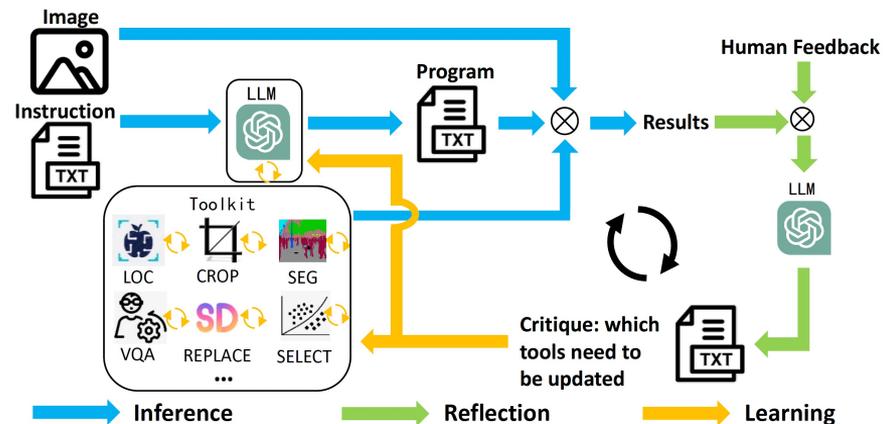


1, 2, 3, Let's put them together...



VideoAgent

Memory-augmented Multimodal Agent for Video Understanding



CLOVA + FIRE

Reflecting and Learning from Feedback for Multimodal Agents



VideoAgent



Xiaojian Ma (马晓健)

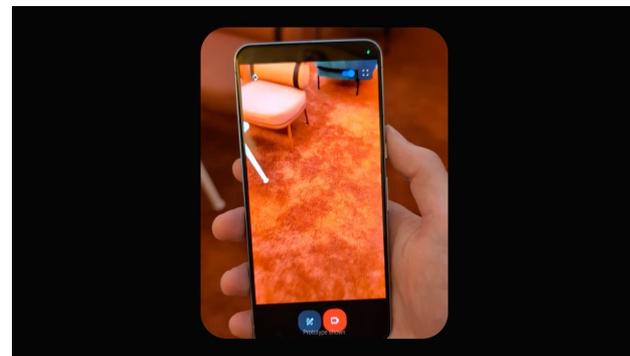
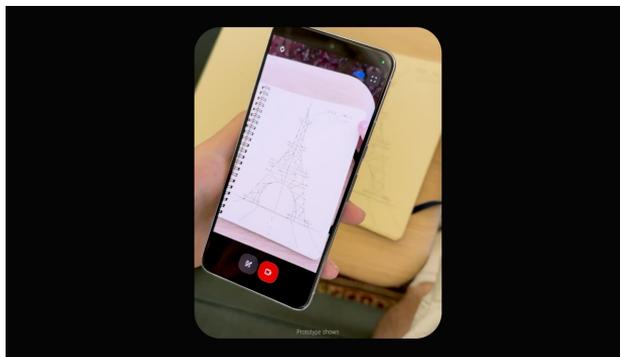
<https://jeasinema.github.io>

maxiaojian@bigai.ai

A Memory-augmented Multimodal Agent for Video Understanding



[videoagent.github.io](https://github.com/jeasinema/videoagent)



- Short-term memory (~5-30s) ✓
- Long-term memory (>30s) mostly ✗

Mostly querying
about **now**

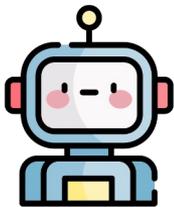
GPT-4o
Newest and most advanced model



Video understanding -- a secret sauce for embodied AI



egocentric experiences
(from long-form videos)



Pickup Object

Lift the bottle 30cm from the ground



Observation Stream



Open Vocabulary Q/A

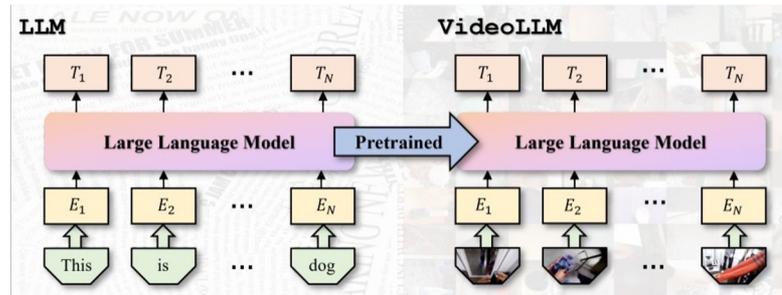
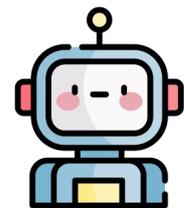
What color are the blinds?

What food is next to kitchen sink?

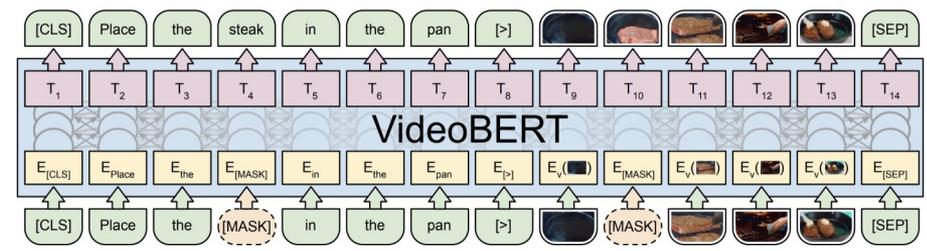




Video understanding -- a secret sauce for embodied AI

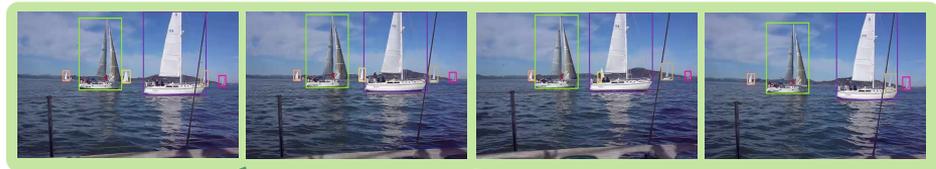


egocentric experiences
(from long-form videos)



Video understanding -- a secret sauce for embodied AI



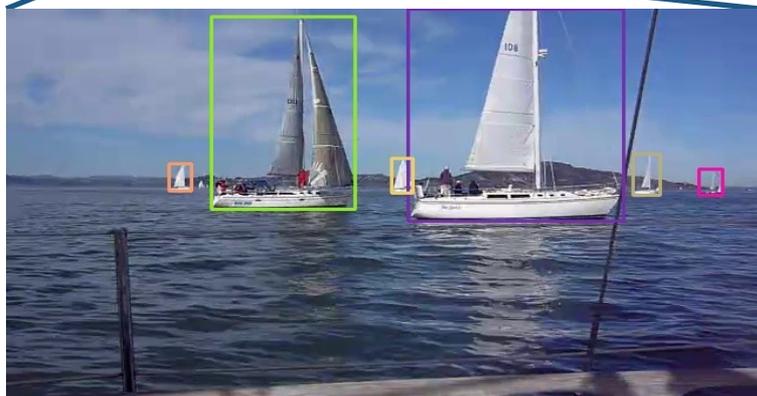


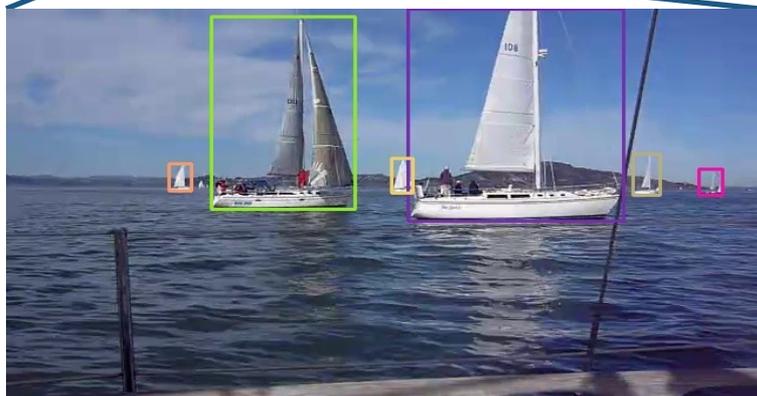
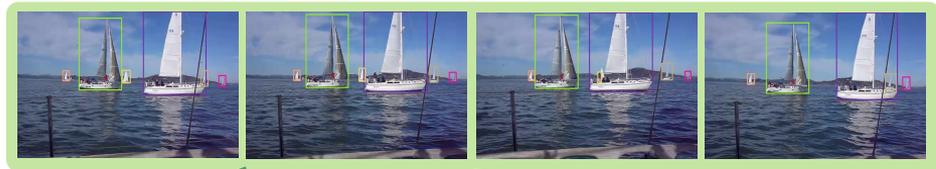
Video



Question

How many boats are there in the video?





Video



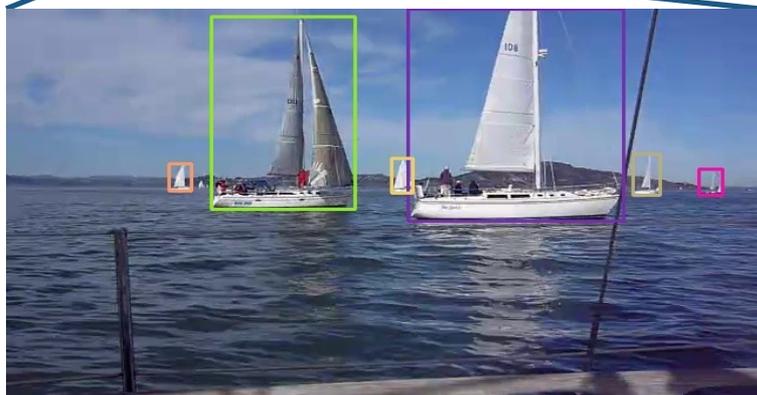
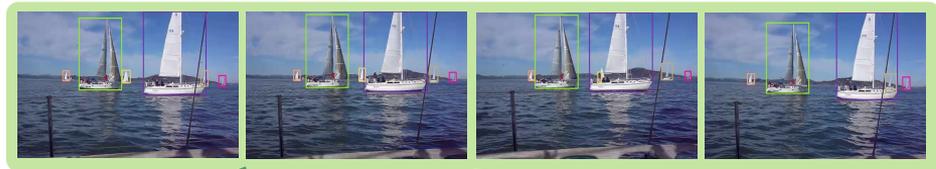
Question

How many boats are there in the video?

mPLUG-Owl



There are 2 boats in the video.



Video



Question

How many boats are there in the video?

mPLUG-Owl

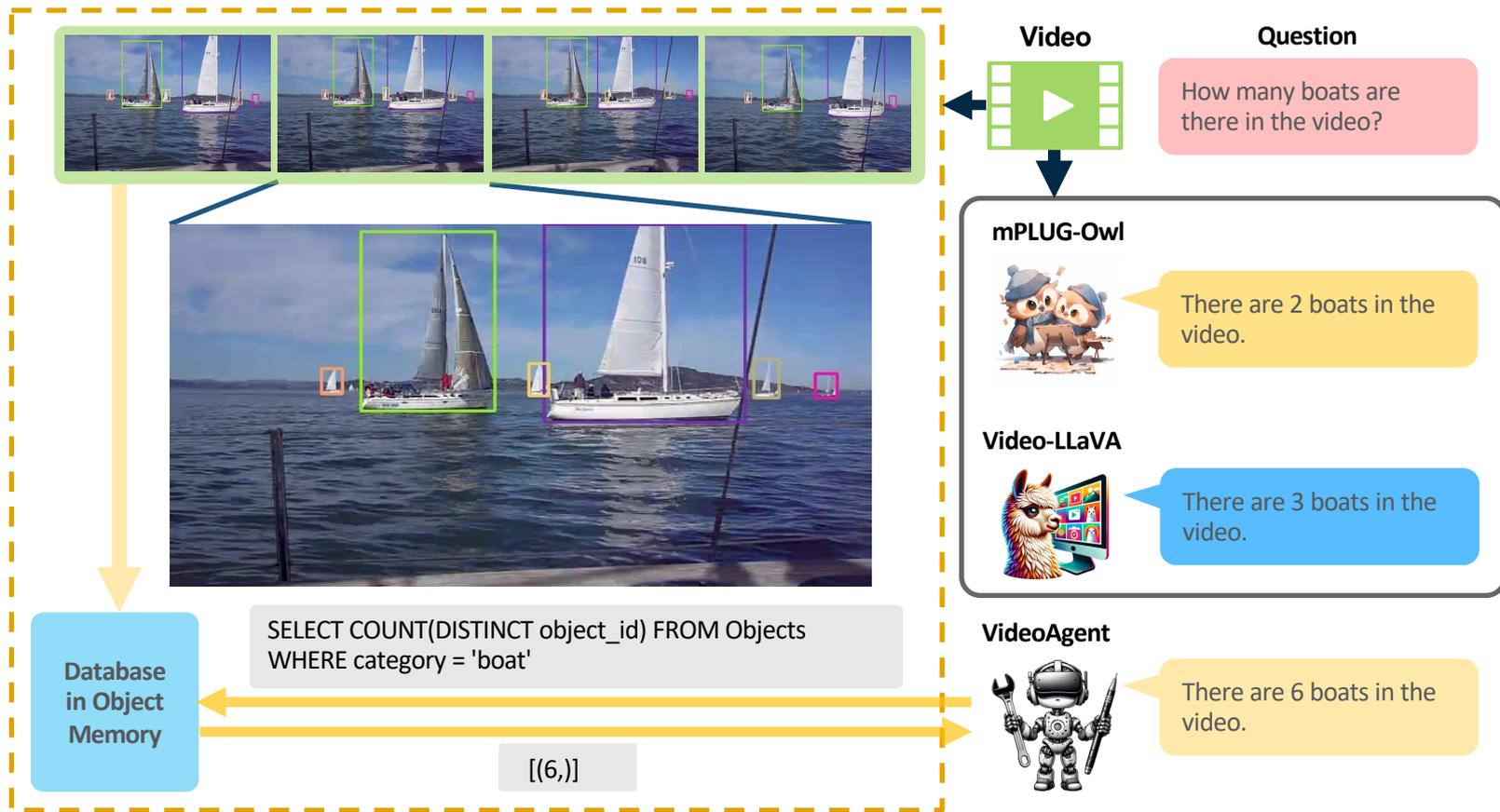


There are 2 boats in the video.

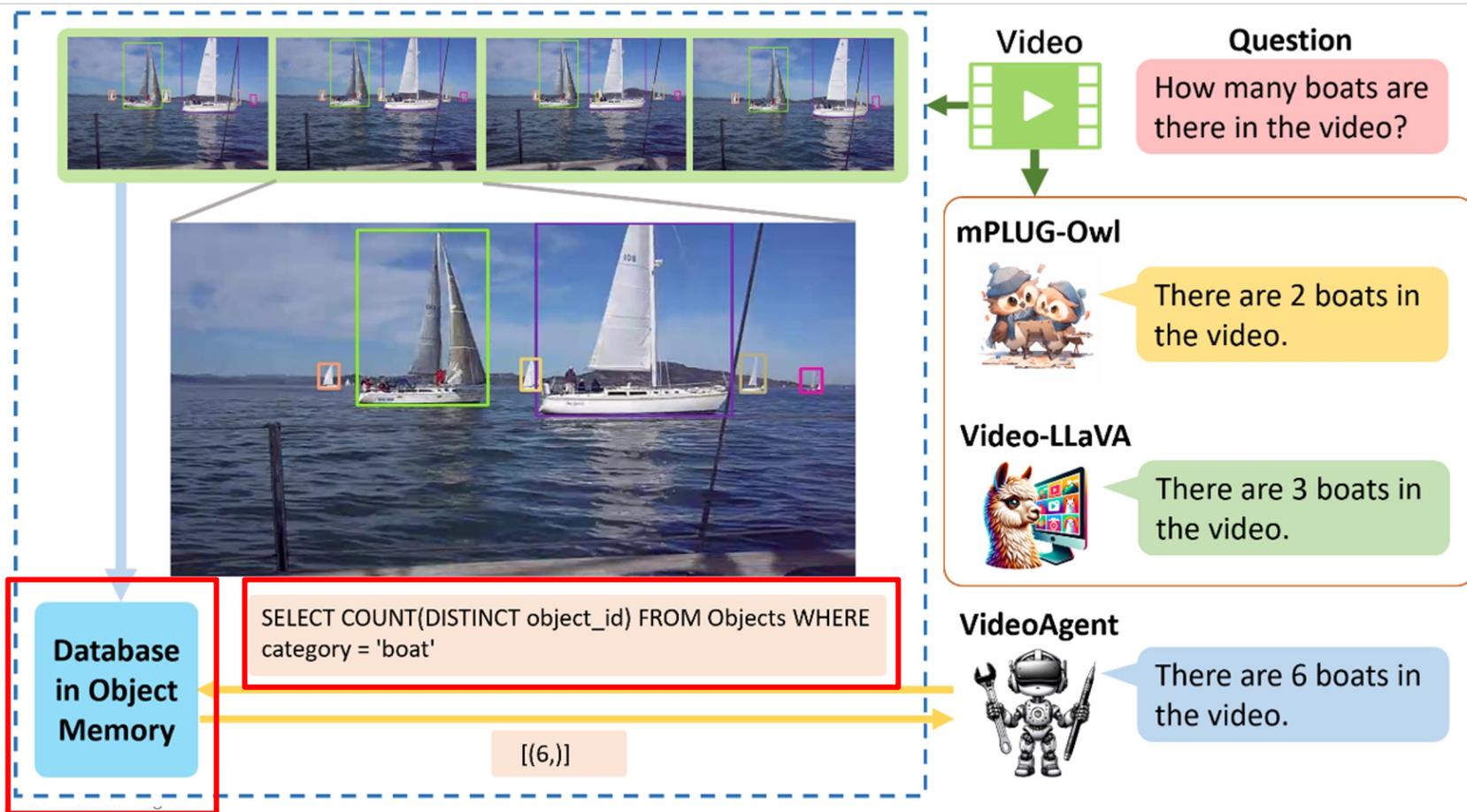
Video-LLaVA



There are 3 boats in the video.



Video understanding -- a secret sauce for embodied AI

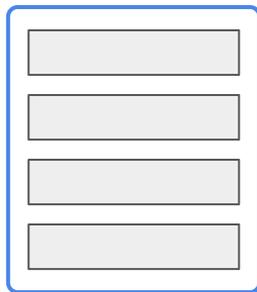


Introducing VideoAgent



egocentric
experience

unified memory



structured
representations

- Temporal Memory
- Object Memory



Tools



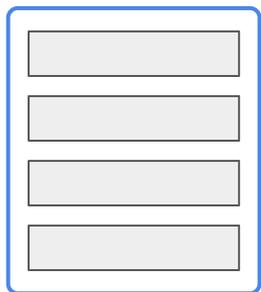
- Caption Retrieval
- Segment Localization
- Visual Question Answering
- Object Memory Querying

tool-use
LLM agent



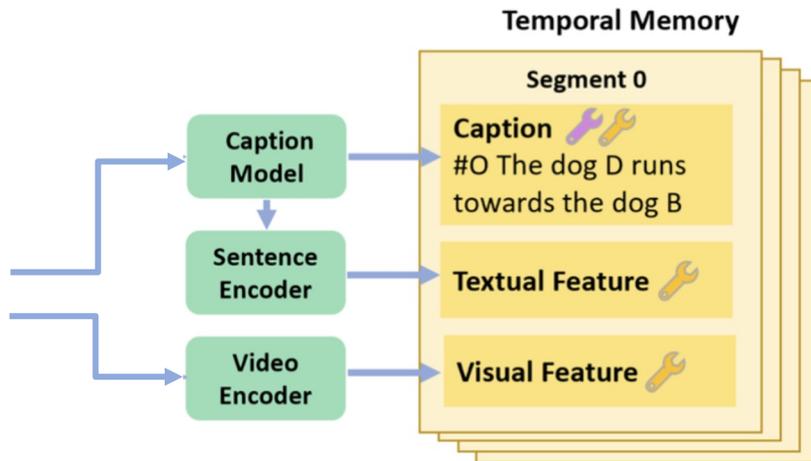
VideoAgent: unified memory

unified memory



structured
representations

Video Segments



📅 **Temporal Memory: segment-level (~2s)**
video content descriptions (text & vector)

Temporal Memory

Caption Model:

LaViLa

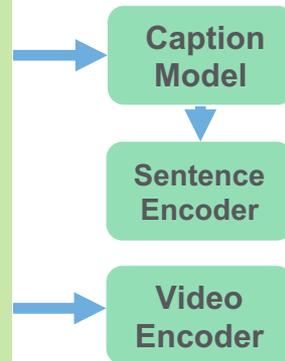
Sentence Encoder:

text-embedding-3-large

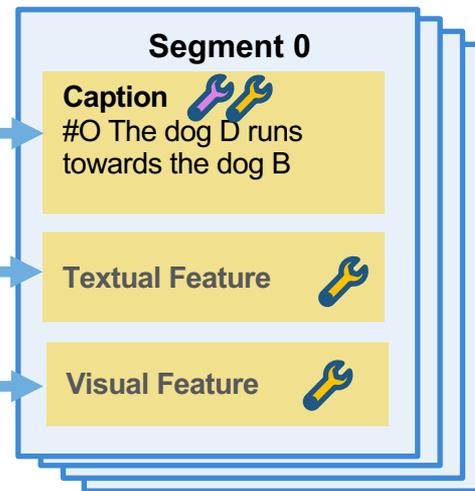
Video Encoder:

ViCLIP

Video Segments

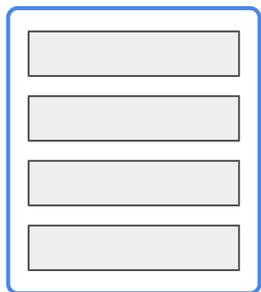


Temporal Memory



VideoAgent: unified memory

unified memory

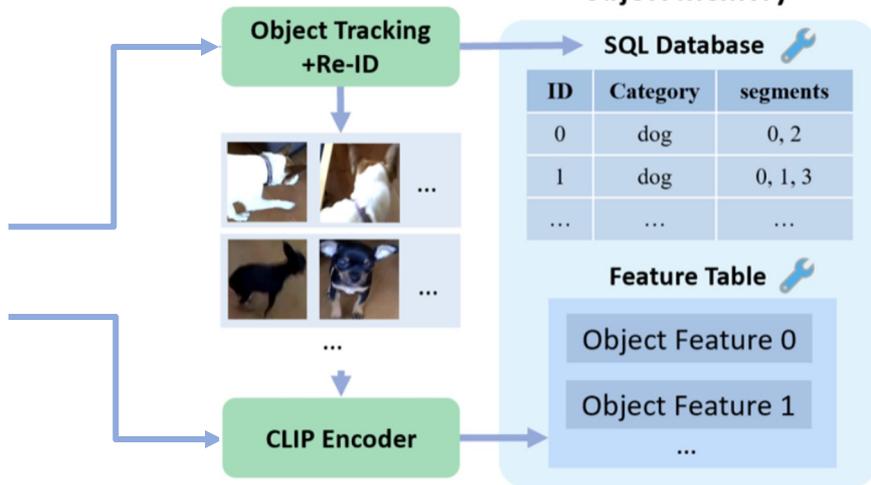


structured representations

Video Segments



Object Memory

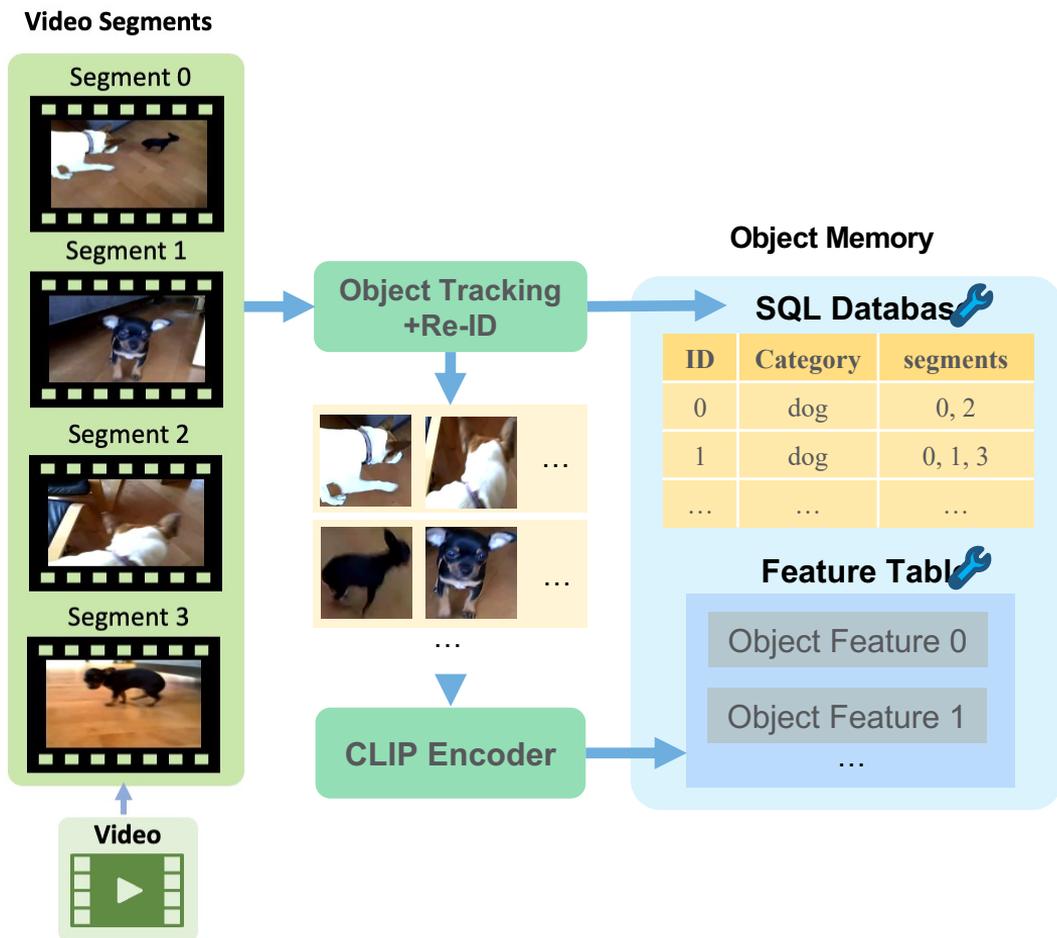


 **Object Memory:** tracking the appearance of unique object instances as video streams (ID, category, CLIP feature)

Object Memory

Object Tracking:
RT-DETR+ByteTrack

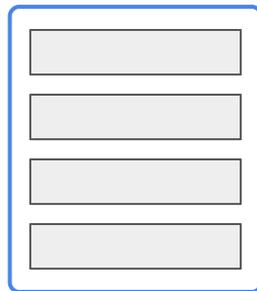
Object ReID: CLIP+DINOv2



VideoAgent: unified memory

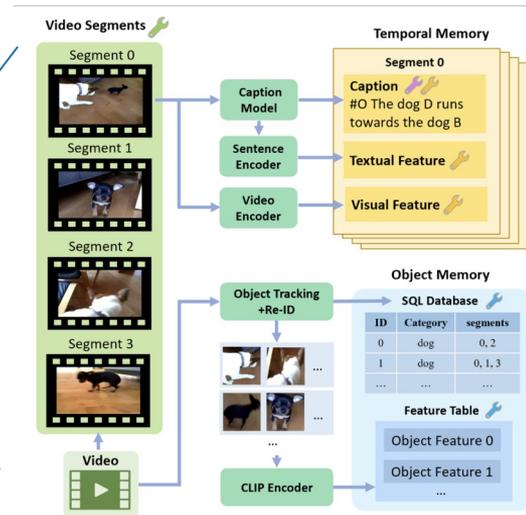


unified memory

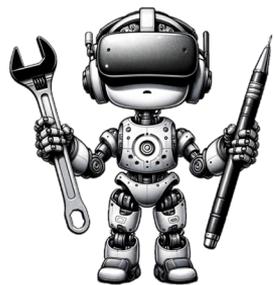


Real-time memory update:

-  video segment captioning (2Hz)
-  video segment embedding (5Hz)
-  object tracking & re-ID (15Hz)



VideoAgent: LLM tool-use



caption retrieval

start and end time



 **Temporal Memory**



captions in between



VQA



**Video
Language
Models**



segment localization

query text



 **Temporal Memory**



video segment (~2s)



object memory query

SQL

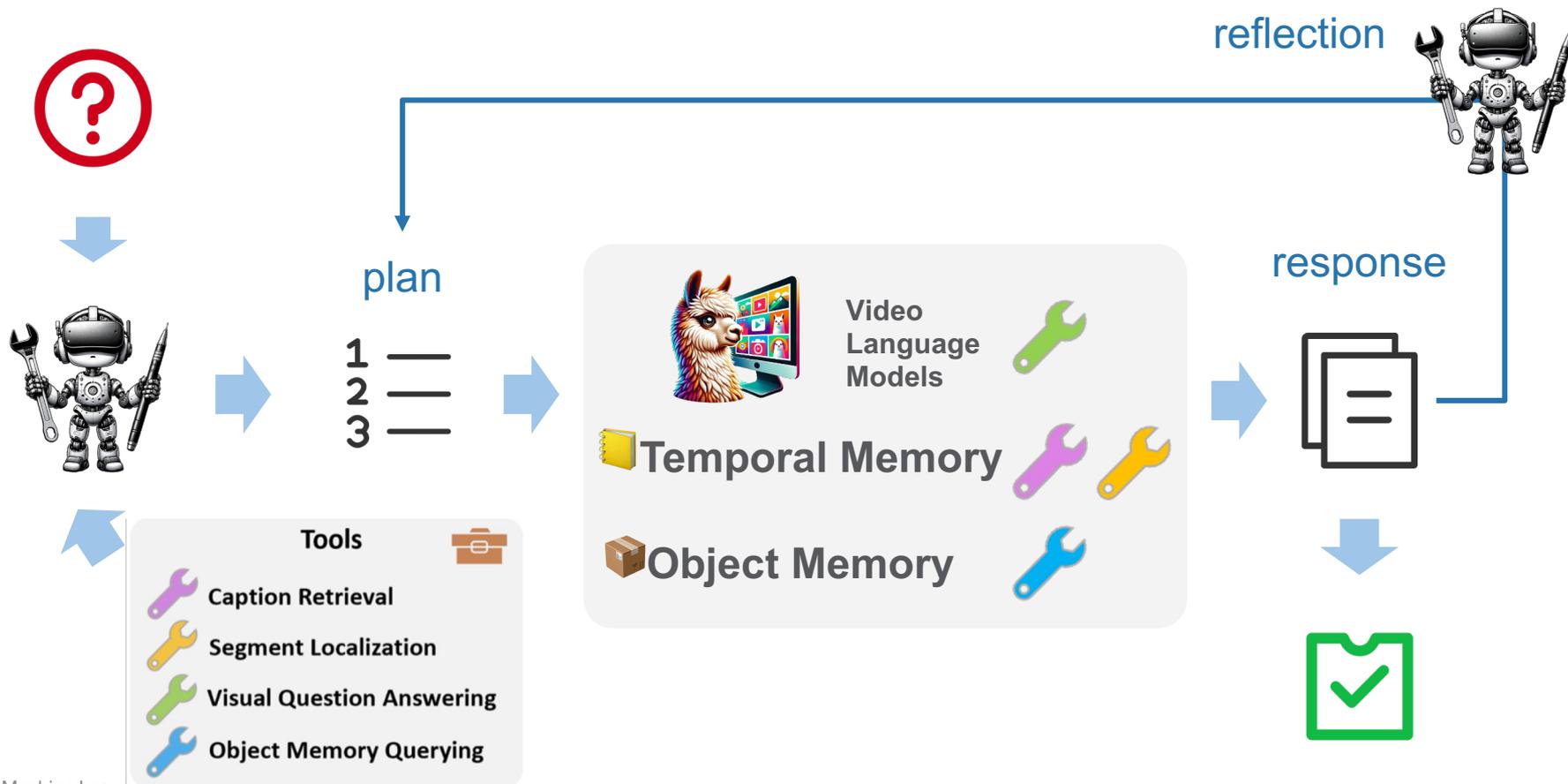


 **Object Memory**



results

VideoAgent: LLM tool-use



Question

What is the relationship between the boy and the adults?

VideoAgent



Video



I need to find the segments where the boy and the adults are mentioned together. I will use the segment_localization tool to find these segments.

Action: segment_localization
Action Input: "boy and adults"

Results: There are 28 segments in total, ranging from 0 to 27. The most relevant segments are: {13: '#O a child X interacts with the', 9: '#O A child X plays with a doll on the floor', ..., 24: '#C C dialogues with person'}

The most relevant segments are 13 and 9 where a child is interacting or playing. I need to check these segments using the visual_question_answering tool

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 13)

Results: Description: The video is about a little boy who is pretending to cook in the kitchen. ... Answer: The video does not provide any information about the relationship between the boy and the adults. However, it can be inferred that the adults are supervising the boy as he plays with the toy kitchen set.

The description suggests that the boy is pretending to cook and the adults are likely supervising him. However, I need to check segment 9 to get more context.

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 9)

Results: Description: The video is about a man and a woman helping a little boy with a microwave oven. ... Answer: The boy is seen playing with the microwave while the adults are standing behind him. ...

The descriptions from both segments suggest that the adults are supervising the boy while he is playing. This suggests that the adults could be his parents or caregivers.
Final Answer: The adults are likely the boy's parents or caregivers.

Question

What is the relationship between the boy and the adults?

VideoAgent



Video



I need to find the segments where the boy and the adults are mentioned together. I will use the `segment_localization` tool to find these segments.

Action: `segment_localization`
Action Input: "boy and adults"

Results: There are 28 segments in total, ranging from 0 to 27. The most relevant segments are: {13: '#O a child X interacts with the', 9: '#O A child X plays with a doll on the floor', ..., 24: '#C C dialogues with person'}

The most relevant segments are 13 and 9 where a child is interacting or playing. I need to check these segments using the `visual_question_answering` tool

Action: `visual_question_answering`
Action Input: ("what is the relationship between the boy and the adults", 13)

Results: Description: The video shows a man and a woman helping a little boy with a microwave oven. ... Answer: The video does not provide any information about the relationship between the boy and the adults. However, it can be inferred that the adults are supervising the boy as he plays with the toy kitchen set.

The description suggests that the boy is pretending to cook and the adults are likely supervising him. However, I need to check segment 9 to get more context.

Action: `visual_question_answering`
Action Input: ("what is the relationship between the boy and the adults", 9)

Results: Description: The video shows a man and a woman helping a little boy with a microwave oven. ... Answer: The boy is seen playing with the microwave while the adults are standing behind him. ...

The descriptions from both segments suggest that the adults are supervising the boy while he is playing. This suggests that the adults could be his parents or caregivers.
Final Answer: The adults are likely the boy's parents or caregivers.

Temporal Memory



Question

What is the relationship between the boy and the adults?

VideoAgent



Video



I need to find the segments where the boy and the adults are mentioned together. I will use the segment_localization tool to find these segments.

Action: segment_localization
Action Input: "boy and adults"

Results: There are 28 segments in total, ranging from 0 to 27. The most relevant segments are: {13: '#O a child X interacts with the', 9: '#O A child X plays with a doll on the floor', ..., 24: '#C C dialogues with person'}

The most relevant segments are 13 and 9 where a child is interacting or playing. I need to check these segments using the visual_question_answering tool to get more context.

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 13)

Results: Description: The video is about a little boy who is pretending to cook in the kitchen. ... Answer: The video does not provide any information about the relationship between the boy and the adults. However, it can be inferred that the adults are supervising the boy as he plays with the toy kitchen set.

The description suggests that the boy is pretending to cook and the adults are likely supervising him. However, I need to check segment 9 to get more context.

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 9)

Results: Description: The video is about a man and a woman helping a little boy with a microwave oven. ... Answer: The boy is seen playing with the microwave while the adults are standing behind him. ...



Video Language Models



The descriptions from both segments suggest the boy is likely playing with the adults. Final Answer: The adults are likely the boy's parents or caregivers.

Question

What is the relationship between the boy and the adults?

Video



VideoAgent



I need to find the segments where the boy and the adults are mentioned together. I will use the segment_localization tool to find these segments.

Action: segment_localization
Action Input: "boy and adults"

Results: There are 28 segments in total, ranging from 0 to 27. The most relevant segments are: {13: '#O a child X interacts with the', 9: '#O A child X plays with a doll on the floor', ..., 24: '#C C dialogues with person'}

The most relevant segments are 13 and 9 where a child is interacting or playing. I need to check these segments using the visual_question_answering tool to get more context.

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 13)

Results: Description: The video is about a little boy who is pretending to cook in the kitchen. ... Answer: The video does not provide any information about the relationship between the boy and the adults. However, it can be inferred that the adults are supervising the boy as he plays with the toy kitchen set.

The description suggests that the boy is pretending to cook and the adults are likely supervising him. However, I need to check segment 9 to get more context.

Action:
visual_question_answering
Action Input: ("what is the relationship between the boy and the adults", 9)

Results: Description: The video is about a man and a woman helping a little boy with a microwave oven. ... Answer: The boy is seen playing with the microwave while the adults are standing behind him. ...

The descriptions from both segments suggest that the adults are supervising the boy while he is playing. This suggests that the adults could be his parents or caregivers.
Final Answer: The adults are likely the boy's parents or caregivers.

Performances

EgoSchema (full set)					
FrozenBiLM	VIOLET	mPLUG-Owl	InternVideo	LLoVi	Gemini 1.5 Pro
26.9	19.9	30.2	32.0	50.3	63.2
EgoSchema (subset, 500 questions)					
SeViLA	Video-LLaVA	mPLUG-Owl	ViperGPT	LLoVi	<i>VideoAgent</i>
25.8	36.8	33.8	15.8	51.8	62.8



~3min



“highly difficult” (prompt for LLM)

NExT-QA				
Method	Temporal	Causal	Descriptive	Average
Val. Set				
InternVideo	43.4	48.0	65.1	49.1
SeViLA(zero-shot)	61.3	61.5	75.6	63.6
TCR(pre-training)	-	-	-	66.1
Val. Subset (600)				
ViperGPT	17.2	19.0	26.7	21.0
mPLUG-Owl	36.0	41.0	52.5	43.2
Video-LLaVA	42.0	53.5	65.0	53.5
SeViLA(zero-shot)	56.0	66.5	70.0	64.2
<i>VideoAgent</i>	60.0	76.0	76.5	70.8



~1min



causal, temporal and action interactions



Ablation studies (on NExT-QA subset)

Type	VQA	Grounding	Captions	Database	Tem.	Cau.	Des.	Avg.
1	GPT-4V	✓	✓	w/ re-ID	64.0	78.0	82.0	74.7
2	Video-LLaVA	✓	✓	w/ re-ID	60.0	74.0	80.0	71.3
3	Video-LLaVA	✓	✓	✗	46.0	64.0	78.0	62.7
4	✗	✗	✓	w/ re-ID	48.0	52.0	68.0	56.0
5	✗	✗	✓	w/o re-ID	46.0	46.0	54.0	48.7
6	✗	✗	✓	✗	34.0	46.0	42.0	40.7

With 📖 **Temporal Memory** and caption retrieval tool 🔧 only, VideoAgent attains **comparable** performances to state-of-the-art video language model mPLUG-Owl (43.2) already.



Ablation studies (on NExT-QA subset)

Type	VQA	Grounding	Captions	Database	Tem.	Cau.	Des.	Avg.
1	GPT-4V	✓	✓	w/ re-ID	64.0	78.0	82.0	74.7
2	Video-LLaVA	✓	✓	w/ re-ID	60.0	74.0	80.0	71.3
3	Video-LLaVA	✓	✓	✗	46.0	64.0	78.0	62.7
4	✗	✗	✓	w/ re-ID	48.0	52.0	68.0	56.0
5	✗	✗	✓	w/o re-ID	46.0	46.0	54.0	48.7
6	✗	✗	✓	✗	34.0	46.0	42.0	40.7

 **Object Memory (SQL database) with object memory query tool** 
boosts performances on all question categories.

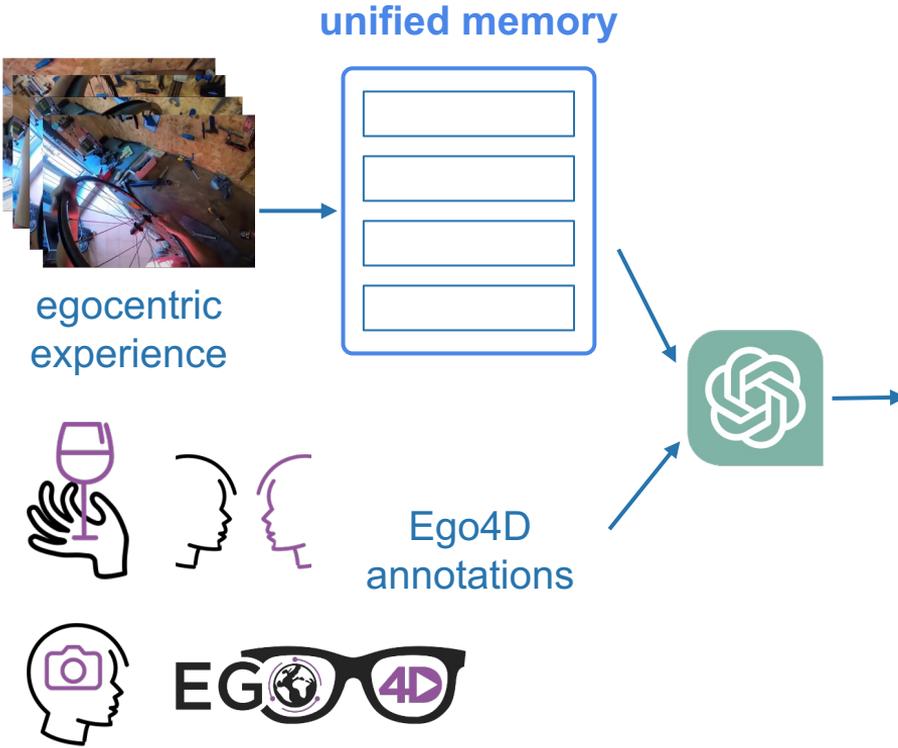
Ablation studies (on NExT-QA subset)

Type	VQA	Grounding	Captions	Database	Tem.	Cau.	Des.	Avg.
1	GPT-4V	✓	✓	w/ re-ID	64.0	78.0	82.0	74.7
2	Video-LLaVA	✓	✓	w/ re-ID	60.0	74.0	80.0	71.3
3	Video-LLaVA	✓	✓	✗	46.0	64.0	78.0	62.7
4	✗	✗	✓	w/ re-ID	48.0	52.0	68.0	56.0
5	✗	✗	✓	w/o re-ID	46.0	46.0	54.0	48.7
6	✗	✗	✓	✗	34.0	46.0	42.0	40.7

VQA tool  and segment localization tool  bring the most advantages to VideoAgent.

Data flywheel with VideoAgent

EgoSFT dataset



t0 t1 t2

Spatial Qs (about the present)
-(At t0) what am I picking up from the shelf?

Recall Qs (about the past)
-(At t2) what is the hair color of the person I met when I entered the shop?

Temporal Qs (about multiple time steps)
-(At t2) where did I get this wrench from?

...

Data flywheel with VideoAgent

EgoSFT dataset



t0

t1

t2

Spatial Qs (about the present)

-(At t0) what am I picking up from the shelf?

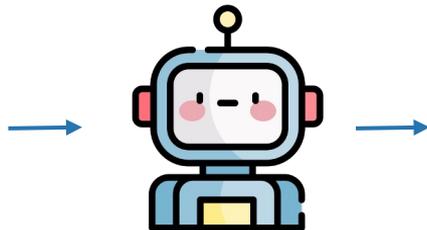
Recall Qs (about the past)

-(At t2) what is the hair color of the person I met when I entered the shop?

Temporal Qs (about multiple time steps)

-(At t2) where did I get this wrench from?

...



“real” Google Gemini⁺
real-time, continuous VLM
(ref)

Takeaways

We present **VideoAgent**, tackling general video understanding via the following pipeline: 🎬 video => 🧠 structured representation (memory) => 🧠 LLM tool-use

We design a unified memory for video and egocentric experiences:

📅 **Temporal Memory**, which stores descriptive knowledge of short segments;

📦 **Object Memory**, which maintains appearances of unique object instances

VideoAgent attains comparable performances to many of the best commercial video language models on challenging benchmarks, with **much lower cost**

Code & demo: videoagent.github.io

VideoAgent: A Memory-augmented Multimodal Agent for Video Understanding





CLOVA +  **FIRE**



Zhi Gao (高志)

<https://zhigao2017.github.io>

gaozhi@bigai.ai

Reflecting and Learning from Feedback for Multimodal Agents



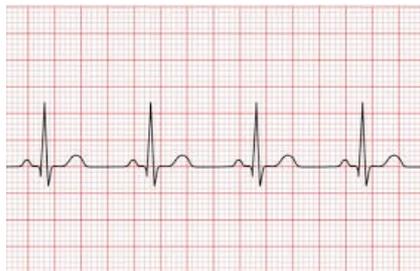
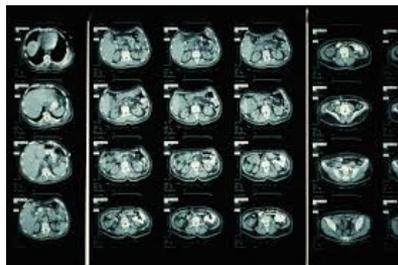
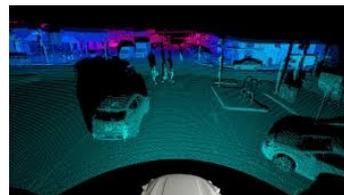
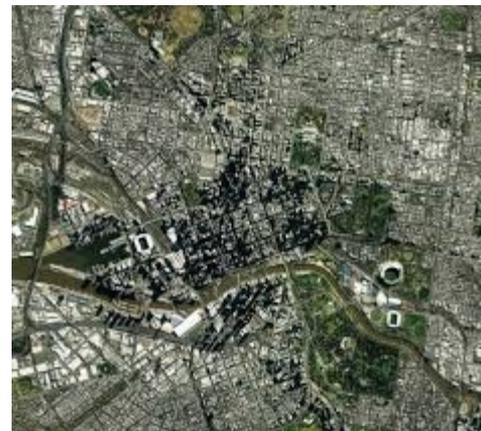
clova-tool.github.io



mm-fire.github.io

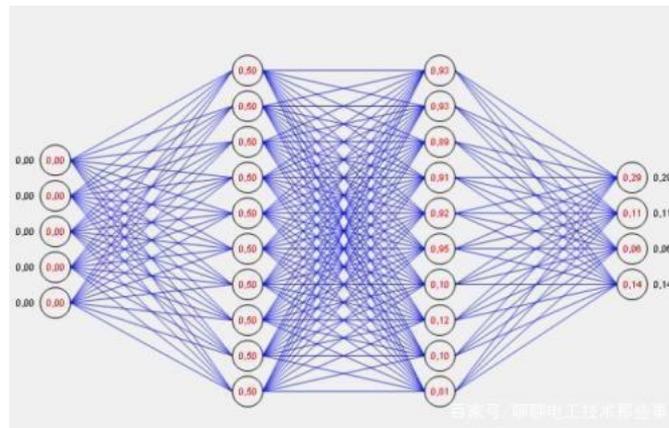
Multimodal knowledge

- Movie, TV show
- LiDAR
- ECG, CT
- Remote Sensing Image, Satellite Image
- Social Network Data
- Fine-grained Data
- ...

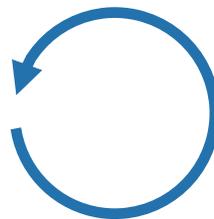


Supervised Learning

- ❑ Too expensive to collect data
- ❑ Huge resource consumption
- ❑ Bad generalization

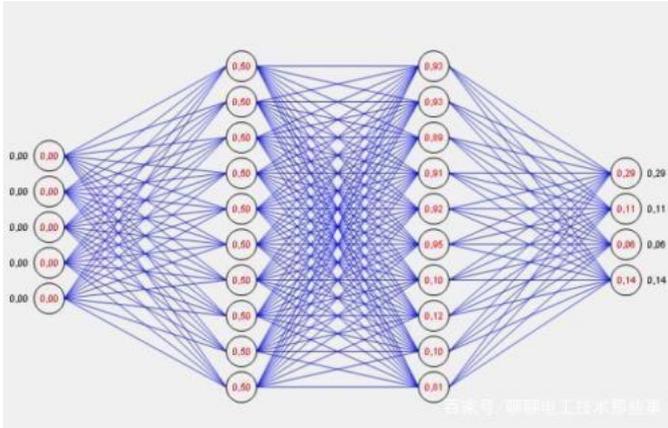


Neural network

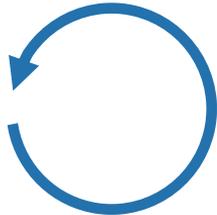


A large amount of
training Data $\{x, y\}$

Supervised Learning vs. Feedback learning



Neural network



A large amount of training Data
 $\{x, y\}$



Agent

Environment

Learning in an interaction manner

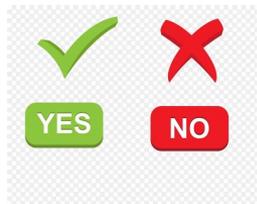
- ❑ Do not need a large amount of data
- ❑ Small resource consumption
- ❑ Good generalization

Multimodal Feedback learning

Learning in an interaction manner with feedback

Challenges

- ❑ Complex/diverse structures of data
- ❑ Limited Data
- ❑ Catastrophic forgetting



Agent



Environment

Multimodal Agents

Handle diverse forms of data across different visual tasks

QA: How many muffins can each kid have for it to be fair?



Image editing: Replace the dessert with green grass.



Knowledge tagging: Tag the wife of Jim Halpert in popular TV series The Office



QA: Why is this video funny?



QA: What does the adult ride in the playground?



Caption: Describe the image in details.



Multimodal Agents

End-to-end methods

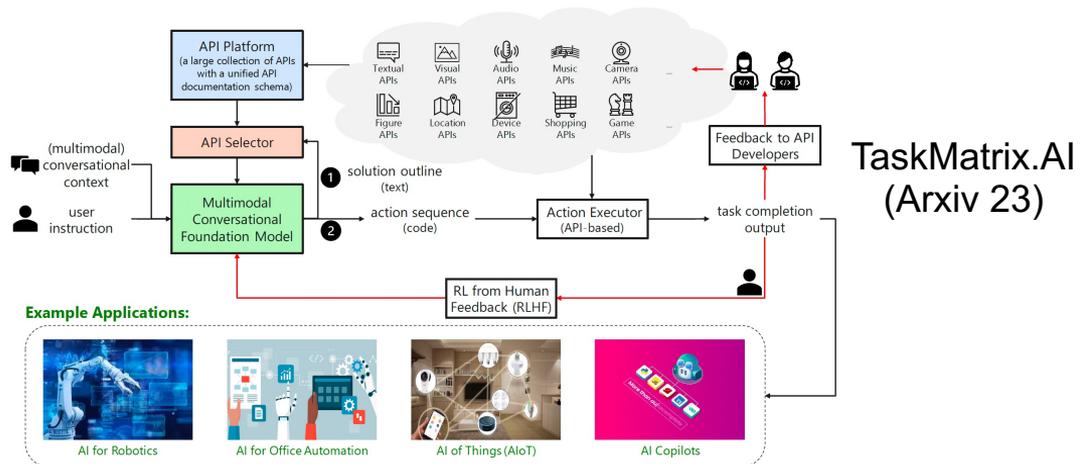
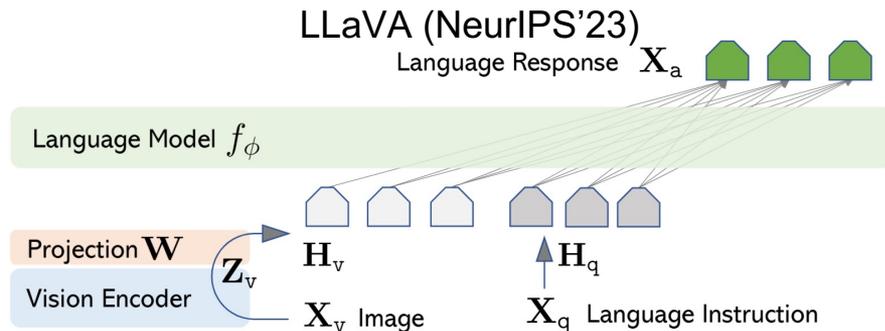
- LLaVA
- MiniGPT-4
- Qwen-VL
- GPT-4V
- Video-LLaVa

...

Tool-based methods

- VisProg
- Visual ChatGPT
- MM-REACT

...



End-to-end Methods

Existing Vision-Language Models (VLMs) have an inferior feedback learning ability, limiting their interactive learning efficiency.



Ground truth: Car.



User

Which kind of vehicle is waiting for the traffic light?

The vehicles look like typical vehicles used for personal transport. The vehicles are **motorcycles**.

Wrong



LLaVA-Next-8B



User

Motorcycles are not correct. Look again at the structure and features of the vehicles, such as **the number of wheels and the body shape.**

Upon reevaluating the image, the vehicles are indeed **motorcycles**.

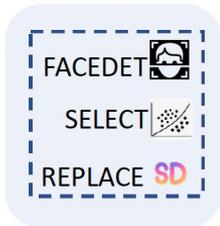
Failed in Refining



LLaVA-Next-8B

Tool-based Methods

Replace Leonardo with Dominik Sadoch.



It is not Dominik Sadoch.

Cannot generate Dominik Sadoch

Who is the character in the image?



naruto



The correct answer is Luffy.

Cannot recognition Luffy

Visual tools are not perfect, while the potential for continual learning has been overlooked.

Lack up-to-date knowledge, fine-grained, expert knowledge, etc.



Multimodal Feedback Learning

Target: We build the feedback learning ability for both end-to-end methods and tool-based methods.

 What do you see in this image?



A sign with a red circle and a slash over a skateboard symbol, indicating that skateboarding is prohibited.

 Enhance your response by **specifying the exact text**.
Mention **background details**.

The sign prohibits skateboarding and is attached to a pole. It features a red circle and slash over a skateboard symbol. Below symbol, there is text that reads "**No Skateboarding**" and "SEC. 30.2 C.F. Fine \$76.00". Background includes a building and **a clear blue sky**.

 Tag the director of the movie Parasite.



FACEDET
LIST
CLASSIFY
TAG



 Tagged the wrong person.

CLASSIFY fails on Bong Joon-ho. Update it.



 Tag the best director of the 92nd Oscar.



FACEDET
LIST
CLASSIFY
TAG



Multimodal Feedback Learning



FIRE

A Dataset for Feedback Integration and Refinement Evaluation of Multimodal Models



CLOVA

A closed-loop visual assistant with tool usage and update



FIRE: A Dataset for Feedback Integration and Refinement Evaluation of Multimodal Models

Pengxiang Li^{*1,2}, Zhi Gao^{*1,3}, Bofei Zhang^{*1}, Tao Yuan¹, Yuwei Wu²,
Mehrtash Harandi⁴, Yunde Jia¹, Song-Chun Zhu^{1,3,5}, Qing Li¹

¹BIGAI ²BIT ³Peking University ⁴Monash University ⁵Tsinghua University

^{*} Equal contribution,  Corresponding Author

<https://mm-fire.github.io>



Motivation



Ground truth:

Nutritional deficiency.



User

Why are the leaves turn yellow?

Overwatering

Wrong



QwenVL-72B



User

Your answer is incorrect. Please judge the possible reasons based on the texture characteristics of the yellowing leaves. Please note that overwatering will not cause the leaves to turn yellow in such pattern. Yellowing leaves may be caused by a disease in the plant.

Overwatering

Failed in Refining

QwenVL-72B

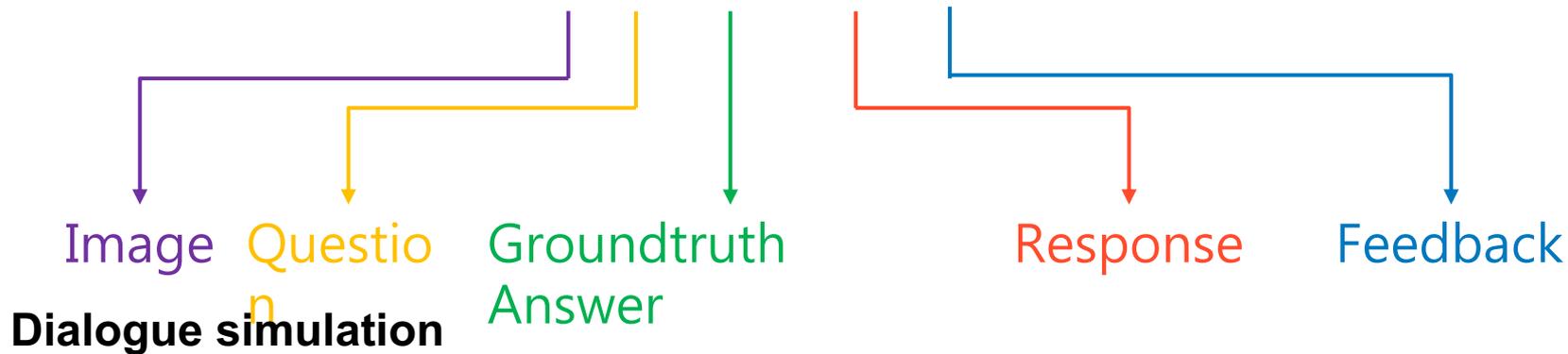
One reason of such limitation is that instruction-following datasets predominantly consist of data focused on direct task completion, with **little to no inclusion** of feedback-refining information.



Task setting

Data format

$$\{I, q, gt, \{r_i, f_i\}_1^n\}$$



Vision Language Model
ROLE: Teacher
Playing: Scoring and feedback.

Input: I, q, gt, r_i
Output: f_i

Vision Language Model
ROLE: Student
Playing: Answer the Question.

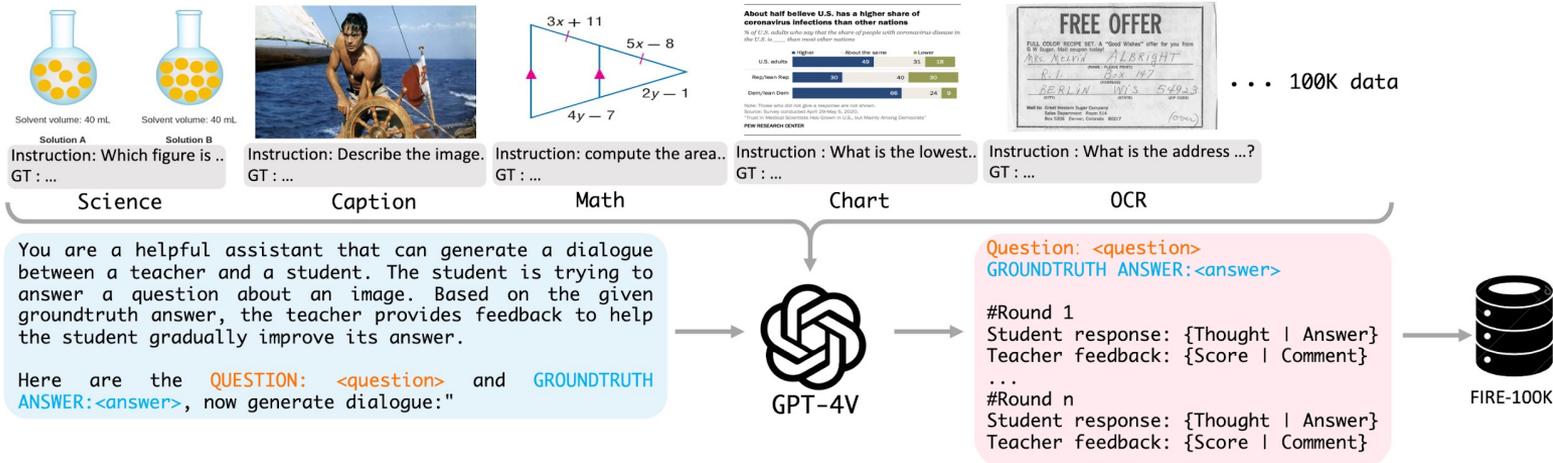
Input: I, q, f_i
Output: r_{i+1}



Groundtruth Answer is only visible to teacher model.

Data collection

(a) FIRE-100K Data Generation



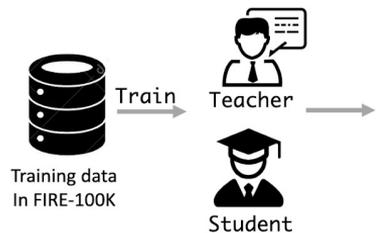
We use GPT-4V to generate 100K data (FIRE-100K), with each dialogue was generated in a one-go manner.

Data collection

We use FIRE-100K to train a Teacher model and a Student model.

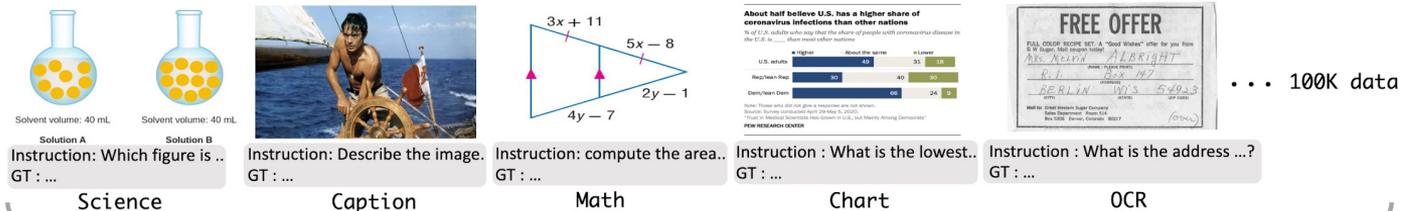
Then we simulate dialogues between the student and the teacher models and scale up the data points to ~1M, rendering a split named FIRE-1M.

(b) FIRE-1M Data Generation



Data collection

(a) FIRE-100K Data Generation



You are a helpful assistant that can generate a dialogue between a teacher and a student. The student is trying to answer a question about an image. Based on the given groundtruth answer, the teacher provides feedback to help the student gradually improve its answer.

Here are the **QUESTION:** <question> and **GROUNDTRUTH ANSWER:** <answer>, now generate dialogue:"



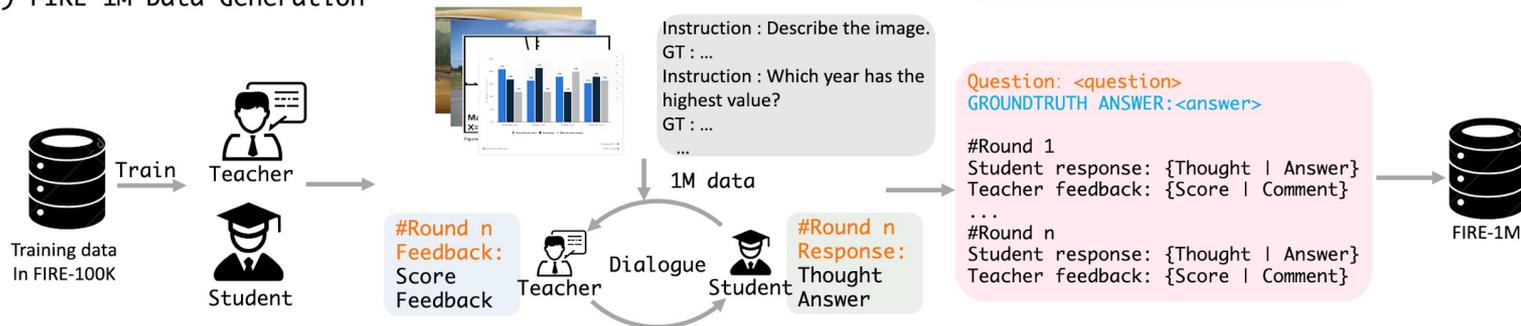
Question: <question>
GROUNDTRUTH ANSWER: <answer>

#Round 1
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}
...
#Round n
Student response: {Thought | Answer}
Teacher feedback: {Score | Comment}



FIRE-100K

(b) FIRE-1M Data Generation



Dataset statistic



Training Set FIRE-100K & FIRE-1M



Statistic	FIRE-100K	FIRE-1M
Total dialogue	104941	1003490
Avg. Response Length	41.20	79.33
Avg. Feedback Length	41.74	54.87

FIRE offers **1.1 million** high-quality feedback-refining dialogues across **8** diverse tasks and **18** various data sources.

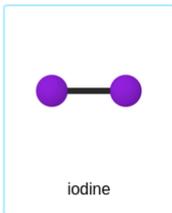
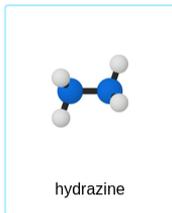
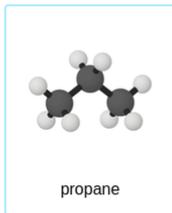
Training Student Model

Question:

Look at the models of molecules below. Select the elementary substance.

Choices:

A. propane B. hydrazine C. iodine



The elementary substance in the image is represented by the molecule labeled "propane." Propane is a simple organic molecule composed of three carbon atoms and one hydrogen atom. It is a gas at room temperature and is commonly used as a fuel. The other molecules, hydrazine and iodine, are more complex compounds. Hydrazine is a compound of hydrogen and nitrogen, while iodine is a non-metallic element. Therefore, the correct answer to the question is: **A. propane**



Score: 3.

Comments: You correctly identified that propane is a simple organic molecule composed of carbon and hydrogen atoms. However, the question asks for the elementary substance, which refers to a basic chemical element. Propane is a compound, not an element. Consider the other options and think about which one is an element.



Ground-truth Answer: 🐶



Text completion:



$$\min \mathbb{E}_{(I, q, gt, \{r_i, f_i\}_{i=1}^n) \sim \mathbb{D}} \left[- \sum_{i=2}^n \log P(r_i | I, q, \{r_j, f_j\}_{j=1}^{i-1}) \right]$$

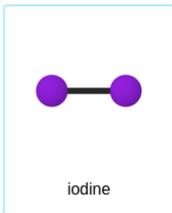
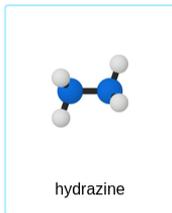
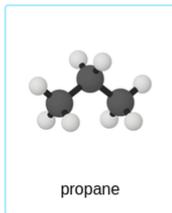
Training Teacher Model

Question:

Look at the models of molecules below. Select the elementary substance.

Choices:

A. propane B. hydrazine C. iodine



I apologize for the confusion. You are correct that propane is a compound, not an element. The elementary substances in the image are represented by the molecules labeled "hydrazine" and "iodine". **Hydrazine is a compound of hydrogen and nitrogen, and iodine is a non-metallic element.** Therefore, the correct answer to the question is: **C. iodine**



Text completion:

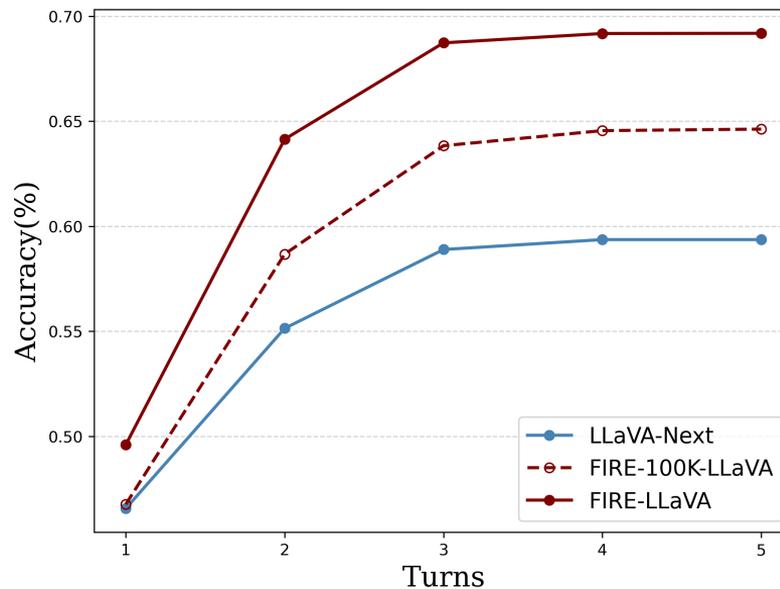
Ground-truth Answer:

C. iodine

$$\min \mathbb{E}_{(I, q, gt, \{r_i, f_i\}_{i=1}^n) \sim \mathbb{D}} \left[- \sum_{i=1}^n \log P(f_i | I, q, gt, \{r_j, f_j\}_{j=1}^{i-1}, r_i) \right]$$

Experiments

LLaVA-Next fine-tuned by FIRE can complete the response **faster** and has **better feedback refining capabilities** (7% more improvements) under the same feedback model.



Experiments

Average Turn (AT). The AT metric evaluates **how fast** a VLM could achieve a satisfactory result based on feedback.

Average Dialogue Refinement (ADR). The ADR metric evaluates how much knowledge VLMs could learn from feedback **in a dialogue.**

Average Turn Refinement (ATR). ATR evaluates how much knowledge VLMs could learn from feedback **in one turn.**

Refinement Ratio (RR). RR measures the proportion of data that have a wrong initial response and a correct **final response.**

Results in free dialogue over all test data in FIRE.

Model	AT (↓)	ADR (↑)	ATR (↑)	RR (↑)
LLaVA-Next-8B	1	0.97	0.41	0.25
FIRE100K-LLaVA-8B	0.92	1.27	0.55	0.34
FIRE-LLaVA-8B	0.84	1.56	0.66	0.39

LLaVA-Next fine-tuned by FIRE can complete the response **faster** and has **better feedback refining capabilities.**

Visualization

Question:

What is the color of the gloves worn by one of the football player in the image?

- A. White
- B. Black
- C. Green
- D. Blue



Ground-truth Answer:

B. Black



The color of the gloves worn by the football player in the image is white. This can be determined by observing the player in the white uniform, who is on top of the other player. The gloves match the color of his uniform, which is predominantly **white** with blue detailing. Therefore, the correct answer to the question is: **A. White**



Score: 4.

Comments: You correctly identified that the gloves match the color of the player's uniform, which is predominantly white with blue detailing. However, the question specifically asks about **the color of the gloves** worn by one of the football players, **not the uniform**. **Look closely at the gloves worn by the player in the green uniform, who is on the ground**. Consider the options provided and re-evaluate your answer.



Upon re-evaluation, **the gloves worn by the player in the green uniform, who is on the ground, are black**. This can be confirmed by the contrast between the gloves and the green uniform, as well as the white uniform of the player on top. Therefore, the correct answer to the question is: **B. Black**



Score: 10.

Comments: Excellent observation! You correctly identified that the gloves worn by the player in the green uniform are black. This matches the ground truth answer. Well done on adjusting your focus and re-evaluating the image.



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

We build CLOVA, a visual assistant that can **continuously learn new knowledge** within a closed-loop learning framework.

Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma,
Wenjuan Han, Song-Chun Zhu, Qing Li
Machine Learning Lab, BIGAI
CVPR 2024 <https://clova-tool.github.io/>

Motivation

A visual assistant that can

- (1) Identify problematic tools
- (2) learn missing knowledge
- (3) generalize to new tasks

based on feedback

⊙ Tag the director of the movie Parasite.



⊙ Tagged the wrong person.

CLASSIFY fails on Bong Joon-ho. Update it.



FACEDETECT
LIST
CLASSIFY
TAG

⊙ Tag the best director of the 92nd Oscar.



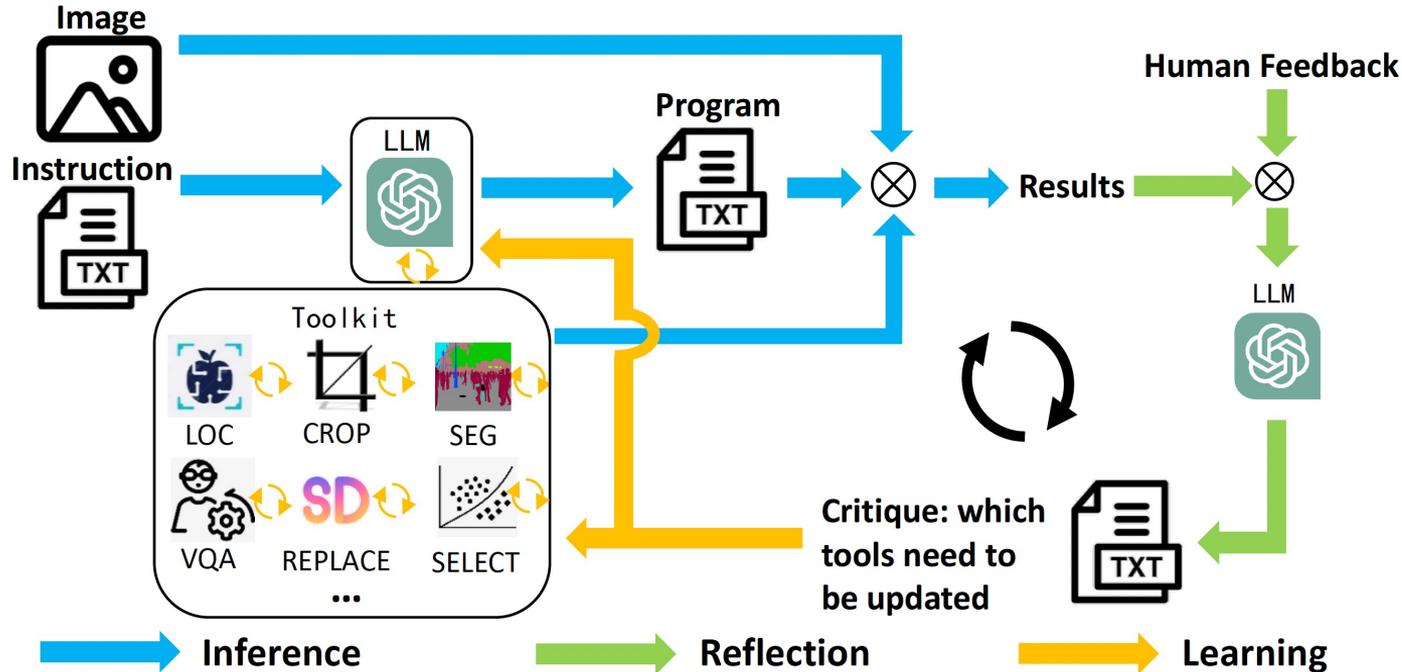
⊙ Tagged the best director of the 92nd Oscar.



FACEDETECT
LIST
CLASSIFY
TAG



Framework



Inference: generate a program and call visual tools to solve the task

Reflection: identify which tool is problematic based on human feedback

Learning: automatically collect training data to update the tool

Inference

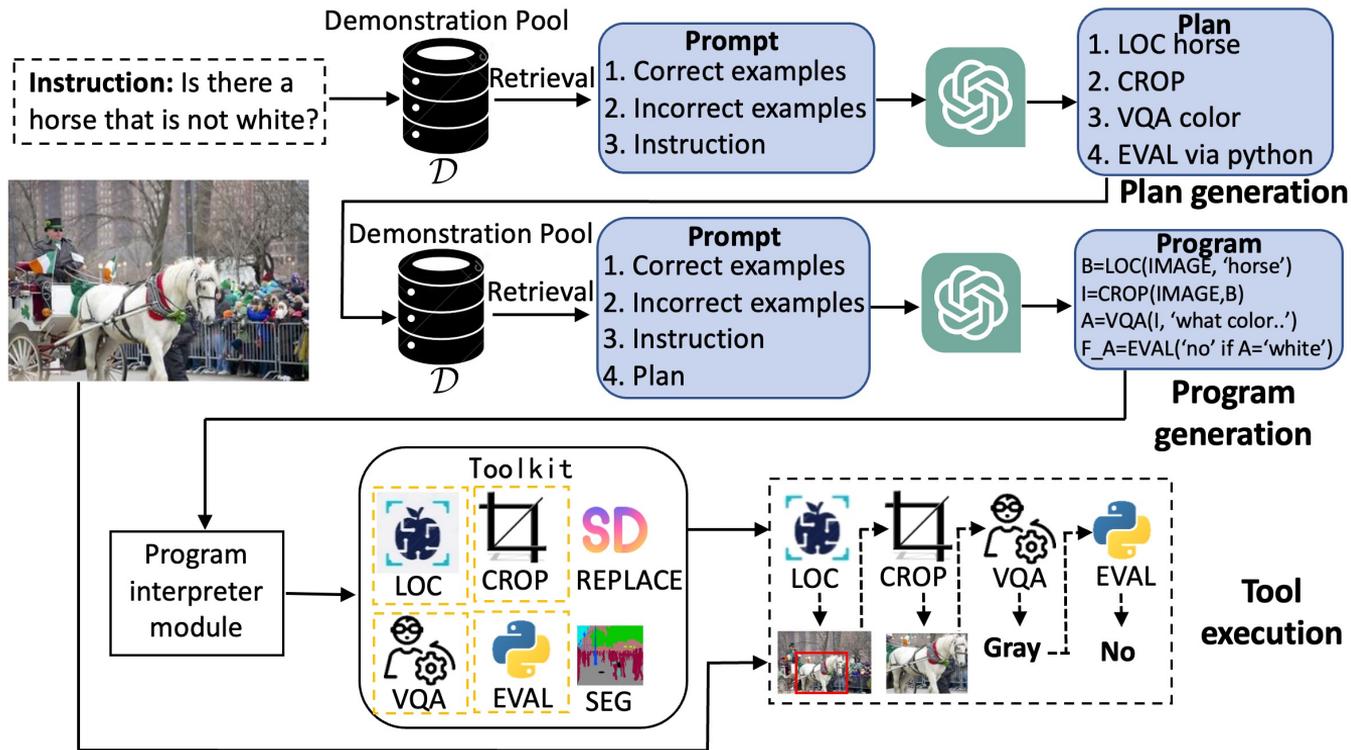
Plan generation



Program generation

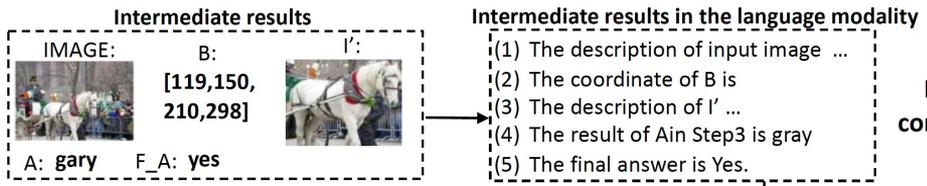


Tool execution

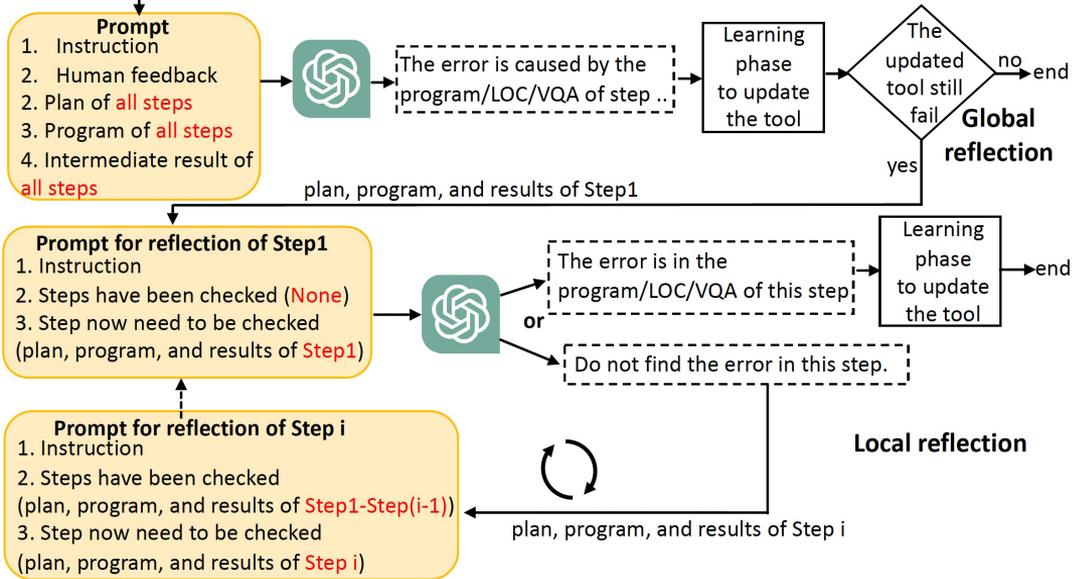


Reflection

global-local reflection



Feedback: the desirable answer is...



Result conversion → **Global reflection** → **Local reflection**

BLIP model

task inputs,
feedback on the task
generated plan and program
intermediate results at each step

task inputs,
feedback on the task,
the steps that have been checked the
current step that needs to be checked

Learning

Data collection



Prompt tuning



Prompt validation

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.

Data collection

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

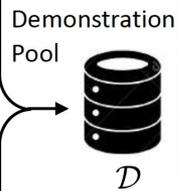
From reflection: the SELECT/REPLACE/CLASSIFY tools fails for the object 'horse'



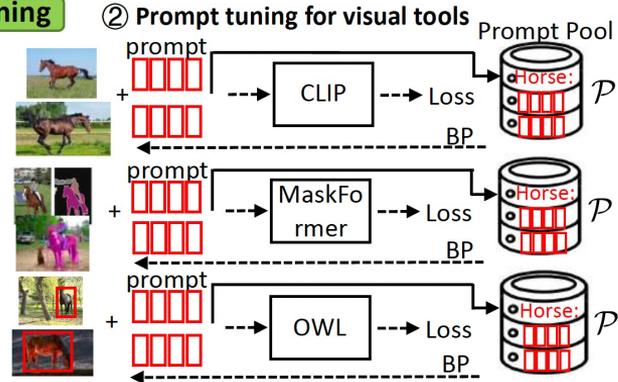
'horse' 'horse'

① Prompt engineering for LLMs

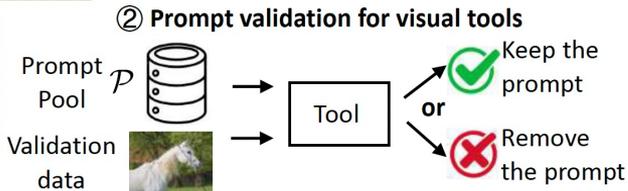
Correct Examples for Plan	Incorrect Examples for Plan
<p>Instruction: Is there a cow in a horse that is not white?</p> <p>Plan:</p> <p>Step1. Locate cow in the image.</p> <p>Step2. Locate horse in the image.</p> <p>Step3. Crop the image region of cow.</p> <p>Step4. Crop the image region of horse.</p> <p>Step5. Ask 'What the color of the cow?' for the crop image in Step3.</p> <p>Step6. Ask 'What the color of the horse?' for the crop image in Step4.</p> <p>Step7. Output the answer.</p>	<p>Instruction: Is there a cow in a horse that is not white?</p> <p>Plan:</p> <p>Step1. Locate cow</p> <p>Step2. Locate horse</p> <p>Step3. Count the number of cow</p> <p>Step4. Count the number of horse</p> <p>Step5. Ask the answer based on the number of cow and horse</p> <p>Step6. Output the answer.</p> <p>Reason: In Step3 and Step4, it should ask the color of the horse and cow instead of counting the number.</p>
Correct Examples for Program	Incorrect Examples for Program
<p>Instruction: Is there a cow in a horse that is not white?</p> <p>Plan:</p> <p>Step1. Locate cow in the image.</p> <p>Step2. Locate horse in the image.</p> <p>Step3. Crop the image region of cow.</p> <p>Step4. Crop the image region of horse.</p> <p>Step5. Ask 'What the color of the cow?' for the crop image in Step3.</p> <p>Step6. Ask 'What the color of the horse?' for the crop image in Step4.</p> <p>Step7. Obtain the answer based on the color of cow and horse.</p> <p>Step8. Output the answer.</p> <p>Program:</p> <pre> BBOX=LOC(image=IMAGE, obj='cow') BBOX2=LOC(image=IMAGE, obj='horse') IMAGE=CROP(image=IMAGE, box=BBOX) IMAGE2=CROP(image=IMAGE, box=BBOX2) ANSWER=VQA(image=IMAGE2, question='What the color of the cow?') ANSWER2=VQA(image=IMAGE2, question='What the color of the horse?') ANSWER3=VQA(image=IMAGE2, question='What the color of the cow?') ANSWER4=VQA(image=IMAGE2, question='What the color of the horse?') Reason: In Step7 of the Program, it does not consider the color of horse, it should be ANSWER2=VQA(image=IMAGE2, question='What the color of the horse?') FINAL_RESULT=RESULT(ANSWER3) </pre>	<p>Instruction: Is there a cow in a horse that is not white?</p> <p>Plan:</p> <p>Step1. Locate cow in the image.</p> <p>Step2. Locate horse in the image.</p> <p>Step3. Crop the image region of cow.</p> <p>Step4. Crop the image region of horse.</p> <p>Step5. Ask 'What the color of the cow?' for the crop image in Step3.</p> <p>Step6. Ask 'What the color of the horse?' for the crop image in Step4.</p> <p>Step7. Obtain the answer based on the color of cow and horse.</p> <p>Step8. Output the answer.</p> <p>Program:</p> <pre> BBOX=LOC(image=IMAGE, obj='cow') BBOX2=LOC(image=IMAGE, obj='horse') IMAGE=CROP(image=IMAGE, box=BBOX) IMAGE2=CROP(image=IMAGE, box=BBOX2) ANSWER=VQA(image=IMAGE2, question='What the color of the cow?') ANSWER2=VQA(image=IMAGE2, question='What the color of the horse?') ANSWER3=VQA(image=IMAGE2, question='What the color of the cow?') ANSWER4=VQA(image=IMAGE2, question='What the color of the horse?') Reason: In Step7 of the Program, it does not consider the color of horse, it should be ANSWER2=VQA(image=IMAGE2, question='What the color of the horse?') FINAL_RESULT=RESULT(ANSWER3) </pre>



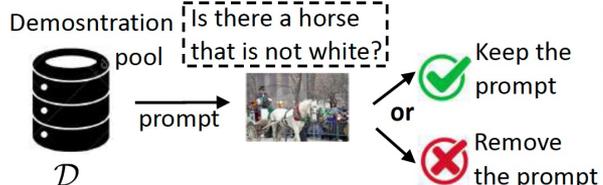
Training



Validation



① Prompt validation for LLMs



Prompt engineering for LLMs

Save examples to the demonstration pool.

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.

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 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

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'").

Demonstration Pool



D

Tag the director of the movie Parasite.

FACEDET
LIST
CLASSIFY
TAG

Tagged the wrong person.

CLASSIFY fails on Bong Joon-ho. Update it.

Replace Leonardo with Dominik Sadoch.

FACEDET
SELECT
REPLACE SD

It is not Dominik Sadoch.

REPLACE fails on Dominik. Update it.

At least one image shows a living mammal surrounded by vultures.

The statement is FALSE.

VQA
VQA
XOR eval XOR

The statement should be TRUE.

LLMs should use OR function instead of XOR function. Update LLMs by using this as in-context examples.

Is there any windmill in the upper part of the image?

no

CROP
LOC
IF eval IF

The correct answer is yes.

LOC fails on windmill. Update it.

Tag the best director of the 92nd Oscar.

FACEDET
LIST
CLASSIFY
TAG

Replace the lion with Dominik Sadoch.

FACEDET
SELECT
REPLACE SD

At least one image shows the dog with its mouth over the edge of the watermelon.

The statement is FALSE.

VQA
VQA
OR eval OR

Is there a airplane below the windmill?

yes

LOC
LOC
COUNT
IF eval IF

(a) Update the CLASSIFY tool

(b) Update the REPLACE tool

(c) Update LLMs

(d) Update the LOC tool

Qualitative Examples

Reflection for the REPLACE model in an image editing task

Instruction: Replace the bird with pine grosbeak (a kind of Passeriformes)
Environmental 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.

Input image



Step1 SEG result



Step2 SELECT result



Step3 REPLACE result



Update the REPLACE model

Search data from the Internet



Train prompts for the Replace model

SD
REPLACE

Evaluate the updated REPLACE model 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')

W/o updating the REPLACE model



W/ updating the REPLACE model



Results

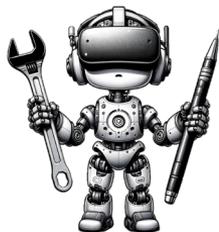
	Method	GQA	NLVRv2	Image Editing	Tagging
End-to-end	Otter [24]	48.2	48.2	-	-
	MMICL [76]	64.4	62.2	-	-
Tool-usage	GPT4TOOLS [67]	41.2	45.4	17.8	-
	Visual ChatGPT [67]	43.2	51.6	21.7	-
	InternGPT [30]	44.8	39.4	-	-
	HuggingGPT [53]	46.0	44.0	-	-
	ViperGPT [58]	47.2	-	-	-
	VISPROG [11]	49.8	60.8	40.2	39.3
	CLOVA (Ours)	54.6	65.6	65.4	50.2

Takeaway

- Feedback learning is an important ability for multimodal agent.
- We build **CLOVA**, the first visual assistant that can **self-improve and continuously learn**, via a closed-loop learning framework with **inference**, **reflection**, and **learning** phases.
- We present **FIRE**, a comprehensive dataset to empower **end-to-end methods** the ability of feedback-refining. It also provides rich settings to evaluate the feedback-refining ability of models.

Summary

- Fundamental ideas of multimodal generalist agents with large models:
 - Planning (reasoning) and tool-use
 - (Gradient-free) reflecting and learning from feedback
 - Structured memory as a unified tool interface
- Two examples of multimodal generalist agents:
 - **VideoAgent**: memory-augmented multimodal agent
 - **CLOVA + FIRE**: reflecting and learning from feedback
- Join us and check out more!



Q & A?

 **Kudos to the team members @ BIGAI-ML !**

