AssistGUI: Task-Oriented PC Graphical User Interface Automation

Supplementary Material

Overview

In the supplementary, we provide additional details for the main paper:

- More comparison with existing and concurrent benchmarks, in Sec.A;
- More details of the AssistGUI Benchmark, in Sec. B;
- More details of our proposed model, AutoPC, in Sec. C.
- More visualized prediction results, in Sec. D

A. Comparison with Previous Benchmarks

We discuss the differences between our approach and existing benchmarks in the following aspects, as shown in Table 1:

Platform: Previous methods [1–3] mainly focused on Web and SmartPhone platforms, such as AndroidEnv, AutoDroid, and WebShop. AssistGUI, however, concentrates on desktop operations. This distinction primarily brings about differences in GUI complexity. The complexity on desktops is significantly higher than on other platforms, mainly reflected in the density of information, the diversity of visual elements, and the diversity of panel layouts.

Task Focus: Exisiting methods [1–3] primarily study two types of tasks. One category is games, for instance, the majority of tasks in AndroidEnv are games, such as FlappyDroid, and Pong. The characteristic of game tasks is that the environment has a clear reward, making it easy to measure the performance of the model. Additionally, for most games, the types of operations are relatively limited. The other category includes web navigation and basic smartphone operations. These tasks have relatively simple operational patterns. For example, web navigation mainly involves buying a series of items according to requirements, with the difficulty lying in planning what to buy. The operations are relatively limited in type.

The distinguishing feature of AssistGUI is its focus on the use of productivity tools. The challenge of this category of tasks lies in the possibility of encountering new types of operations with different software. For instance, with After Effects, one might need to perform some drawing on the material. This presents a more formidable challenge for the model's understanding of the GUI and the generation of actions.

Dataset Scale and Annotation: Previous benchmarks [1–3] mainly involved about a hundred tasks. WebShop is somewhat unique; it primarily consists of one task, which is purchasing items, but it comes with different instructions specifying various purchasing requirements. The dataset scale of our benchmark is similar. However, a distinctive feature of our tasks is the use of professional software to modify documents or materials. Therefore, we also provide some project files to ensure that all methods start from the same initial state.

B. AssistGUI Benchmark

We provide more details of the state observation, action Space, and evaluation in this section.

State observation. The observation consists of two main types of information: one part is the GUI's meta-data, which is obtained by using the PywinAuto Library and the other part is screenshots of the GUI. In Figure 1, we present an example of the metadata and screenshot. It's worth mentioning that for software like Premier Pro, it is challenging to obtain meta-data that encompasses all information of the software. The main information obtainable is about large panels, while specific texts and buttons are almost impossible to extract from the meta-data. Therefore, the model must rely on visual perception capabilities to process screenshots.

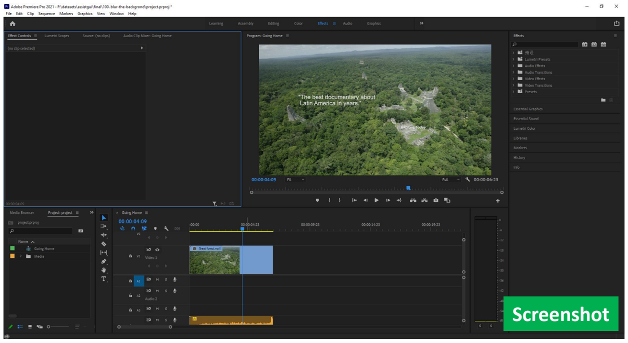
Action Space. We provide more examples of actions that are supported in AssistGUI:

- **Mouse Movement:** Move the mouse cursor to a specific position on the screen.
  Example: `moveTo(100, 150)`
- **Mouse Clicks:** Automate mouse clicks at a specified location.
  Example: `click(200, 220)`
- **Typing and Sending Keystrokes:** Simulate typing text or pressing keys.
  Example: `write('Hello, world!')`
- **Keyboard Hotkey Combinations:** Press and release keyboard shortcuts or hotkeys.
  Example: `hotkey('ctrl', 'c')`
- **Scrolling the Mouse:** Automate mouse scrolling up or down.
  Example: `scroll(-200)` for scrolling down.
- **Drag and Drop:** Automate drag and drop actions.
  Example: `dragTo(100, 200, duration=2)`
- **Mouse Down and Mouse Up:** Hold down and release the mouse button.
  Examples: `mouseDown(); mouseUp()`
- **Press and Release Keys:** Press and release individual keyboard keys.
  Examples: `press('enter')`
- **Key Down and Key Up:** Hold down and release a keyboard key.
  Examples: `keyDown('shift')`

Evaluation. To ensure a fair comparison among models,

Table 1. **Comparison of related benchmarks.** ASSISTGUI is unique in its platform and task focus. It additionally provides project files for each task.

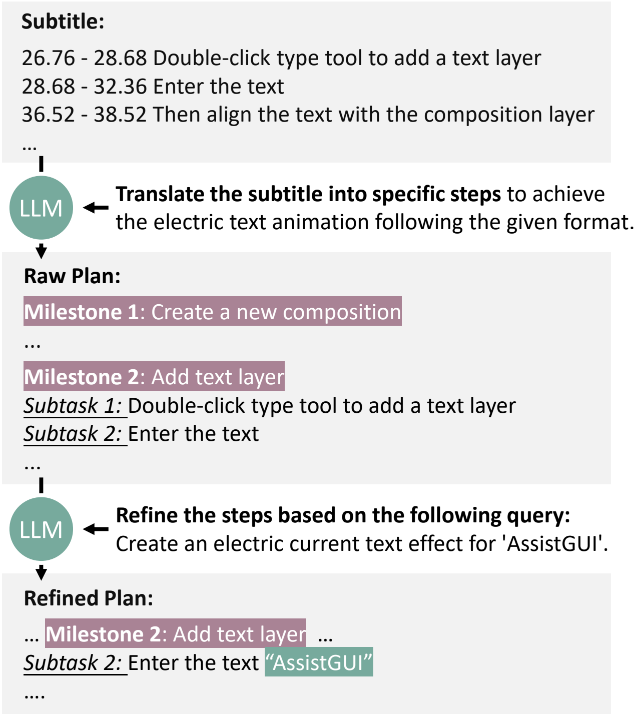
Benchmark	# APPs	# Tasks	Platform	Task Focus	Project File
AndroidEnv [1]	~30	>100	Android OS	Game & App Usage	✗
WebShop [3]	1	1 task, 12K instructions	OpenAI Gym	Web-based e-commerce	✗
AutoDroid [2]	13	158	Android OS	App Usage	✗
AssistGUI	9	100	Windows	Productivity Software Usage	✓



```
{'Adobe Premiere Pro 2021 - F:\\datasets\\assistgui\\100. blur-the-
backgrond\\project.prproj *': [{'properties': {'friendly_class_name': 'Pane',
   'texts': ['DroverLord - Frame Window'],
   'rectangle': [0, 87, 737, 631]},
  'children': [{'properties': {'friendly_class_name': 'Pane',
     'texts': ['OS_ViewContainer'],
     'rectangle': [2, 115, 734, 628]},
    'children': []}]},
 {'properties': {'friendly_class_name': 'Pane',
   'texts': ['DroverLord - Frame Window'],
   'rectangle': [737, 87, 1562, 631]},
  'children': [{'properties': {'friendly_class_name': 'Pane',
     'texts': ['OS_ViewContainer'],
     'rectangle': [740, 115, 1559, 628]},
    'children': []}]},
 {'properties': {'friendly_class_name': 'Pane',
   'texts': ['DroverLord - Frame Window'],
   'rectangle': [0, 631, 284, 1020]},
  'children': [{'properties': {'friendly_class_name': 'Pane',
     'texts': ['OS_ViewContainer'],
     'rectangle': [2, 661, 281, 1017]},
    'children': []}]},
 {'properties': {'friendly_class_name': 'Pane',
   'texts': ['DroverLord - Frame Window'],
   'rectangle': [284, 631, 336, 1020]},
  'children': [{'properties': {'friendly_class_name': 'Pane',
     'texts': ['OS_ViewContainer'],
     'rectangle': [287, 644, 333, 1017]},
    'children': []}]},
 {'properties': {'friendly_class_name': 'Pane',
   'texts': ['DroverLord - Frame Window'],
   'rectangle': [336, 631, 1443, 1020]},
  'children': [{'properties': {'friendly_class_name': 'Pane',
     'texts': ['OS_ViewContainer'],
     'rectangle': [339, 661, 1440, 1017]},
    ‘children’: []}]},
...
```

Figure 1. One example of screenshot and metadata.

we standardized the specific version numbers and languages for each software. We also provided corresponding scripts for software opening and saving, allowing the algorithms to focus solely on the operational aspects.



Subtitle:

26.76 - 28.68 Double-click type tool to add a text layer
28.68 - 32.36 Enter the text
36.52 - 38.52 Then align the text with the composition layer
...

Translate the subtitle into specific steps to achieve the electric text animation following the given format.

Raw Plan:

Milestone 1: Create a new composition
...
Milestone 2: Add text layer
Subtask 1: Double-click type tool to add a text layer
Subtask 2: Enter the text
...

Refine the steps based on the following query:
Create an electric current text effect for 'AssistGUI'.

Refined Plan:

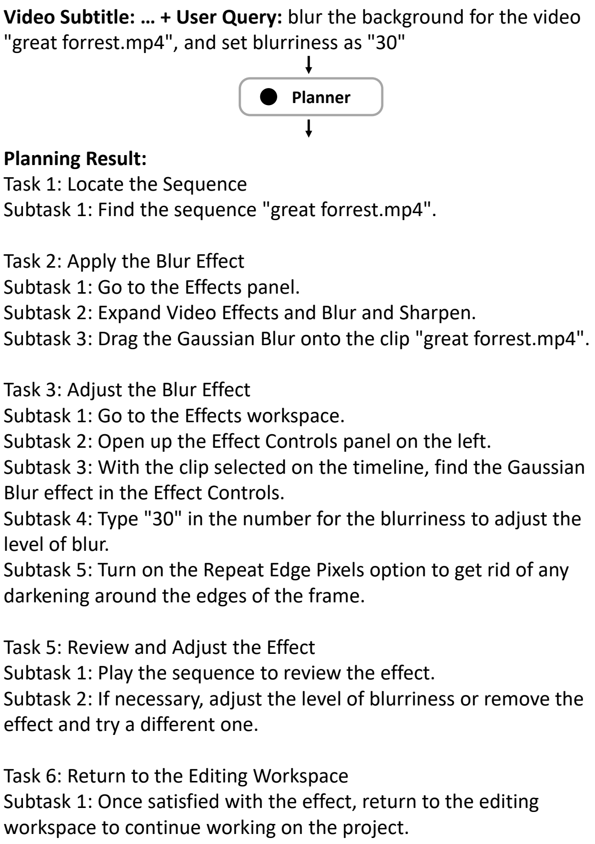
... **Milestone 2**: Add text layer ...
Subtask 2: Enter the text "AssistGUI"
....

Figure 2. **Diagram Illustration of Planner.** The Planner first translates video subtitles into a structured raw plan with milestones and subtasks. It then refines this plan by specifying the user-provided query.

For tasks that require output files, such as video editing, Word, PowerPoint, etc., we convert the output into images, for example, sample video at 1 fps, PowerPoint saved as PNG. Then, we compare the differences between the ground truth and the prediction. They assess normalized pixel-level differences against a specified threshold: $\epsilon = \frac{1}{T}\sum_{t=1}^{T}\mathrm{MAE}(V_p^t - V_{gt}^t)$ and $score = 1, \epsilon < \tau, otherwise, 0$, where τ varies based on the specific task.

C. Method

We elaborate on the Planner in this section. The Planner operates in two stages to produce structured outcomes. Initially, it processes video subtitles, generating a preliminary



Video Subtitle: ... + User Query: blur the background for the video "great forrest.mp4", and set blurriness as "30"

Planning Result:

Task 1: Locate the Sequence
Subtask 1: Find the sequence "great forrest.mp4".

Task 2: Apply the Blur Effect
Subtask 1: Go to the Effects panel.
Subtask 2: Expand Video Effects and Blur and Sharpen.
Subtask 3: Drag the Gaussian Blur onto the clip "great forrest.mp4".

Task 3: Adjust the Blur Effect
Subtask 1: Go to the Effects workspace.
Subtask 2: Open up the Effect Controls panel on the left.
Subtask 3: With the clip selected on the timeline, find the Gaussian Blur effect in the Effect Controls.
Subtask 4: Type "30" in the number for the blurriness to adjust the level of blur.
Subtask 5: Turn on the Repeat Edge Pixels option to get rid of any darkening around the edges of the frame.

Task 5: Review and Adjust the Effect
Subtask 1: Play the sequence to review the effect.
Subtask 2: If necessary, adjust the level of blurriness or remove the effect and try a different one.

Task 6: Return to the Editing Workspace
Subtask 1: Once satisfied with the effect, return to the editing workspace to continue working on the project.

Figure 3. **Planning Results.** The UI elements are organized panel by panel.

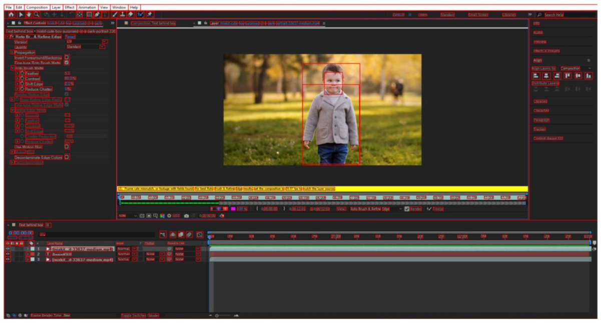
structured plan featuring milestones and subtasks. Presently, we solely rely on video subtitles for this task, as recognizing specific steps visually remains a formidable challenge in video understanding. Subsequently, the Planner refines the plan, tailoring it to the user's specific query, as illustrated in Figure 2. In both stages, the Planner leverages a Large Language Model (LLM) to perform these functions.

Additionally, it's important to note that every module within our model works in a zero-shot manner.

D. Experimental results

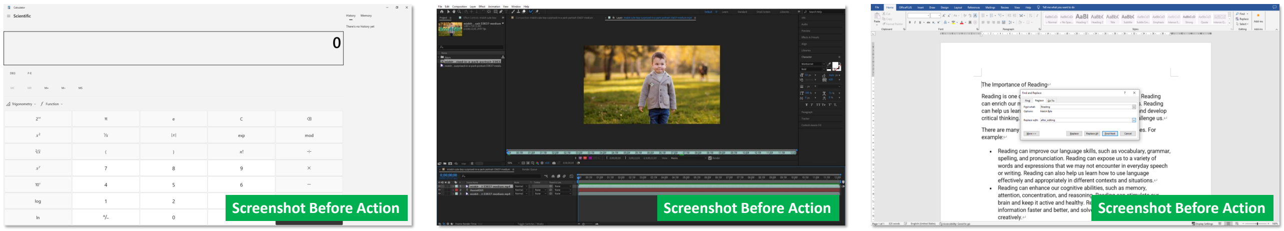
Planner. In Figure 3, we present an example of a Planner prediction. The results show that, despite the strength of GPT-4, the predictions still have some flaws, such as including redundant operations. For instance, Task 6 does not actually correspond to any specific action. This issue mainly arises from the fact that these steps are included in the instructional video, and GPT-4 cannot definitively determine whether to exclude them.

GUI Parser. In Figure 4, we show one example of the outputs of GUI Parser. The model can detect most UI ele-



```
panel:
  name:
    Effect Controls
  rectangle: [0, 68, 363, 721]
  elements:
    Effect Controls [100, 82], mixkit [155, 82], cute [184, 82], boy [208, 82], surprised [243, 83], in [274, 83], a [286, 83], park [304, 83]
    Text behind boy [49, 109], mixkit cute boy surprised [172, 109], in [250, 109], a [262, 109], park portrait 336 [315, 109]
    expand_icon [12, 126], fx Roto Br ... & Refine Edge [101, 126], Reset [214, 127]
    Version [54, 144], 20 [210, 142], expand_icon [326, 142]
    expand_icon [318, 159]
    ...
  name:
    Layer
  rectangle: [363, 68, 1699, 721]
  elements:
    Composition Text behind boy [482, 84], Layer mixkit cute [710, 83], boy [768, 83], surprised [803, 83], in [834, 83], a [846, 83], park [864, 83], portrait 33637 medium.mp4 [950, 83]
    media_asset [1031, 394], Boy [965,  214, 1151,  543], Clothing [966,  282, 1149,  539], Human Face [1036, 242, 1094,  314]
    A Frame rate mismatch , or footage with fields found . [492, 619], For best Roto [647, 619], Brush & Refine [713, 619], Edge [760, 619], results , [790, 619], set the composition to [858, 619], 29.97 fps to [937, 619], match the layer source . [1019, 619]
    Of [383, 646], 00 : 15f [426, 646], 01 : 00f [481, 647], 01 : 15f [535, 647], 02 : 00f [589, 648], 02 : 15f [644, 647], 03 : 00f [698, 648], 03 : 15f [752, 647], 04 : 00f [806, 648], ...
    ☆☆☆ * [694, 684], 100 % [770, 684], 00 00 00 [862, 684], 00 12 01 [953, 683], A [995, 684], 00 12 03 [1031, 684], View : [1091, 684], Roto Brush & Refine [1169, 684], Edge [1235, 684], Render [1332, 684], Freeze [1404, 684]
    50 % [382, 706], +0.0 [556, 706], 00 00 00 [662, 706], . [708, 707]
  name:
    Align
  rectangle: [1699, 68, 1920, 721]
  elements:
    Info [1722, 84]
    Audio [1727, 113]
    Preview [1733, 144]
    Effects & Presets [1755, 173]
    Align [1726, 204]
    Align Layers to : [1745, 230], Composition [1832, 230]
    align_left_icon [1720, 254], align_Horizontally_icon [1754, 255], align_right_icon [1789, 254], align_top_icon [1823, 252], align_Vertically_icon [1857, 252], align_bottom_icon [1891, 254]
    Distribute Layers [1750, 278], : [1794, 278]
    Libraries [1735, 336]
    Character [1737, 366]
    Paragraph [1738, 397]
    Tracker [1732, 425]
    Content Aware Fill [1760, 455]
  name:
    Timeline
  rectangle: [0, 721, 1920, 1040]
  elements:
    Text behind boy [87, 736], setting_icon [143, 737]
    timeline-1 [58, 761], composition_mini_flowchart_icon [514, 767], enables_frame_blending_icon [573, 767], enable_motion_blur_icon [599, 767]
    hide_all_layers_icon [546, 768], graph_editor_icon [633, 768]
    00000 [26, 777], : 00 : 00 : 00 [60, 761], 30.00 fps [70, 777], lo [124, 766], Of [673, 773], 05f [731, 773], 10f [800, 773], 15f [868, 773], 20f [936, 773], 25f [1004, 773], ...
    hide_icon [12, 796], OB [34, 795], solo_icon [42, 796], lock_icon [57, 796], label_icon [89, 795], Layer Name [164, 796], Mode [374, 796], .TrkMat [470, 796], Parent & Link [556, 795]
    expand_icon [422, 813], expand_icon [628, 813]
    hide_icon [12, 830], expand_icon [77, 830], T_icon [142, 830], expand_icon [422, 830], expand_icon [511, 830], parent_pick_whip_icon [534, 830], expand_icon [628, 830]
    hide_icon [12, 847], expand_icon [77, 847], ▶ ... [143, 849], ▸ .. [144, 814], AssistGUI [187, 830], mixkit ... it 33637 medium.mp4 [254, 814], mixkit ... it 33637 medium.mp4 [254, 849], Normal [387, 814], Normal [387, 830], Normal [386, 846], expand_icon [422, 847], None [473, 830], None [473, 847], expand_icon [511, 847], parent_pick_whip_icon [534, 847], None [567, 813], None [567, 831], None [567, 847], expand_icon [628, 847]
    Frame Render Time : Oms [149, 1028], Toggle Switches [419, 1029], Modes [484, 1028]
  name:
    Accessory
  rectangle: [0, 43, 1920, 68]
  elements:
    Home_icon [24, 55], Selection Tool_activated_icon [56, 55], Hand Tool_icon [81, 54], Zoom Tool_icon [105, 55], Pan Behind Tool_icon [243, 54], Rectangle_Tool_icon [278, 56], pen_tool_icon [302, 56], horizontal_type_tool_icon [327, 56], Brush Tool_icon [358, 55], Clone Stamp Tool_icon [385, 55], Eraser Tool_icon [409, 55], Roto Brush Tool_icon [440, 55], Puttet Position Pin Tool_icon [472, 56], Default [1276, 56], Learn [1354, 56], Standard [1434, 55], Small Screen [1533, 56], Libraries [1633, 56], Search Help [1779, 56]
  name:
    Application
  rectangle: [0, 23, 1920, 42]
  elements:
    File [16, 32], Edit [49, 32], Composition [107, 32], Layer [170, 32], Effect [213, 32], Animation [270, 32], View [324, 32], Window [373, 32], Help [421, 32]
```

Figure 4. **Parsed GUI Results.** The UI elements are organized panel by panel.



Input Subtask:
Enter the number 100 into the calculator by pressing the buttons "1", "0", "0".

Output Action:

```
# Click on the button "1"
click(481, 924)

# Click on the button "0"
click(800, 1000)

# Click on the button "0" again
click(800, 1000)
```

Input Subtask:
Draw a line on the boy using the roto brush.

Output Action:

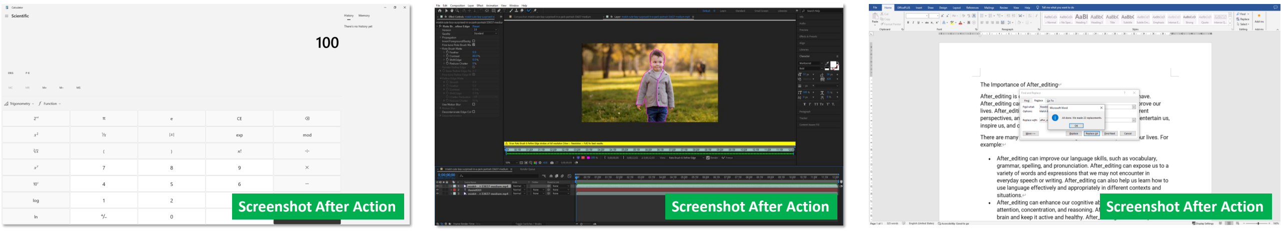
```
# Move to the start position
moveTo(1036, 239, duration=1)

# Draw a line vertically downwards
dragTo(1036, 530, duration=2)
```

Input Subtask:
Click Replace All to change every instance of "Reading" in the document to "after_editing".

Output Action:

```
# Click "Replace All" button
click(1038, 618)
```



Critic Result:
{"Reason": "Actions correctly followed to display 100 on the calculator.",
"Success": True,
"Finished": True}

Critic Result:
{"Reason": "Correct action. The line is drawn vertically as instructed, starting from the correct position.",
"Success": True,
"Finished": True}

Critic Result:
{"Reason": "Action successful with 22 replacements confirmed; close the confirmation pop-up window next.",
"Success": True,
"Finished": False}

Figure 5. **Prediction Results of Actor and Critic Module.** We show the prediction results of one specific subtask in solving a query.

ments, but there are still some flaws. 1) There are still errors in text detection. For example, there are some issues with the detection of timestamps. The timestamp in the lower-left corner should be 0:00:00:00, but it is detected as : 00 : 00 : 00. The numbers in the Timeline in the lower-left corner are not detected. 2) Some visual elements are still difficult to recognize, such as the long bars on the timeline corresponding to each layer. Additionally, the current method is unable to understand some curves and figures, and it might be necessary to leverage the capabilities of GPT-4V in the future.

Actor and Critic. In Figure 5 we showcase prediction examples from the Actor and Critic modules. It is evident that the model is capable of not only producing individual step actions but also generating a continuous action sequence. Additionally, for the Critic module, current models can effectively judge the outcomes of some simple actions, such as clicking action, as demonstrated in the left and right examples. However, for more complex scenarios, such as determining whether an object has been completely cropped out, as seen in the middle case, the model still lacks the capability to perceive this accurately.

Predicted Action Sequence. In Figure 6 - 12, we display some example tasks for each software and then visualize the model's predicted action trajectories.

References

[1] Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. Androidenv: A reinforcement learning platform for android. *arXiv preprint arXiv:2105.13231*, 2021.

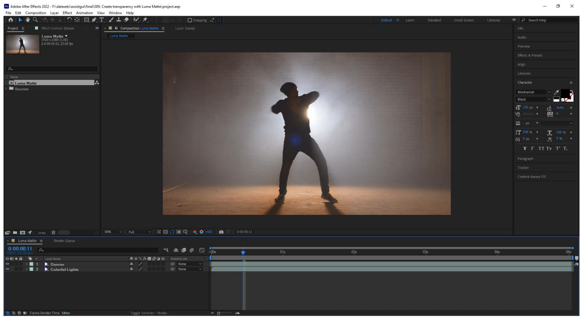
[2] Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. Empowering llm to use smartphone for intelligent task automation. *arXiv preprint arXiv:2308.15272*, 2023.

[3] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

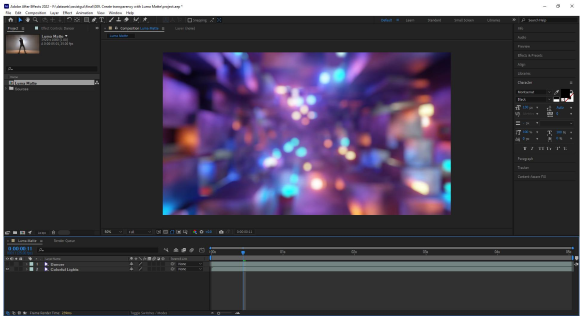
After Effect Task

Query: Creating transparency with luminance to blend dancer and colorful lights layers in video editing.

Initial Screenshot

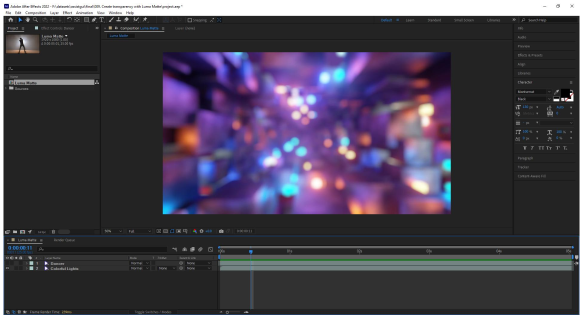


Task 1: Turn off the visibility of the dancer layer



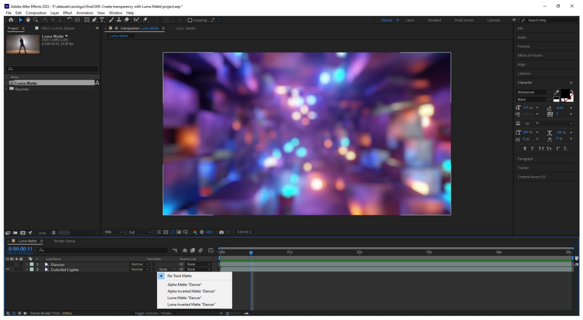
Subtask 1: Click on the eye icon next to the dancer layer in the timeline to turn off its visibility

Task 2: Toggle to 'Modes' in the timeline



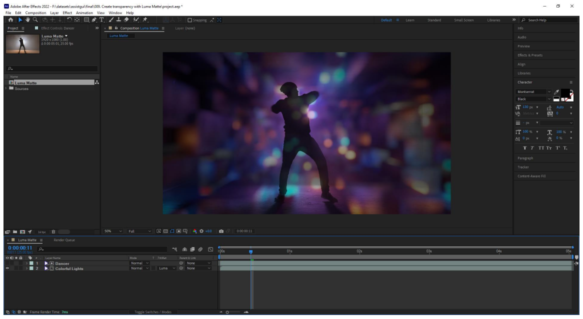
Subtask 1: Click on the 'Toggle Switches/Modes' button at the bottom of the timeline until you're in 'Modes'

Task 3: Set the track matte for the background layer



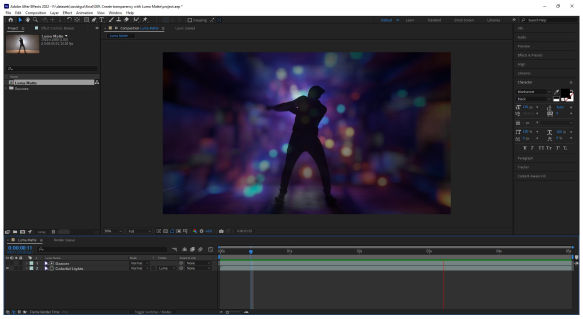
Subtask 1: Click on the 'Track Matte' dropdown menu for the colorful lights layer in the timeline

Task 3: Set the track matte for the background layer



Subtask 2: Click to select 'Luma Matte' from the dropdown menu

Task 4: Apply the Luma Matte to the layer

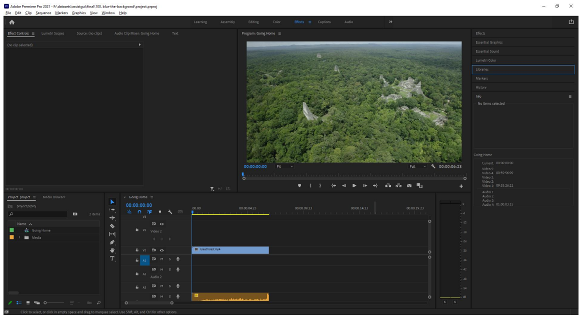


Subtask 1: Press the space bar to preview the result and adjust as necessary

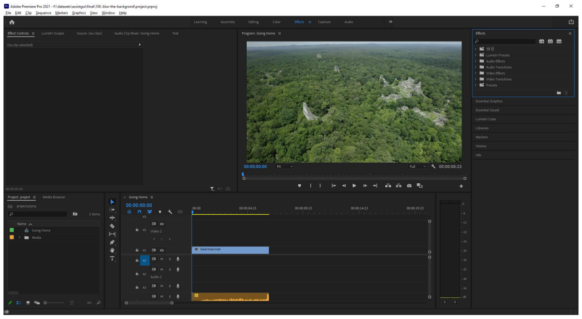
Figure 6. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

Premiere Pro Task **Query: blur the background of the "great forrest.mp4", and set blurriness as "30" by typing the number in it.**

Initial Screenshot

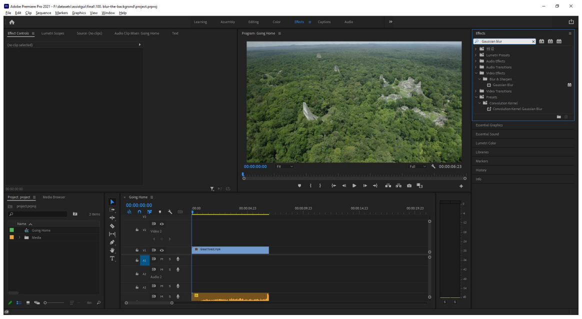


Task 1: Search for the Gaussian blur effect



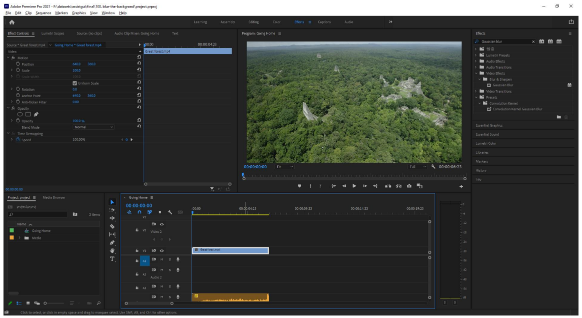
Subtask 1: Click on the effects panel

Task 1: Search for the Gaussian blur effect



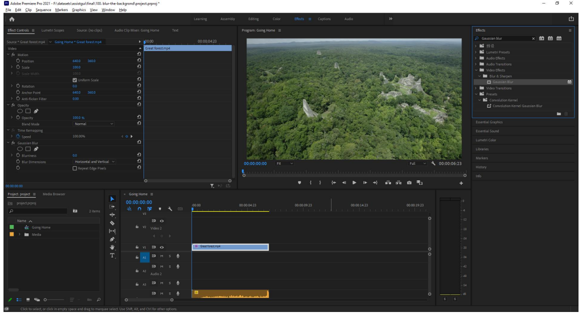
Subtask 2: Type "Gaussian blur" in the search box at the top of the effects panel

Task 2: Apply the Gaussian blur effect to the "great forrest.mp4" clip



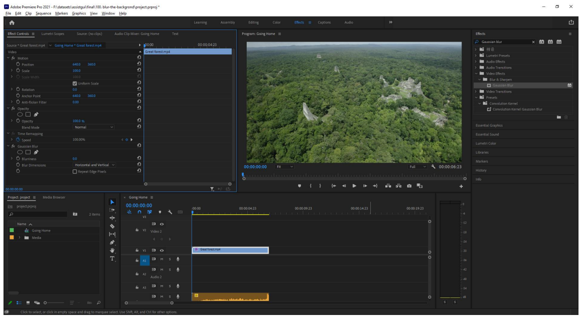
Subtask 1: Click on the "great forrest.mp4" clip in the timeline

Task 2: Apply the Gaussian blur effect to the "great forrest.mp4" clip



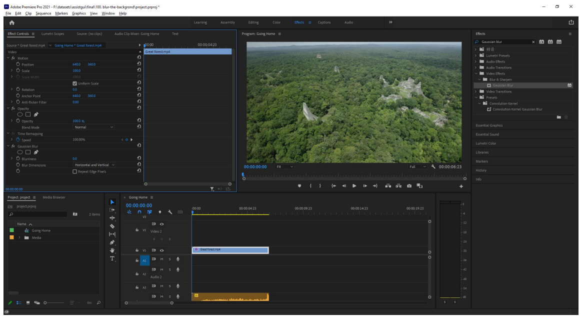
Subtask 2: Drag the Gaussian blur effect onto the "great forrest.mp4" clip

Task 3: Set the blurriness of the Gaussian blur effect



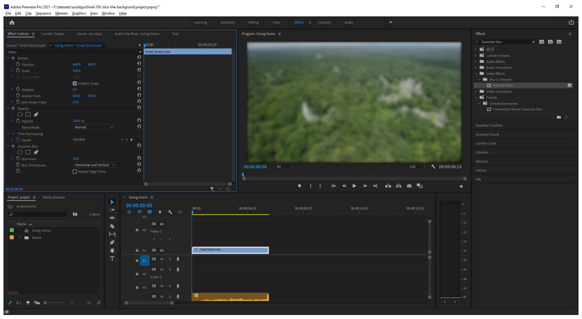
Subtask 1: Click on the effects workspace

Task 3: Set the blurriness of the Gaussian blur effect



Subtask 2: Click on the Effect Controls panel

Task 3: Set the blurriness of the Gaussian blur effect

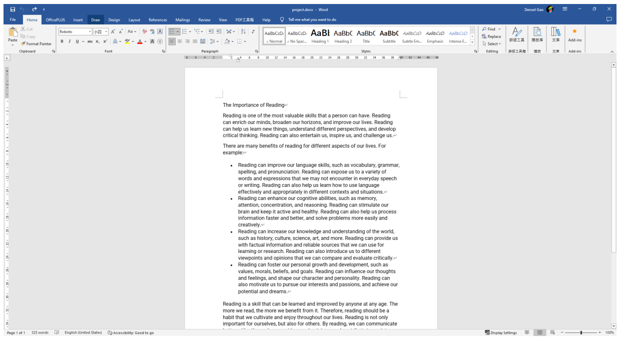


Subtask 3: Type "30" in the blurriness control of the Gaussian blur effect

Figure 7. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

Word Task **Query: Replace all the 'Reading' in the text with 'after_editing'**

Initial Screenshot



The Importance of Reading

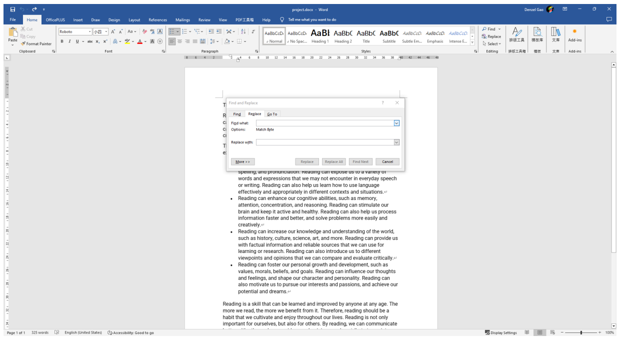
Reading is one of the most valuable skills that a person can have. Reading can enrich our minds, broaden our horizons, and improve our lives. Reading can help us learn new things, understand different perspectives, and develop critical thinking. Reading can also entertain us, inspire us, and challenge us.

There are many benefits of reading for different aspects of our lives. For example:

- Reading can improve our language skills, such as vocabulary, grammar, spelling, and pronunciation. Reading can expose us to a variety of words and expressions that we may not encounter in everyday speech or writing. Reading can also help us learn how to use language effectively and appropriately in different contexts and situations.
- Reading can enhance our cognitive abilities, such as memory, attention, concentration, and reasoning. Reading can stimulate our brain and keep it active and healthy. Reading can also help us process information faster and better, and solve problems more easily and creatively.
- Reading can increase our knowledge and understanding of the world, such as history, culture, science, art, and more. Reading can provide us with factual information and reliable sources that we can use for learning or research. Reading can also introduce us to different viewpoints and opinions that we can compare and evaluate critically.
- Reading can foster our personal growth and development, such as values, morals, beliefs, and goals. Reading can influence our thoughts and feelings, and shape our character and personality. Reading can also motivate us to pursue our interests and passions, and achieve our potential and dreams.

Reading is a skill that can be learned and improved by anyone at any age. The more we read, the more we benefit from it. Therefore, reading should be a habit that we cultivate and enjoy throughout our lives. Reading is not only important for ourselves, but also for others. By reading, we can communicate

Task 1: Use the Replace feature to replace all instances of 'Reading' with 'after_editing'



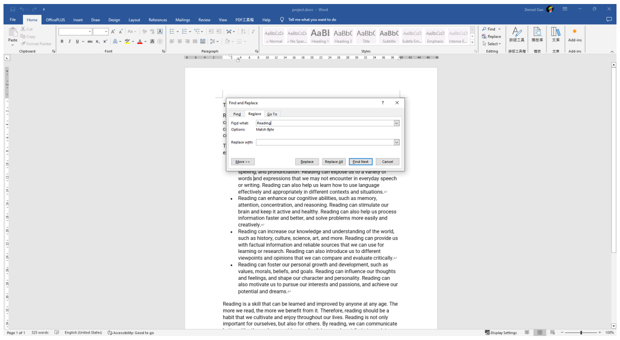
Find what:

Options: Match Byte

Replace with:

Subtask 1: Click the Replace command on the Home tab in the Editing group.

Task 1: Use the Replace feature to replace all instances of 'Reading' with 'after_editing'



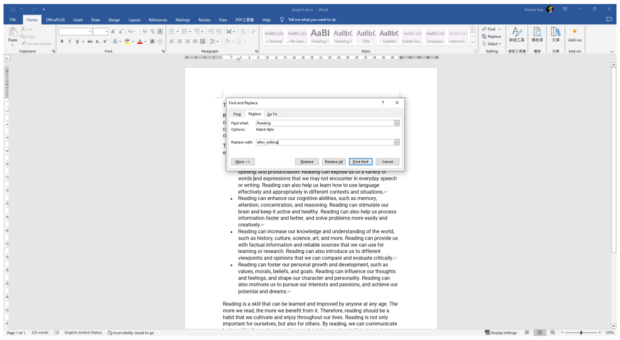
Find what: Reading

Options: Match Byte

Replace with:

Subtask 2: Type 'Reading' in the "Find what" field.

Task 1: Use the Replace feature to replace all instances of 'Reading' with 'after_editing'



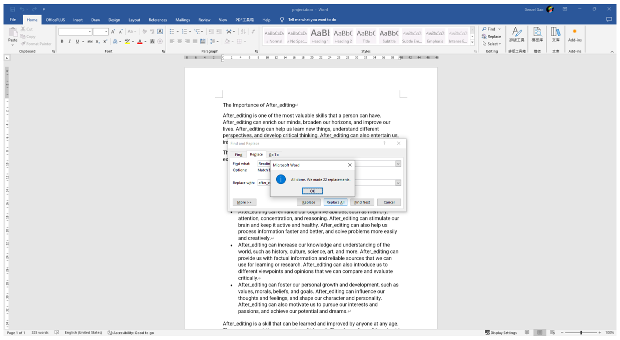
Find what: Reading

Options: Match Byte

Replace with: after_editing

Subtask 3: Type 'after_editing' in the "Replace with" field.

Task 1: Use the Replace feature to replace all instances of 'Reading' with 'after_editing'



The Importance of After_editing

After_editing is one of the most valuable skills that a person can have. After_editing can enrich our minds, broaden our horizons, and improve our lives. After_editing can help us learn new things, understand different perspectives, and develop critical thinking. After_editing can also entertain us,

All done. We made 22 replacements.

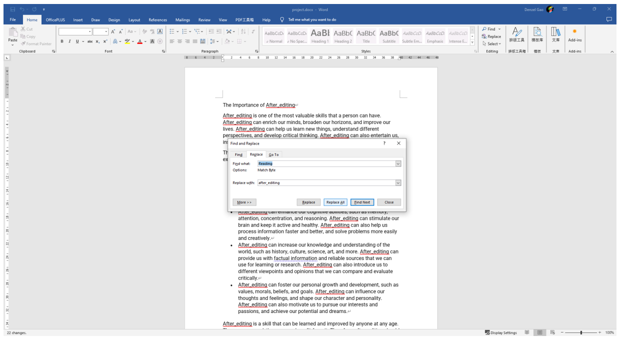
attention, concentration, and reasoning. After_editing can stimulate our brain and keep it active and healthy. After_editing can also help us process information faster and better, and solve problems more easily and creatively.

- After_editing can increase our knowledge and understanding of the world, such as history, culture, science, art, and more. After_editing can provide us with factual information and reliable sources that we can use for learning or research. After_editing can also introduce us to different viewpoints and opinions that we can compare and evaluate critically.
- After_editing can foster our personal growth and development, such as values, morals, beliefs, and goals. After_editing can influence our thoughts and feelings, and shape our character and personality. After_editing can also motivate us to pursue our interests and passions, and achieve our potential and dreams.

After_editing is a skill that can be learned and improved by anyone at any age.

Subtask 4: Click the 'Replace All' button to replace all instances of 'Reading' with 'after_editing'.

Task 1: Use the Replace feature to replace all instances of 'Reading' with 'after_editing'



Find what: Reading

Options: Match Byte

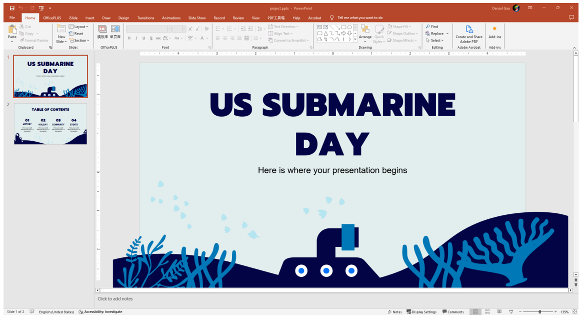
Replace with: after_editing

Subtask 4: Click the 'Replace All' button to replace all instances of 'Reading' with 'after_editing'.

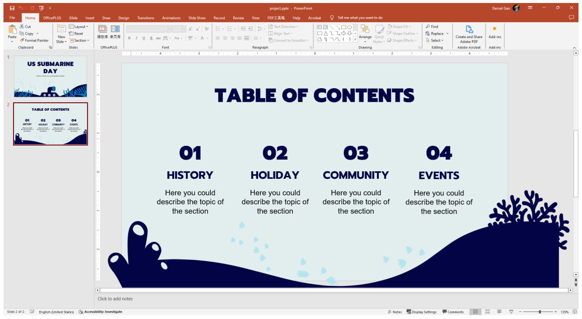
Figure 8. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

PowerPoint Task **Query: Set the transition of the second slide to Push**

Initial Screenshot

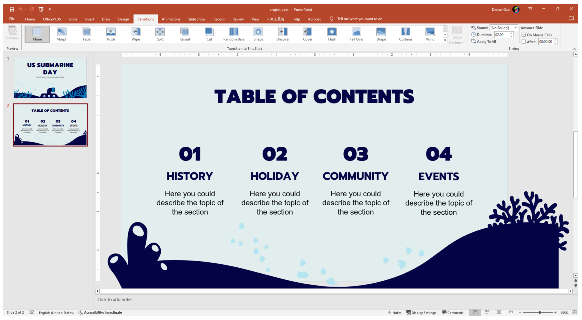


Task 1: Select the second slide for transition



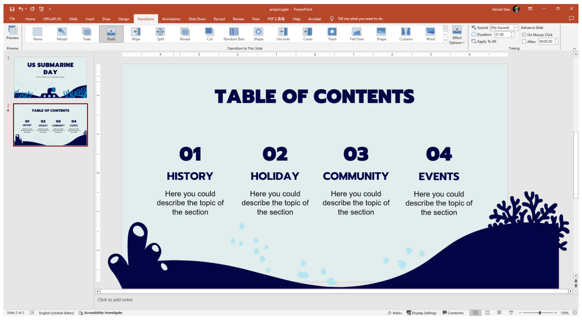
Subtask 1: Click on the second slide in the slide navigation pane

Task 2: Apply the transition effect



Subtask 1: Click on the 'Transitions' tab in the top menu bar

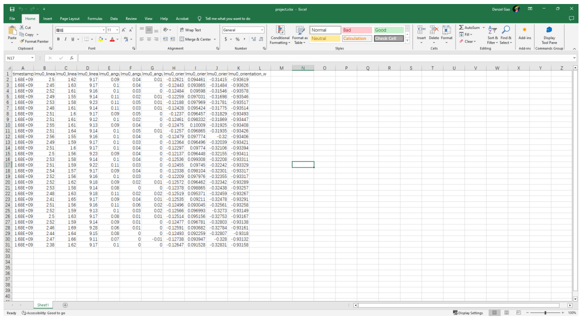
Task 2: Apply the transition effect



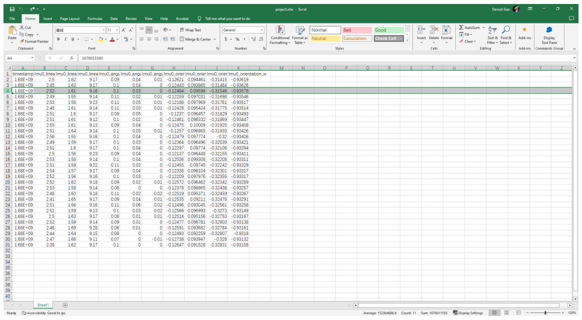
Subtask 2: Click on the 'Push' transition effect in the transitions menu

Excel Task **Query: Delete the entire 4-th row**

Initial Screenshot

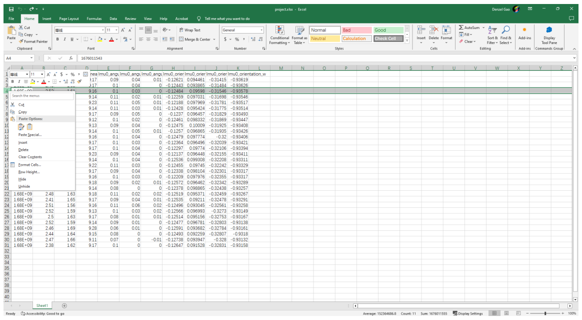


Task 1: Select the 4th row



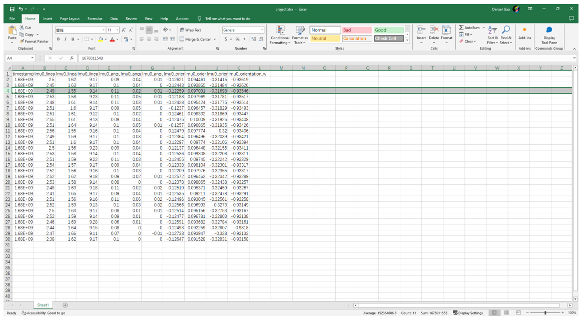
Subtask 1: Click on the row header of the 4th row to select the entire row.

Task 2: Delete the 4th row



Subtask 1: Right-click on the selected 4th row.

Task 2: Delete the 4th row

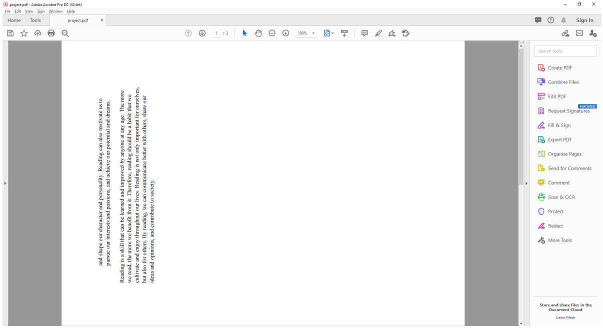


Subtask 2: Click on 'Delete' from the context menu to delete the entire row.

Figure 9. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

Adobe Acrobat Task **Query: Rotate the first page clockwise**

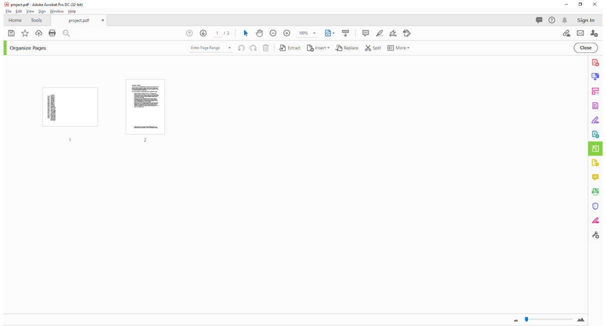
Initial Screenshot



and shape our character and personality. Reading can also motivate us to pursue our interests and passions, and achieve our potential and dreams.

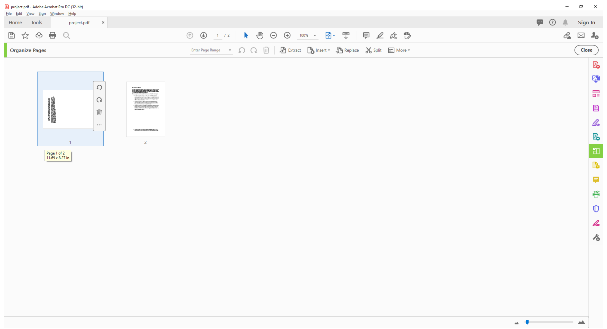
Reading is a skill that can be learned and improved by anyone at any age. The more we read, the more we benefit from it. Therefore, reading should be a habit that we cultivate and enjoy throughout our lives. Reading is not only important for ourselves, but also for others. By reading, we can communicate better with others, share our ideas and opinions, and contribute to society.

Task 1: Access the Organised Pages option



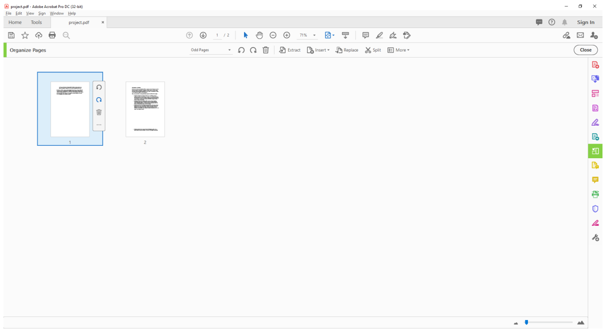
Subtask 1: Click on the Organised Pages option on the right hand side

Task 2: Rotate the first PDF page clockwise



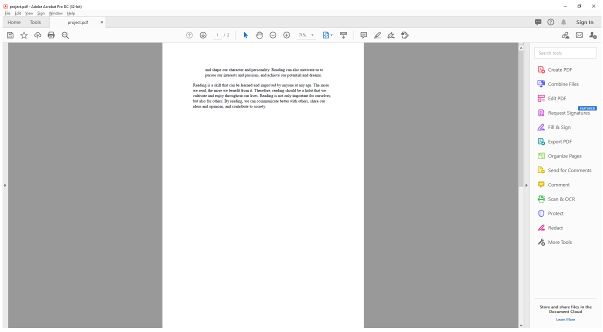
Subtask 1: Hover your cursor over the thumbnail of the first page

Task 2: Rotate the first PDF page clockwise



Subtask 2: Click on the clockwise arrow icon next to the first page

Task 3: Return to normal view



and shape our character and personality. Reading can also motivate us to pursue our interests and passions, and achieve our potential and dreams.

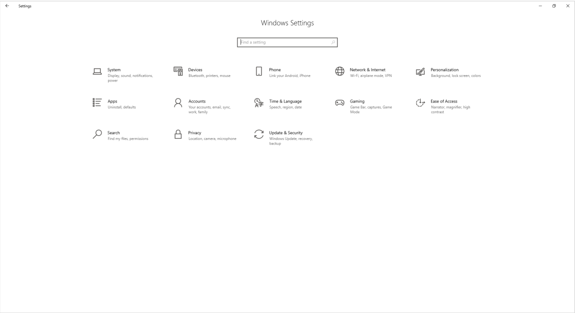
Reading is a skill that can be learned and improved by anyone at any age. The more we read, the more we benefit from it. Therefore, reading should be a habit that we cultivate and enjoy throughout our lives. Reading is not only important for ourselves, but also for others. By reading, we can communicate better with others, share our ideas and opinions, and contribute to society.

Subtask 1: Click on close in your Organised Pages taskbar

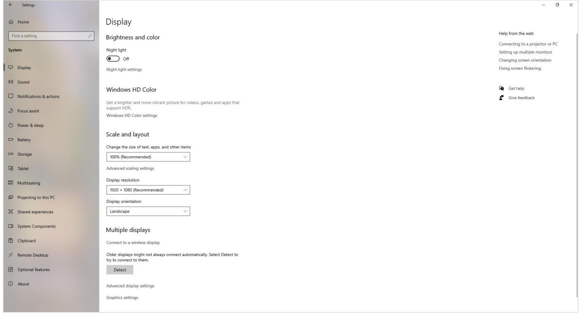
Figure 10. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

System Setting Task **Query: Disable all notifications**

Initial Screenshot

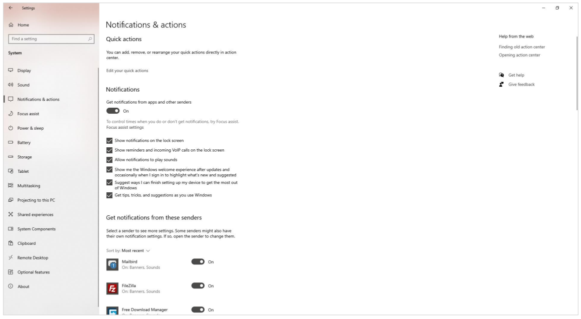


Task 1: Navigate to Notifications



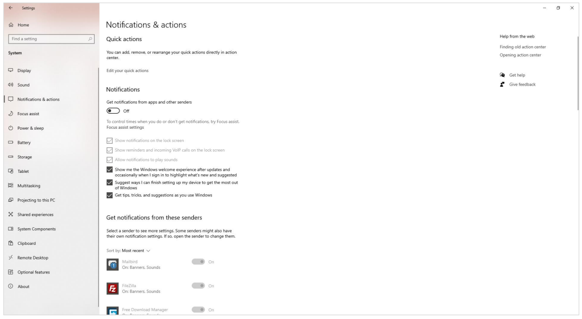
Subtask 1: Click on the 'System' option in the settings window

Task 1: Navigate to Notifications



Subtask 2: Click on the 'Notifications' option

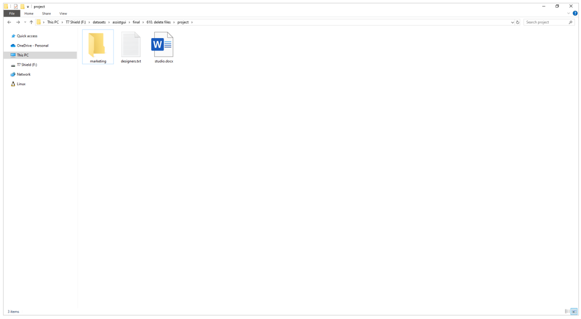
Task 2: Disable All Notifications



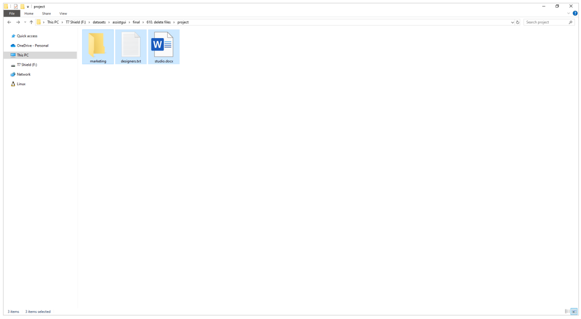
Subtask 1: Click on the toggle at the top of the notifications window to disable all notifications

File Manipulation Task **Query: Delete all files under the current folder**

Initial Screenshot

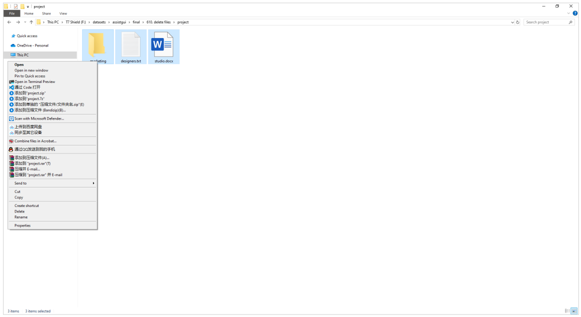


Task 1: Delete All Files in the Current Folder



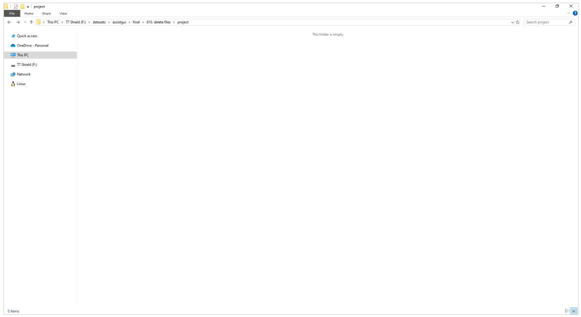
Subtask 1: Press 'Ctrl' + 'A' to select all files within the folder

Task 1: Delete All Files in the Current Folder



Subtask 2: Right-click on any of the highlighted files

Task 1: Delete All Files in the Current Folder

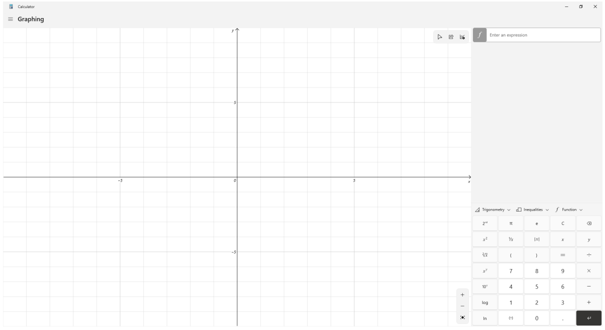


Subtask 3: Choose 'Delete' from the context menu to remove all selected files from the folder

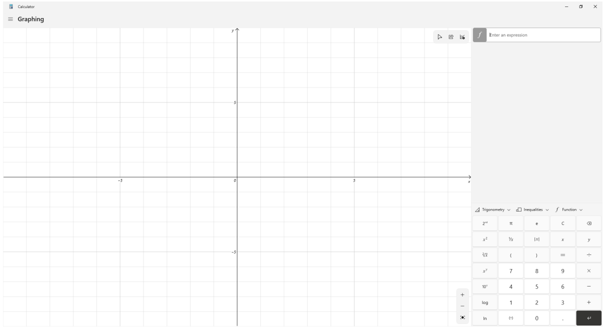
Figure 11. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.

Calculator Task **Query: Use the calculator to draw a log(x) in graphing mode**

Initial Screenshot

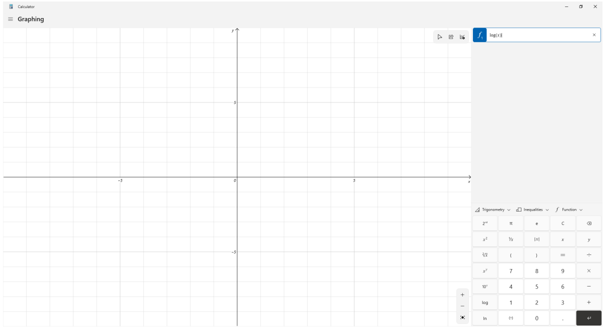


Task1: Enter equations and functions



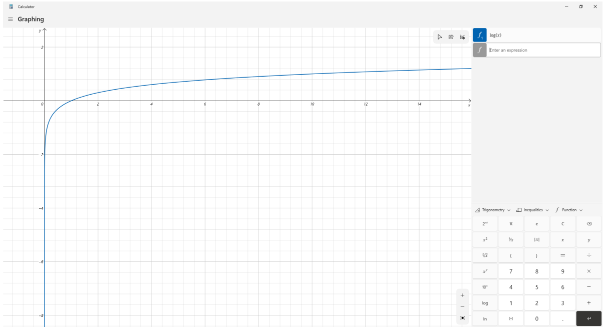
Subtask 1: Click on the input field located on the right side of the screen

Task1: Enter equations and functions



Subtask 2: Type 'log(x)' into the input field to enter the logarithmic function

Task 2: Visualize the graph



Subtask 1: Press 'Enter' to observe the function graphed visually on the graph interface to the left

Figure 12. **Visualized Action Trajectory.** We present the query, planned results, and corresponding screenshots to illustrate our model's trajectory in solving the example tasks.