



# VideoGUI:

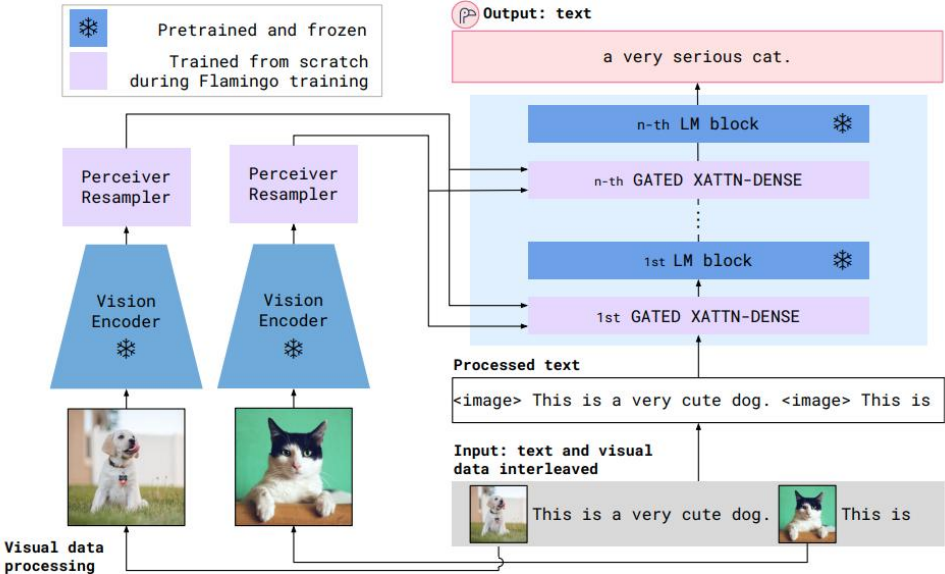
A Benchmark for GUI Automation  
from Instructional Videos



# Motivation: Simple Vision Task 👉 Complex Computer Task

## 1. VLMs (Vision-Language-Models)

- CLIP, Flamingo, LLaVA



*Experts in specific vision tasks  
with Large-scale pretraining*

Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

# Motivation: Simple Vision Task 👉 Complex Computer Task

## 1. VLMs

## 2. Multi-modal Agent

- MM-ReAct, ViperGPT, AssistGPT

🗣️ I am trying to create a short version of the video, can you suggest a shorter segment for each step?

Here's a suggested shorter segment for each step:



1. Heat a little olive oil in a pan and grill the tomatoes: 00:00:24 - 00:00:30
2. Cook the bacon in a frying pan: 00:00:53 - 00:00:59
3. Place the cooked bacon on a paper towel: 00:01:39 - 00:01:41
4. Spread mayo and add Worcestershire sauce: 00:02:00 - 00:02:08
5. Assemble the sandwich (lettuce, tomatoes, bacon): 00:02:44 - 00:03:01
  - a. Layer lettuce: 00:02:44 - 00:02:47
  - b. Add grilled tomatoes: 00:02:55 - 00:02:59
  - c. Place bacon on top: 00:03:15 - 00:03:20
6. Sprinkle salt and pepper: 00:03:08 - 00:03:11
7. Top the sandwich with the other slice of bread: 00:03:29 - 00:03:34



*Solve complex vision tasks  
with LLM reasoning*

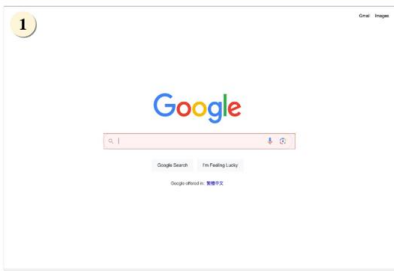
# Motivation: Simple Vision Task Complex Computer Task

- 1. VLMs
- 2. Multi-modal Agent
- 3. Actionable Model / Agent
  - o CogAgent, MM-Navigator, AssistGUI

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


**Computer Agent**

**1**



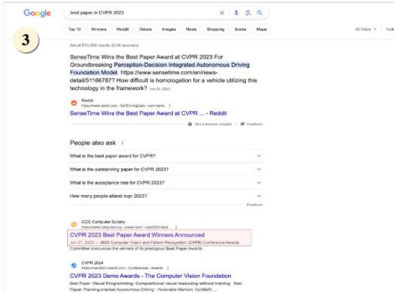
**CogAgent:** Move the cursor to the Google Search box located approximately at the center of the screen, then type 'best paper in CVPR 2023'.  
[Operation]: [searchbox] Search -> TYPE: best paper in CVPR 2023 at the box [[250,479,750,540]]

**2**




**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

**3**



**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 [[168,775,656,829]]

**4**



**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 reasoning without training... -> CLICK at the box [[087,675,817,695]]

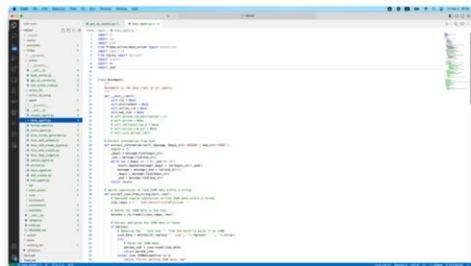
*Interact with environment by actions*

GUI (Graphical user interface) being the most representative environment

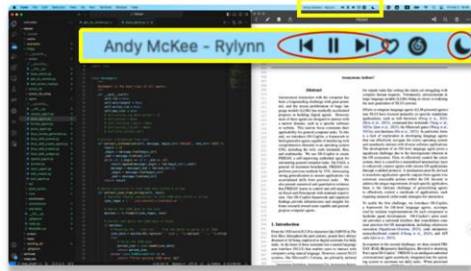


# Motivation: Existing GUI benchmarks

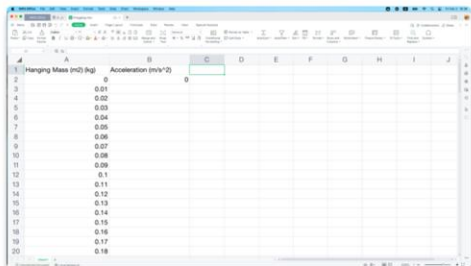
- Existing GUI benchmarks
  - Simple Task *can be clearly described by textual query*



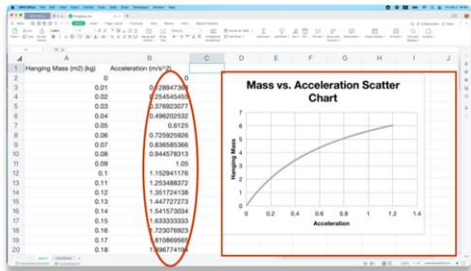
 : Enter focused mode.



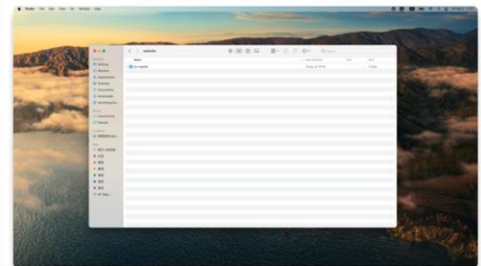
- Adjust work layout and theme.
- Play music.



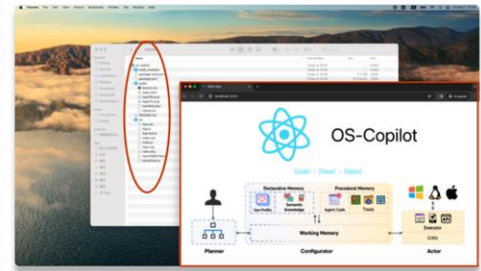
 : Fill in sheet, draw chart.



- Calculate and fill out the spreadsheet.
- Draw a bar chart.



 : Create a webpage.



- Generate the files required.
- Compile and build to the webpage.

Credit to OS-copilot (2024)

# Motivation: Existing GUI benchmarks

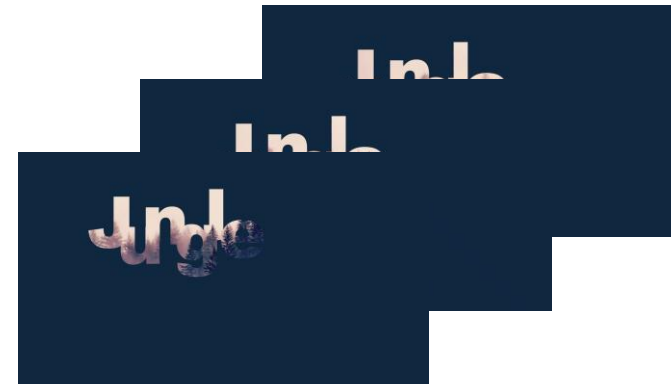
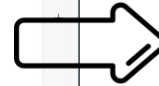
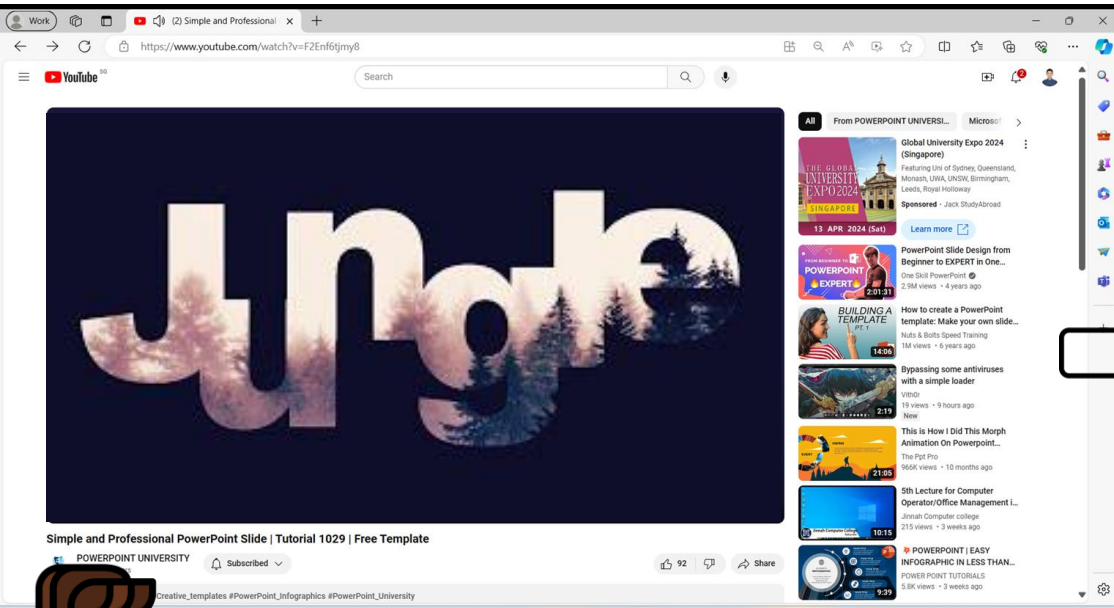
- **Complex Task** by
  - **Visual Query** e.g., How to create such effect in Powerpoint?



***Challenging 1: Reverse Engineering from Vision Preview***

# Motivation: Existing GUI benchmarks

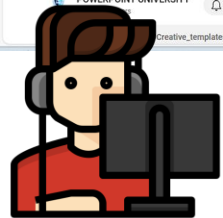
- Advanced Need
  - Visual-centric Query *e.g., How to create such effect in Powerpoint?*



Human reproduced result

**Human Learning from Web Instructional Videos**

***Challenging 2: Long Procedural & Multiple Actions***

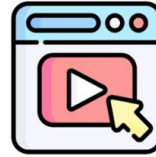


# What's New on VideoGUI?

- **Creation** e.g., Runway




*Text2Video*  
+  
*a dolly zoom effect.*



What's new on  
VideoGUI?

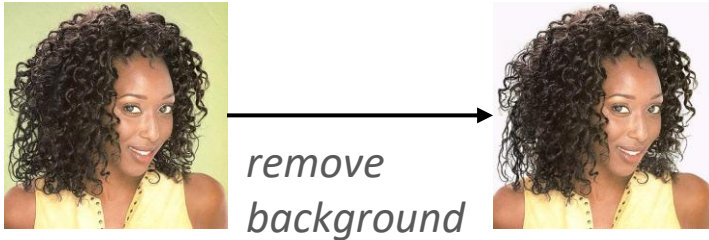
Visual-centric Software



runway Stable Diffusion

Media Creation, Editing, AI tools

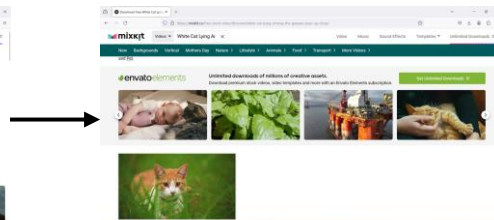
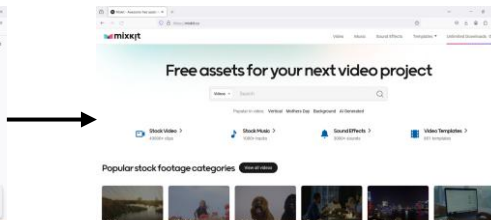
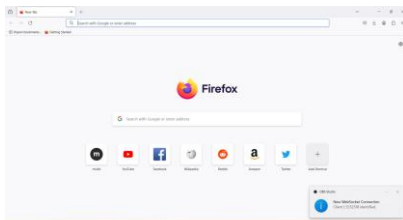
- **Editing** e.g., PhotoShop



## 1. Novel Softwares

- Media creation, editing, browsing
- AI Tools

- **Browsing** e.g., Download a video from mixkit.co



mixkit-white-cat-  
lying-among-the-  
grasses-seen-up-  
-close-22732

# What's New on VideoGUI?

## Tasks from Instructional Videos



Tutor's Instructions

Goal: How to create this effect in PPT?



What's new on VideoGUI?

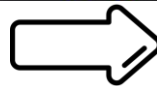
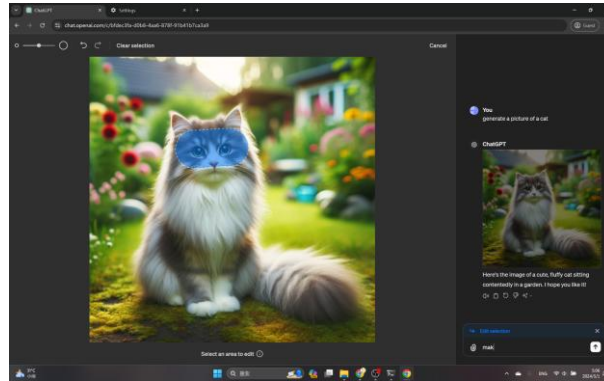
## Visual-centric Software



Media Creation, Editing, AI tools

## 2. Novel Tasks

- o Source from Inst. Video e.g., interactive editing in DALLE-3?





# What's New on VideoGUI?

## Tasks from Instructional Videos



Tutor's Instructions

Goal: How to create this effect in PPT?



What's new on VideoGUI?

## Visual-centric Software



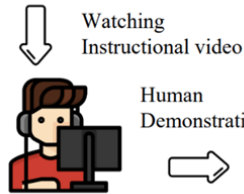
Media Creation, Editing, AI tools

## 3. Hierarchical Annotations

### High-level Planning

### Mid-level Planning

### Atomic-action Exec.



### [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

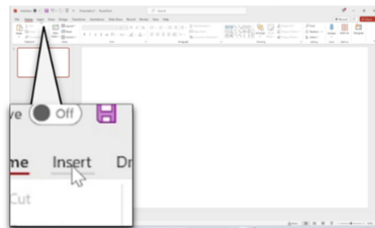
- A-33. Key board Type Ctrl + A
- A-34. Click on Shape Format
- A-35. Click on Merge Shapes
- A-36. Click on Union

- 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
- ...

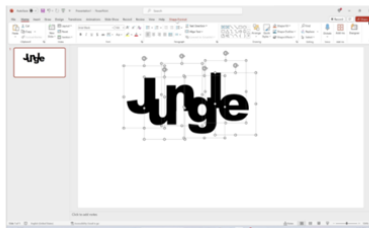
- C-28. Drag slider to decrease Smooth start duration

↑ Reproduced results

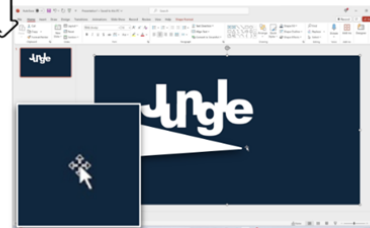


### [Atomic-action Execution]

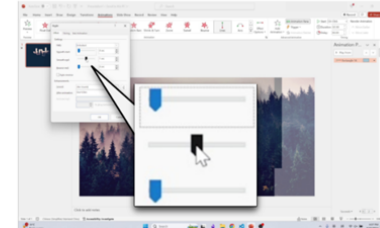
Action: Click Element: Insert  
Coordinate: [208, 100]



Action: Type / Press  
Element: Ctrl + A



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



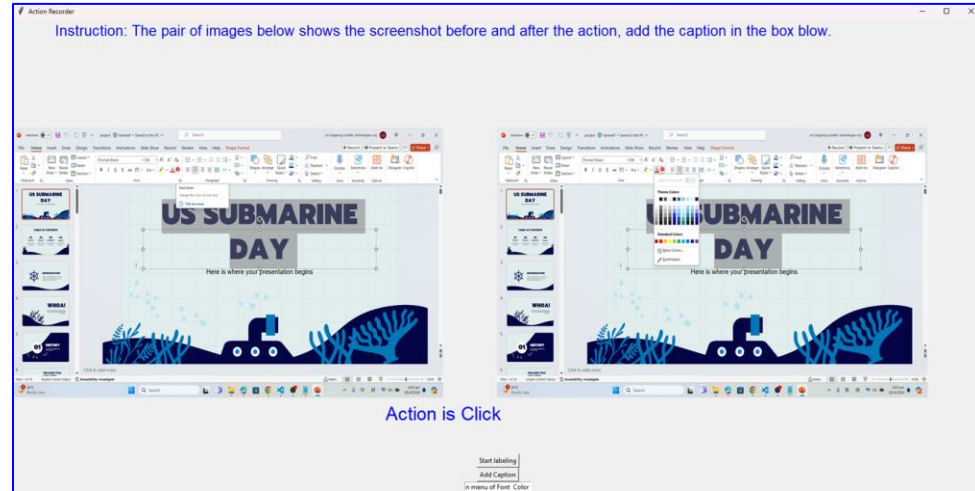
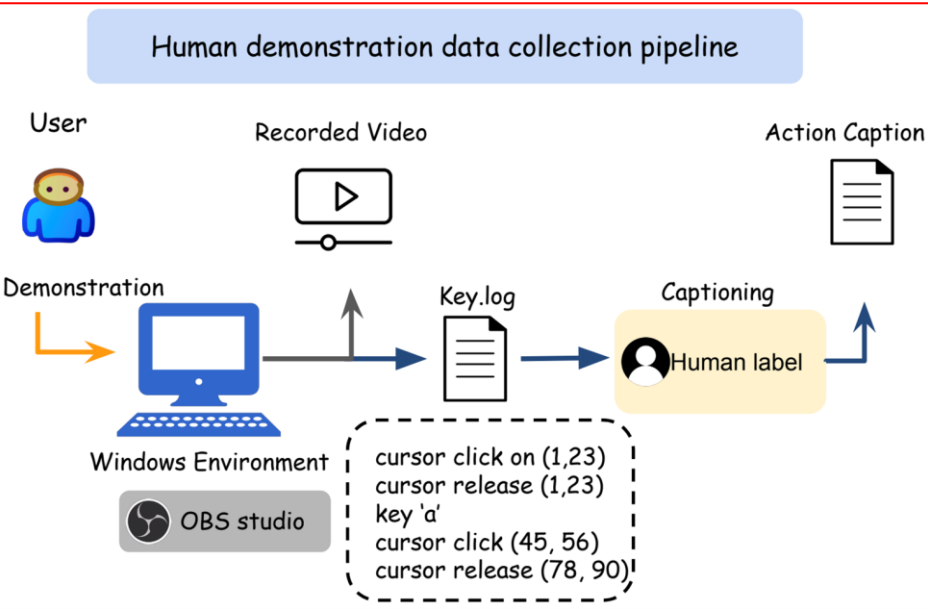
Action: Drag  
Element: slider of Smooth end  
Coordinate: [468, 344] to [281, 346]



# How we VideoGUI collect



**Figure 2: Illustration of pipeline**, encompassing four phases: **(i)** High-quality instruction videos are manually selected. **(ii)** Participants replicate skills demonstrated in videos. **(iii)** Participants annotate task elements and procedures. **(iv)** Annotated data is validated manually for use.





# VideoGUI Hierarchical Annotations

## Video preview



Start



End



### Full task

**Visual query:** How to transform from [start] to [end] in Premiere Pro?

**Textual query:** Change the blue sky in the background of the picture to the red color at sunset.

### ● High-level Plans

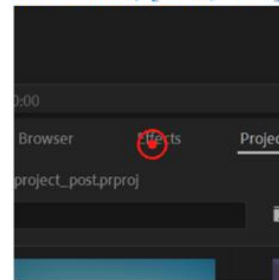
- a. Add ultra key effect to the video
- b. Get the color of the background
- c. Adjust the track order to the second track
- d. Add the new background photo to the first video track;

### ● Mid.-level Plans

- a1. Click on Effects panel
- a2. Click on search bar in Effects panel
- a3. Key board Type 'ultra'
- a4. Click on 'Ultra Key'
- a5. Drag Ultry key effect from effects panel to the video. (Purpose: add ultra key to the video)

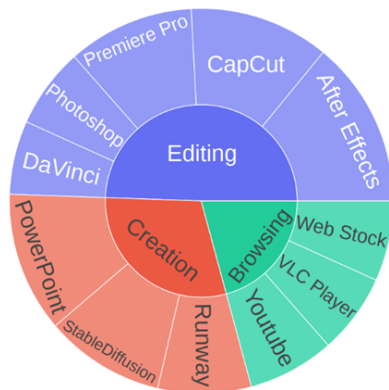
### ● Atomic Actions

- a1. Click, [216, 996]

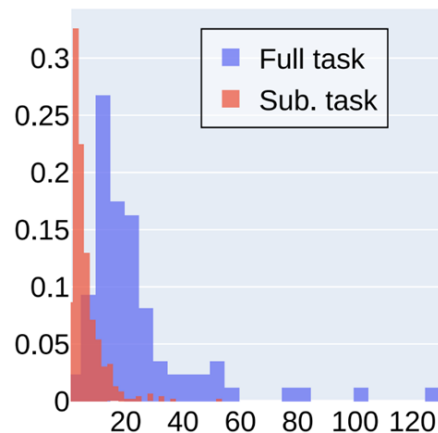


# VideoGUI Statistics

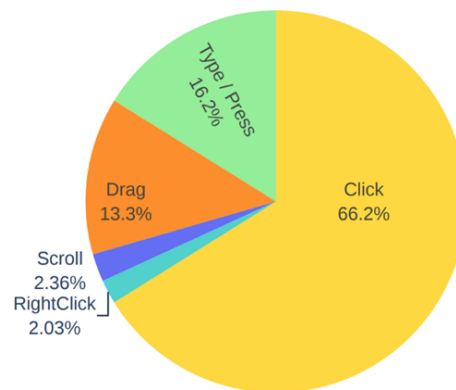
Benchmark	# Task	Platform	Source	Query format			# Avg. Action	Eval. dimension		
				Text	Image	Video		Task SR.	Hier. Plan.	Action Exec.
Mind2Web [6]	2350	Web	Screenshot	✓			7.3	✓		✓
PixelHelp [20]	187	Android	Emulator	✓			4.2	✓		✓
AITW [10]	30K	Android	Emulator	✓			6.5	✓		✓
AssistGUI [21]	100	Windows	Web Video	✓			–	✓		
OSWorld [22]	369	Win.+Ubuntu	Emulator	✓			–	✓		
V-WebArena [23]	910	Web	Screenshot	✓	✓		–	✓		
VideoGUI	SUBTASK	463	Win. +Web	Video + Human Demonstration			5.6	✓		
	FULLTASK	86			✓	✓	22.7		✓	✓



(a) Dist. of Software taxonomy.



(b) Num. of Action per task.

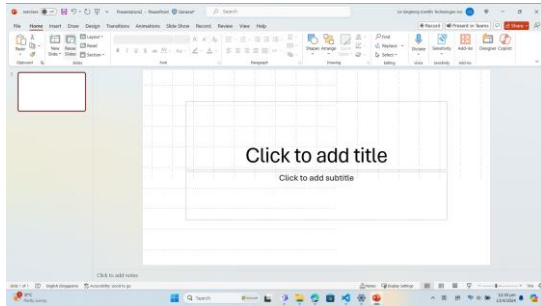


(c) Dist. of Atomic actions.

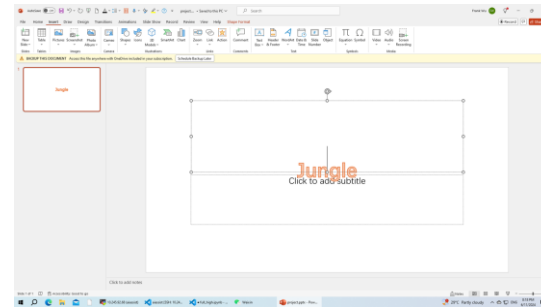
Figure 3: Data statistics of VideoGUI.

# How to Evaluate such Complex Task?

- 0 / 1 Success Rate?
  - *Easy to be 0 and fail to receive enough signal*



*Success*



*Fail*

# How to Evaluate such Complex Task?

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

# How to Evaluate such Complex Task?

## ● Hierarchical Assessment & Multiple Settings

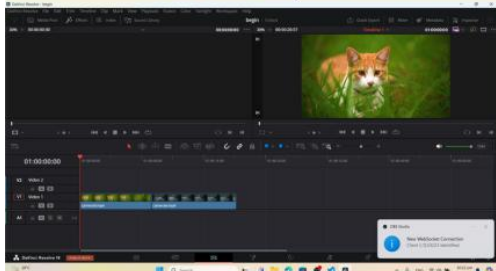
- Procedural Planning – **LLM Eval** (e.g., MM-Vet style) to overcome subjectivity
  - **High-level:** Visual Preview 👉 Milestones
  - **Mid.-level:** Milestone 👉 Action Narration Sequence



[Visual input]



[Visual output]



[Visual init state]

### Textual query:

Insert the title effect ‘Clean and Simple Lower Third’ with text ‘The cat scared the dog away’ at the beginning of the “cameraB” video;

### High-level Planning:

How [Visual input] to [Visual output]?

### GT:

A. Insert the title effect ‘Clean and Simple Lower Third’ with text ‘The cat scared the dog away’ at the beginning of the first clip.

B. Make the second clip color warmer by moving the gamma value closer to orange;

### Middle-level Planning: [Init state] + Text

#### GT:

A1. Click on Effects

A2. Click on Titles

A3. Drag Clean and Simple Lower Third from original position to the beginning of the cameraB.mp4. (Purpose: insert the title effect)

A4. Click on Inspector

A5. scroll up 5

A6. Drag SAMPLE TEXT from the last letter to the first letter. (Purpose: select all text)

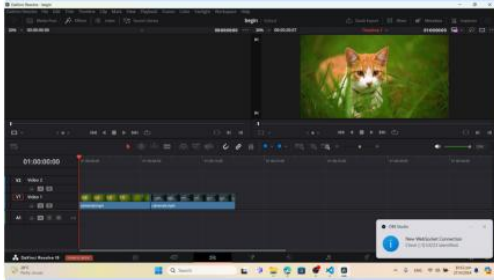
A7. Key board Type ‘The cat scared the dog away’

# How to Evaluate such Complex Task?

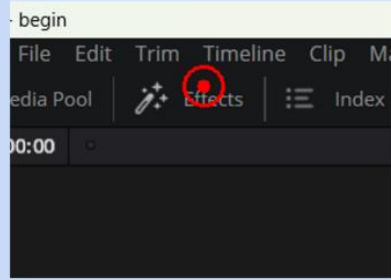
- **Hierarchical Assessment & Multiple Settings**

- **Action Execution (4 type)**

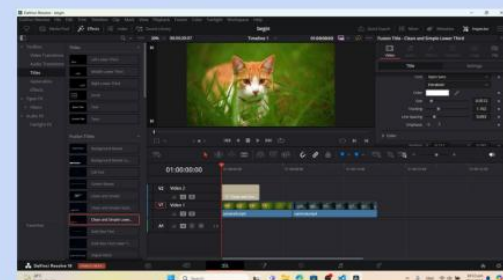
- Click; Drag; Type / Press; Scroll



1. **Click:** Where is the 'Effects'?



GT: [289, 66]



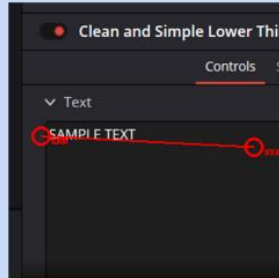
2. **Scroll:** Should I scroll up / down / not to find the text box?

Action-level

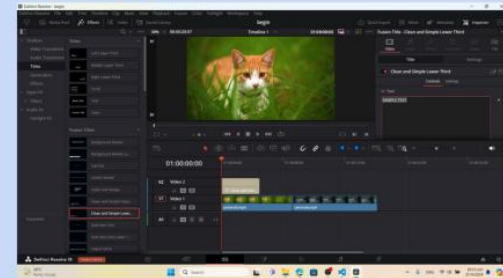
GT: UP



3. **Drag:** How to drag SAMPLE TEXT from the last letter to the first letter.



GT\_Start: [1645, 393]  
GT\_End: [1415, 381]



4. **Type / Press:** Please type "The cat scared the dog away"

GT:  
The cat scared the dog away



Env. Monitor



# How to Evaluate such Complex Task?

- **Hierarchical Assessment & Multiple Settings**

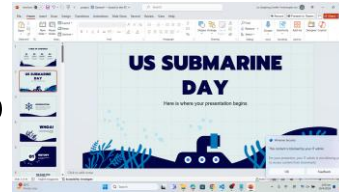
- Procedural Planning

- **High-level:** Visual Preview 🖱 Milestones
- **Mid.-level:** Milestone 🖱 Action Narration Sequence

**Visual [Hardest]:** How



to



?

**Visual + Text:**

Based



, *Swap the first ppt and the second ppt*

**Text:**

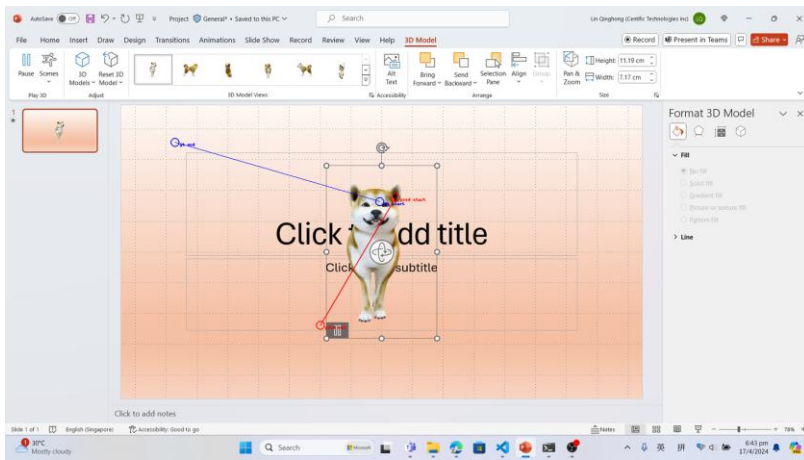
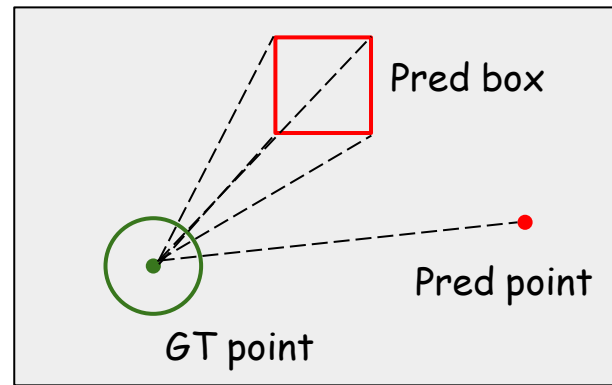
*Swap the first ppt and the second ppt*



# Action Metrics – Click / Drag – Grounding

- Human GT: **point** [x,y]
- Model Predictions:
  - **Boxes:** CogAgent, Qwen-VL-Max
  - **Points:** GPT-4-SoM
- Metrics
  - Distance
  - Recall @ r

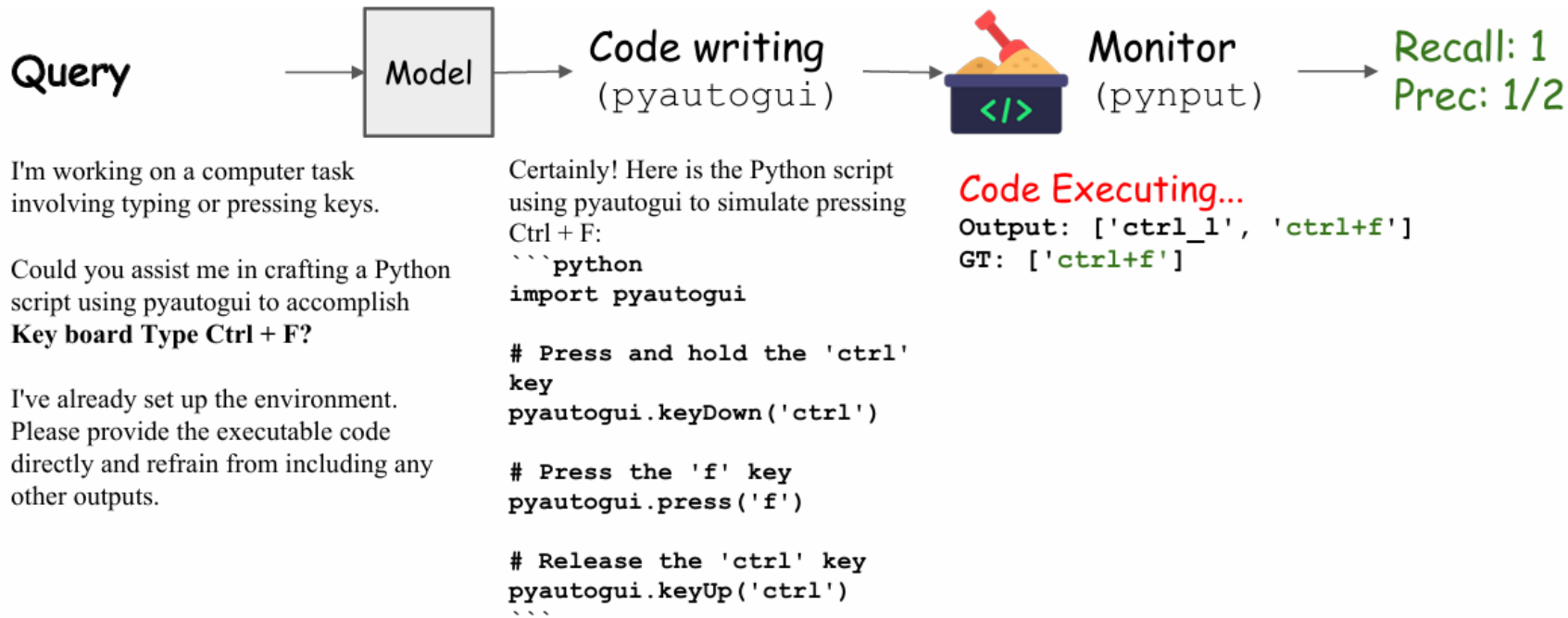
**Click**



*e.g., drag the the dog to top-left*

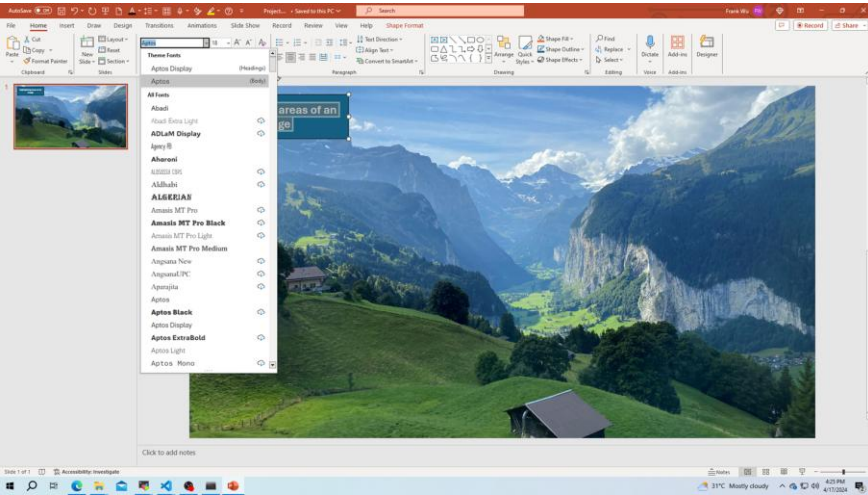
***Drag** is more strict and requires both start and end are satisfied*

# Action Metrics – Key / Press – Coding

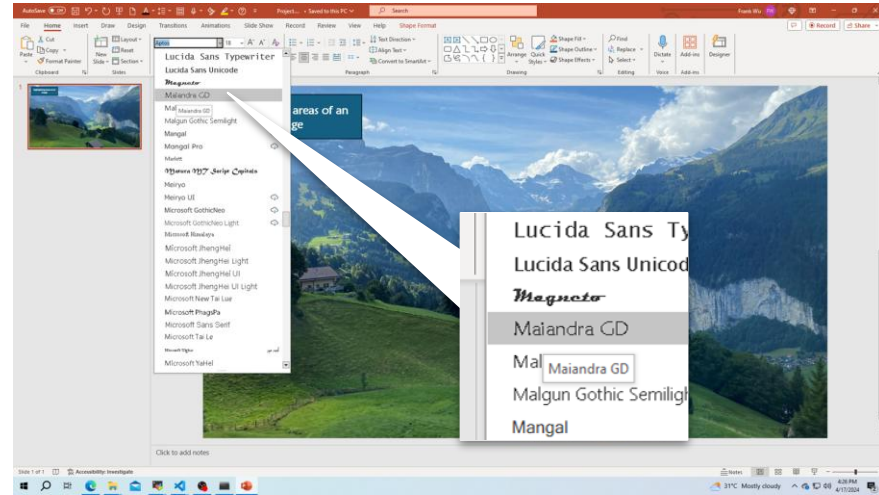


**Figure 2:** Illustration of how we evaluate the key / press action.

# Action Metrics – Scroll – Multiple-Choice Question



Scroll down



Before Scrolling, we assume that [the target element] is not within the screenshot.

After Scrolling, we assume that [the target element] should be inside the screenshot.

Next Action: Click [the Maiandra GD button]

Question: Should I scroll to find the maiandra GD button?

Question: Should I scroll to find the maiandra GD button?

Answer: Scroll Up / [✓] Scroll Down / No need

Answer: Scroll Up / Scroll Down / [✓] No need

# Baseline Models

- Model scope
  - Text LLM / Multi-modal LLM (1f or >1f)
- Interleaved Instructions
  - Text / Image / Image+Text / Image Pairs / Video / Video+Text / Video Pairs

Model	Support Interleaved Instructions?			VideoGUI Evaluation (%)			
	Text	Image (1f)	Media (> 1f)	High Plan	Mid. Plan	Action	Overall
LLama3-70B [43]	✓			–	40.5	20.3	20.3
Mixtral-8x22B [44]	✓			–	36.0	19.6	18.6
GPT-3.5-Turbo [42]	✓			–	49.1	22.3	23.8
CogAgent [19]	✓	✓		4.4	21.8	7.4	17.6
Qwen-VL-Max [41]	✓	✓	✓	5.1	35.7	28.9	27.7
Gemini-Pro-V [40]	✓	✓	✓	7.9	28.6	23.8	14.8
Claude-3-Opus [39]	✓	✓	✓	9.7	45.6	39.4	38.7
GPT-4-Turbo [36]	✓	✓	✓	14.3	52.9	34.4	34.9
GPT-4o [36]	✓	✓	✓	17.1	53.5	47.6	50.6

# Results on Procedural Planning

*2. With clear text description, gpt3.5 or open-source is sufficient for planning*

Model	High-level Planning (0 – 5)			Middle-level Planning (0 – 5)		
	Vision	Text	Vision & Text	Vision	Text	Vision & Text
LLama3-70B [43]	–	2.62	–	–	2.02	–
Mixtral-8x22B [44]	–	2.43	–	–	1.80	–
GPT-3.5-Turbo [42]	–	2.67	–	–	2.46	–
CogAgent [19]	0.22	1.12	1.23	–	1.32	1.09
Qwen-VL-Max [41]	0.25	2.30	1.96	0.70	1.72	1.79
Gemini-Pro-Vision [40]	0.39	2.35	1.45	0.34	1.61	1.43
Claude-3-Opus [39]	0.48	2.54	2.17	0.66	2.26	2.28
GPT-4-Turbo [36]	0.71	2.57	<b>2.55</b>	1.49	<b>2.57</b>	2.65
GPT-4o [36]	<b>0.86</b>	<b>2.68</b>	2.46	<b>1.78</b>	2.45	<b>2.68</b>
<b>Avg. by models</b>	<b>0.49</b>	<b>2.37</b>	<b>1.97</b>	<b>0.99</b>	<b>2.02</b>	<b>1.98</b>

*1. Query by **visual-preview** is extremely challenging; High-level is harder than middle-level*

*3. Vision & Text is even poor than Text-only, Calling stronger Interleaved Multi-Modal understanding*

# Results on Action Execution

Model	Grd.?	1. Click		2. Drag		3. Type / Press		4. Scroll	Action
		Dist. ↓	Recall ↑	Dist. ↓	Recall ↑	Recall	Prec.	Acc.	Full
Random	–	49.9	0.7	47.2	0.0	–	–	31.3	8.0
<i>LLMs</i>									
LLama3-70B [43]	–	–	–	–	–	84.9	81.3	–	20.3
Mixtral-8x22B [44]	–	–	–	–	–	82.6	78.5	<i>Keyboard</i>	19.6
GPT-3.5-Turbo [42] [42]	–	–	–	–	–	<b>93.1</b>	<b>89.5</b>	<i>is easy,</i>	22.4
<i>Multi-modal LLMs</i>									
CogAgent [19]	✓	30.9	3.4	44.7	0.0	–	–	26.6	7.5
Qwen-VL-Max [41]	✓	46.8	0.0	42.0	0.3	84.3	73.0	42.2	28.9
Gemini-Pro-Vision [40]		40.7	5.0	40.8	0.0	86.4	82.2	7.8	23.8
Claude-3-Opus [39]		30.7	7.0	30.6	1.7	92.5	88.1	60.9	39.4
GPT-4-Turbo [36]		23.8	10.0	31.3	1.4	92.3	88.8	37.5	34.4
GPT-4o [36]		16.6	<b>17.7</b>	21.9	2.5	92.3	89.0	<b>81.3</b>	47.6
<i>Modular methods: LLMs + Tools</i>									
GPT-3.5 + OCR [42]	✓	16.8	48.7	36.4	5.5	93.1	89.5	56.3	50.0
GPT-4o + OCR [42]	✓	<b>12.0</b>	<b>60.1</b> (+42.4)	25.7	<b>11.3</b> (+8.8)	92.3	88.8	82.8 (+1.5)	<b>56.3</b> (+8.7)
GPT-4o + SoM [33]	✓	15.7	35.9 (+18.2)	<b>22.9</b>	3.0 (+0.5)	92.3	88.8	<b>89.0</b> (+7.7)	54.3 (+6.7)

*GPT-4o is strong on visual perception, with resolution hint prompting even better than grounding model*



# Results on Action Execution

Model	Grd.?	1. Click		2. Drag		3. Type / Press		4. Scroll	Action
		Dist. ↓	Recall ↑	Dist. ↓	Recall ↑	Recall	Prec.	Acc.	Full
Random	–	49.9	0.7	47.2	0.0	–	–	31.3	8.0
<i>LLMs</i>									
LLama3-70B [43]	–	–	–	–	–	84.9	81.3	–	20.3
Mixtral-8x22B [44]	–	–	–	–	–	82.6	78.5	–	19.6
GPT-3.5-Turbo [42] [42]	–	–	–	–	–	<b>93.1</b>	<b>89.5</b>	–	22.4
<i>Multi-modal LLMs</i>									
CogAgent [19]	✓	30.9	3.4	44.7	0.0	–	–	26.6	7.5
Qwen-VL-Max [41]	✓	46.8	0.0	42.0	0.3	84.3	73.0	42.2	28.9
Gemini-Pro-Vision [40]		40.7	5.0	40.8	0.0	86.4	82.2	7.8	23.8
Claude-3-Opus [39]		30.7	7.0	30.6	1.7	92.5	88.1	60.9	39.4
GPT-4-Turbo [36]		23.8	10.0	31.3	1.4	92.3	88.8	37.5	34.4
GPT-4o [36]		16.6	17.7	21.9	2.5	92.3	89.0	81.3	47.6
<i>Modular methods: LLMs + Tools</i>									
GPT-3.5 + OCR [42]	✓	16.8	48.7	36.4	5.5	93.1	89.5	56.3	50.0
GPT-4o + OCR [42]	✓	<b>12.0</b>	<b>60.1</b> (+42.4)	25.7	<b>11.3</b> (+8.8)	92.3	88.8	82.8 (+1.5)	<b>56.3</b> (+8.7)
GPT-4o + SoM [33]	✓	15.7	35.9 (+18.2)	<b>22.9</b>	3.0 (+0.5)	92.3	88.8	<b>89.0</b> (+7.7)	54.3 (+6.7)

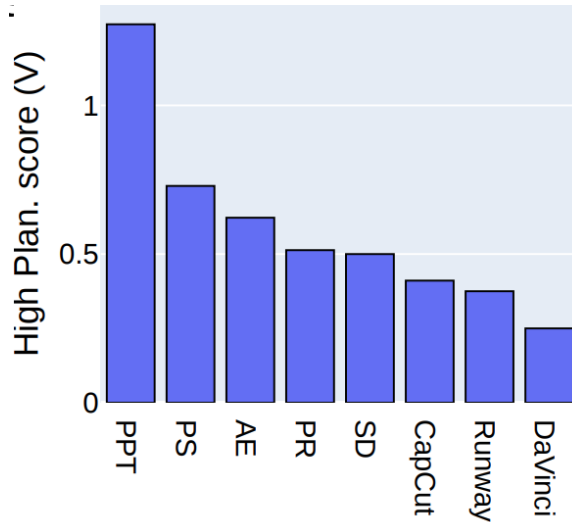
Both OCR / SoM boost performance,

Drag is the most



# Ablations

- High-level planning scores by *different software*
- Mid.-level planning scores by *action number*



- *PPT is the easiest one;*
- *Professional software like PS / AE are more challenging;*
- *Runway and CapCut, which are web-based, still relatively difficult to master.*



- *Planning becomes more difficult as the number of actions increases.*

# Qualitative Results



## High-level Planning (Final visual effect)

How to create such effect in Powerpoint?

### GT:

1. Set up a blank slide with a **gradient fill background**
2. Insert a **3D model of a shiba lun dog** and adjust its size and position
3. **Add a text 'Shiba Inu'** in title box
4. Duplicate the first slide
5. Drag the title box to bottom with text **'Side view'**
6. Drag the dog to center and **rotate to right**
7. Apply the **Morph transition** to all slides;

### GPT-4o [4]

- 1: Insert a **3D model of a Shiba Inu**.
- 2: **Add text 'Shiba Inu'** and position it next to the 3D model.
- 3: Apply the **'Fly-in' animation** to the 3D model from the left.
- 4: Add a **second slide** with a **side view** of the Shiba Inu 3D model.
- 5: **Add text 'Side view'** and position it below the 3D model.
- 6: Apply the **'Morph' transition** between the two slides.

### Gemini-Pro-V [2]

1. Insert a **3D model of a Shiba Inu dog**.
2. **Apply a "Rotate" animation** to the dog, specifying the "Side" direction.
3. Insert a text box with the word **"Side view"** and position it below the dog.

### Claude-3-Opus [2]

1. **Insert an image of a Shiba Inu dog**.
2. **Add the text "Shiba Inu"** next to the image.
3. Apply a **3D rotation effect** to the dog image, rotating it from a side view to a front view.

# Qualitative Results



## High-level Planning (Final visual effect)

How to create such effect in Powerpoint?

### GT:

1. Set up a blank slide with a **gradient fill background**
2. Insert a **3D model of a shiba lun dog** and adjust its size and position
3. **Add a text 'Shiba Inu'** in title box
4. Duplicate the first slide
5. Drag the title box to bottom with text **'Side view'**
6. Drag the dog to center and **rotate to right**
7. Apply the **Morph transition** to all slides;

### GPT-4o [4]

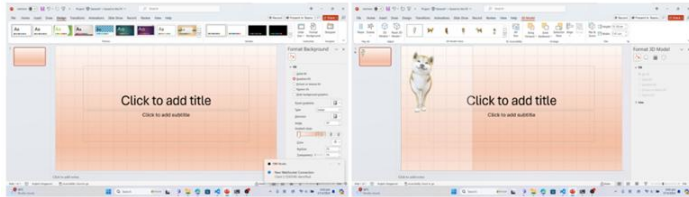
- 1: Insert a **3D model of a Shiba Inu**.
- 2: **Add text 'Shiba Inu'** and position it next to the 3D model.
- 3: Apply the **'Fly-in' animation** to the 3D model from the left.
- 4: Add a **second slide with a side view** of the Shiba Inu 3D model.
- 5: **Add text 'Side view'** and position it below the 3D model.
- 6: Apply the **'Morph' transition** between the two slides.

### Gemini-Pro-V [2]

1. Insert a **3D model of a Shiba Inu dog**.
2. **Apply a "Rotate" animation** to the dog, specifying the "Side" direction.
3. Insert a text box with the word **"Side view"** and position it below the dog.

### Claude-3-Opus [2]

1. **Insert an image of a Shiba Inu dog**.
2. **Add the text "Shiba Inu"** next to the image.
3. **Apply a 3D rotation effect** to the dog image, rotating it from a side view to a front view.



Init. visual state

Final visual state

### GPT-4o [5]

1. Click the **'Insert'** tab in the ribbon at the top of the screen.
2. Click the **'3D Models'** button in the 'Illustrations' group.
3. Select **'From Online Sources...'** from the dropdown menu.
4. In the search bar, type **'shiba inu dog'** and press Enter.
5. Browse the results and click on the desired 3D model of a shiba inu dog.
6. Click the **'Insert'** button to add the 3D model to the slide.
7. Click on the 3D model to select it.
8. Use the corner handles to **adjust the size of the 3D model by clicking and dragging**.
9. Click and drag the 3D model to position it in the desired location on the slide.

### Gemini-Pro-V [3]

1. Click the **'Insert'** tab.
2. In the **'3D Models'** group, click the **'From File'** button.
3. In the 'Insert 3D Model' dialog box, navigate to the folder containing the shiba inu dog model.
4. Select the 'shiba\_inu.fbx' file and click the **'Insert'** button.
5. The shiba inu dog model will be inserted into the slide.
6. Click and drag the model to reposition it.
7. Use the **'Scale' tool** to resize the model.
8. Click the **'Format'** tab.
9. **In the 'Shape Styles' group, click the '3-D Model Styles' button.**
10. In the **'3-D Model Styles' gallery**, select the desired style for the model.
11. Click the **'Close' button to close the 'Format Shape' dialog box.**
12. The shiba inu dog model will be updated with the selected style.

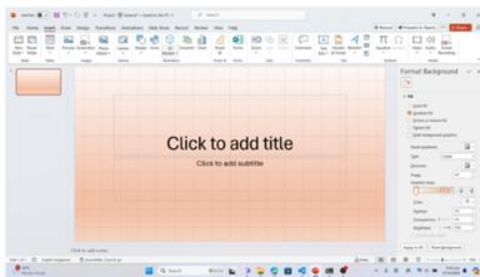
## Middle-level Planning (Init. visual state + Text)

Insert a 3D model of a shiba lun dog and adjust its size and position

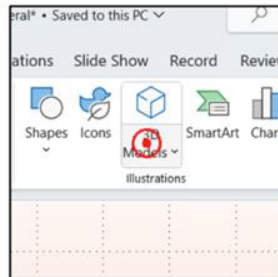
### GT:

1. Click on **Insert**
2. Click on **3D Models**
3. Click on **Stock 3D Models**
4. Keyboard **Type dog** then Enter
5. Click on Shiba-Inu dog with motion
6. Click on **Insert**
7. Drag the the dog to top-left
8. Drag the the lower right corner to enlarge the dog

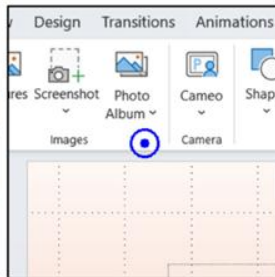
# Qualitative Results



**Action Execution (Click)**  
Click on '3D Models'

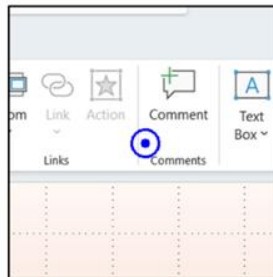


GT



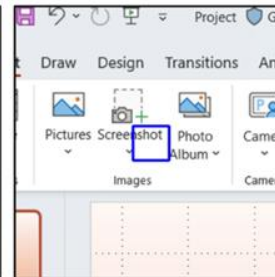
GPT-4V

Dist: 0.16, Recall: 0



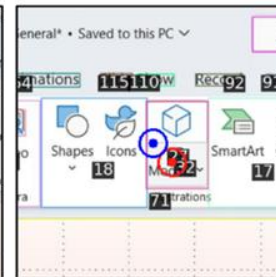
Gemini-Pro-V

Dist: 0.32, Recall: 0



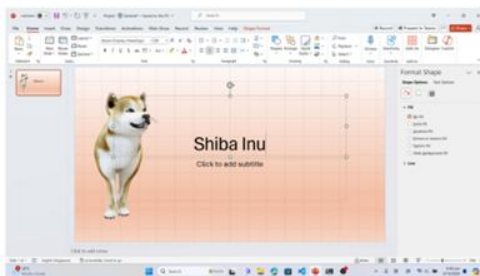
CogAgent

Dist: 0.2, Recall: 0



GPT-4V-SoM

Dist: 0.02, Recall: 1



**Action Execution (Drag)**  
Drag to select the text 'Shiba Inu' from right to left

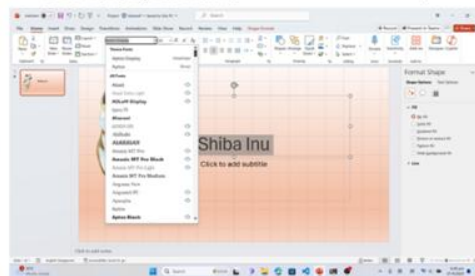


GT



GPT-4V-SoM

Dist: 0.19, Recall: 0



**Action Execution (Scroll)**

Should I scroll up / down / not to find the 'Calibri font type'?



GT: Down

GPT-4o: Down

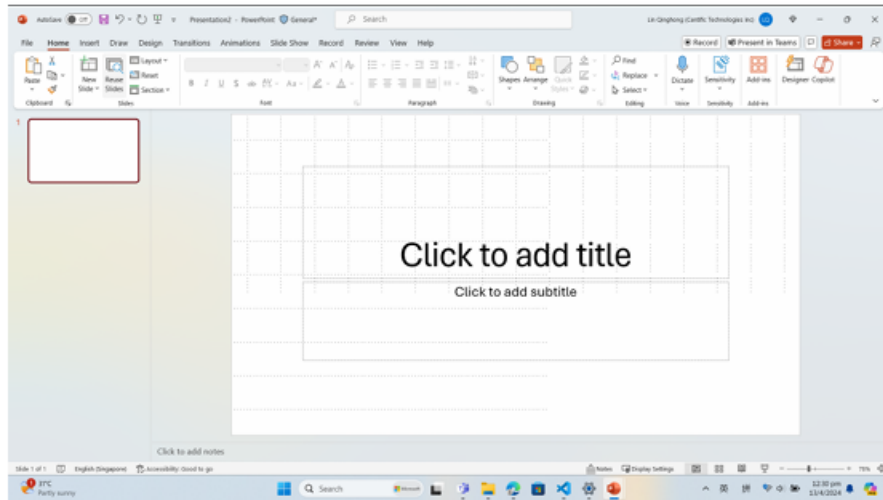
GPT-4V: Up

Gemini-Pro-V: No need

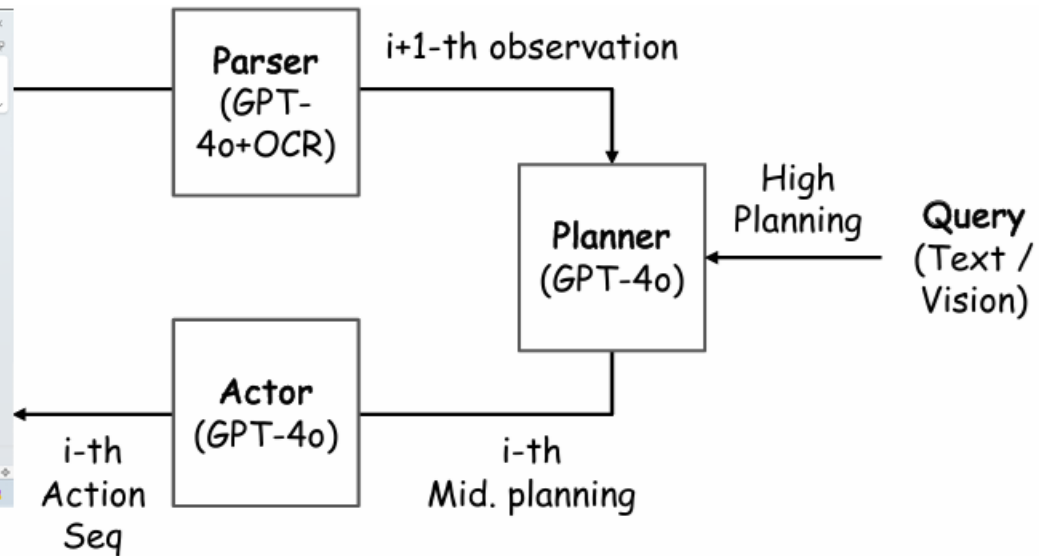
Claude-3-Opus: No need



# Using GPT-4o to build a mini Agent



i+1-th State (screenshot)



# Real-world Simulator Result

Model	Settings	VideoGUI Eval.			Full task Eval.	
		High Plan.	Mid Plan.	Action	Success Rate	Rank (Arena) ↓
GPT-4o [9]	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	<b>1.38</b>

Table 13: Simulator Evaluation on VideoGUI's PPT *full tasks*.

*Full tasks are extremely hard, SR. cannot provide enough feedback.* *human planning provides meaningful assistance*

Model	Settings	VideoGUI Eval.		Subtask Eval.	
		Mid Plan.	Action	Success Rate (%)	Avg. Round ↓
GPT-4o [9]	Orig. Query (V+T)	53.5	56.3	20.0	5.4
	w. GT Mid Plan.	100	56.3	<b>50.0</b>	<b>3.3</b>

Table 14: Simulator Evaluation on VideoGUI's PPT *subtasks*.

*Mid-level planning improve subtask significantly*

*Agent planning is usually more redundant than human*



# Comparison

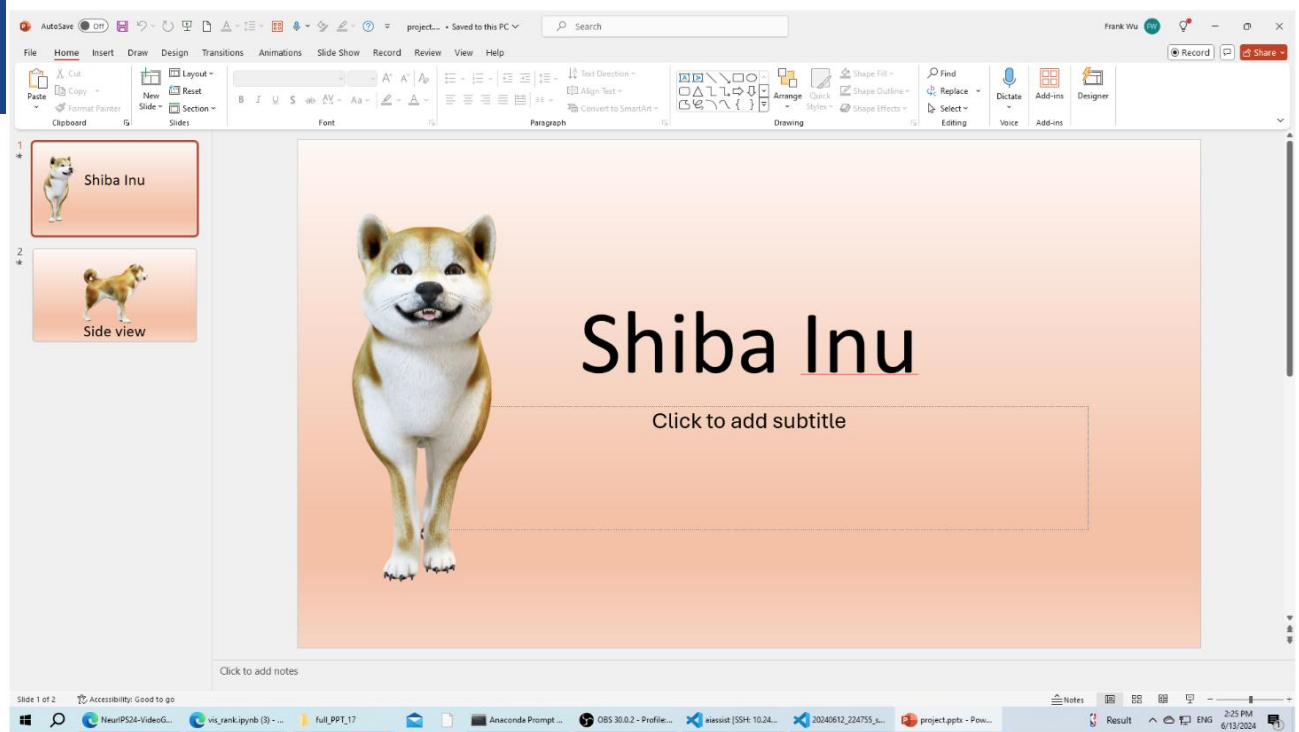
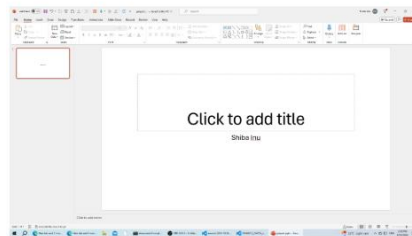
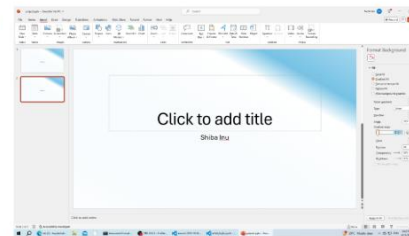


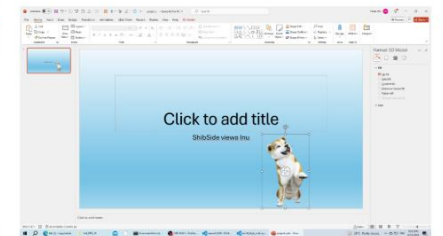
Figure 8: Final effect in Powerpoint files.



(a) GPT-4o



(b) GPT-4o w. GT High Plan



(c) GPT-4o w. GT High+Mid. Plan



# Comparison

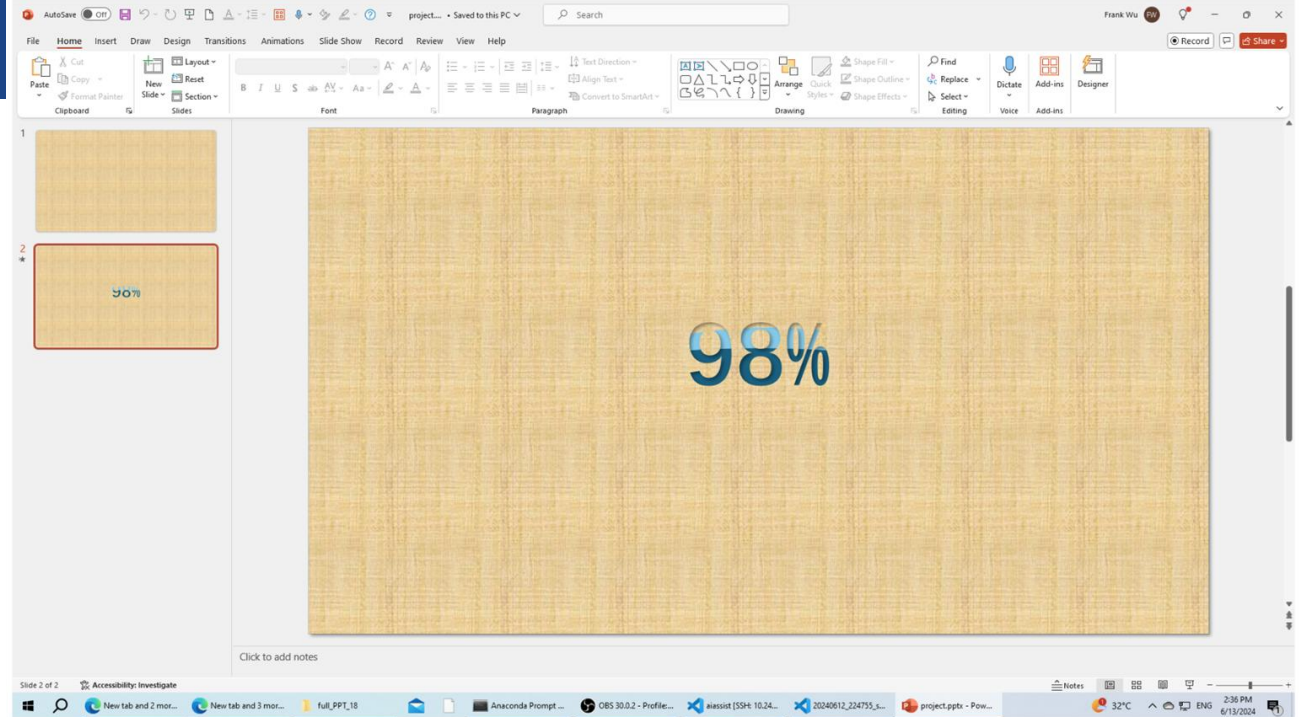
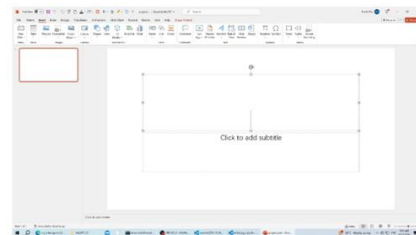
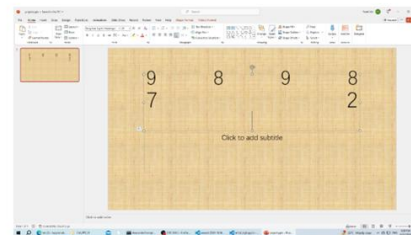


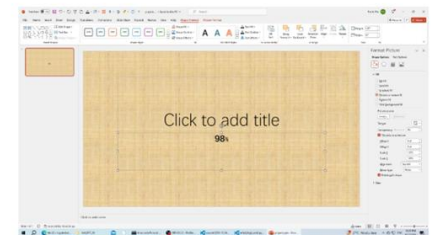
Figure 10: Final effect in Powerpoint files.



(a) GPT-4o



(b) GPT-4o w. GT High Plan

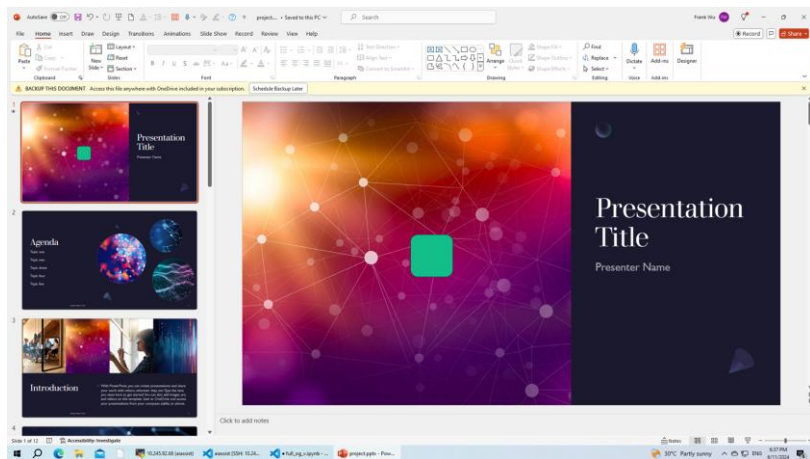
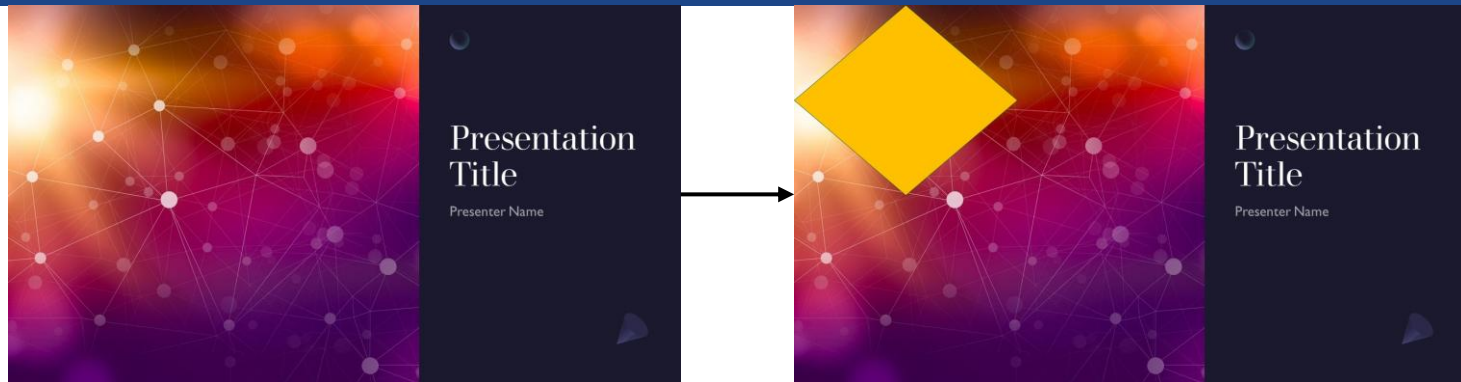


(c) GPT-4o w. GT High+Mid. Plan

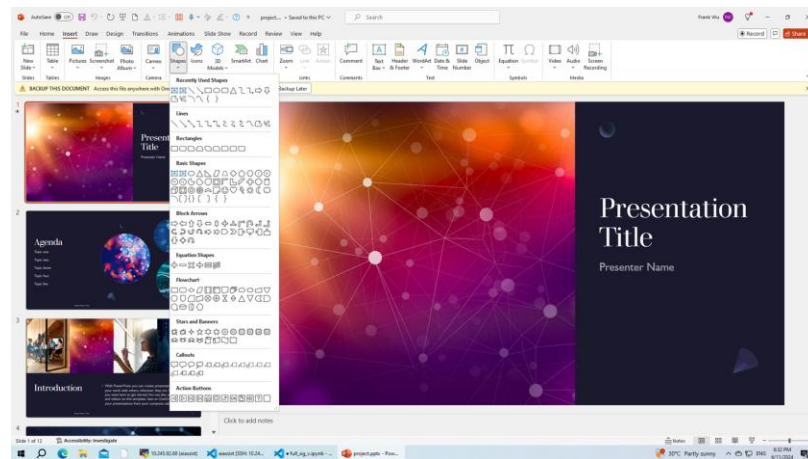
# Error Analysis (Subtask)

## Subtask:

On the first slide, draw an orange diamond on the top-left side;



GPT-4o: Final output



Failed to select the correct shape.