CVPR
JUNE 17-21, 2024
SEATTLE, WA

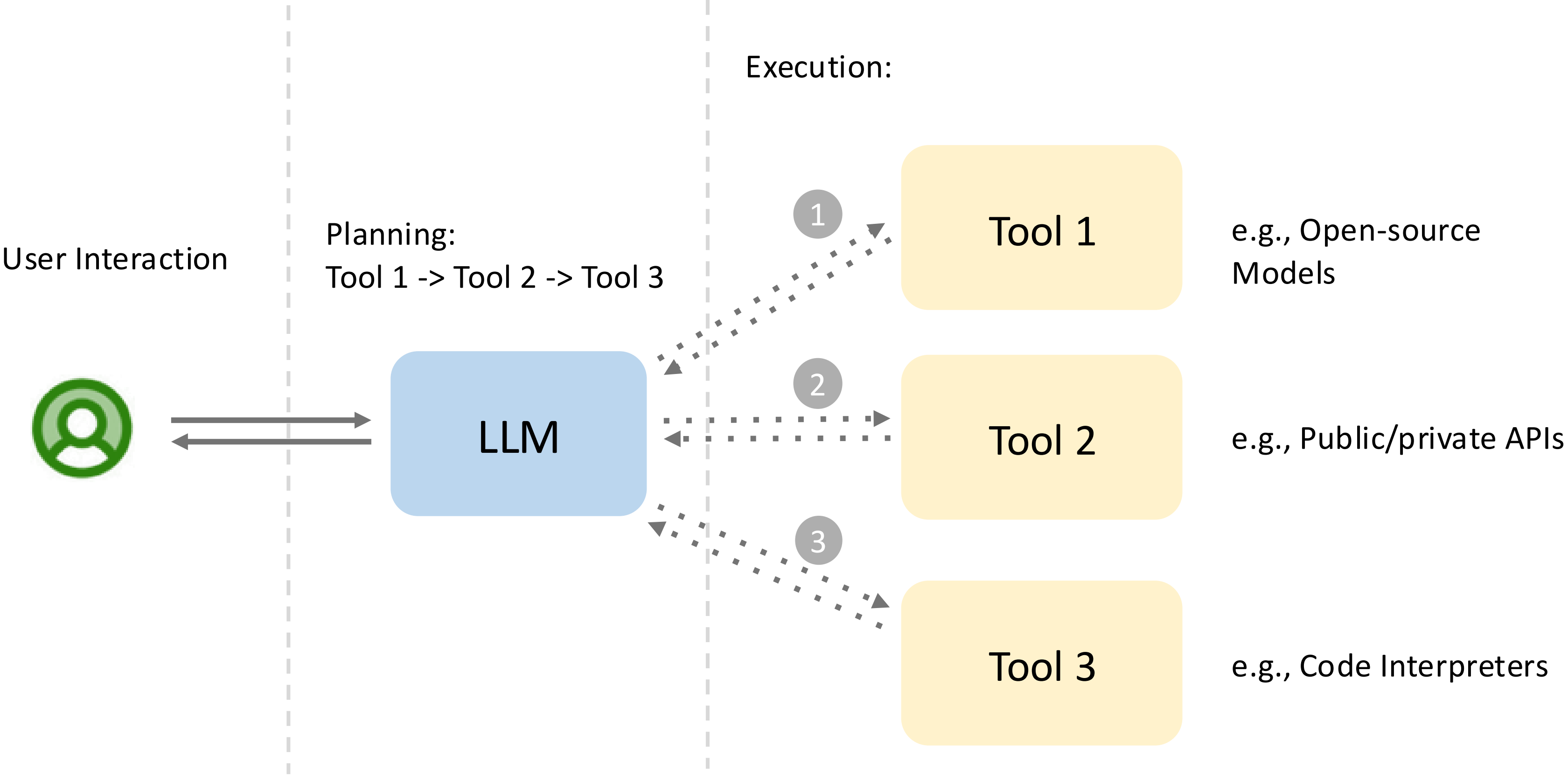
Multimodal Agents

Linjie Li

06/17/2024

Microsoft

CVPR 2023 **Multimodal Agents**: Chaining Multimodal Experts with **LLMs**



User Interaction

Planning:
Tool 1 -> Tool 2 -> Tool 3

LLM

Execution:

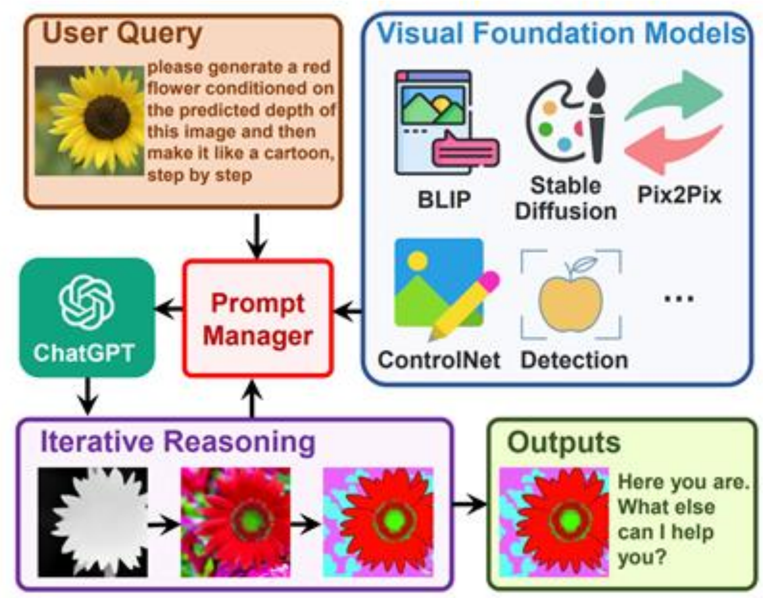
① Tool 1 — e.g., Open-source Models

② Tool 2 — e.g., Public/private APIs

③ Tool 3 — e.g., Code Interpreters

Credit: Multimodal Agent, CVPR 2023 Tutorial on Vision Foundation Models

CVPR 2023 **Multimodal Agents**: Chaining Multimodal Experts with **LLMs**


User Query
please generate a red flower conditioned on the predicted depth of this image and then make it like a cartoon, step by step
Visual Foundation Models
BLIP
Stable Diffusion
Pix2Pix
ControlNet
Detection
...
ChatGPT
Prompt Manager
Iterative Reasoning
Outputs
Here you are. What else can I help you?

Visual ChatGPT

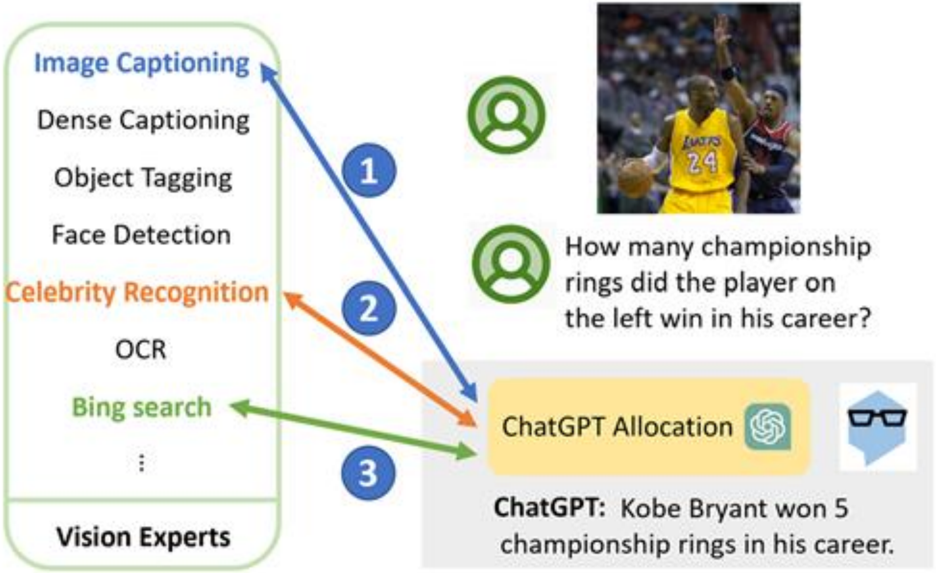
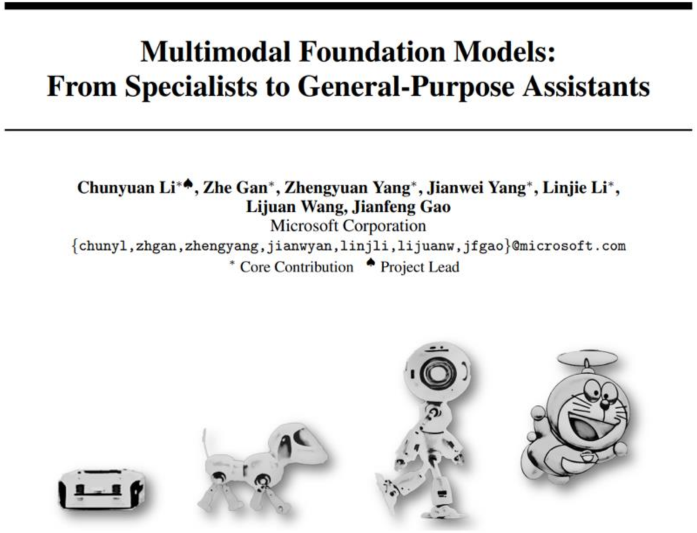


Image Captioning
Dense Captioning
Object Tagging
Face Detection
Celebrity Recognition
OCR
Bing search
⋮
Vision Experts
1
2
3
How many championship rings did the player on the left win in his career?
ChatGPT Allocation
ChatGPT: Kobe Bryant won 5 championship rings in his career.

MM-ReAct

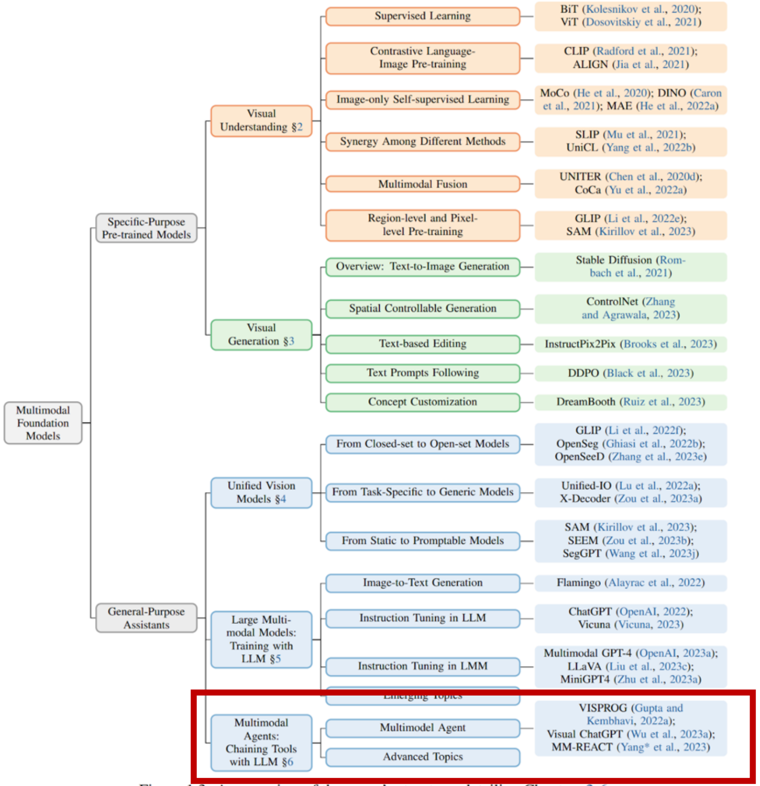
[1] Wu, Chenfei, et al. "Visual chatgpt: Talking, drawing and editing with visual foundation models." arXiv preprint arXiv:2303.04671 (2023).
[2] Yang, Zhengyuan, et al. "Mm-react: Prompting chatgpt for multimodal reasoning and action." arXiv preprint arXiv:2303.11381 (2023).

CVPR 2023 **Multimodal Agents**: Chaining Multimodal Experts with **LLMs**



Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Chunyuan Li*♠, Zhe Gan*, Zhengyuan Yang*, Jianwei Yang*, Linjie Li*, Lijuan Wang, Jianfeng Gao
Microsoft Corporation
{chunyl,zhgan,zhengyang,jianwyan,linjli,lijuanw,jfgao}@microsoft.com
* Core Contribution ♠ Project Lead



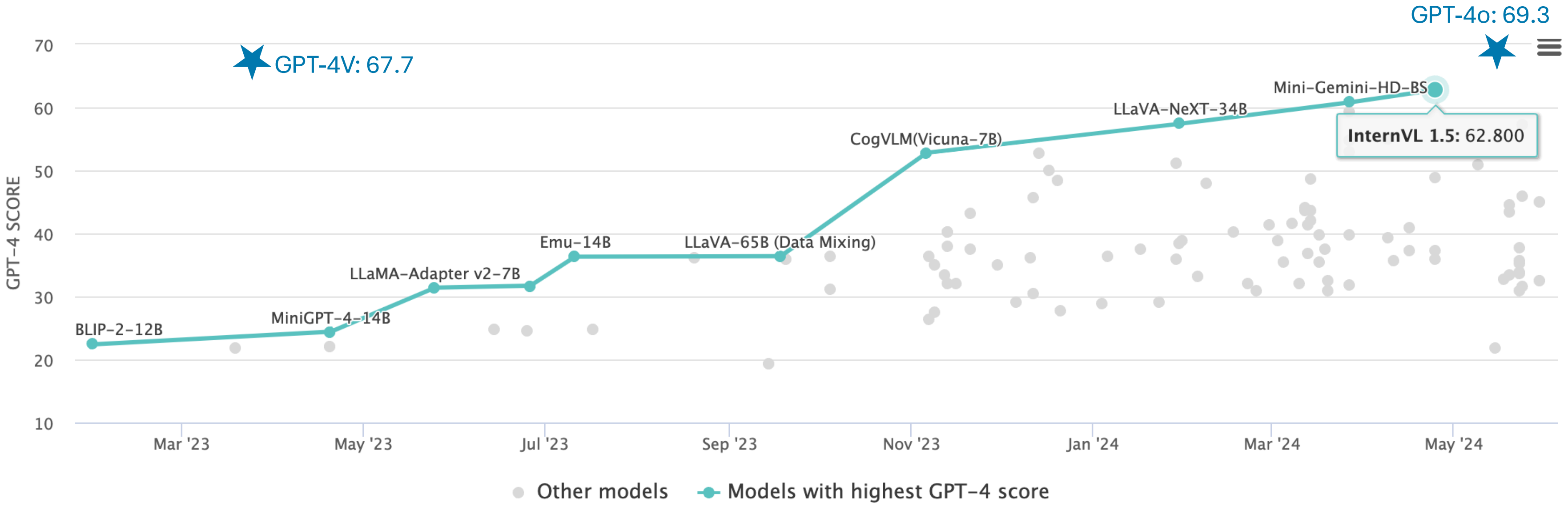
- Multimodal Foundation Models
  - Specific-Purpose Pre-trained Models
    - Visual Understanding §2
      - Supervised Learning: BiT (Kolesnikov et al., 2020); ViT (Dosovitskiy et al., 2021)
      - Contrastive Language-Image Pre-training: CLIP (Radford et al., 2021); ALIGN (Jia et al., 2021)
      - Image-only Self-supervised Learning: MoCo (He et al., 2020); DINO (Caron et al., 2021); MAE (He et al., 2022a)
      - Synergy Among Different Methods: SLIP (Mu et al., 2021); UniCL (Yang et al., 2022b)
      - Multimodal Fusion: UNITER (Chen et al., 2020d); CoCa (Yu et al., 2022a)
      - Region-level and Pixel-level Pre-training: GLIP (Li et al., 2022e); SAM (Kirillov et al., 2023)
    - Visual Generation §3
      - Overview: Text-to-Image Generation: Stable Diffusion (Rombach et al., 2021)
      - Spatial Controllable Generation: ControlNet (Zhang and Agrawala, 2023)
      - Text-based Editing: InstructPix2Pix (Brooks et al., 2023)
      - Text Prompts Following: DDPO (Black et al., 2023)
      - Concept Customization: DreamBooth (Ruiz et al., 2023)
  - General-Purpose Assistants
    - Unified Vision Models §4
      - From Closed-set to Open-set Models: GLIP (Li et al., 2022f); OpenSeg (Ghiasi et al., 2022b); OpenSeeD (Zhang et al., 2023e)
      - From Task-Specific to Generic Models: Unified-IO (Lu et al., 2022a); X-Decoder (Zou et al., 2023a)
      - From Static to Promptable Models: SAM (Kirillov et al., 2023); SEEM (Zou et al., 2023b); SegGPT (Wang et al., 2023j)
    - Large Multi-modal Models: Training with LLM §5
      - Image-to-Text Generation: Flamingo (Alayrac et al., 2022)
      - Instruction Tuning in LLM: ChatGPT (OpenAI, 2022); Vicuna (Vicuna, 2023)
      - Instruction Tuning in LMM: Multimodal GPT-4 (OpenAI, 2023a); LLaVA (Liu et al., 2023c); MiniGPT4 (Zhu et al., 2023a)
      - Emerging Topics
    - Multimodal Agents: Chaining Tools with LLM §6
      - Multimodel Agent: VISPROG (Gupta and Kembhavi, 2022a); Visual ChatGPT (Wu et al., 2023a); MM-REACT (Yang* et al., 2023)
      - Advanced Topics

Figure 1.3: An overview of the paper's structure, detailing Chapters 2-6.

[1] Li, Chunyuan, et al. "Multimodal foundation models: From specialists to general-purpose assistants." *Foundations and Trends® in Computer Graphics and Vision* 16.1-2 (2024): 1-214.

Rapid Progress in LMMs

MM-Vet : Evaluating integrated vision-language capabilities



[1] Yu, Weihao, et al. "Mm-vet: Evaluating large multimodal models for integrated capabilities." ICML 2024

https://github.com/yuweihao/MM-Vet

Do We Still Need Multimodal Agents?

Agents with Multimodal Memory

MM-Narrator MM-Vid

Actionable Agents

MM-Navigator VideoGUI

Agent with Feedback

Idea2Img

Audio Description

GUI Navigation

Visual Design & Creation

Do We Still Need Multimodal Agents?

Agents with Multimodal Memory

MM-Narrator MM-Vid

Audio Description

Actionable Agents

MM-Navigator VideoGUI

GUI Navigation

Agent with Feedback

Idea2Img

Visual Design & Creation

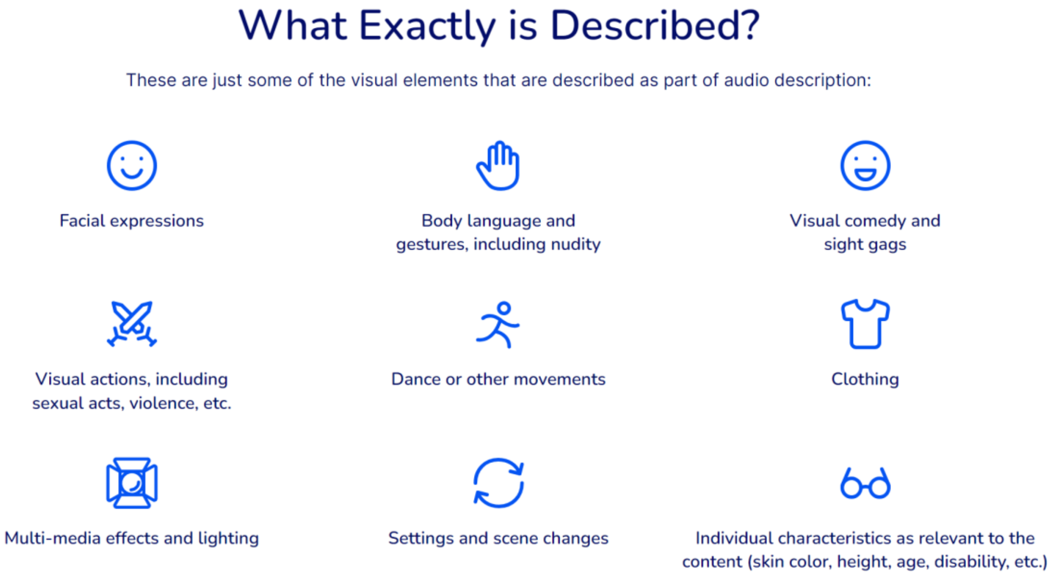
Why Agents Need Multimodal Memory?

Source: https://www.youtube.com/watch?v=XB4cjbYywqg&ab_channel=Apple

Audio Description

What is Audio Description?

Audio description is the verbal depiction of key visual elements in media and live productions. AD is meant to provide information on visual content that is considered essential to the comprehension of the program. In these cases, not providing AD would inhibit blind and visually impaired individuals from gaining a complete understanding of the given program.

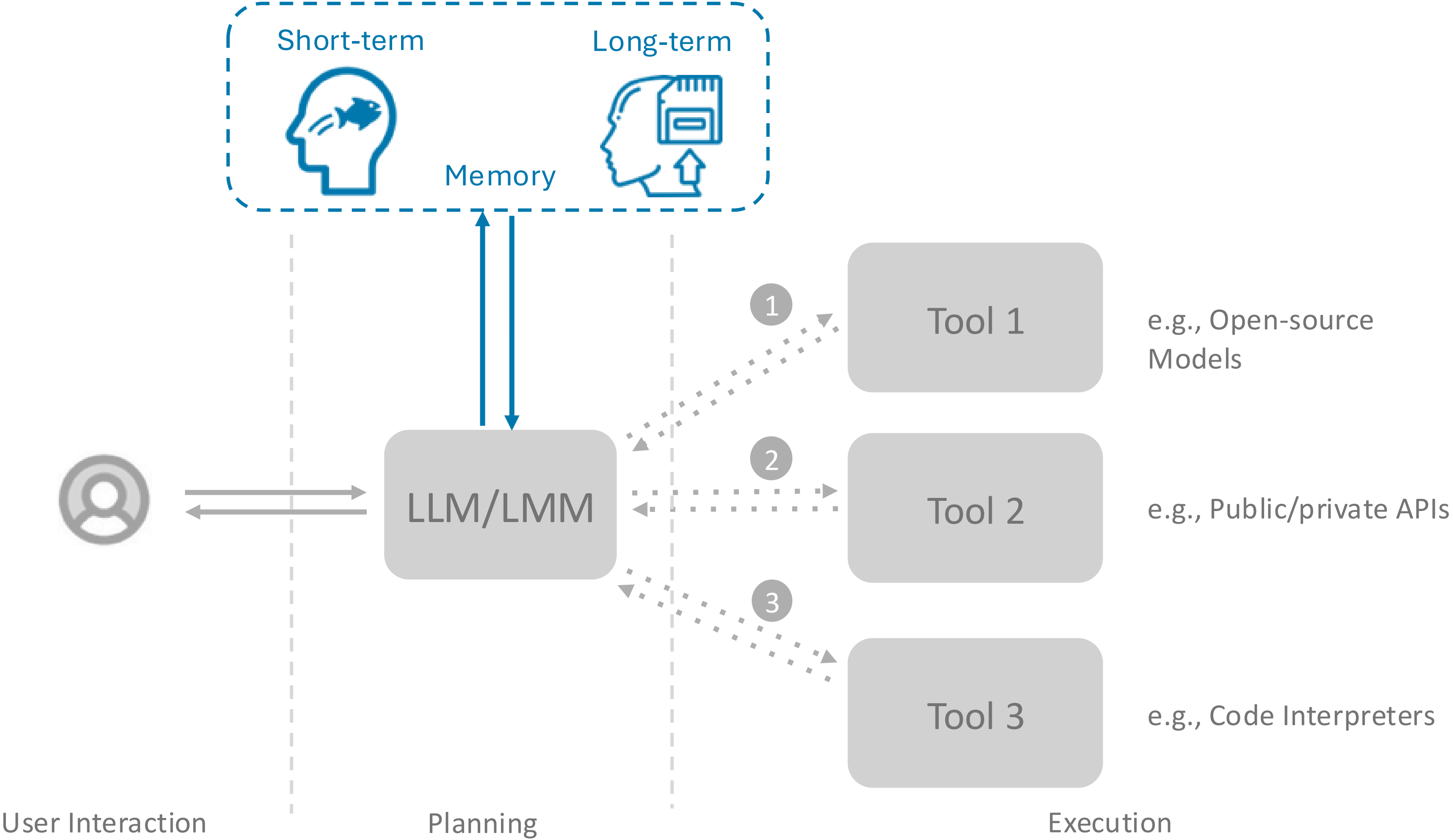


What Exactly is Described?

These are just some of the visual elements that are described as part of audio description:

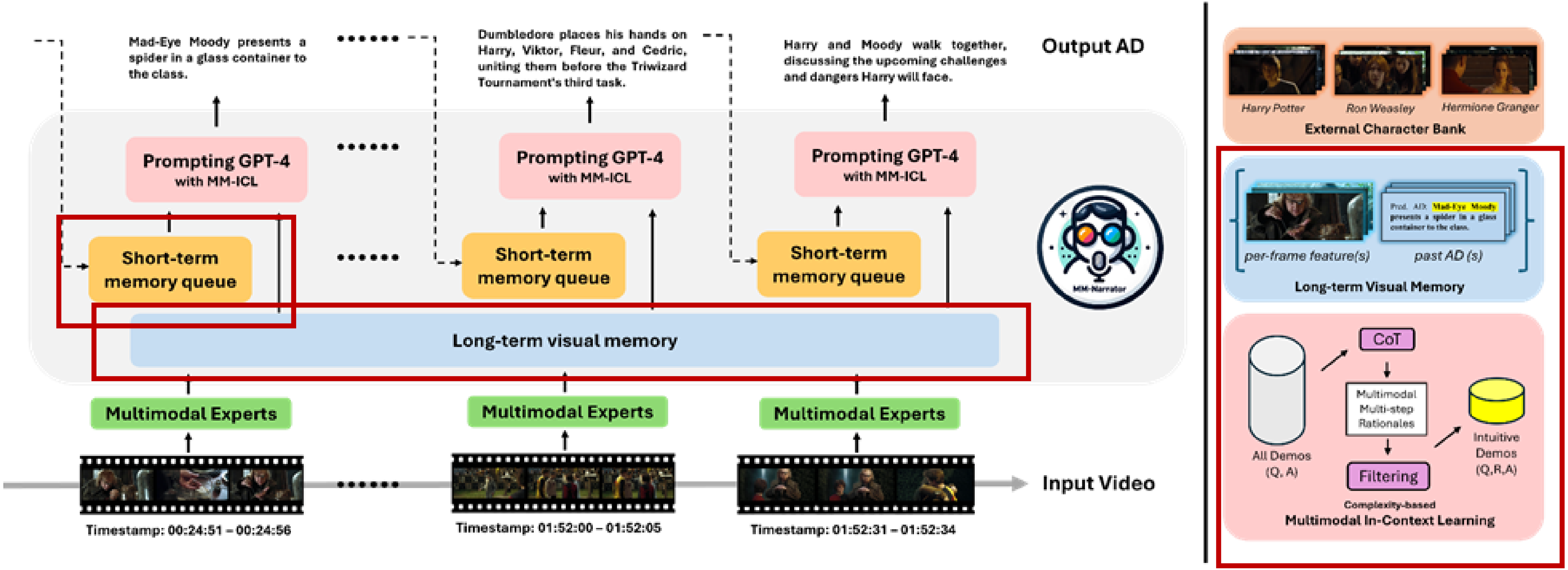
- Facial expressions
- Body language and gestures, including nudity
- Visual comedy and sight gags
- Visual actions, including sexual acts, violence, etc.
- Dance or other movements
- Clothing
- Multi-media effects and lighting
- Settings and scene changes
- Individual characteristics as relevant to the content (skin color, height, age, disability, etc.)

Source: https://vitac.com/audio-description/


Short-term
Long-term
Memory
Tool 1
e.g., Open-source Models
LLM/LMM
1
2
3
Tool 2
e.g., Public/private APIs
Tool 3
e.g., Code Interpreters
User Interaction
Planning
Execution

Narrating Long-form Videos with Long-term and Short-term Memories

MM-Narrator


Mad-Eye Moody presents a spider in a glass container to the class.

Dumbledore places his hands on Harry, Viktor, Fleur, and Cedric, uniting them before the Triwizard Tournament's third task.

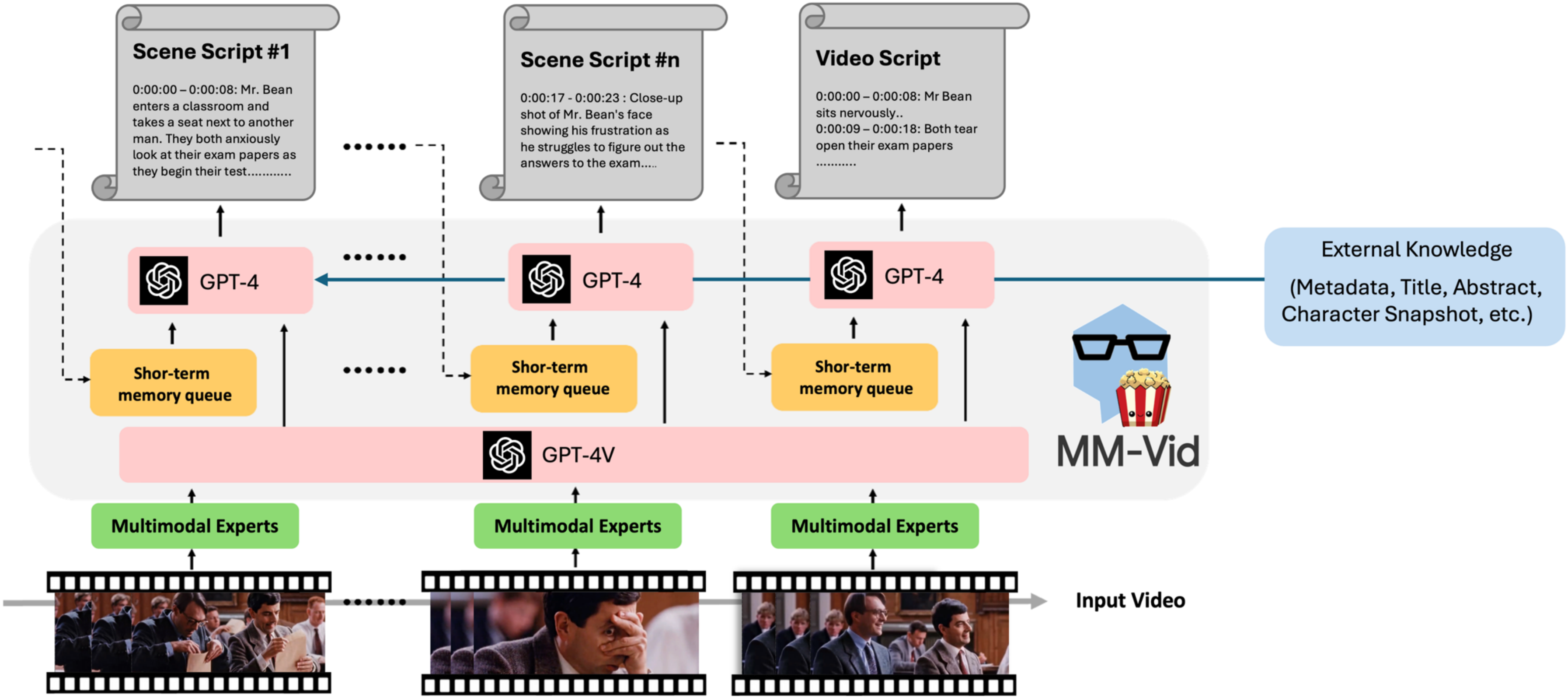
Harry and Moody walk together, discussing the upcoming challenges and dangers Harry will face.

[1] Zhang, Chaoyi, et al. "Mm-narrator: Narrating long-form videos with multimodal in-context learning." CVPR 2024

https://mm-narrator.github.io/

Advancing Video Understanding with GPT-4V

MM-Vid



Scene Script #1

0:00:00 – 0:00:08: Mr. Bean enters a classroom and takes a seat next to another man. They both anxiously look at their exam papers as they begin their test...........

Scene Script #n

0:00:17 - 0:00:23 : Close-up shot of Mr. Bean's face showing his frustration as he struggles to figure out the answers to the exam.....

Video Script

0:00:00 – 0:00:08: Mr Bean sits nervously..
0:00:09 – 0:00:18: Both tear open their exam papers
....

GPT-4

GPT-4

GPT-4

External Knowledge

(Metadata, Title, Abstract, Character Snapshot, etc.)

Shor-term memory queue

Shor-term memory queue

Shor-term memory queue

GPT-4V

MM-Vid

Multimodal Experts

Multimodal Experts

Multimodal Experts

Input Video

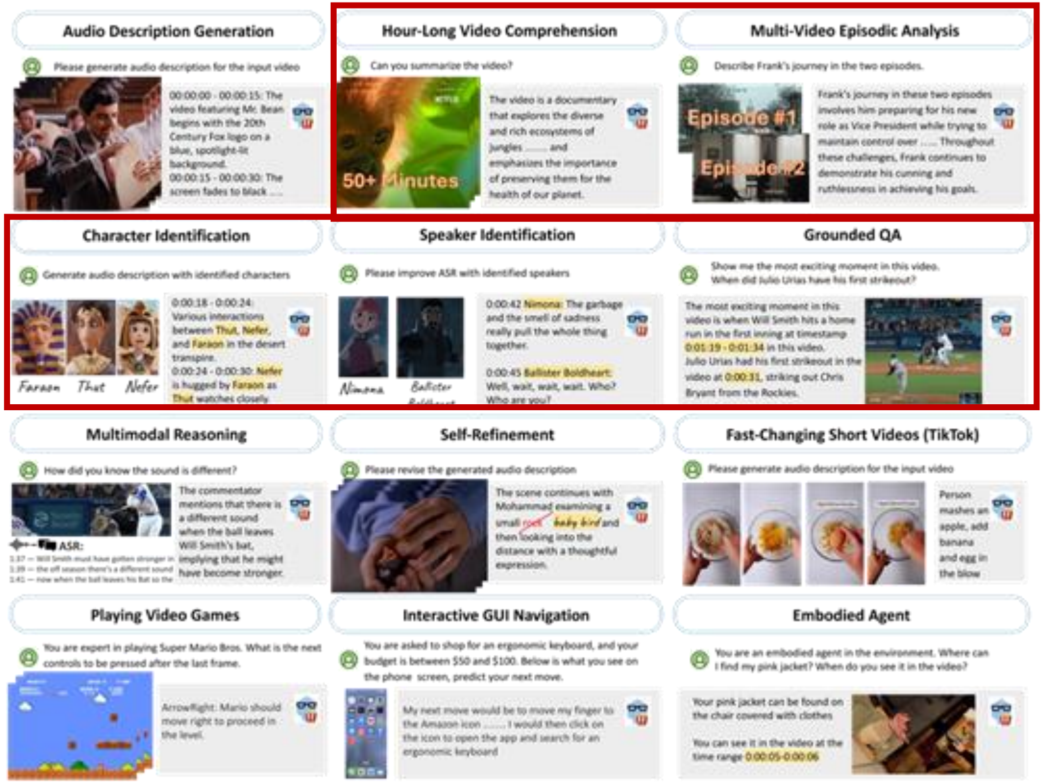
[1] Lin, Kevin, et al. "Mm-vid: Advancing video understanding with gpt-4v (ision)."

https://multimodal-vid.github.io/

Advancing Video Understanding with GPT-4V

MM-Vid

- Can handle various video lengths and domains
- Showcase for various granularities of video understanding
- Demonstrate potentials for actionable agents



[1] Lin, Kevin, et al. "Mm-vid: Advancing video understanding with gpt-4v (ision)."

https://multimodal-vid.github.io/

Case Study I: Egocentric Video Understanding

MM-Vid

Source: Ego4D [2]

Loading video ...
Video length is 1 minute 48 seconds.

Script Generation:

0:00:00-0:00:04: The living room is filled with a mix of furniture and items.

....
....

Where can I find my pink jacket?

Your pink jacket can be found on the chair covered with clothes.

When did I pass by it?

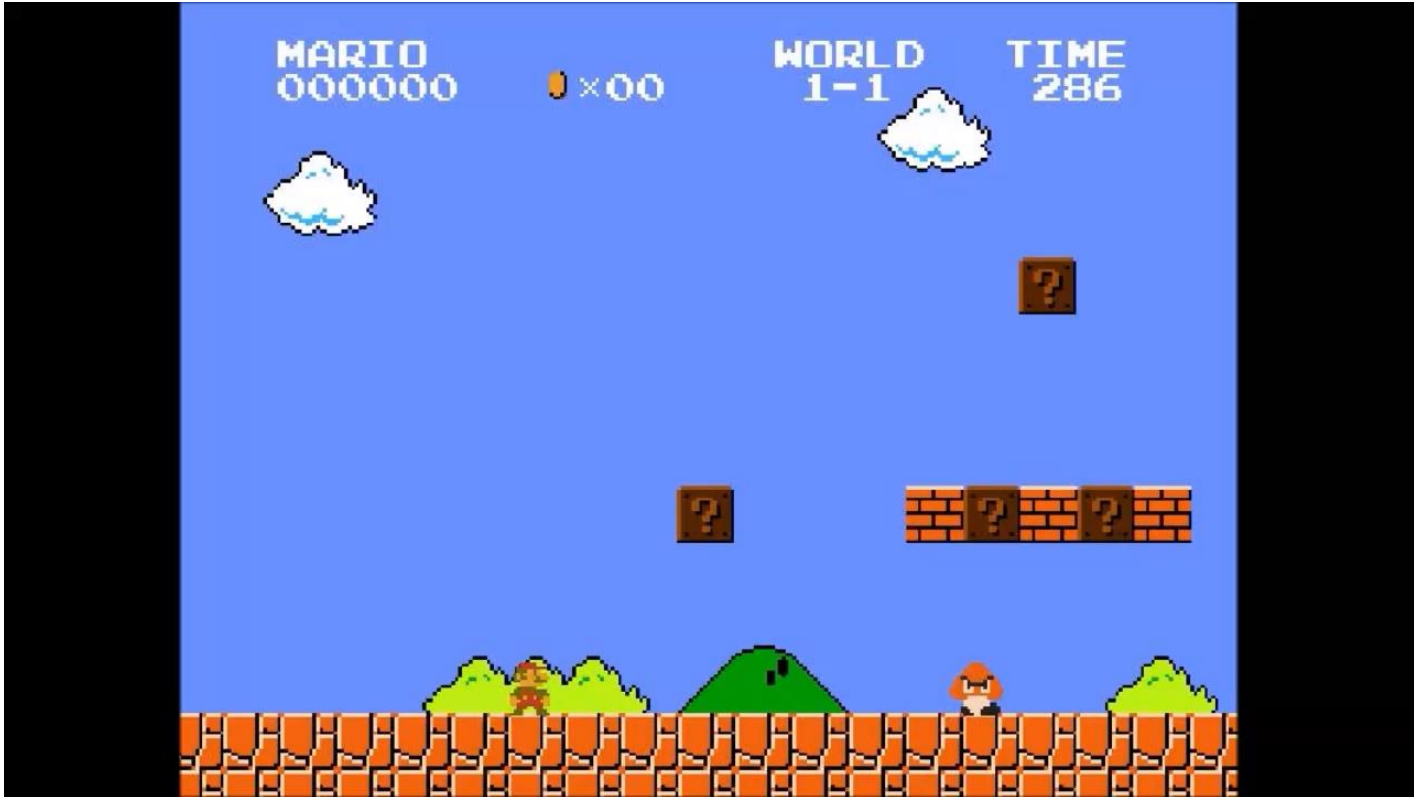
You can see it in the video at the time range 0:00:05 – 0:00:06.

[1] Lin, Kevin, et al. "Mm-vid: Advancing video understanding with gpt-4v (ision)."
[2] Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." CVPR 2022.

https://multimodal-vid.github.io/

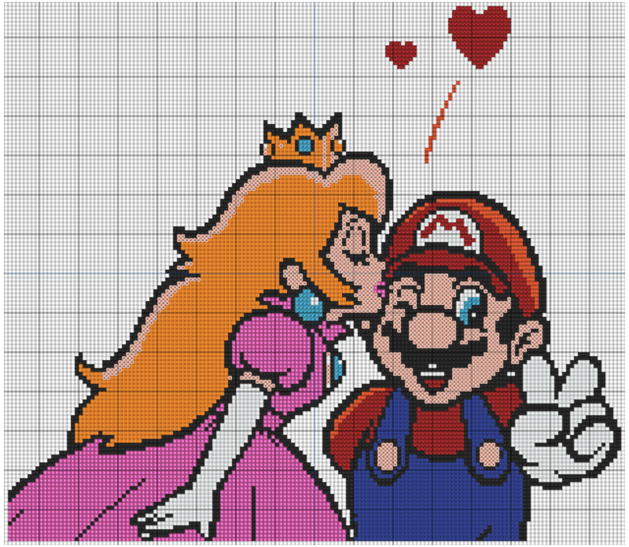
Case Study II: Perception 👉 Interaction with Environment

MM-Vid

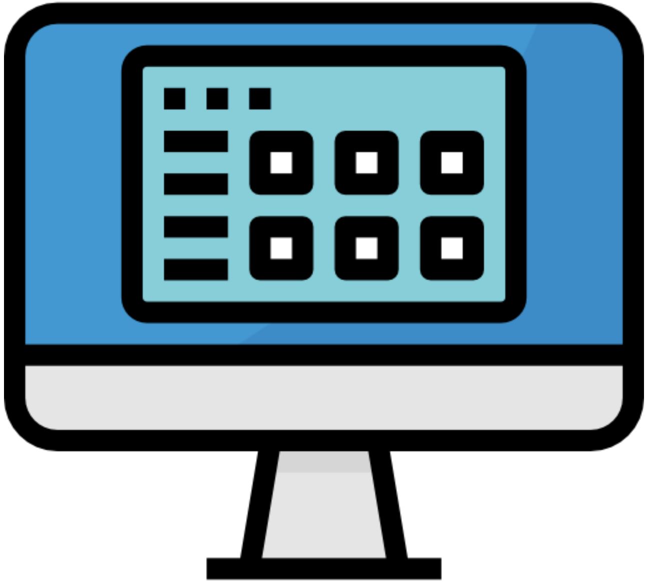


[1] Lin, Kevin, et al. "Mm-vid: Advancing video understanding with gpt-4v (ision)."

https://multimodal-vid.github.io/

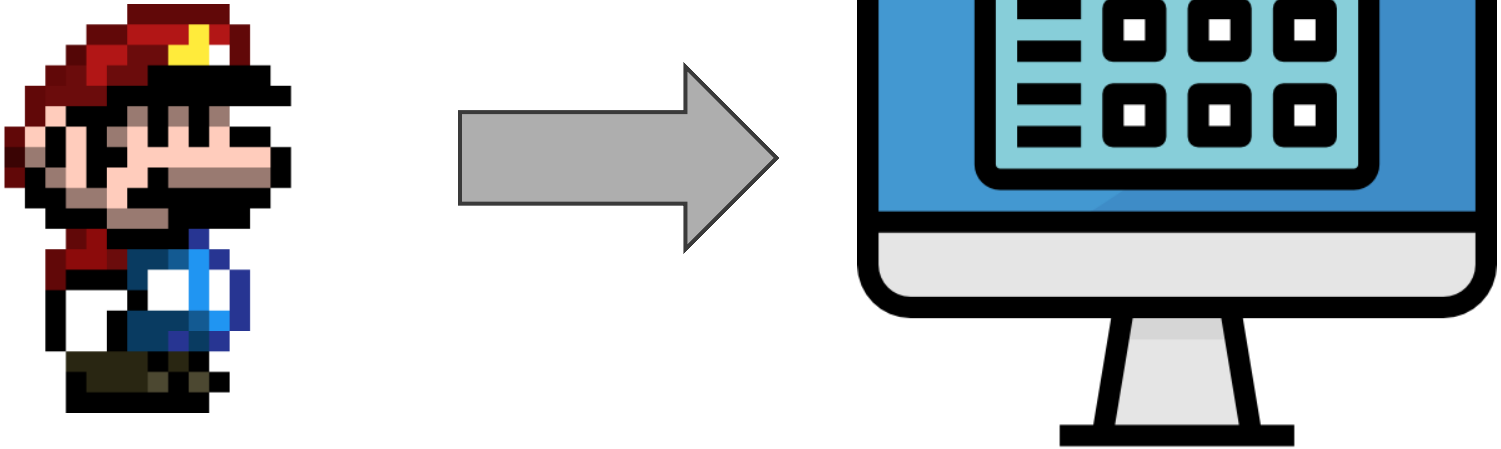


Task: Rescue Princess!



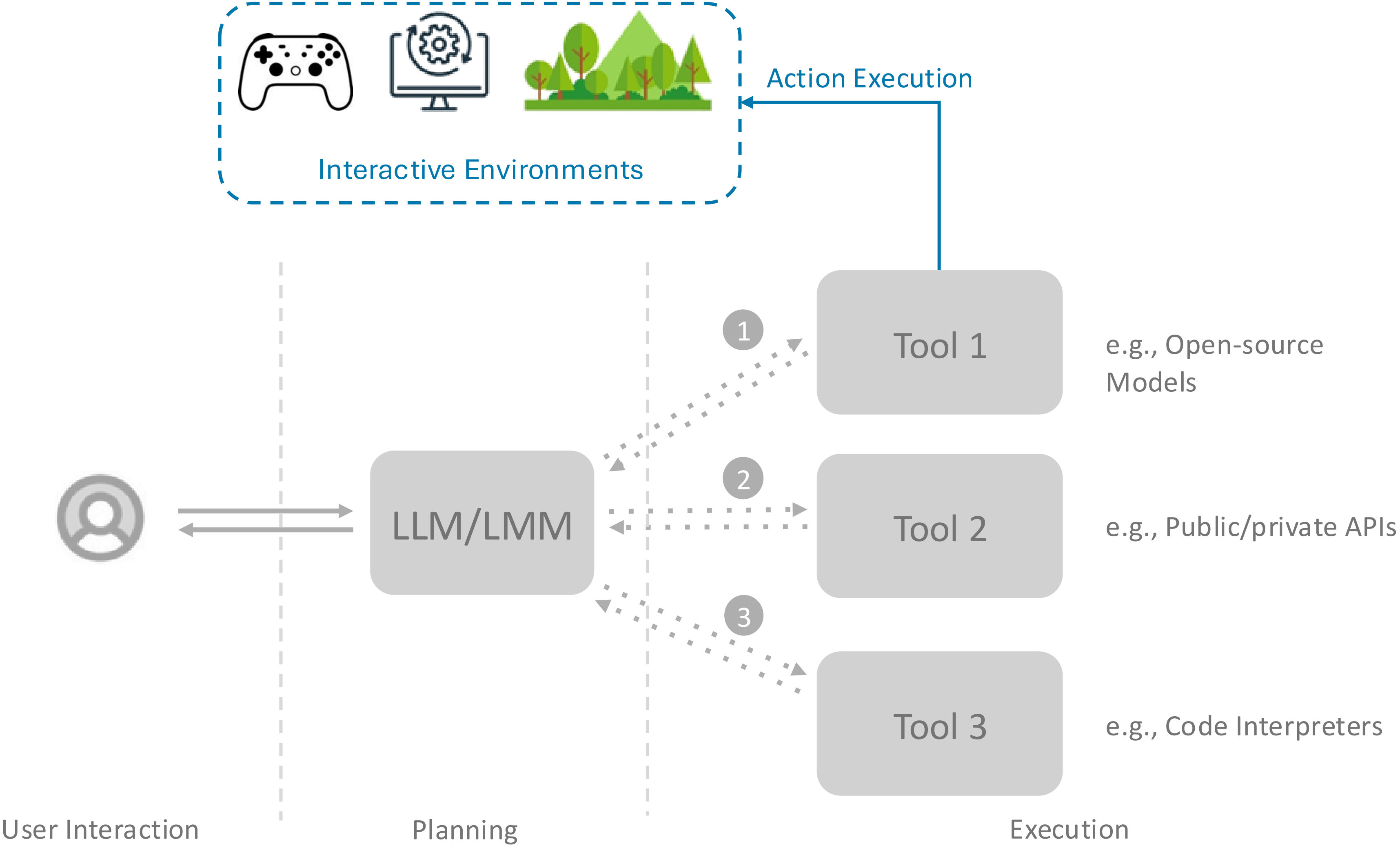
Tasks:

- AI Companions (e.g., Microsoft CoPilot)
- Personal AI Assistant
- Office Automation
- ...



- Jump
- Left
- Right
- …

- Click, what and where?
- Type, what and where?
- Select, what and where?
- Drag, what and from where to where?
- …


Interactive Environments
Action Execution
Tool 1
e.g., Open-source Models
1
2
LLM/LMM
Tool 2
e.g., Public/private APIs
3
Tool 3
e.g., Code Interpreters
User Interaction
Planning
Execution

Do We Still Need Multimodal Agents?

Agents with Multimodal Memory

MM-Narrator MM-Vid

Audio Description

Actionable Agents

MM-Navigator VideoGUI

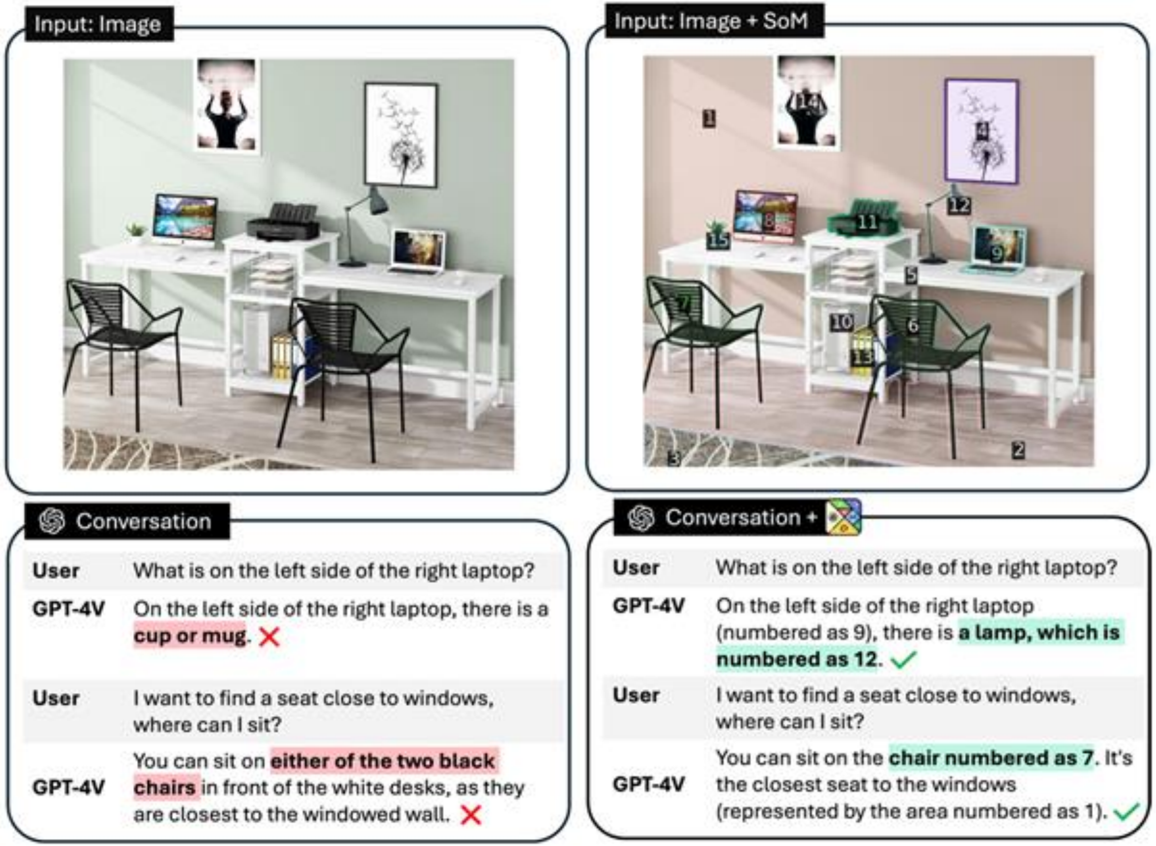
GUI Navigation

Agent with Feedback

Idea2Img

Visual Design & Creation

ReCap: SoM Prompting



Conversation

User	What is on the left side of the right laptop?
GPT-4V	On the left side of the right laptop, there is a **cup or mug**. ✗
User	I want to find a seat close to windows, where can I sit?
GPT-4V	You can sit on **either of the two black chairs** in front of the white desks, as they are closest to the windowed wall. ✗

Conversation +

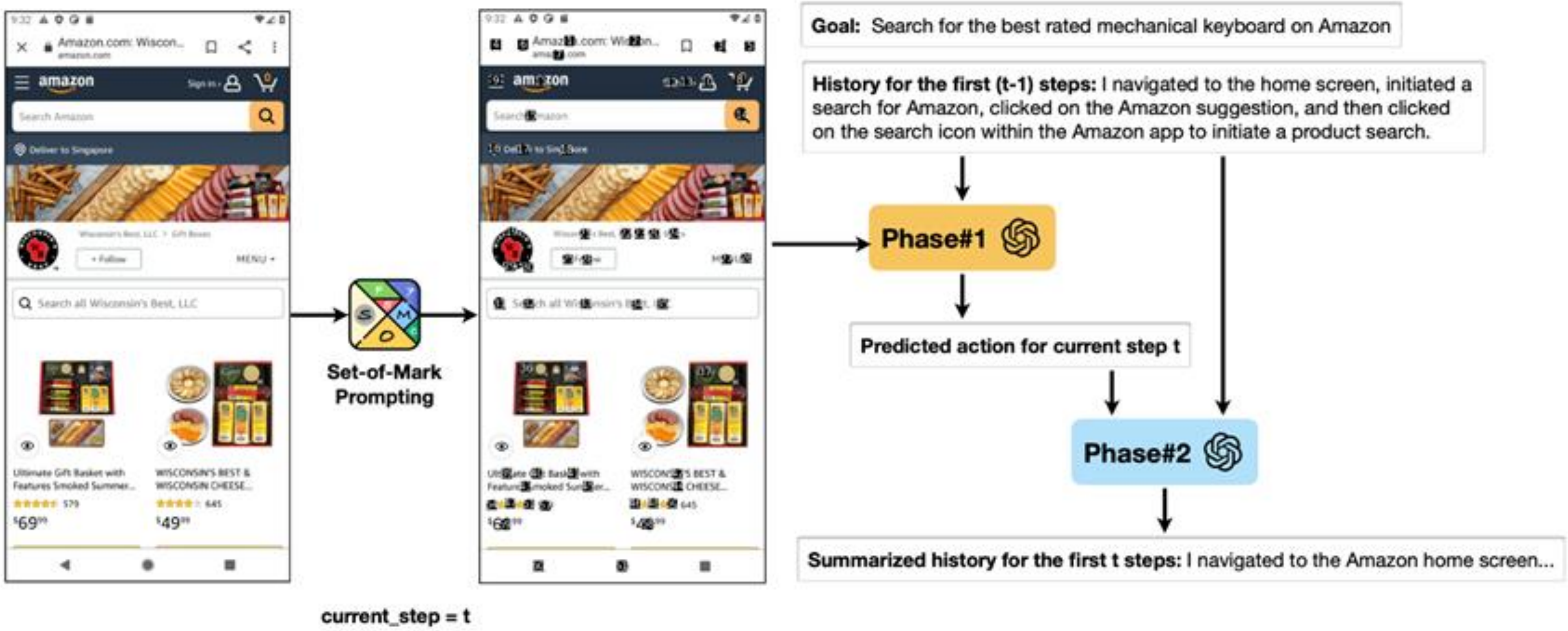
User	What is on the left side of the right laptop?
GPT-4V	On the left side of the right laptop (numbered as 9), there is **a lamp, which is numbered as 12**. ✓
User	I want to find a seat close to windows, where can I sit?
GPT-4V	You can sit on the **chair numbered as 7**. It's the closest seat to the windows (represented by the area numbered as 1). ✓

[1] Yang, Jianwei, et al. "Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V"

https://som-gpt4v.github.io/

LMM Agents for GUI Navigation

MM-Navigator



Goal: Search for the best rated mechanical keyboard on Amazon

History for the first (t-1) steps: I navigated to the home screen, initiated a search for Amazon, clicked on the Amazon suggestion, and then clicked on the search icon within the Amazon app to initiate a product search.

Predicted action for current step t

Summarized history for the first t steps: I navigated to the Amazon home screen...

- Human evaluation on iOS (single-step)
  - Intended Action Description: 90.9%
  - Localized Action Execution: 74.5%

[1] Yan, An, et al. "Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation."

https://github.com/zzxslp/MM-Navigator

Demo with A Sequence of Actions

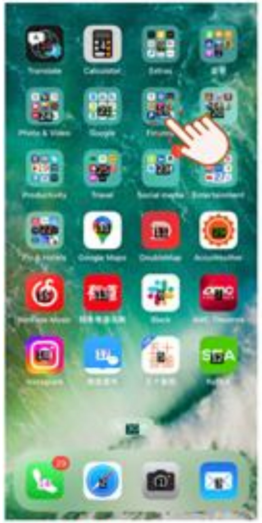
MM-Navigator

Instruction query:

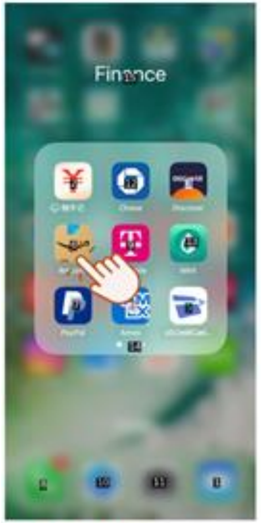
You are asked to shop for a milk frother, your budget is between $50 and $100.

Prompt for all steps:

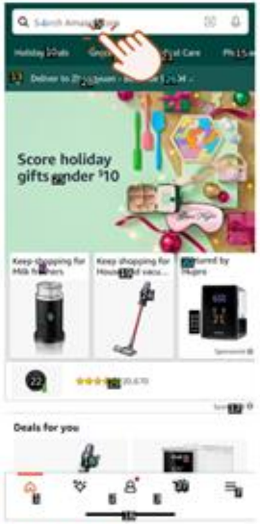
Imagine that you are a robot operating an iPhone.
<u>You are asked to shop for a milk frother, your budget is between $50 and $100.</u>
Your previous action is:
….
Below is what you see on the screen, with numeric IDs added. For the next action, which ID would you click on the screen next?



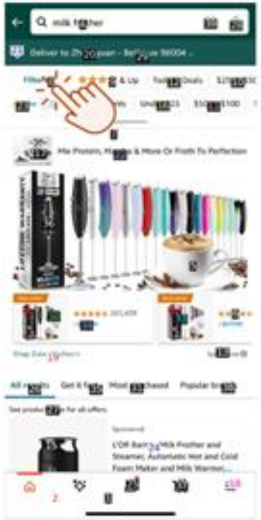
(1) GPT-4V:
To shop for a milk frother within the specified budget using the iPhone interface shown in the image, I would select an application that is capable of online shopping. In this scenario, I would click on ID 16 to use the Amazon app, as it is a comprehensive and user-friendly platform for purchasing a wide variety of products, including kitchen appliances like milk frothers.



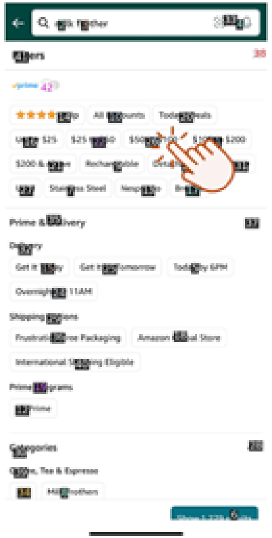
(2) GPT-4V:
To proceed with shopping for a milk frother within the specified budget using the Amazon app, the next step would be to click on ID 6, which represents the Amazon app on the screen. This will open the app and allow me to search for milk frothers within the budget range.



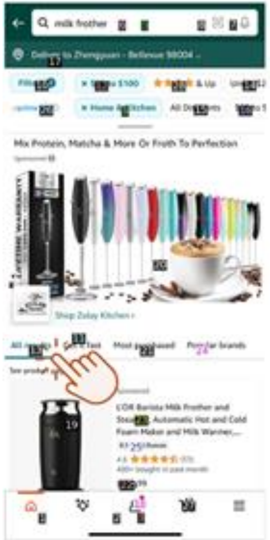
(3) GPT-4V:
I would click on the search bar at the top of the Amazon app interface, which is typically where you would enter a search query. This search bar is labeled with ID 5. By clicking on this, I can then type in "milk frother" and set the price filter to show results that are between $50 and $100.



(4) GPT-4V:
The next step would be to refine the search results by setting the price filter. In this interface, the ID most likely to represent the filter or sorting options is 8 labeled "Filters," which seems to be the right place to start refining the search results according to the budget. So the next action would be to click on ID 8 to access filtering options, including setting the desired price range.



(5) GPT-4V:
The logical next step would be to click on the price filter that matches your budget range. Looking at the screen, you would choose the price filter that includes $50 to $100. The closest available options are ID 25 for "$25 to $50" and ID 26 for "$50 to $100". Since your budget starts at $50, you would click on ID 26 to filter the search results to show only milk frothers that are priced between $50 and $100.



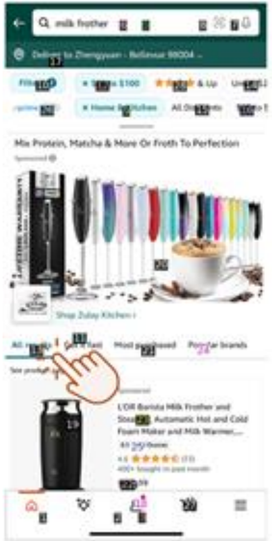
(6) GPT-4V:
The next action would be to click on ID 13, "All results," to view all available milk frothers within the selected price range of $50 to $100.

[1] Yan, An, et al. "Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation."

https://github.com/zzxslp/MM-Navigator

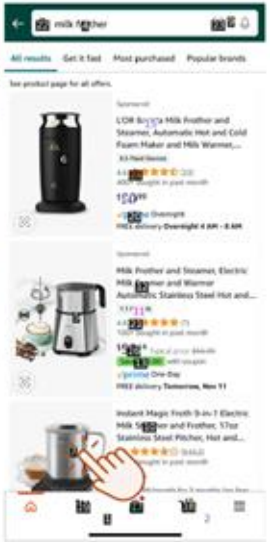
Demo with A Sequence of Actions

MM-Navigator



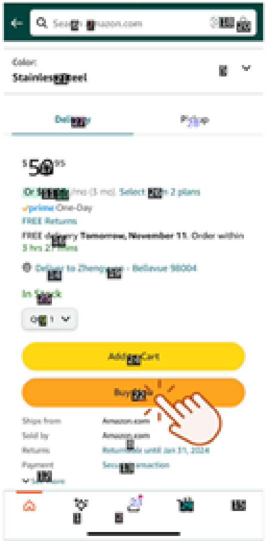
(6) GPT-4V:

The next action would be to click on ID 13, "All results," to view all available milk frothers within the selected price range of $50 to $100.



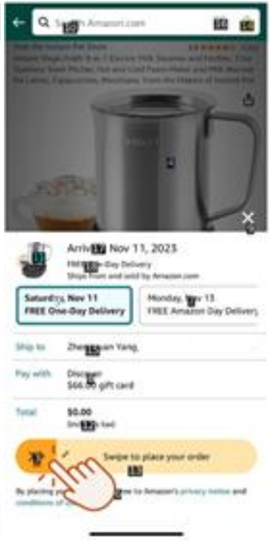
(7) GPT-4V:

The next action would be to click on ID 7, which shows the "Instant Magic Froth 9-in-1 Electric Milk Steamer and Frother" listed at $59.95. This item is within the specified budget of $50 to $100.



(8) GPT-4V:

The next action would be to click on ID 22, which is the "Buy Now" button for the item that was selected in the previous action, the "Instant Magic Froth 9-in-1 Electric Milk Steamer and Frother" priced at $59.95.



(9) GPT-4V:

The next action would be to click on ID 3, which is the button to "Swipe to place your order". This would finalize the purchase of the milk frother selected.

[1] Yan, An, et al. "Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation."

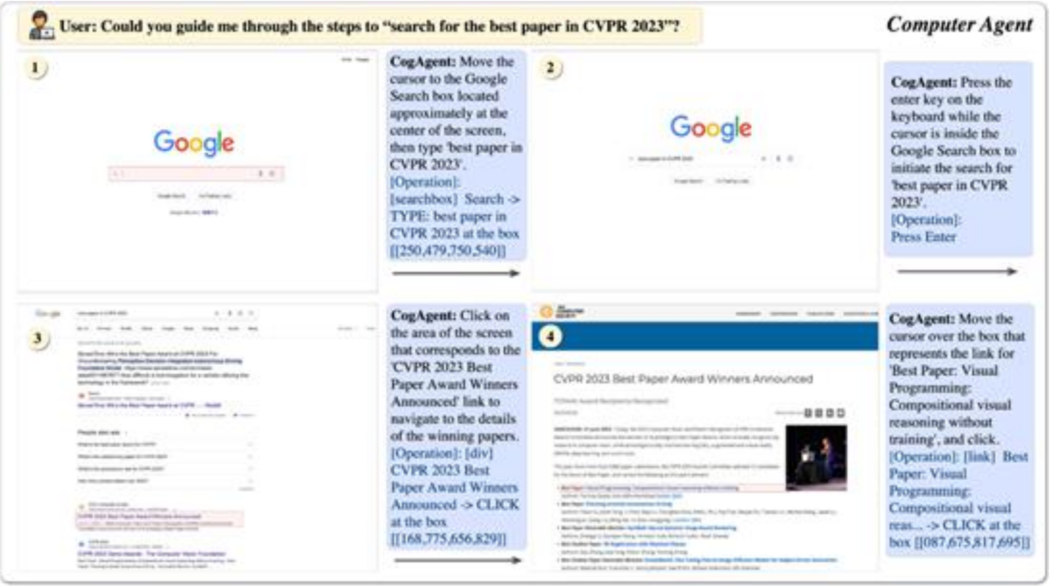
https://github.com/zzxslp/MM-Navigator

Beyond **MM-Navigator**

- Building Agents with (Open-source) Small LMMs
  - Lower cost, lower latency

[1] Lee, Kenton, et al. “Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding.” ICML2023
[2] Gur, Izzeddin, et al. "A real-world webagent with planning, long context understanding, and program synthesis." ICLR2024
[3] Furuta, Hiroki, et al. "Multimodal web navigation with instruction-finetuned foundation models." ICLR2024
[4] Cheng, Kanzhi, et al. "Seeclick: Harnessing gui grounding for advanced visual gui agents." ACL2024
[5] Hong, Wenyi, et al. "Cogagent: A visual language model for gui agents." CVPR 2024.
[6] You, Keen, et al. "Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs.”

Empowering Open-source Models for GUI Navigation



User: Could you guide me through the steps to "search for the best paper in CVPR 2023"?

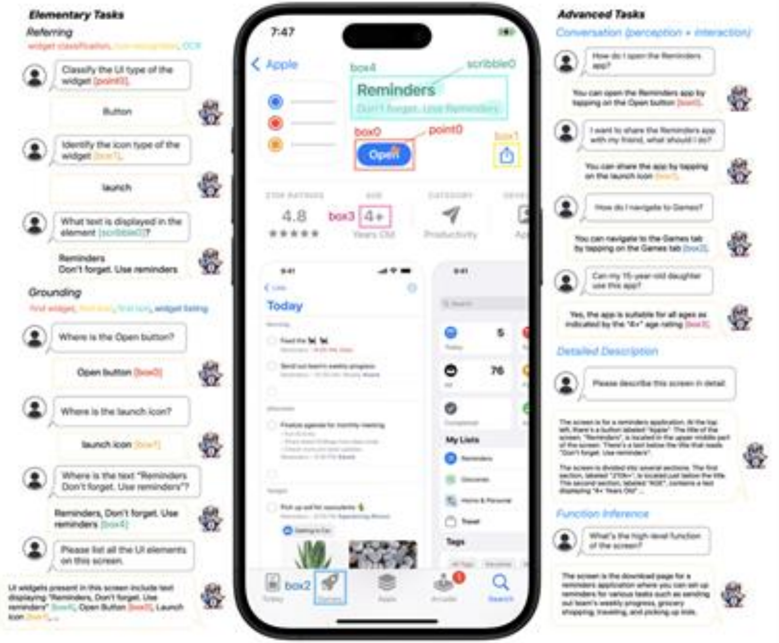
CogAgent: Move the cursor to the Google Search box located approximately at the center of the screen, then type 'best paper in CVPR 2023'. [Operation]: [searchbox] Search -> TYPE: best paper in CVPR 2023 at the box [[250,479,750,540]]

CogAgent: Press the enter key on the keyboard while the cursor is inside the Google Search box to initiate the search for 'best paper in CVPR 2023'. [Operation]: Press Enter

CogAgent: Click on the area of the screen that corresponds to the 'CVPR 2023 Best Paper Award Winners Announced' link to navigate to the details of the winning papers. [Operation]: [div] CVPR 2023 Best Paper Award Winners Announced -> CLICK at the box [[168,775,656,829]]

CogAgent: Move the cursor over the box that represents the link for 'Best Paper: Visual Programming: Compositional visual reasoning without training', and click. [Operation]: [link] Best Paper: Visual Programming: Compositional visual reas... -> CLICK at the box [[087,675,817,695]]

CogAgent



Elementary Tasks

Referring

widget classification, icon recognition, OCR

Classify the UI type of the widget [point0].

Button

Identify the icon type of the widget [box1].

launch

What text is displayed in the element [scribble0]?

Reminders
Don't forget. Use reminders

Grounding

find widget, find icon, find text, widget listing

Where is the Open button?

Open button [box0]

Where is the launch icon?

launch icon [box1]

Where is the text "Reminders Don't forget. Use reminders"?

Reminders, Don't forget. Use reminders [box4]

Please list all the UI elements on this screen.

UI widgets present in this screen include text displaying "Reminders, Don't forget. Use reminders" [box4], Open Button [box0], Launch icon [box1], ...

Advanced Tasks

Conversation (perception + interaction)

How do I open the Reminders app?

You can open the Reminders app by tapping on the Open button [box0].

I want to share the Reminders app with my friend, what should I do?

You can share the app by tapping on the launch icon [box1].

How do I navigate to Games?

You can navigate to the Games tab by tapping on the Games tab [box2].

Can my 15-year-old daughter use this app?

Yes, the app is suitable for all ages as indicated by the "4+" age rating [box3].

Detailed Description

Please describe this screen in detail.

The screen is for a reminders application. At the top left, there's a button labeled "Apple". The title of the screen, "Reminders", is located in the upper middle part of the screen. There's a text below the title that reads "Don't forget. Use reminders".

The screen is divided into several sections. The first section, labeled "210k+", is located just below the title. The second section, labeled "AGE", contains a text displaying "4+ Years Old"...

Function Inference

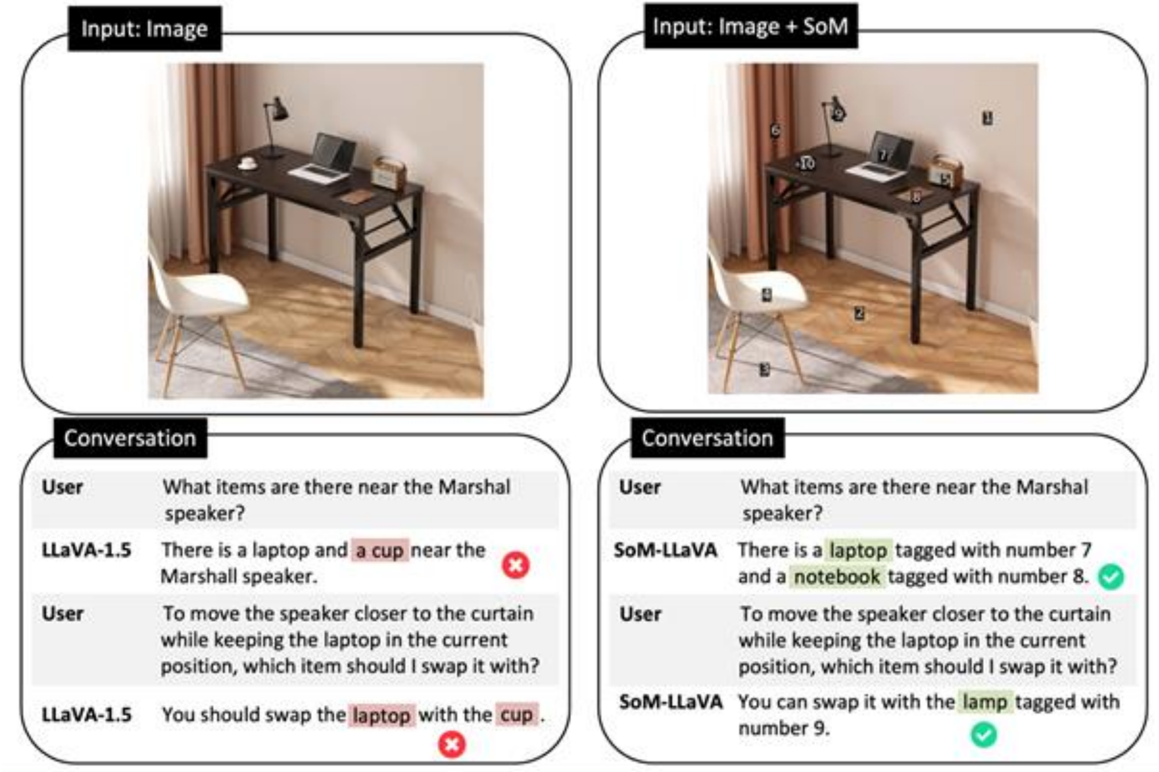
What's the high-level function of the screen?

The screen is the download page for a reminders application where you can set up reminders for various tasks such as sending out team's weekly progress, grocery shopping, traveling, and picking up kids.

FerretUI

[1] Hong, Wenyi, et al. "Cogagent: A visual language model for gui agents." CVPR 2024.
[2] You, Keen, et al. "Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs."

Distilling SoM prompting: SoM-LLaVA



Conversation

User	What items are there near the Marshal speaker?
LLaVA-1.5	There is a laptop and a cup near the Marshall speaker. ✗
User	To move the speaker closer to the curtain while keeping the laptop in the current position, which item should I swap it with?
LLaVA-1.5	You should swap the laptop with the cup. ✗

Conversation

User	What items are there near the Marshal speaker?
SoM-LLaVA	There is a laptop tagged with number 7 and a notebook tagged with number 8. ✓
User	To move the speaker closer to the curtain while keeping the laptop in the current position, which item should I swap it with?
SoM-LLaVA	You can swap it with the lamp tagged with number 9. ✓

[1] Yan, An, et al. "List Items One by One: A New Data Source and Learning Paradigm for Multimodal LLMs."

Beyond **MM-Navigator**

- Building Agents with (Open-source) Small LMMs
    - Lower cost, lower latency

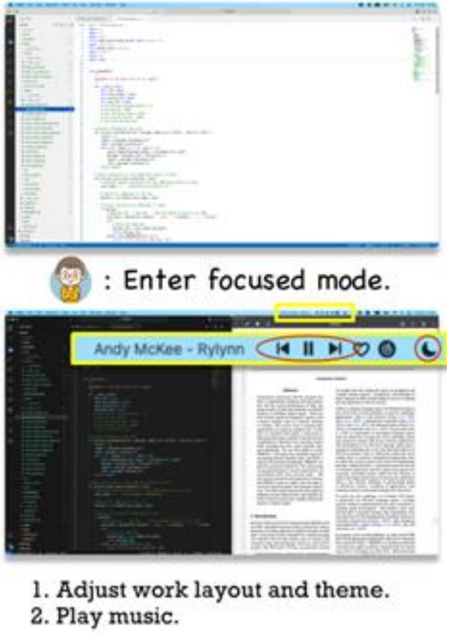
- Evaluation Benchmarks
    - Long procedural planning
    - Visual-centric tasks
    - Interactive Environments

VideoGUI

[1] Xiang Deng, et al. "MIND2WEB: Towards a Generalist Agent for the Web" NeurIPS 2023.
[2] Christopher Rawles, et al. "Android in the Wild: A Large-Scale Dataset for Android Device Control" NeurIPS 2023.
[3] Jing Yu Koh, et al. "VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks" ACL 2024.
[4] Difei Gao, et al. "ASSISTGUI: Task-Oriented Desktop Graphical User Interface Automation" CVPR 2024.
[5] Tianbao Xie, et al. "OS-World: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments" 2024.
[6] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

Existing GUI Benchmarks

Simple Task *can be clearly described by textual query*

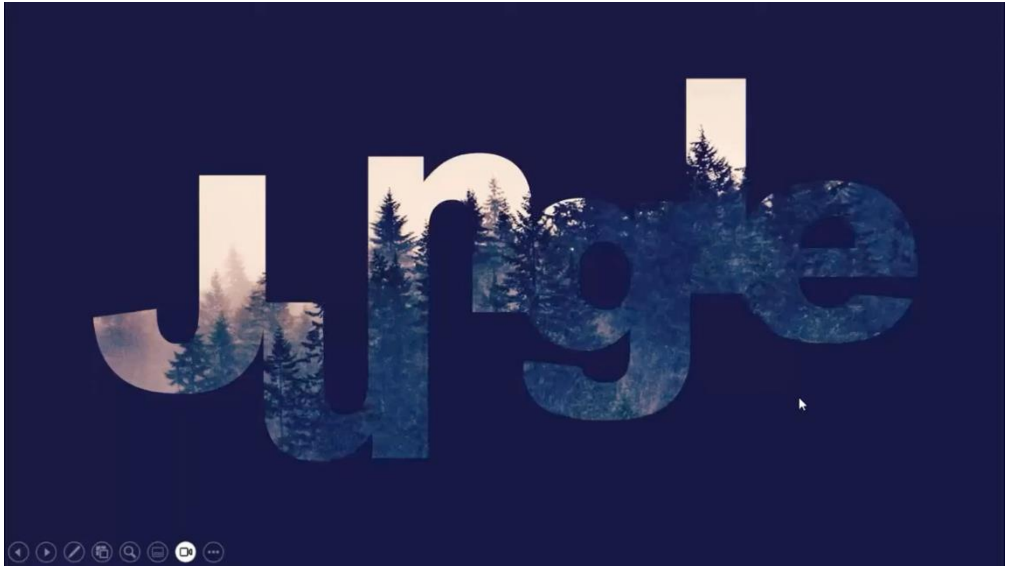


: Enter focused mode.

1. Adjust work layout and theme.
2. Play music.

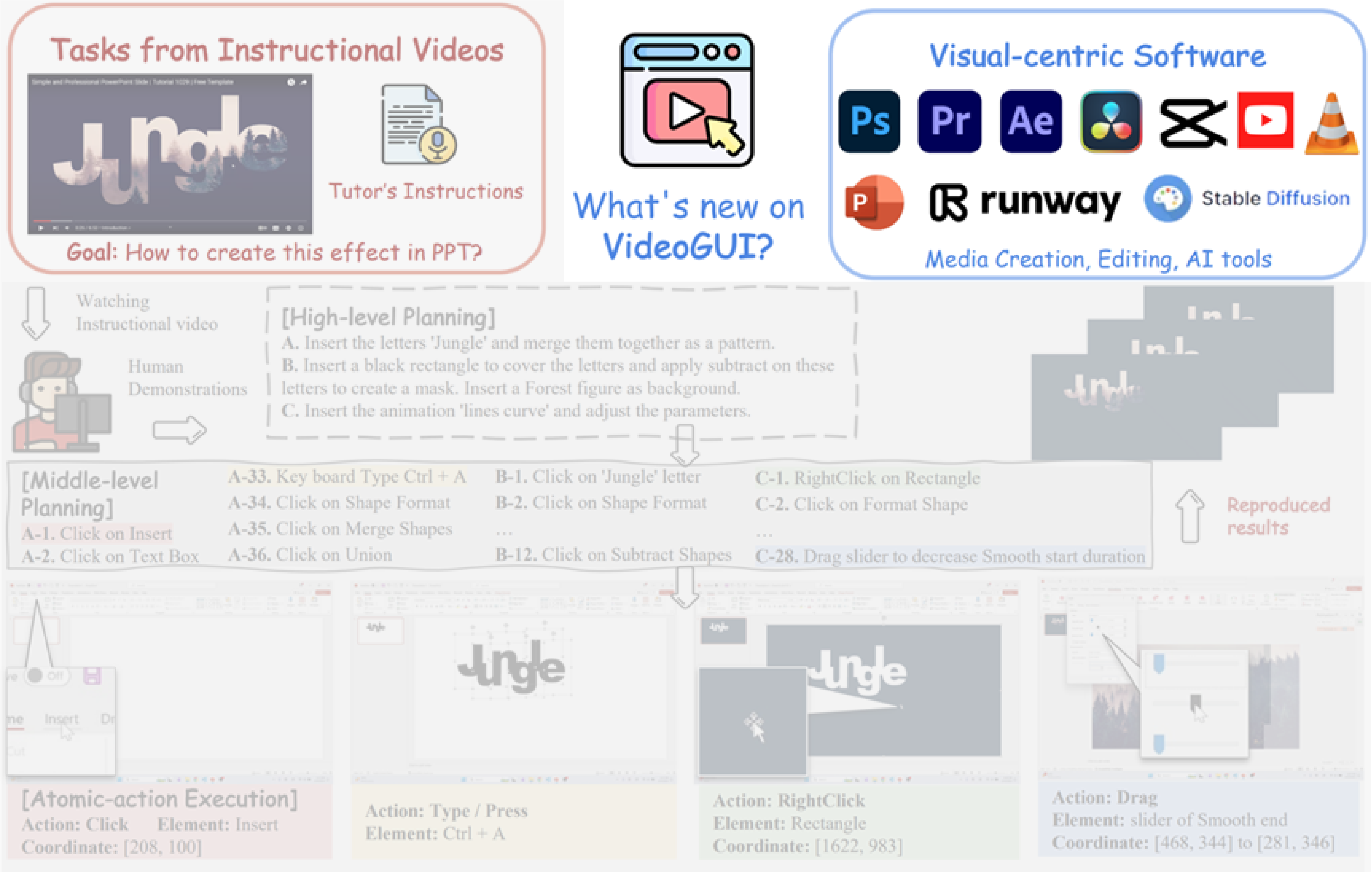
VideoGUI :Visual-centric GUI Task

Create the below effect in PowerPoint.



---

[1] Wu, Zhiyong, et al. "Os-copilot: Towards generalist computer agents with self-improvement."
[2] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."


Tasks from Instructional Videos
Tutor's Instructions
Goal: How to create this effect in PPT?
What's new on VideoGUI?
Visual-centric Software
runway
Stable Diffusion
Media Creation, Editing, AI tools
Watching Instructional video
Human Demonstrations
Reproduced results

[High-level Planning]
A. Insert the letters 'Jungle' and merge them together as a pattern.
B. Insert a black rectangle to cover the letters and apply subtract on these letters to create a mask. Insert a Forest figure as background.
C. Insert the animation 'lines curve' and adjust the parameters.

[Middle-level Planning]
A-1. Click on Insert
A-2. Click on Text Box
A-33. Key board Type Ctrl + A
A-34. Click on Shape Format
A-35. Click on Merge Shapes
A-36. Click on Union
B-1. Click on 'Jungle' letter
B-2. Click on Shape Format
…
B-12. Click on Subtract Shapes
C-1. RightClick on Rectangle
C-2. Click on Format Shape
…
C-28. Drag slider to decrease Smooth start duration

[Atomic-action Execution]
Action: Click Element: Insert
Coordinate: [208, 100]

Action: Type / Press
Element: Ctrl + A

Action: RightClick
Element: Rectangle
Coordinate: [1622, 983]

Action: Drag
Element: slider of Smooth end
Coordinate: [468, 344] to [281, 346]

GUI Navigation Evaluation

Existing Benchmarks

- 0 / 1 Success Rate
  - Easy to be 0
  - Fail to give enough feedback

VideoGUI

- Hierarchical Assessment

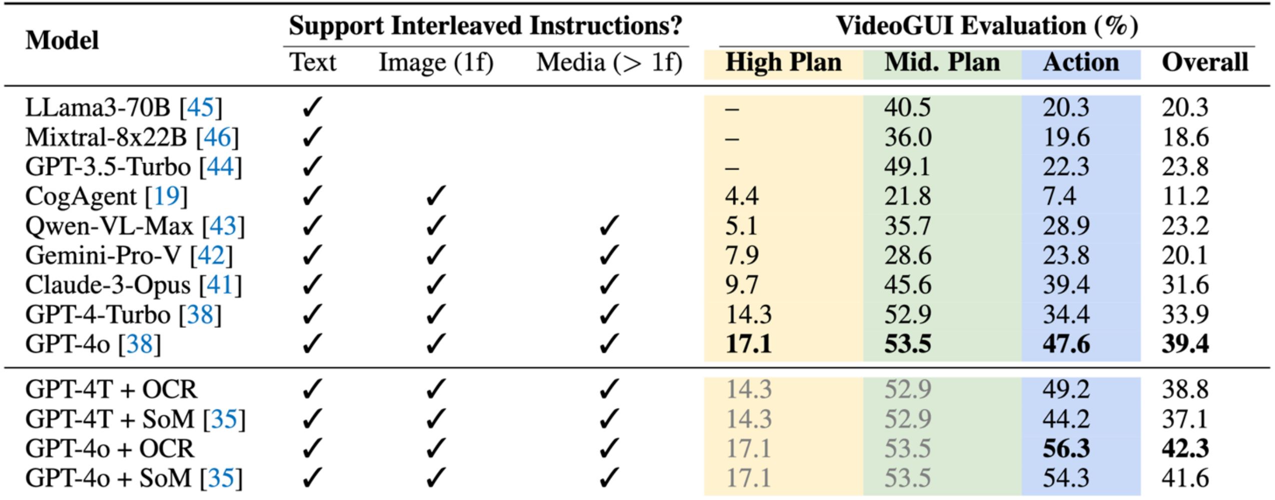
- Procedural Planning
  - High-level
  - Mid.-level
- Atomic Action Execution
  - Click, Drag, Type / Press, Scroll

[1] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

https://showlab.github.io/videogui

Benchmarking LMMs and LMM Agents

VideoGUI



Model	**Support Interleaved Instructions?**			**VideoGUI Evaluation (%)**			
	Text	Image (1f)	Media (> 1f)	**High Plan**	**Mid. Plan**	**Action**	**Overall**
LLama3-70B [45]	✓			–	40.5	20.3	20.3
Mixtral-8x22B [46]	✓			–	36.0	19.6	18.6
GPT-3.5-Turbo [44]	✓			–	49.1	22.3	23.8
CogAgent [19]	✓	✓		4.4	21.8	7.4	11.2
Qwen-VL-Max [43]	✓	✓	✓	5.1	35.7	28.9	23.2
Gemini-Pro-V [42]	✓	✓	✓	7.9	28.6	23.8	20.1
Claude-3-Opus [41]	✓	✓	✓	9.7	45.6	39.4	31.6
GPT-4-Turbo [38]	✓	✓	✓	14.3	52.9	34.4	33.9
GPT-4o [38]	✓	✓	✓	**17.1**	**53.5**	**47.6**	**39.4**
GPT-4T + OCR	✓	✓	✓	14.3	52.9	49.2	38.8
GPT-4T + SoM [35]	✓	✓	✓	14.3	52.9	44.2	37.1
GPT-4o + OCR	✓	✓	✓	17.1	53.5	**56.3**	**42.3**
GPT-4o + SoM [35]	✓	✓	✓	17.1	53.5	54.3	41.6

Table 3: Full evaluation on VideoGUI with Baselines and their supported *interleaved* instructions, which might be a text query, an image (1 frame), or a media (more than 1 frame) such as two photos, one or two videos.

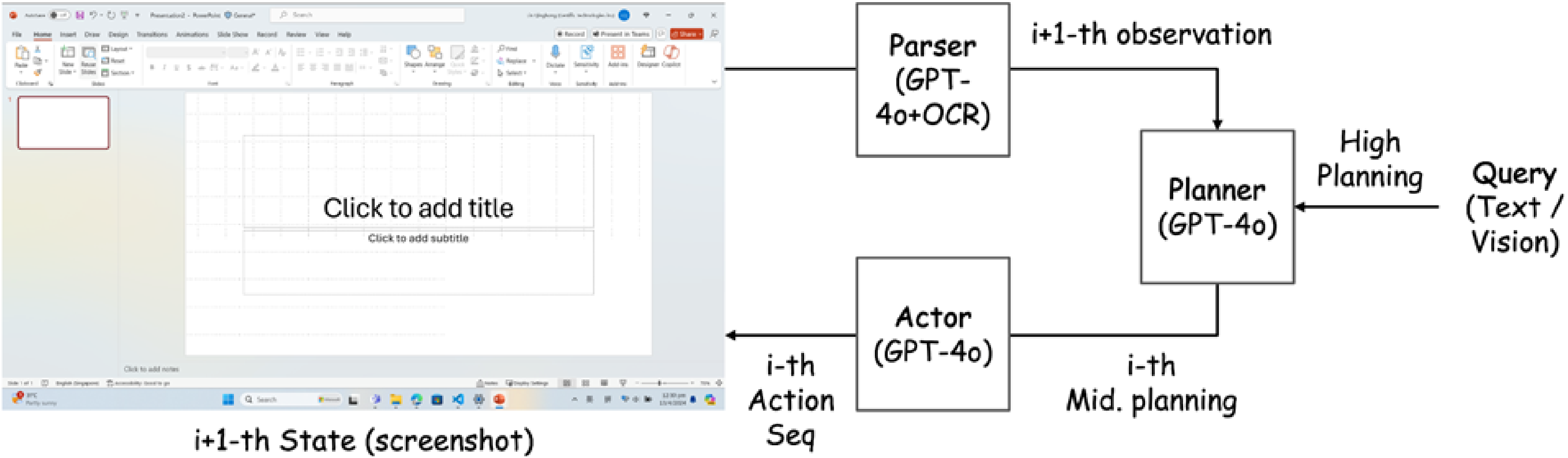
- Planning is way harder than execution!
- SoTA LMM (e.g., GPT-4o) still need tools (e.g., OCR or SoM) to perform better on action execution.

[1] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

https://showlab.github.io/videogui

A Simple Agent Baseline with GPT-4o

VideoGUI



i+1-th State (screenshot)

Parser (GPT-4o+OCR)

i+1-th observation

Planner (GPT-4o)

High Planning

Query (Text / Vision)

Actor (GPT-4o)

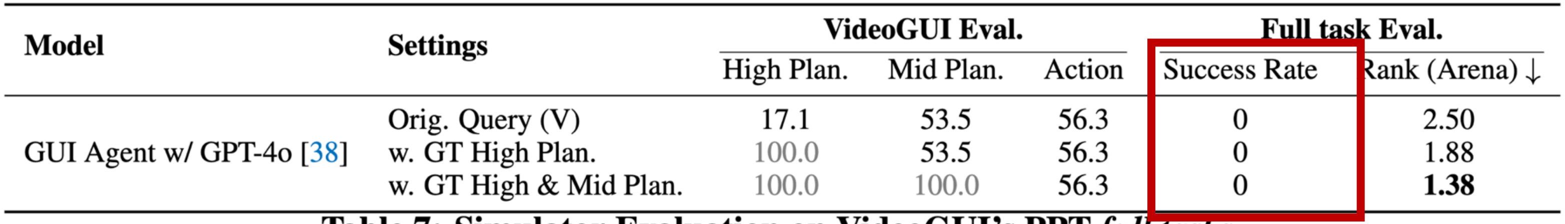
i-th Mid. planning

i-th Action Seq

[1] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

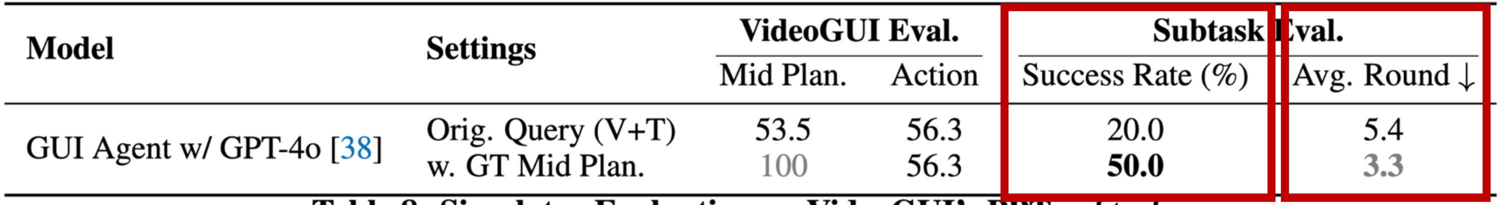
https://showlab.github.io/videogui

Model Performance on VideoGUI



Model	Settings	VideoGUI Eval.			Full task Eval.	
		High Plan.	Mid Plan.	Action	Success Rate	Rank (Arena) ↓
GUI Agent w/ GPT-4o [38]	Orig. Query (V)	17.1	53.5	56.3	0	2.50
	w. GT High Plan.	100.0	53.5	56.3	0	1.88
	w. GT High & Mid Plan.	100.0	100.0	56.3	0	**1.38**

Table 7: Simulator Evaluation on VideoGUI's PPT *full tasks*.



Model	Settings	VideoGUI Eval.		Subtask Eval.	
		Mid Plan.	Action	Success Rate (%)	Avg. Round ↓
GUI Agent w/ GPT-4o [38]	Orig. Query (V+T)	53.5	56.3	20.0	5.4
	w. GT Mid Plan.	100	56.3	**50.0**	3.3

Table 8: Simulator Evaluation on VideoGUI's PPT *subtasks*.

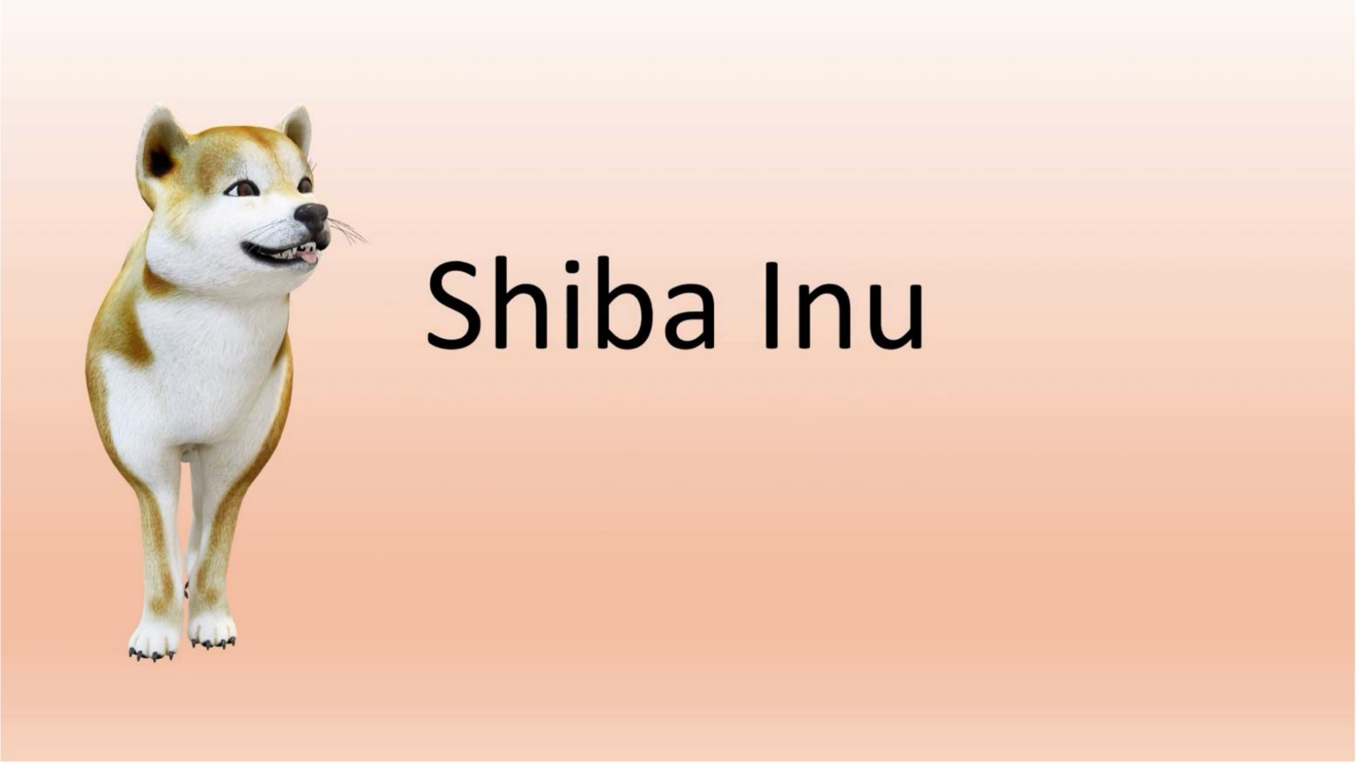
GPT-4o Agent fails to complete the full task, even with GT high-level / mid-level planning

[1] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

https://showlab.github.io/videogui

Qualitative Results

Goal:



Shiba Inu

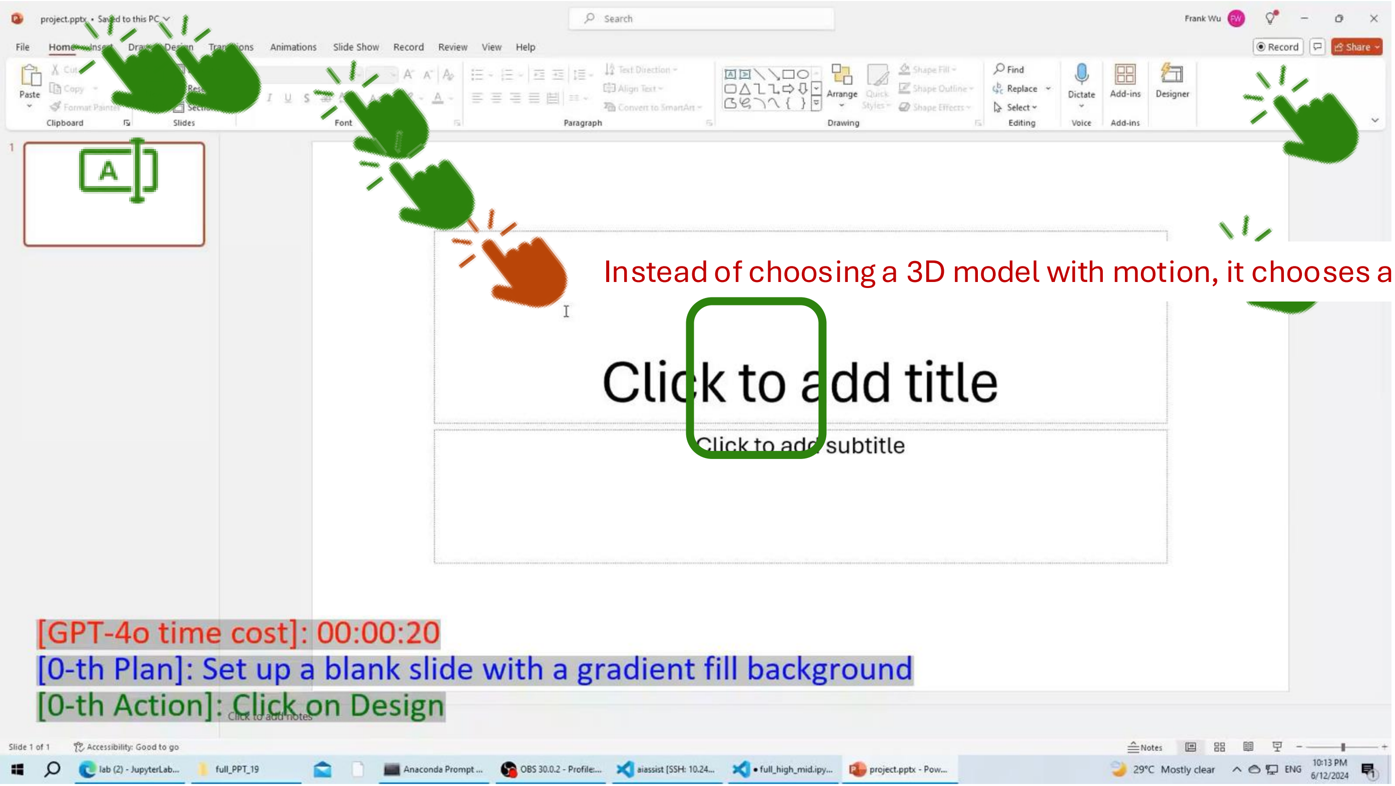
Final product Slideshow Preview (4x speed up)

[1] Lin, Kevin Qinghong, et al. ”VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

https://showlab.github.io/videogui

Qualitative Results

VideoGUI



Instead of choosing a 3D model with motion, it chooses a static 3D model

[GPT-4o time cost]: 00:00:20

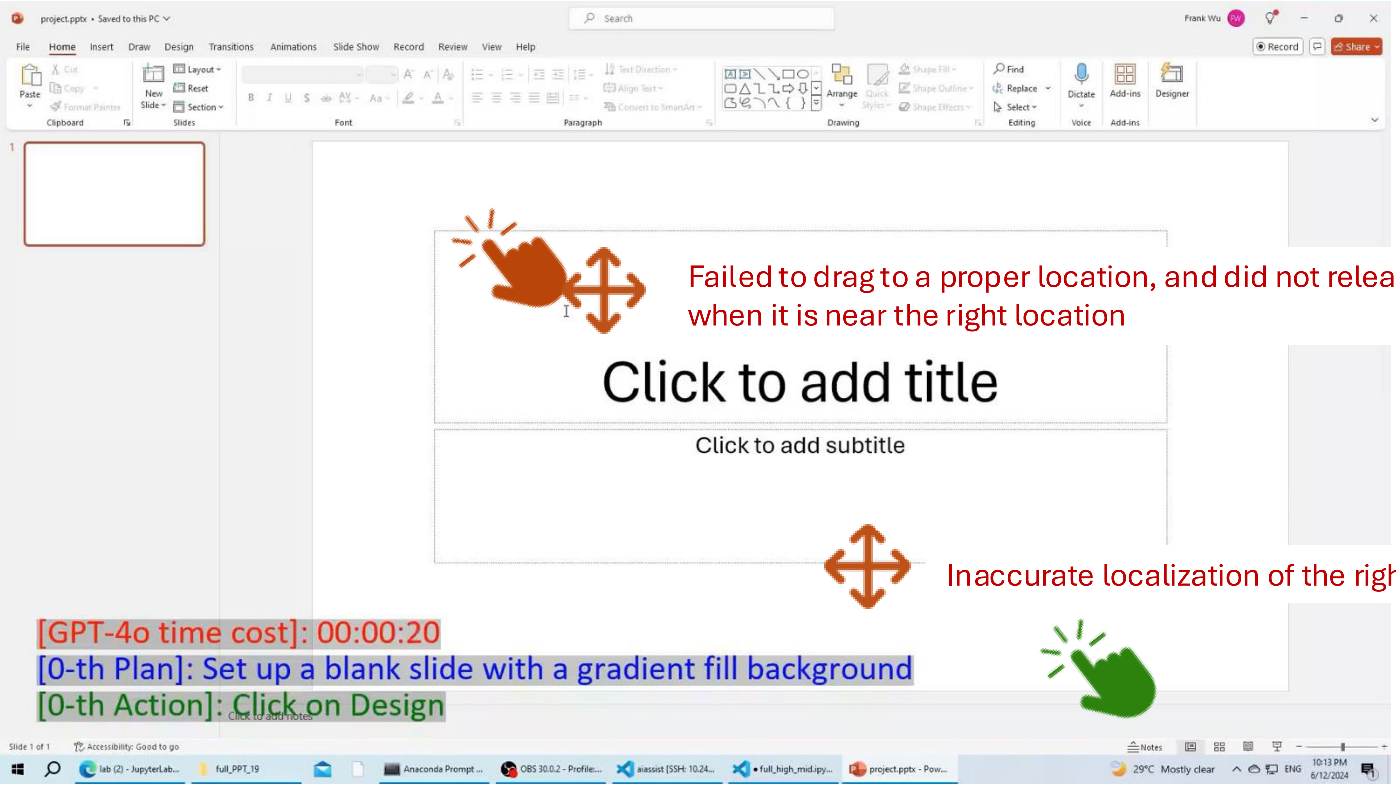
[0-th Plan]: Set up a blank slide with a gradient fill background

[0-th Action]: Click on Design

GPT-4o GUI Agent with GT planning (3x speed up)

Qualitative Results

VideoGUI



Failed to drag to a proper location, and did not release it when it is near the right location

Inaccurate localization of the right corner

[GPT-4o time cost]: 00:00:20

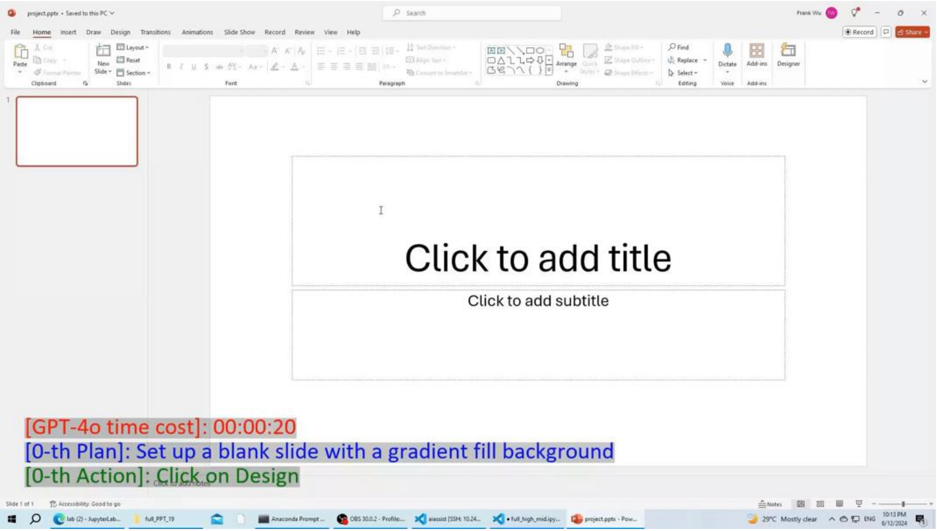
[0-th Plan]: Set up a blank slide with a gradient fill background

[0-th Action]: Click on Design

GPT-4o GUI Agent with GT planning (3x speed up)

Qualitative Results

VideoGUI



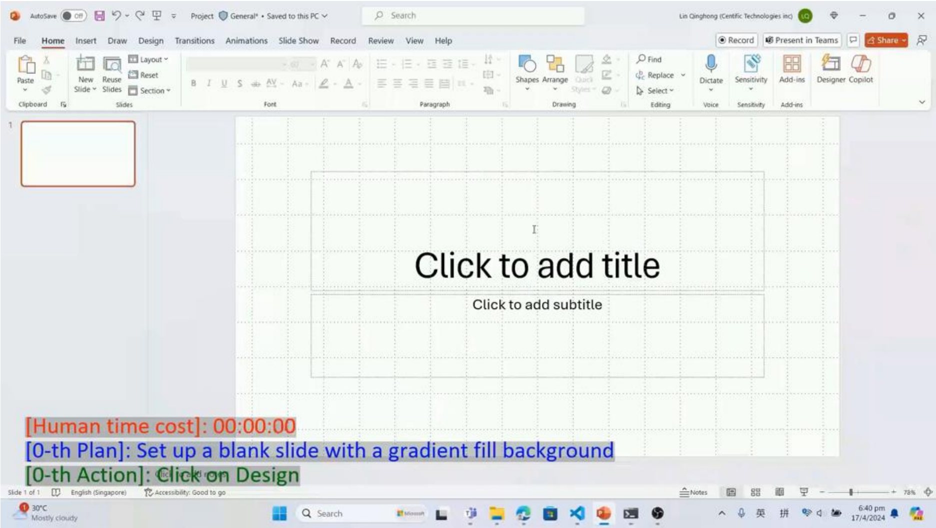
[GPT-4o time cost]: 00:00:20

[0-th Plan]: Set up a blank slide with a gradient fill background

[0-th Action]: Click on Design

GPT-4o GUI Agent w/ GT planning (10x speed up)

Total time: ~ 10 min



[Human time cost]: 00:00:00

[0-th Plan]: Set up a blank slide with a gradient fill background

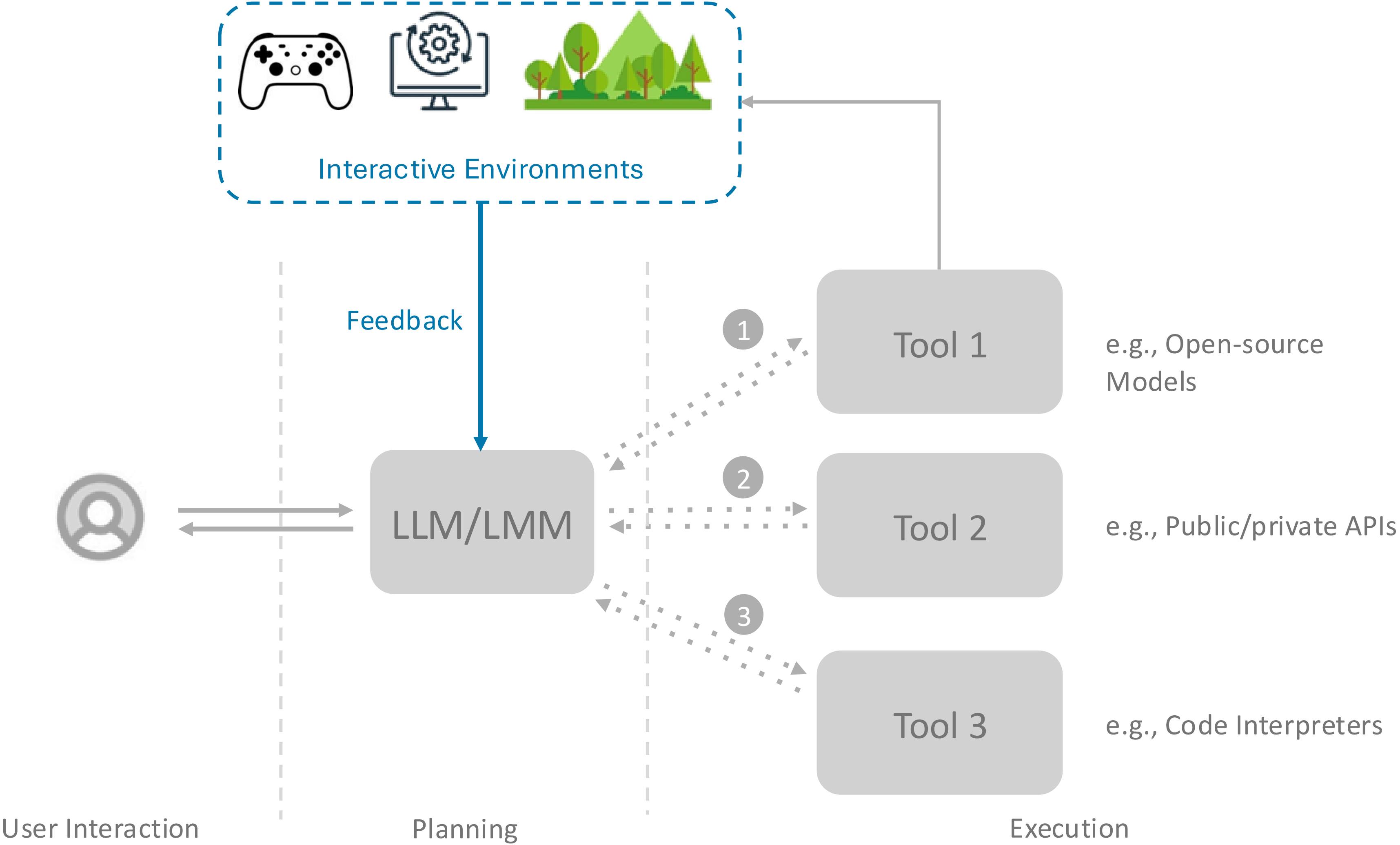
[0-th Action]: Click on Design

Human w/ GT planning (10x speed up)

Total time: ~ 4 min

[1] Lin, Kevin Qinghong, et al. "VideoGUI: A Benchmark for GUI Automation from Instructional Videos."

https://showlab.github.io/videogui


Interactive Environments
Feedback
1
2
3
LLM/LMM
Tool 1
Tool 2
Tool 3
e.g., Open-source Models
e.g., Public/private APIs
e.g., Code Interpreters
User Interaction
Planning
Execution

Do We Still Need Multimodal Agents?

Agents with Multimodal Memory

MM-Narrator MM-Vid

Actionable Agents

MM-Navigator VideoGUI

Agent with Feedback

Idea2Img

Audio Description

GUI Navigation

Visual Design & Creation

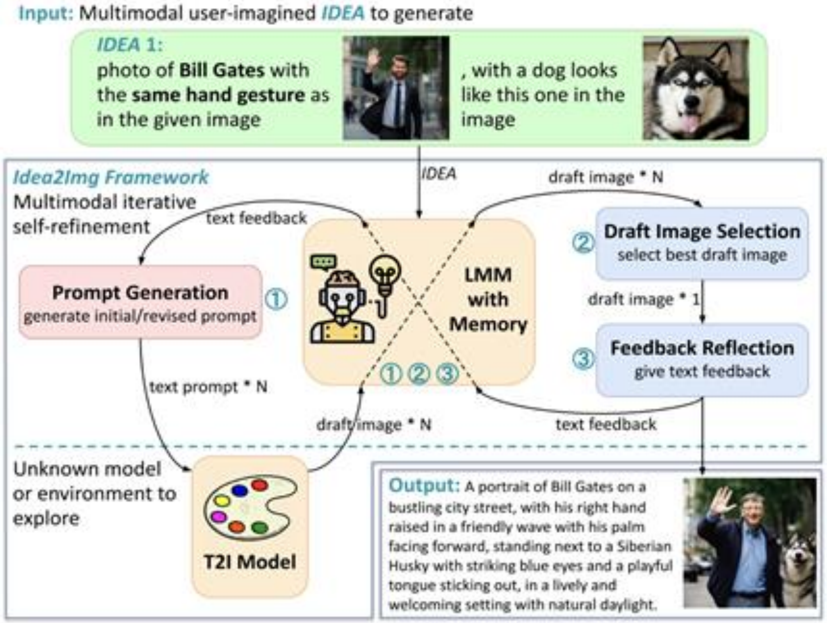
Why Agents Need Feedback?

- Humans exploring unknown environments/problems
- Iterative self-refinement in exploration:
  - Proposal
  - Selection
  - Feedback

* GIF source: giphy.com

Idea2Img

- A GPT-4V powered agent systems with iterative self-refinement for automatic visual design and creation


Input: Multimodal user-imagined *IDEA* to generate
IDEA 1:
photo of **Bill Gates** with the **same hand gesture** as in the given image
, with a dog looks like this one in the image
Idea2Img Framework
Multimodal iterative self-refinement
IDEA
draft image * N
text feedback
Prompt Generation
generate initial/revised prompt
①
LMM with Memory
① ② ③
② Draft Image Selection
select best draft image
draft image * 1
③ Feedback Reflection
give text feedback
text prompt * N
draft image * N
text feedback
Unknown model or environment to explore
T2I Model
Output:

A portrait of Bill Gates on a bustling city street, with his right hand raised in a friendly wave with his palm facing forward, standing next to a Siberian Husky with striking blue eyes and a playful tongue sticking out, in a lively and welcoming setting with natural daylight.

[1] Yang, Zhengyuan, et al. "Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation."

https://idea2img.github.io/

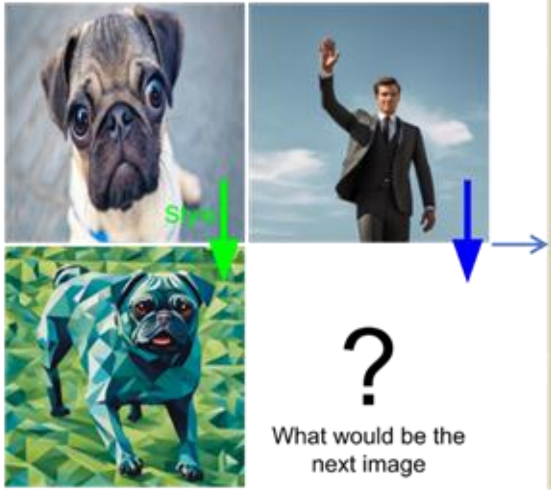
Visual Design and Creation

Idea2Img

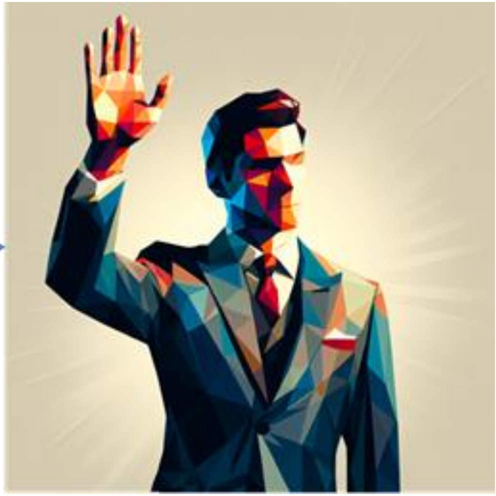
- Visual creation with multimodal idea inputs, instead of detailed text descriptions

A man in a formal suit is captured mid-motion, his hand raised in a wave, each finger distinctly visible. The entire scene is reimagined in a low-poly art style, where each facet of the man and his attire is defined by geometric shapes filled with a palette of bold, contrasting colors. The background remains a solid, untextured expanse, free of additional elements like sun rays or clouds, ensuring the focus remains solely on the colorful, stylized figure of the man waving. The lighting is neutral, casting no shadows or highlights, preserving the simplicity of the scene.

Detailed Text Description



Input IDEA

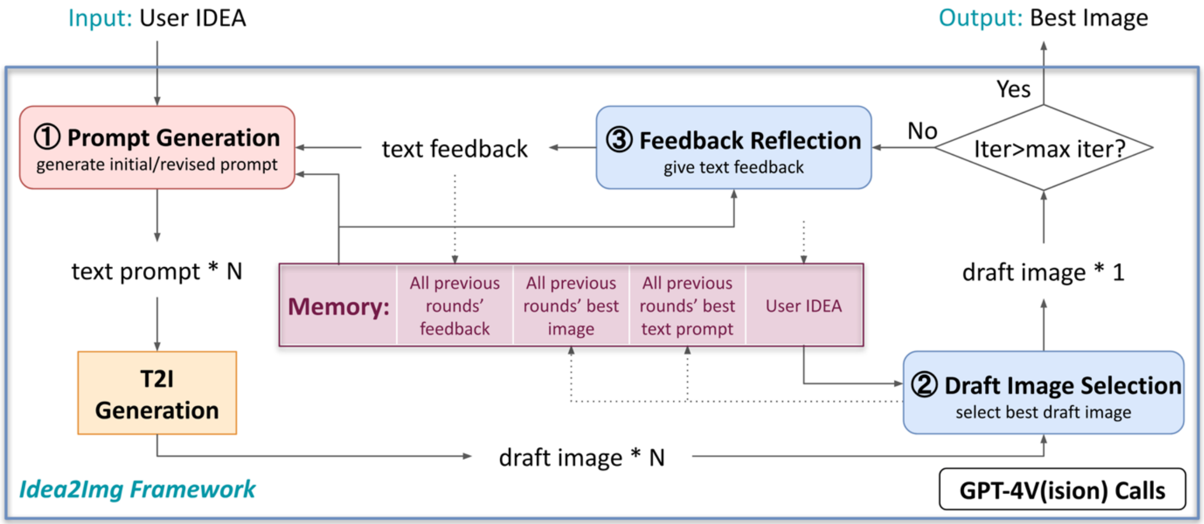


Output Design

[1] Yang, Zhengyuan, et al. "Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation."

https://idea2img.github.io/

Idea2Img Execution Flow



1. Initial or **Revised** prompt generation (Improving) for the T2I model
2. Draft image selection (Assessing)
3. Provide feedback on the errors and revision directions (Verifying)

- Memory

[1] Yang, Zhengyuan, et al. "Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation."

https://idea2img.github.io/

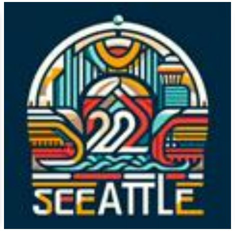
Input user *IDEA*

Single-round Human Prompting

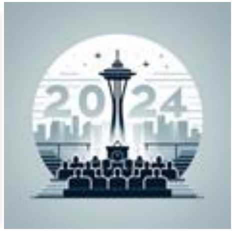
Iterative self-refined *Idea2Img* prompting

A logo for a 2024 conference held in Seattle. Please provide a textual description of the design before generating

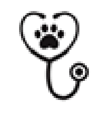
a logo for a 2024 conference held in Seattle. Please provide a textual description of the design before generating



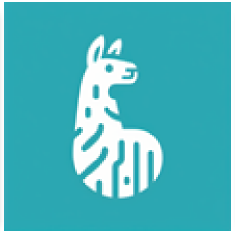
An elegant logo for the 2024 conference, where the Space Needle forms part of the "2024", adorned with subtle imagery of a speaker's podium and audience, set against a simplified Seattle skyline, with a dean, monochromatic palette, precise detailing, and a soft lighting to accentuate the conference theme and professionalism.



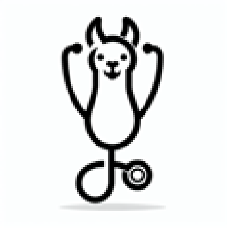
A logo of llama with the exact same style as in the given image



A logo of llama with the exact same style as in the given image that has a minimalist design



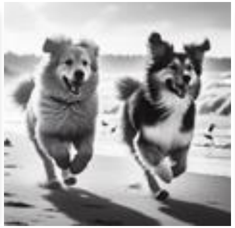
A black and white logo where a stethoscope seamlessly forms a llama, the tubing outlines the body and legs, and the earpieces and diaphragm shape the head, akin to the minimalist style of a paw print, no facial features, embodying pure simplicity, elegance, and clean lines.



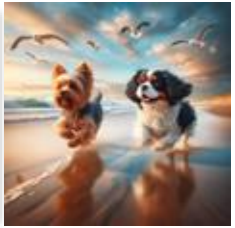
Two dogs running on the beach, the left one looks like

and the right one looks like the left dog in the image

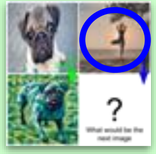
Two dogs running on the beach, the left one looks like a fluffy golden dog and the right one looks like the left dog in the image that is black and white



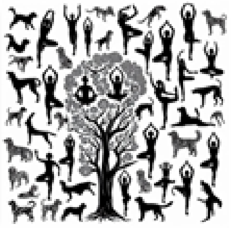
Two dogs, a Yorkshire Terrier with detailed brown and white fur that glistens in the sunlight, on the left of a black and white Cavalier King Charles Spaniel, are immersed in a game of chase along a scenic beach. The ocean waves, with their rhythmic motion, complement the dogs' lively energy. The sky, a masterpiece of soft, blended colors, serves as a canvas for the seagulls that glide gracefully, casting intricate shadows that dance on the sandy surface, amplifying the scene's vibrancy.



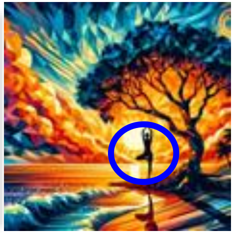
Find the the image style pattern in the left two dog images and apply it on the top right people in tree pose image. Provide a textual description that keeps the content in the people in tree pose image, with the correct style pattern.



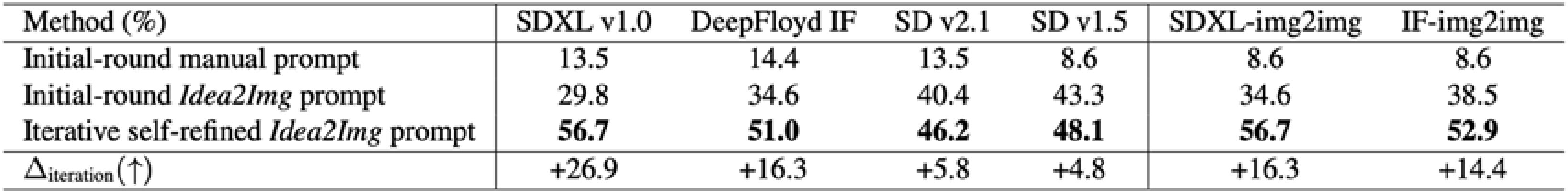
Find the the image style pattern in the left two dog images and apply it on the top right people in tree pose image



Amidst the gentle waves lapping at the shore and the sun bidding adieu, a person is encapsulated in the serene act of performing the tree pose yoga. A tree, grand and robust, shares the spotlight, its leaves whispering tales of nature's splendor. The scene is reimagined, ... more distinct geometric shapes, echoing the style of the polygonal dog images. Each shape, bold and less complex, is a canvas capturing the sunset's myriad hues, crafting a cohesive, visual narrative.



Human Evaluation



Method (%)	SDXL v1.0	DeepFloyd IF	SD v2.1	SD v1.5	SDXL-img2img	IF-img2img
Initial-round manual prompt	13.5	14.4	13.5	8.6	8.6	8.6
Initial-round *Idea2Img* prompt	29.8	34.6	40.4	43.3	34.6	38.5
Iterative self-refined *Idea2Img* prompt	**56.7**	**51.0**	**46.2**	**48.1**	**56.7**	**52.9**
$\Delta_{\text{iteration}}(\uparrow)$	+26.9	+16.3	+5.8	+4.8	+16.3	+14.4

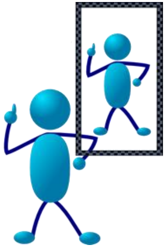
- Higher user preference score on all generation models
- Stronger generation models benefit more from Idea2Img iteration

[1] Yang, Zhengyuan, et al. "Idea2Img: Iterative Self-Refinement with GPT-4V(ision) for Automatic Image Design and Generation."

https://idea2img.github.io/

Beyond Idea2Img

- Environment interaction tasks beyond image generation
    - GUI navigation, Embodied AI, Image2code
- Self-Reflection
    - Think twice/self-evaluate its own output



[1] Gao, Difei, et al. "Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn."
[2] Wang, Junyang, et al. "Mobile-Agent-v2: Mobile Device Operation Assistant with Effective Navigation via Multi-Agent Collaboration."
[3] Chi, Zhang, et al. "AppAgent: Multimodal Agents as Smartphone Users"
[4] Yang, Yijun, et al. " Embodied Multi-Modal Agent trained by an LLM from a Parallel TextWorld"
[5] Zhang, Wenqi, et al. "Agent-Pro: Learning to Evolve via Policy-Level Reflection and Optimization", ACL 2024
[6] Wang, Guanzhi, et al. "VOYAGER: An Open-Ended Embodied Agent with Large Language Models", NeurIPS 2023 Workshop
[7] Si, Chenglei, et al. "Design2Code: How Far Are We From Automating Front-End Engineering?"

Do We Still Need Multimodal Agents?

YES

Agents with Multimodal Memory

MM-Narrator MM-Vid

Actionable Agents

MM-Navigator VideoGUI

Agent with Feedback

Idea2Img

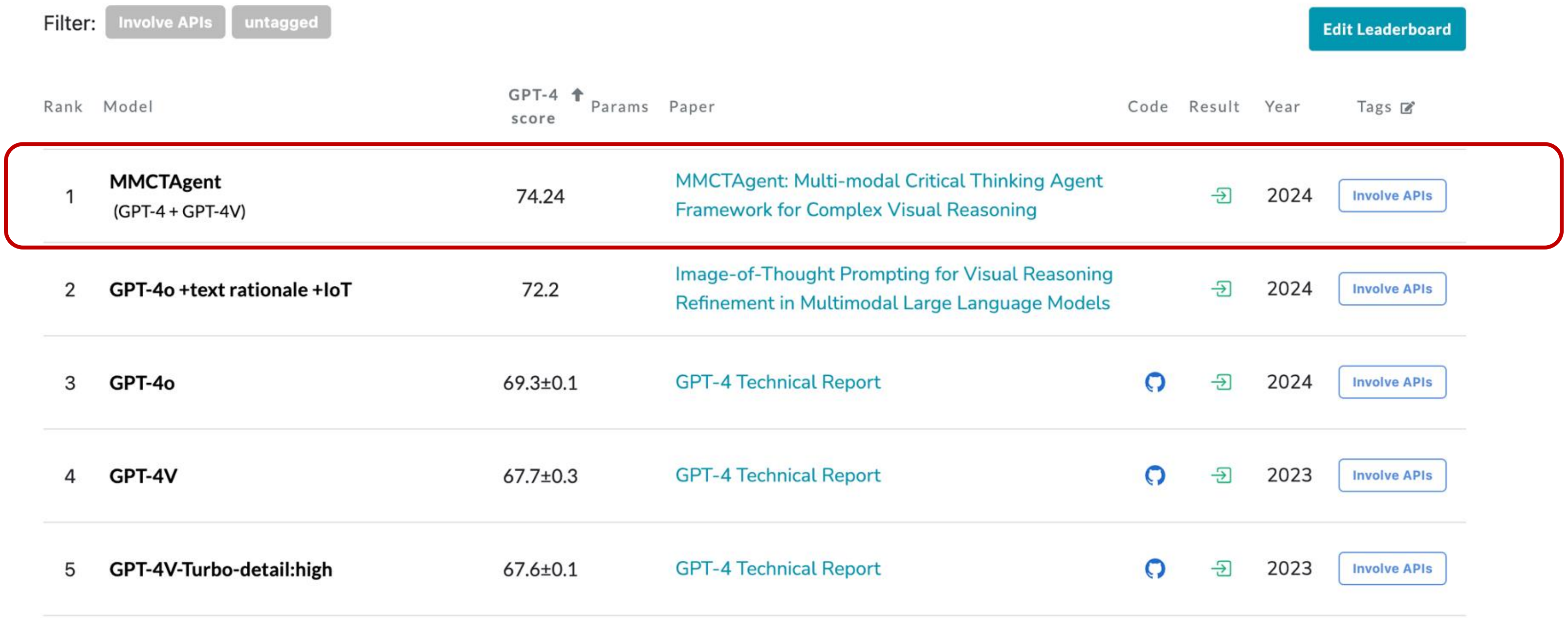
Audio Description

GUI Navigation

Visual Design & Creation

Do We Still Need Multimodal Agents?

MM-Vet : Evaluating integrated vision-language capabilities



Filter: Involve APIs untagged

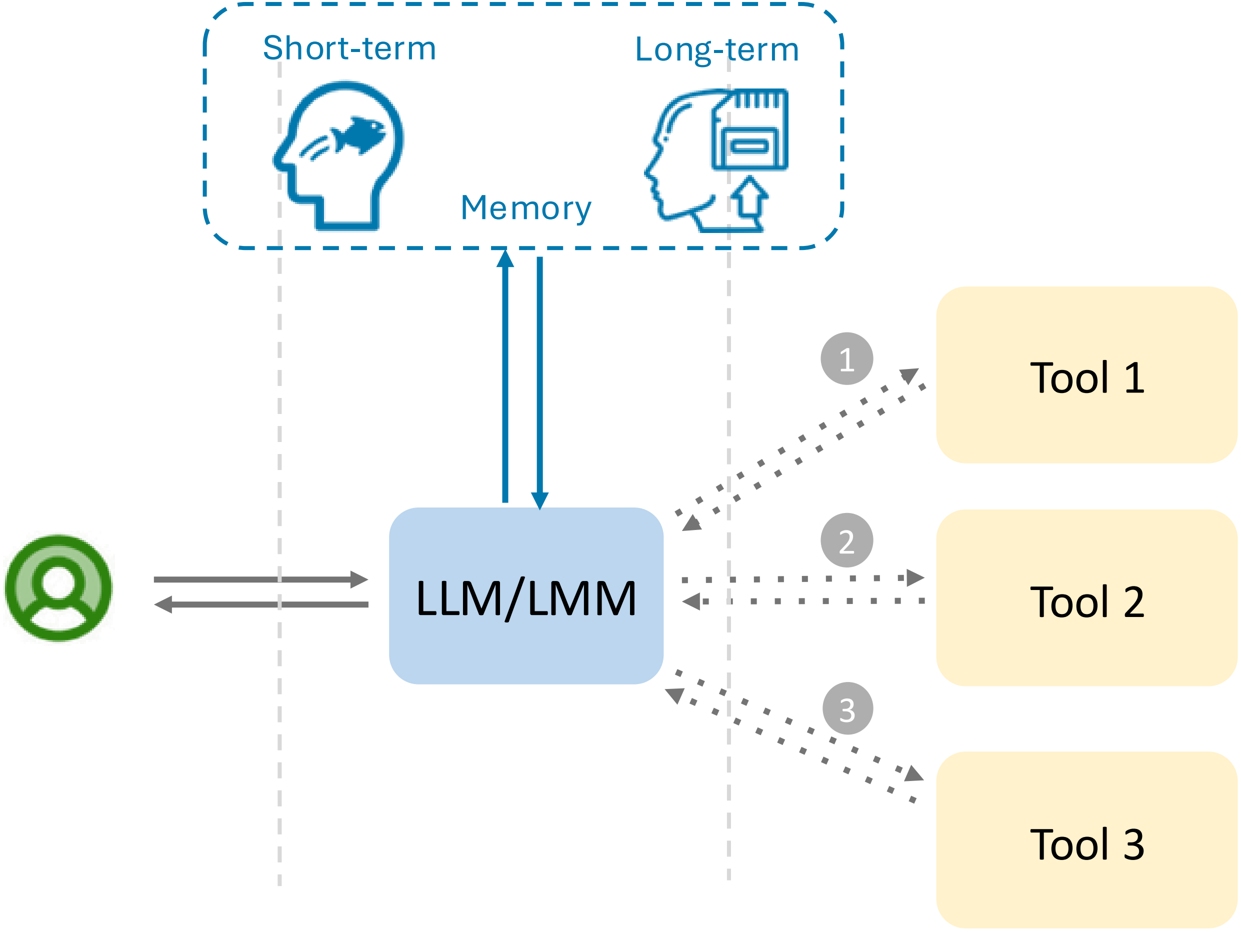
Edit Leaderboard

Rank	Model	GPT-4 score	Params	Paper	Code	Result	Year	Tags
1	**MMCTAgent** (GPT-4 + GPT-4V)	74.24		MMCTAgent: Multi-modal Critical Thinking Agent Framework for Complex Visual Reasoning		✓	2024	Involve APIs
2	**GPT-4o +text rationale +IoT**	72.2		Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models		✓	2024	Involve APIs
3	**GPT-4o**	69.3±0.1		GPT-4 Technical Report	✓	✓	2024	Involve APIs
4	**GPT-4V**	67.7±0.3		GPT-4 Technical Report	✓	✓	2023	Involve APIs
5	**GPT-4V-Turbo-detail:high**	67.6±0.1		GPT-4 Technical Report	✓	✓	2023	Involve APIs

[1] Yu, Weihao, et al. "Mm-vet: Evaluating large multimodal models for integrated capabilities." ICML 2024

https://github.com/yuweihao/MM-Vet

Discussion

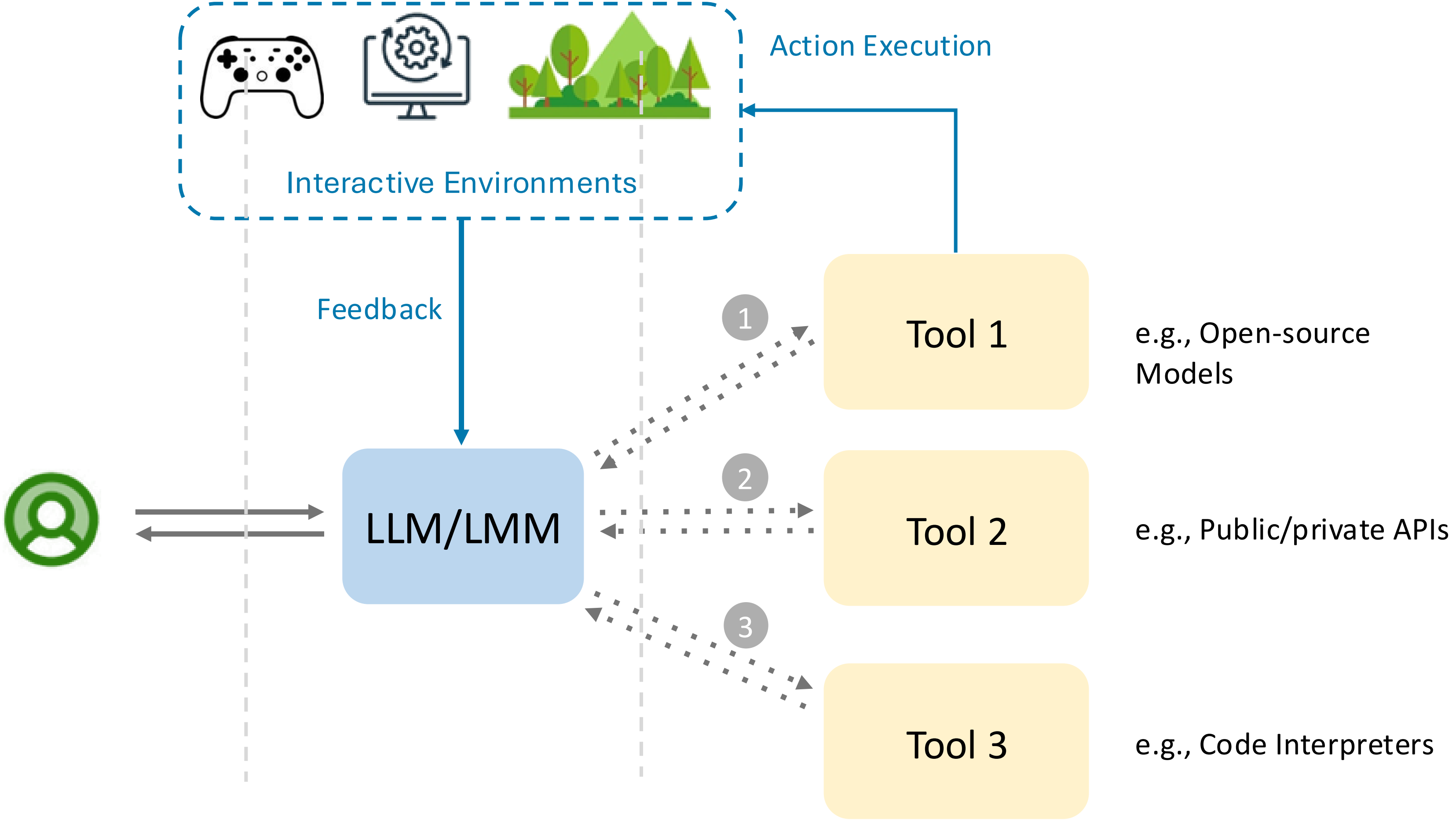


e.g., Open-source Models

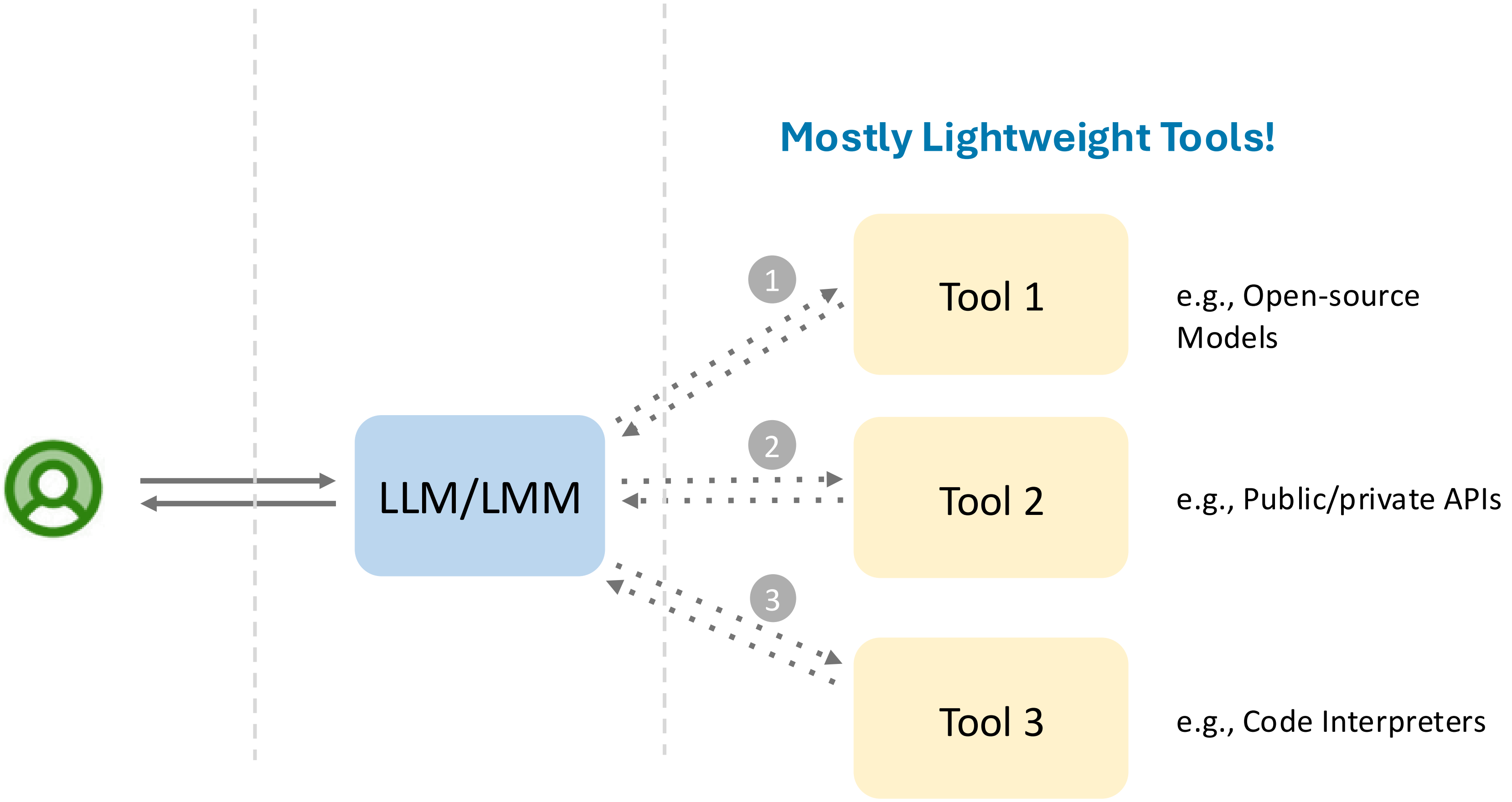
e.g., Public/private APIs

e.g., Code Interpreters

Discussion


Action Execution
Interactive Environments
Feedback
1
2
3
LLM/LMM
Tool 1
e.g., Open-source Models
Tool 2
e.g., Public/private APIs
Tool 3
e.g., Code Interpreters

Discussion



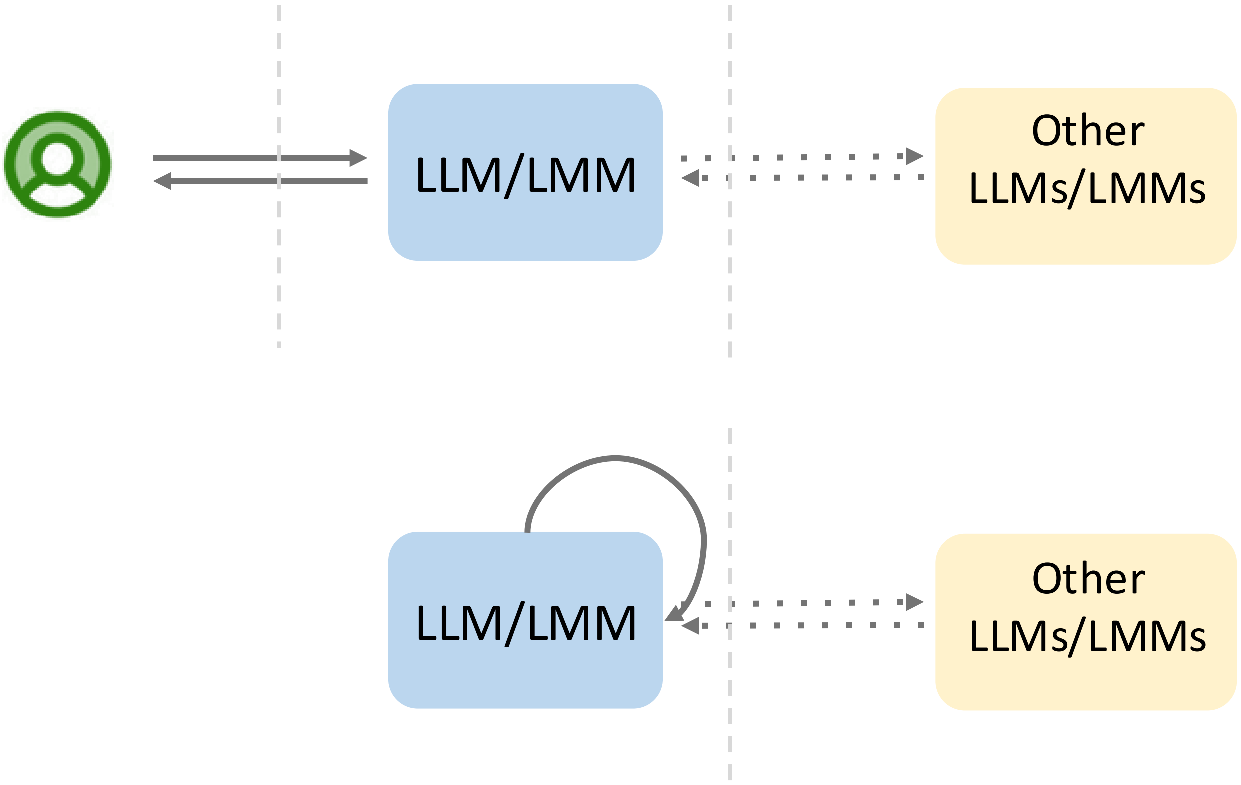
Mostly Lightweight Tools!

e.g., Open-source Models

e.g., Public/private APIs

e.g., Code Interpreters

Discussion



LLM/LMM

Other LLMs/LMMs

LLM/LMM

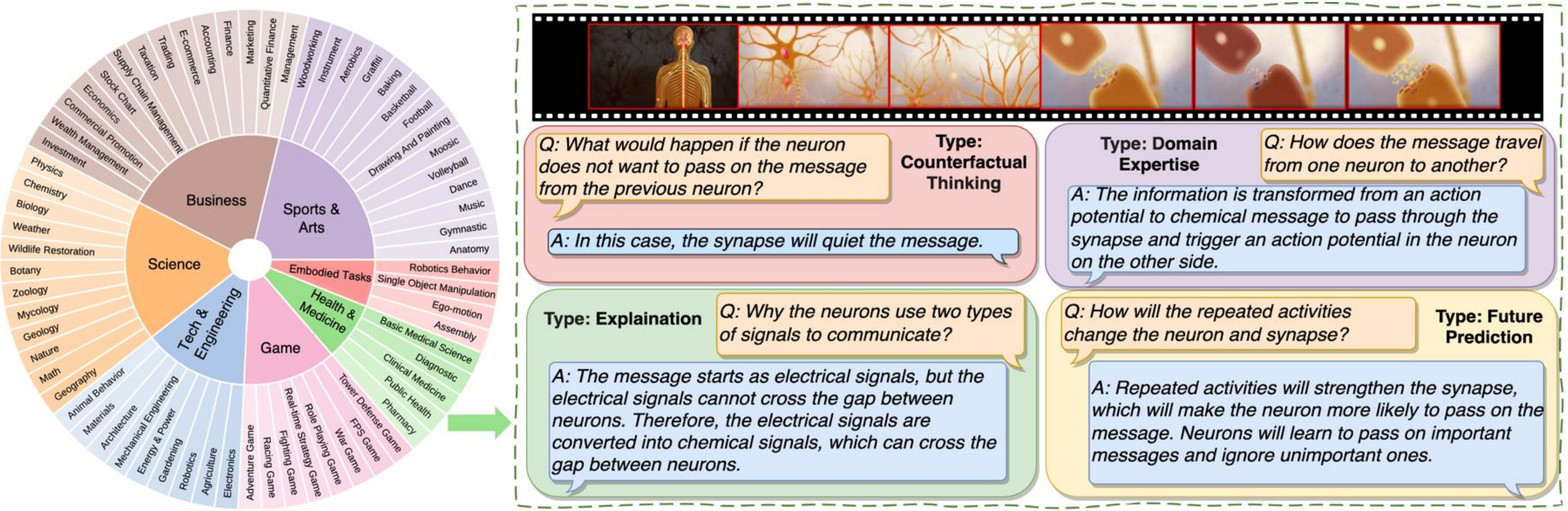
Other LLMs/LMMs

[1] Kumar, Somnath, et al. "MMCTAgent: Multi-modal Critical Thinking Agent Framework for Complex Visual Reasoning."
[2] Li, Guohao, et al. "Camel: Communicative agents for" mind" exploration of large scale language model society." NeurIPS 2023
[3] Karpathy, Andrej. “LLM OS”. Link
[4] Hong, Sirui, et al. “MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework”, ICLR 2024

Discussion

- Evals! Evals! Evals!

MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos



Type: Counterfactual Thinking

Q: What would happen if the neuron does not want to pass on the message from the previous neuron?

A: In this case, the synapse will quiet the message.

Type: Domain Expertise

Q: How does the message travel from one neuron to another?

A: The information is transformed from an action potential to chemical message to pass through the synapse and trigger an action potential in the neuron on the other side.

Type: Explaination

Q: Why the neurons use two types of signals to communicate?

A: The message starts as electrical signals, but the electrical signals cannot cross the gap between neurons. Therefore, the electrical signals are converted into chemical signals, which can cross the gap between neurons.

Type: Future Prediction

Q: How will the repeated activities change the neuron and synapse?

A: Repeated activities will strengthen the synapse, which will make the neuron more likely to pass on the message. Neurons will learn to pass on important messages and ignore unimportant ones.

[1] He, Xuehai, et al. "MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Video."

https://mmworld-bench.github.io

Discussion

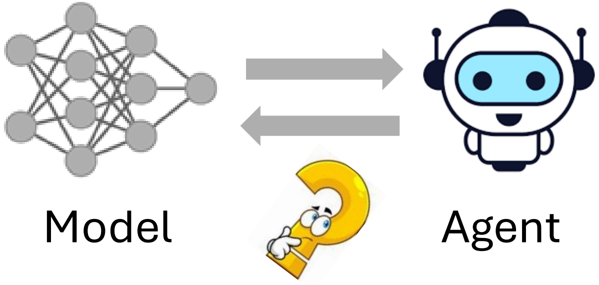
- Evals! Evals! Evals!
- Agents are far from "perfect", require a lot of system optimization to be actually useful

Latency

Accuracy

Privacy

- Can Agent help improve LLM/LMMs?



Model

Agent