

Appendix

This is the Appendix for the paper “Boosting Virtual Agent Learning and Reasoning: A Step-wise, Multi-dimensional, and Generalist Reward Model with Benchmark”.

Overview

In this supplementary material we present:

- Detailed prompts for testing general MLLMs on *SRMEval* and for building *SRM* are provided in Section A.
- Pseudocode for the **Similar** pipeline, including key algorithmic steps, is described in Section B.
- Human evaluation details and human acceptance of 5-dimension data from *SRM* are illustrated in Section C.
- Case studies demonstrating the applications of **Similar** are presented in Section D.
- Comprehensive ablation experiments are conducted and analyzed in Section E.
- Additional visualizations of *SRMEval*, showcasing task performance, are provided in Section F.

A. Detailed Prompt Design

This section provides a comprehensive overview of the prompt designs used in our experiments. We detail the prompts for testing general MLLMs on *SRMEval* and the prompts for building the *SRM* model, highlighting their structure and purpose.

A.1. Prompt for testing general MLLMs on *SRMEval*

This subsection describes the prompts used to evaluate general MLLMs on the *SRMEval* benchmark.

Prompt for *SRMEval* (main part)

You are an expert in evaluating the performance of a Virtual Agent.

The Virtual Agent is designed to help a human user complete specified tasks (such as app usage, web navigation, web content Q&A, etc.) on various platform applications (such as websites, mobile devices, operation systems, etc.) based on given instructions. Given the user’s **INSTRUCTION**, the **OBSERVATION** of current platforms, the action **TRAJECTORY** of the

agent, the two **ACTION_X** and **ACTION_Y** predicted by the agent, and the current action step number **STEP_IDX**. Your **GOAL** is to help me complete step-wise evaluation, that is, evaluate the quality of the Agent’s **ACTION** in a specific dimension. Choose the better action (**ACTION_X** or **ACTION_Y**) based on the given (**EVALUATION DIMENSION**). Output ‘Y’ and the reason if **ACTION_X** is better, or ‘X’ and the reason if **ACTION_Y** is better. Do not output responses like ‘two actions are similar’.

Word Meaning

1. **INSTRUCTION**: refers to the command of human users to the Agent, which is the specific content that the Agent needs to complete the task on a specific platform, that is, the ultimate **GOAL** of the Agent.
2. **OBSERVATION**: refers to the specific information of the current platform that an agent can observe on the platform where the task needs to be completed, which is the environment in which the agent is currently located. In our task, observations are presented in the form of images, known as screenshots.
3. **TRAJECTORY**: refers to the action prediction made by an agent in the past to complete the **INSTRUCTION**, which records all actions taken by the agent from the first step to the current step. If this is the first step, then the trajectory is empty.
4. **ACTION**: refers to the predicted operation of the Agent in the current state to complete the **INSTRUCTION** in the current step. This operation generally refers to a simple action command, such as ‘CLICK’, ‘TYPE’, etc. Note that **ACTION** is the result predicted by the agent after observing the current **OBSERVATION**, and the Agent often cannot complete the task in one step.
5. **STEP_IDX**: refers to the sequence number of the Agent executing the current **ACTION** to complete the **INSTRUCTION**.

Here is the evaluation dimension part of the prompts.

Helpfulness.

Prompt for SRMEval (Helpfulness)

1. [HELPLESSNESS]

1.1 Meaning: It indicates the degree to which this step contributes to the completion of the final task. There are good and bad contributions, the correct steps will give a positive contribution, and the wrong steps will give a negative contribution.

1.2 Design motivation: Different steps contribute differently to the completion of the final task, with good steps helping to accomplish the task and bad steps hindering it. Good steps should be rewarded positively, while bad steps should be punished negatively. If each step is correct and the total number of steps is 5, then the contribution of each step can be considered as $1/5$, meaning that each step completes $1/5$ of the final task. If 4 more steps are needed from the current step and the current step is incorrect, then the contribution of the current step is $-1/4$, indicating that it hinders $1/4$ of the final task progress.

Odds of Success.

Prompt for SRMEval (Odds of Success)

2. [ODDS OF SUCCESS]

2.1 Meaning: It indicates the potential of the step to complete the task, which is the probability of a step reaching the completion of the task. 2.2 Design

motivation: The more correct steps lead to a higher probability of success in the final task, and the more incorrect steps lead to a higher probability of failure in the final task. Different steps have different potential to complete the task. If one step of the agent is to follow the Instructions to complete the task, then this step generally has high potential. We can derive the probability of a step leading to success from the N paths generated by that step, which serves as the potential for that step to complete the task which is crucial for evaluating.

Efficiency.

Prompt for SRMEval (Efficiency)

3. [EFFICIENCY]

3.1 Meaning: It indicates whether this step is efficient in completing the task. We calculate this metric as the difference between 'the number of steps required to complete the final task after the current step' and 'the number of steps required to complete the final task after the previous step', divided by 'the total number of steps required to complete the task'. This indicates the degree of efficiency improvement in completing tasks after the current step is executed.

3.2 Design motivation: A basic assumption is that the fewer steps the Agent operates, the more efficient it is, because the consumption of these paths (time consumption, hardware consumption) can be considered to be the least and the efficiency is the highest. Therefore, if the operation of a step can reduce the number of steps required to complete the task as a whole, then it can be considered that the operation of this step is very efficient. For example, after the previous step, it takes 7 steps to complete the task, but after the current step, it only takes 4 steps to complete the task. The difference of $7-4=3$ is the efficiency improvement of the current step in completing the final task.

Task Relevance.

Prompt for SRMEval (Task Relevance)

4. [TASK RELEVANCE]

4.1 Meaning: It indicates whether the operation of the Agent is related to achieving the **INSTRUCTION**.

4.2 Design motivation: Some operational steps may prevent the task from being completed, but they are related to the task (for example, we need to ask the agent to take notes, and the agent takes notes, which is related to the task, but the recorded note content is incorrect, indicating that this is an incorrect step). Some operational steps may be

meaningless, but they can still lead to task completion (such as clicking on a blank screen without generating any response, which is unrelated to the task, but the agent’s subsequent actions can still result in task success). Therefore, an indicator is needed to identify whether the current step of operation is related to the task.

4.3 Range of values after mapping: {0, 1}. The larger the value, the greater the correlation between the step and the task.

Coherence.

Prompt for SRMEval (Coherence)

5. [COHERENCE]

5.1 Meaning: It represents the compactness and coherence between the current step and the previous step.

5.2 Design motivation: Some operations, although task-related, not inefficient, and highly likely to lead to success, lack coherence with the previous step. For example, the task is to ‘‘query the Lakers’ game results and record them in the Note’’. The Agent operations are as follows: a Open the browser; b. Open Note; c. Create new notes; d. Search for Lakers games; e. Query the results of the competition; f. Record the results of the competition in your notes. It can be found that the operations of a and b lack coherence, and it is more in line with human preferences to directly search for competition results after opening the browser instead of simultaneously opening Note.

5.3 Range of values after mapping: {0, 1}. The larger the value, the greater the coherence of the step.

Total dimension and Trajectory-level dimension.

Prompt for SRMEval (Total and Trajectory-level)

6. [TOTAL]

Meaning: Integrated decision-making based on the 5 dimensions mentioned earlier.

7. [TRAJECTORY]

Meaning: Represents the quality of the entire trajectory, which can be

expressed as the average total score of all steps in the trajectory.

A.2. Prompt for building SRM

This subsection outlines the prompts designed for constructing the *SRM* model.

Prompt for building SRM

You are a virtual agent.

The Virtual Agent is designed to help a human user complete specified tasks (such as app usage, web navigation, web content Q&A, etc.) on various platform applications (such as websites, mobile devices, operation systems, etc.) based on given instructions.

You will predict the next action based on the following content [INSTRUCTION], [OBSERVATION], [REASON_STEPS]:

1. [INSTRUCTION]: It is your ultimate **GOAL**, and all your actions are aimed at completing this task.
2. [OBSERVATION]: It is an observation of an image, which is the screenshot of the platform (such as a computer screen).
3. [REASON_STEPS]: They are the trajectory of the actions you performed in the past to complete the instruction, from which you can understand how you thought in order to complete the instruction. If it is empty, it means it is currently the first step.

B. Similar Pipeline Pseudocode

The rapid advancement of MLLMs (Li et al., 2020; 2022; Yu et al., 2023; 2024; Pan et al., 2024; Wu et al., 2024; Fei et al., 2024; Miao et al., 2024), which excel at integrating text, vision, and other modalities, has enabled the development of GVAs (Gao et al., 2024; Zhang et al., 2024; Shen et al., 2023), which also inspires us to study Reward Models for GVAs with its Benchmark.

In this section, we present the pipeline pseudocode for training our proposed **Similar** model. The pipeline consists of three main components: 1) a five-dimensional process supervision framework to evaluate agent steps, 2) an automatic generalist dataset collecting process, and 3) a Triple-M strategy for reward model training.

B.1. Five-Dimensional Process Supervision Framework

The five-dimensional process supervision framework systematically evaluates the quality of an agent’s steps using

Algorithm 1 Five-Dimensional Process Supervision

```
1: Input: Step  $S_i$ , ground truth  $a^*$ , number of simulations  $N$ 
2: Output: Five-dimensional scores ( $H_i, OS_i, E_i, TR_i, C_i$ )
3: Compute Helpfulness (H):
4:  $H_i = \frac{1-AC_{i-1}}{M-i+1} (1-2r_i)$ 
5:  $AC_i = \max(AC_{i-1} + H_i, 0)$ 
6: Compute Odds of Success (OS):
7:  $OS_i = \frac{\sum_{j=1}^N \mathbb{I}(a_{i,j}=a^*)}{N}$ 
8: Compute Efficiency (E):
9:  $E_i = \frac{Len_{i-1} - Len_i}{Len_0}$ 
10:  $Len_i = \text{avg}(Len(S_{i,j}))$ 
11: Compute Task Relevance (TR) and Coherence (C):
12:  $TR_i = \text{MLLM.Evalute}(S_i, \text{instruction})$ 
13:  $C_i = \text{MLLM.Evalute}(S_i, S_{i-1})$ 
14: return ( $H_i, OS_i, E_i, TR_i, C_i$ )
```

Algorithm 2 Automatic Generalist Dataset Collecting

```
1: Input: Task instruction  $q$ , platforms  $\mathcal{P} = \{\text{Web, Android, Linux, Windows}\}$ 
2: Output: Annotated dataset  $\mathcal{D}$ 
3: Initialize empty dataset  $\mathcal{D}$ 
4: for each platform  $p \in \mathcal{P}$  do
5:   Initialize MCTS-P tree  $T_q$  for task  $q$  on platform  $p$ 
6:   for each node  $S_{i,j}$  in  $T_q$  do
7:     Calculate minimum steps  $M$  to reach a correct answer
8:     Simulate  $N$  trajectories to compute basic reward  $r_i$ 
9:     Compute  $H_i, OS_i, E_i$  using formulas from Algorithm 1
10:    Evaluate  $TR_i$  and  $C_i$  using MLLM (e.g., GPT-4)
11:    if node  $S_{i,j}$  leads to a complete trajectory then
12:      Verify correctness using platform-specific evaluation methods
13:      Add annotated step ( $S_{i,j}, H_i, OS_i, E_i, TR_i, C_i$ ) to  $\mathcal{D}$ 
14:    end if
15:  end for
16: end for
17: Prune incomplete branches from  $T_q$ 
18: return Annotated dataset  $\mathcal{D}$ 
```

five distinct dimensions: *Helpfulness (H)*, *Odds of Success (OS)*, *Efficiency (E)*, *Task Relevance (TR)*, and *Coherence (C)*. The following pseudocode outlines the computation of these dimensions for a given step S_i , providing a comprehensive assessment of step quality.

B.2. Automatic Generalist Dataset Collecting

The automatic generalist dataset collecting process leverages the MCTS-P algorithm to collect annotated step-wise data across multiple platforms, including Web, Android, Linux, and Windows.

B.3. Triple-M Strategy for Reward Model Training

The Triple-M strategy integrates multi-step, multi-dimensional, and multi-modal data for reward model training, ensuring high flexibility, robustness, and adaptability. The following pseudocode outlines the two-stage training process, which includes regression layer optimization and dynamic gating network adjustment.

C. Human Evaluation Details and Human Acceptance of SRM

To ensure high-quality annotations, we collaborated with a professional commercial data labeling team. The process included: **1) Training Phase:** Annotators underwent three

Algorithm 3 Triple-M Strategy for Reward Model Training

```
1: Input: Training dataset  $\mathcal{D}$ , pre-trained MLLM  $f_\theta$ , gating network  $g_\phi$ 
2: Output: Trained reward model  $R$ 
3: Stage 1: Regression Layer Training
4: for each batch  $(x, y, r) \in \mathcal{D}$  do
5:   Extract hidden state  $h = f_\theta(x \oplus y)$ 
6:   Compute predicted scores  $\hat{r} = W^\top h$ 
7:   Update  $\theta, W$  using  $L_{RG} = \|\hat{r} - r\|_2^2$ 
8: end for
9: Stage 2: Gating Network Training
10: for each batch  $(x, y_{\text{chosen}}, y_{\text{rejected}}) \in \mathcal{D}$  do
11:   Compute coefficients  $w = g_\phi(f_\theta(x))$ 
12:   Compute preference scores  $R_{\text{chosen}} = w^\top r_{\text{chosen}}$ 
13:   Compute preference scores  $R_{\text{rejected}} = w^\top r_{\text{rejected}}$ 
14:   Update  $\phi$  using  $L_{BT} = -\log \frac{\exp(R_{\text{chosen}})}{\exp(R_{\text{chosen}}) + \exp(R_{\text{rejected}})}$ 
15: end for
16: return Trained reward model  $R = g_\phi(f_\theta(x))^\top r$ 
```

rounds of iterative “label-review-feedback” cycles to clarify ambiguities of annotation (e.g., the complexity of UI interaction tasks). Only after achieving $> 95\%$ accuracy on validation samples did formal annotation begin. **2) Formal Annotation:** Each test sample in *SRMEval* was independently labeled by three annotators and three checkers. The final data in test set required $> 99\%$ accuracy.

To validate the quality of our five-dimensional assessment data and ensure alignment with human preferences, we randomly sampled a batch of data from the *SRM* Benchmark. The sample size varied across dimensions due to differences in score distributions. This stems from their fundamental design - for instance, *Task Relevance* and *Coherence* are binary values, which naturally yield fewer possible preference pairs. Human annotators were then asked to select the better action from candidate action pairs in the sampled data, based on specific evaluation types. If the annotator considers a sample correct, mark it as 1; otherwise, mark it as 0, and calculate Accuracy as Human Acceptance.

The results, as shown in Table 1, demonstrate that the human acceptance rate for all five dimensions exceeds 78.8%, strongly indicating the superiority of our designed annotation dimensions and the high quality of the collected data.

The evaluation process was further enhanced by incorporating a rigorous double-blind annotation protocol, where neither the annotators nor the analysts were aware of the origin or automated scores of the candidate actions.

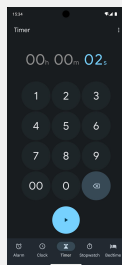
Table 1: Sample size and human acceptance rate for each dimension in *SRM*.

Dimension	Sample Size	Human Acceptance
Helpfulness	6000	87.9%
Odds of Success	2000	78.8%
Efficiency	6000	82.6%
Task Relevance	1000	84.7%
Coherence	2000	93.5%

D.1. Helpfulness

► **Input:**

[OBS]



[/OBS]

[TRAJ]
Step 1: Click Search.

Step 2: Click Clock.

Step 3: Click 2.

[/TRAJ]

► Output :

[ACTION_X]
Click Backspace 2.

[/ACTION_X]

[SCORE_X]

H: 0.72

[/SCORE X]

[ACTION_Y]

Click 1.

[/ACTION_Y]

[SCORE_Y]

H: -0.37

[/SCORE_Y]

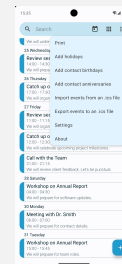
D.2. Odds of Success

Case of Odds of Success

► **Input:**

```
[INST]
In Simple Calendar Pro, delete all the
calendar events on 2023-10-27.
[/INST]
```

[OBS]



[/OBS]

[TRAJ]

Step 1: Click Search.

Step 2: Click Calendar.

Step 3: Scroll down.

Step 4: Long press 27 Friday.

Step 5: Click More options.

[/TRAJ]

► Output :

```
[ACTION_X]
Navigate back.
```

[/ACTION_X]

[SCORE X]

OS: 0.75

[/SCORE_X]

[ACTION_Y]

Click Import events from a .ics file.

[/ACTION_Y]

[SCORE_Y]

OS: 0.13

[/SCORE Y]

5

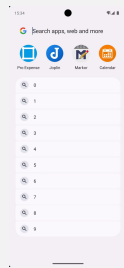
D.3. Efficiency

Case of Efficiency

► Input:

```
[INST]
In Simple Calendar Pro, delete all the
events.
[/INST]
```

[OBS]



[/OBS]

```
[TRAJ]
Step 1: Click Search.
[/TRAJ]
```

► Output:

```
[ACTION_X]
Click Calender.
[/ACTION_X]
[SCORE_X]
E: 0.88
[/SCORE_X]

[ACTION_Y]
Input text ``Simple Calendar Pro``.
[/ACTION_Y]
[SCORE_Y]
E: 0.65
[/SCORE_Y]
```

In this case, since the search interface displays past search history, including “Calendar”, ACTION_X (Click “Calendar”) is more efficient in the *Efficiency* dimension. However, ACTION_Y is also a correct approach, so its *Efficiency* score remains relatively high.

D.4. Task Relevance

Case of Task Relevance

► Input:

```
[INST]
Add these recipes to the Broccoli app:

1. Chicken Alfredo Pasta
   - Description: A healthy, delicious
```

```
meal.
  - Servings: 2
  - Prep Time: 10 mins
  - Ingredients: As desired
  - Directions: Cook pasta, toss with
Alfredo sauce and grilled chicken. Top
with Parmesan and spices.
```

```
2. Quinoa Salad with Vegetables
  - Description: Quick and easy for
busy days.
  - Servings: 8
  - Prep Time: 30 mins
  - Ingredients: To your liking
  - Directions: Mix quinoa, diced
veggies, feta, and lemon olive oil
dressing. Add spices for flavor.
```

```
3. Butternut Squash Soup
  - Description: A healthy, delicious
choice.
  - Servings: 1
  - Prep Time: 45 mins
  - Ingredients: Per taste
  - Directions: Saute onions and
garlic, add squash and broth. Puree and
season with nutmeg, salt, and pepper.
Substitute as needed.
[/INST]
```

[OBS]



[/OBS]

```
[TRAJ]
Step 1: Click Search.

Step 2: Input text Broccoli.

Step 3: Click Broccoli.

Step 4: Click New Recipe.

Step 5: Input text Chicken Alfredo
Pasta.

Step 6: Input text A delicious and
healthy choice for any time of the day.

Step 7: Input text 2 servings.

Step 8: Input text 10 mins.
```

Step 9: Input text as desired.

Step 10: Scroll down.

Step 11: Input text Cook fettuccine pasta, toss with Alfredo sauce and grilled chicken strips. Serve with a sprinkle of Parmesan cheese. Try adding a pinch of your favorite spices for extra flavor.

Step 12: Click SAVE.
[/TRAJ]

► **Output:**

```
[ACTION_X]
Click Cook.
[/ACTION_X]
[SCORE_X]
C: 1
[/SCORE_X]

[ACTION_Y]
Navigate back.
[/ACTION_Y]
[SCORE_Y]
C: 0
[/SCORE_Y]
```

In this case, the task is to create a recipe. The first 12 steps have completed the recipe creation process. In the Broccoli app, the next action should be to directly click “Cook” (i.e., ACTION_X), which is highly relevant to the instruction. In contrast, “Navigate back” (i.e., ACTION_Y) is not directly related to the instruction. Therefore, ACTION_X receives a higher *Task Relevance* score.

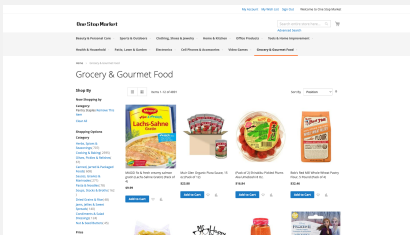
D.5. Coherence

Case of Coherence

► **Input:**

```
[INST]
Add this exact product to my shopping
cart. I think it is in the "Herbs,
Spices \& Seasonings" category.
[/INST]
```

[OBS]



[/OBS]

[TRAJ]

Step 1: Click menuitem '\ue622 Grocery & Gourmet Food' hasPopup: menu.

Step 2: Click link 'Pantry Staples (4891 item)'.
[/TRAJ]

► **Output:**

```
[ACTION_X]
Click link 'Herbs, Spices \&
Seasonings ( 707 item )'.
[/ACTION_X]
[SCORE_X]
C: 1
[/SCORE_X]

[ACTION_Y]
Click menuitem '\ue622 Grocery &
Gourmet Food' hasPopup: menu.
[/ACTION_Y]
[SCORE_Y]
C: 0
[/SCORE_Y]
```

In this case, the task is to find relevant products in the “Herbs, Spices & Seasonings” category. The previous step involved the agent entering a link, and the current step should logically advance the task. Clearly, ACTION_X (clicking the “Herbs, Spices & Seasonings” link) is a coherent and logical continuation, while ACTION_Y (repeating the action from step 1) is unintelligible and incoherent. Therefore, ACTION_X receives a higher *Coherence* score.

E. Comprehensive Ablation Experiments

The extended ablation study, presented in Table 2, provides a more comprehensive analysis of the impact of each dimension in the **Similar** model. The results confirm the trends observed in the main experiments and offer additional insights into the contributions of the five dimensions—*Helpfulness (H)*, *Odds of Success (OS)*, *Efficiency (E)*, *Task Relevance (TR)*, and *Coherence (C)*—across three benchmarks: Android World, WebArena, and OSWorld.

E.1. Impact of Individual Dimensions

The results demonstrate that the *Helpfulness (H)* dimension has the most significant impact on performance, consistent with the findings in the main experiments. For example, adding *H* alone improves the success rate on Android World from 30.7% to 32.5%, on WebArena from 20.6% to 26.1%, and on OSWorld from 14.6% to 15.8%. This aligns with our hypothesis that the quality of a step is primarily reflected in its contribution to task completion, which *H* effectively

Table 2: Ablation study (inference experiments). **Similar** in table represents **Similar**-TM-Llama.

MODEL	DIMENSION					SUCCESS RATE		
	H	OS	E	TR	C	ANDROID WORLD	WEBARENA	OSWORLD
BACKBONE						30.4	20.6	14.3
+H	✓					32.5	26.1	15.8
+OS		✓				31.9	24.7	15.4
+E			✓			31.6	23.3	15.2
+TR				✓		31.1	21.6	14.9
+C					✓	30.9	21.0	14.8
+H,OS	✓	✓				33.4	31.4	16.7
+H,E	✓		✓			33.1	29.8	16.5
+H,TR	✓			✓		32.8	28.5	16.3
+H,C	✓				✓	32.6	28.2	16.2
+OS,E		✓	✓			32.7	27.5	16.3
+OS,TR		✓		✓		32.3	26.8	16.0
+OS,C		✓			✓	32.1	26.5	15.9
+E,TR			✓	✓		31.8	25.9	15.7
+E,C			✓		✓	31.7	25.7	15.6
+TR,C				✓	✓	31.5	22.5	15.1
+H,OS,E	✓	✓	✓			34.3	35.9	17.2
+H,OS,TR	✓	✓		✓		34.0	34.5	17.0
+H,OS,C	✓	✓			✓	33.8	34.2	16.9
+H,E,TR	✓		✓	✓		33.5	33.8	16.7
+H,E,C	✓		✓		✓	33.4	33.6	16.6
+H,TR,C	✓			✓	✓	33.2	33.1	16.5
+OS,E,TR		✓	✓	✓		32.9	32.8	16.4
+OS,E,C		✓	✓		✓	32.8	32.7	16.3
+OS,TR,C		✓		✓	✓	32.6	32.5	16.2
+E,TR,C			✓	✓	✓	32.4	32.3	16.1
+H,OS,E,TR	✓	✓	✓	✓		35.1	37.7	17.8
+H,OS,E,C	✓	✓	✓		✓	34.7	37.2	17.6
+H,OS,TR,C	✓	✓		✓	✓	34.2	36.5	17.3
+H,E,TR,C	✓		✓	✓	✓	33.9	35.7	17.1
+OS,E,TR,C		✓	✓	✓	✓	33.1	33.9	16.9
SIMILAR	✓	✓	✓	✓	✓	35.4	38.2	17.8

captures. The *Odds of Success* (OS) dimension follows, with improvements of 31.9%, 24.7%, and 15.4% on the respective benchmarks, indicating its importance in guiding the agent. The *Efficiency* (E) dimension also contributes positively, though to a lesser extent, while *Task Relevance* (TR) and *Coherence* (C) show more modest improvements, consistent with their secondary roles.

E.2. Combined Impact of Multiple Dimensions



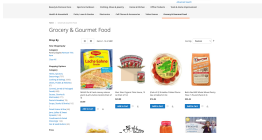
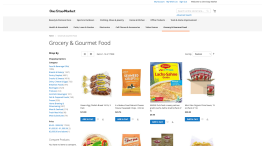
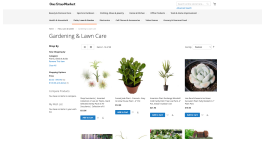
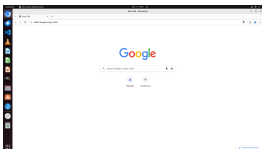
The extended results further highlight the synergistic effects of combining multiple dimensions. For instance, the combination of *H* and *OS* achieves success rates of 33.4%, 31.4%, and 16.7% on Android World, WebArena, and OSWorld, respectively, outperforming models with only one of these dimensions. Similarly, the combination of *H*, *OS*, and *E*

yields even higher success rates (34.3%, 35.9%, and 17.2%), demonstrating the cumulative benefits of integrating complementary dimensions. These results reinforce the importance of fine-grained rewards over coarse-grained ones, as models with partial-dimensional rewards consistently underperform compared to the full **Similar** model.

F. More Visualizations of *SRMEval*

As depicted in Table 3, we present additional visualizations of *SRMEval*. From the data in the table, we can more clearly understand the content of *SRMEval*, which is the first benchmark in the virtual agent domain designed for step-wise, multi-dimensional, and multi-platform evaluation of reward models. It comprehensively tests the ability of reward models to assess the quality of agent actions, as well as the degree of preference alignment.

Table 3: Cases of *SRMEval*.

Instruction	Observation	Step Idx	Trajectory	Type	Candidate Action Pair	
In Simple Calendar Pro, delete all the events.		2	Step 1: Click Search.	E	Click Calender.	Input text "Simple Calendar Pro".
In Simple Calendar Pro, delete all the calendar events on 2023-10-27.		6	Step 1: Click Search. Step 2: Click Calendar. Step 3: Scroll down. Step 4: Long press 27 Friday. Step 5: Click More options.	OS	Navigate back.	Click Import events from an.ics file.
Add this exact product to my shopping cart. I think it is in the "Herbs, Spices & Seasonings" category.		3	Step 1: Click menuItem "Grocery & Gourmet Food" hasPopup: menu. Step 2: Click link "Pantry Staples(4891 item)".	C	Click link "Herbs, Spices & Seasonings(707 item)".	Click menuItem "Grocery & Gourmet Food" hasPopup: menu.
Add a exact product to my shopping cart.		2	Step 1: Click menuItem "Grocery & Gourmet Food" hasPopup: menu.	H	Scroll down.	Click menuItem "Grocery & Gourmet Food" hasPopup: menu.
Can you add the red flower seeds with around 4 stars to my cart?		4	Step 1: Scroll down. Step 2: Click link "Page 2". Step 3: Click link "Plants, Seeds & Bulbs(59 item)" hasPopup: menu.	TR	Click link "Page Next".	Hover menuItem "Patio, Lawn & Garden" hasPopup: menu.
Can you make Bing the main search thingy when I look stuff up on the internet?		3	Step 1: Coordinates for "Customize Chromium". Step 2: Click on "Customize Chromium".	Total	Click on the address and search bar Type the URL to access search engine settings.	Click on the "Account Settings" button coordinates for "Account Settings".

References

- Fei, H., Wu, S., Zhang, H., Chua, T.-S., and Yan, S. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.
- Gao, M., Bu, W., Miao, B., Wu, Y., Li, Y., Li, J., Tang, S., Wu, Q., Zhuang, Y., and Wang, M. Generalist virtual agents: A survey on autonomous agents across digital platforms, 2024. URL <https://arxiv.org/abs/2411.10943>.
- Li, J., Wang, X., Tang, S., Shi, H., Wu, F., Zhuang, Y., and Wang, W. Y. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12123–12132, 2020.
- Li, J., He, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., and Tang, S. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022.
- Miao, B., Zhang, W., Li, J., Tang, S., Li, Z., Shi, H., Xiao, J., and Zhuang, Y. Radar: Robust two-stage modality-incomplete industrial anomaly detection, 2024. URL <https://arxiv.org/abs/2410.01737>.
- Pan, K., Fan, Z., Li, J., Yu, Q., Fei, H., Tang, S., Hong, R., Zhang, H., and Sun, Q. Towards unified multimodal editing with enhanced knowledge collaboration. *Advances in Neural Information Processing Systems*, 37:110290–110314, 2024.
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., and Zhuang, Y. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/77c33e6a367922d003ff102ffb92b658-Abstract-Conference.html.
- Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pp. 53366–53397, 2024.
- Yu, Q., Li, J., Wu, Y., Tang, S., Ji, W., and Zhuang, Y. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21560–21571, October 2023.
- Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., and Zhuang, Y. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12944–12953, June 2024.
- Zhang, C., Li, L., He, S., Zhang, X., Qiao, B., Qin, S., Ma, M., Kang, Y., Lin, Q., Rajmohan, S., Zhang, D., and Zhang, Q. Ufo: A ui-focused agent for windows os interaction, 2024. URL <https://arxiv.org/abs/2402.07939>.