

Optimus-2 : Multimodal Minecraft Agent with Goal-Observation-Action Conditioned Policy

Supplementary Material

The supplementary document is organized as follows:

- Sec. A: Limitation and Future Work.
- Sec. B: Minecraft Environment.
- Sec. C: MGOA Dataset.
- Sec. D: Training Details.
- Sec. E: Evaluation Benchmark.
- Sec. F: Experimental Results.
- Sec. G: Case Study.

A. Limitation and Future Work

In this paper, we aim to explore how agents can mimic human behavior patterns in Minecraft to accomplish various tasks. Experimental results demonstrate that Optimus-2 performs exceptionally well in both atomic tasks and long-horizon tasks. However, due to the lack of sufficient high-quality data for open-ended tasks (such as “building a house” and “defeating the Ender Dragon”), there remains significant room for improvement. Once such datasets are available, the ability of Optimus-2 to complete open-ended tasks will be enhanced. Moreover, despite showing promising performance in Minecraft, we have not yet extended our exploration to other simulation platforms, which represents a potential direction for future research.

B. Minecraft

Minecraft is an extremely popular sandbox video game developed by Mojang Studios¹. It allows players to explore a blocky, procedurally generated 3D world with infinite terrain, discover and extract raw materials, craft tools and items, and build structures or earthworks. In this environment, AI agents need to face situations that are highly similar to the real world, making judgments and decisions to deal with various environments and problems. As shown in Figure 1, both agents and humans are required to receive natural language instructions and current observations as input, and then output low-level actions, such as mouse and keyboard control commands. Therefore, Minecraft serves

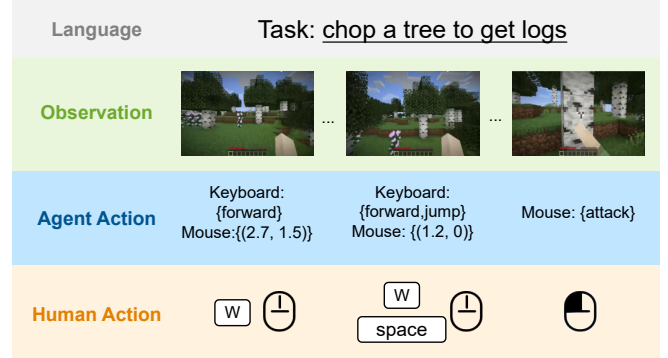
















Figure 1. Illustration of behavior patterns of both human and agents in Minecraft.


as an ideal open-world environment for training agent that can learn human behavior patterns.

B.1. Basic Rules

Biomes. The Minecraft world is divided into different areas called “biomes”. Different biomes contain different blocks and plants and change how the land is shaped. There are 79 biomes in Minecraft 1.16.5, including ocean, plains, forest, desert, etc. Diverse environments have high requirements for the generalization of agents.

Item. In Minecraft 1.16.5, there are 975 items can be obtained, such as wooden pickaxe , iron sword . Item can be obtained by crafting or destroying blocks or attacking entities. For example, agent can attack cows  to obtain leather  and beef . Agent also can use 1 stick  and 2 diamonds  to craft diamond sword .









Technology Tree. In Minecraft, the technology hierarchy comprises six levels: wood , stone , iron , gold , diamond , and redstone . Each tool level corresponds to specific mining capabilities. Wooden tools can mine stone-level blocks but are incapable of mining iron-level or higher-level blocks. Stone tools can mine iron-level blocks but cannot mine diamond-level or higher-level blocks. Iron tools are capable of mining diamond-level blocks. Finally, diamond tools can mine blocks of any level, including redstone-level.

Gameplay progress. Progression in Minecraft primarily involves discovering and utilizing various materials and resources, each unlocking new capabilities and opportunities. For instance, crafting a wooden pickaxe  enables players

¹<https://www.minecraft.net/en-us/article/meet-mojang-studios>

Table 1. Action space of agent in Minecraft.

Index	Agent Action	Human Action	Description
1	Forward	key W	Move forward.
2	Back	key S	Move back.
3	Left	key A	Strafe left.
4	Right	key D	Strafe right.
5	Jump	key Space	Jump. When swimming, keeps the player afloat.
6	Sneak	key left Shift	Slowly move in the current direction of movement.
7	Sprint	key left Ctrl	Move quickly in the direction of current movement.
8	Attack	left Button	Destroy blocks (hold down); Attack entity (click once).
9	Use	right Button	Place blocks, entity, open items or other interact actions defined by game.
10	hotbar [1-9]	keys 1-9	Selects the appropriate hotbar item.
11	Open/Close Inventory	key E	Opens the Inventory. Close any open GUI.
12	Yaw	move Mouse X	Turning; aiming; camera movement.Ranging from -180 to +180.
13	Pitch	move Mouse Y	Turning; aiming; camera movement.Ranging from -180 to +180.
14	Craft	-	Execute a crafting recipe to obtain new item
15	Smelt	-	Execute a smelting recipe to obtain new item.

to mine stone , which can then be used to create a stone pickaxe  and a furnace . These tools allow for the mining and smelting of iron ore . Subsequently, crafting an iron pickaxe  enables the extraction of diamonds , while a diamond pickaxe  can mine virtually any block in the game. Similarly, cultivating crops facilitates breeding various animals, each providing unique resources beyond sustenance. Drops from enemies also serve specific purposes, with some offering greater utility than others. By integrating resources from mining, farming, and breeding, players can enchant their equipment, further enhancing their capabilities. Additionally, collecting and crafting materials support construction, enabling players to create diverse structures. Beyond practical functions, such as building secure bases or farms, constructing personalized structures forms a significant aspect of the Minecraft experience. Figure 5 illustrates an example of progression: crafting an iron sword .

B.2. Observation and Action Spaces

Observation. In this paper, observation space of agent is completely consistent with human players. The agent only receives an RGB image with dimensions of 640×360 during the gameplay process, including the hotbar, health indicators, food saturation, and animations of the player’s hands. It is worth helping the agent see more clearly in extremely dark environments, we have added a night vision effect for the agent, which increases the brightness of the environment during the night.

Action Spaces. In MineRL [5] environment, agent’s action space is almost similar to human players. It consists of two parts: the mouse and the keyboard. The keypresses are re-

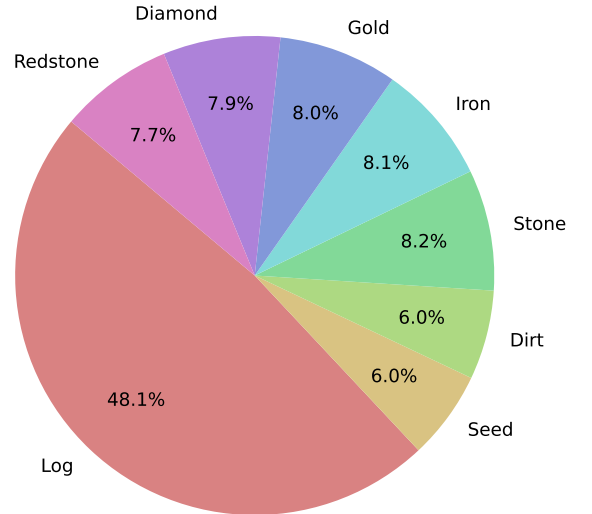









Figure 2. Statistical information on MGOA dataset. It contains 8 Atomic Tasks: ‘Log ’, ‘Seed ’, ‘Dirt ’, ‘Stone ’, ‘Iron ’, ‘Gold ’, ‘Diamond ’, ‘Redstone .

sponsible for controlling the movement of agents, such as jumping, forward, back, etc. The mouse movements are responsible for controlling the perspective of agents and the cursor movements when the GUI is opened. The left and right buttons of the mouse are responsible for attacking and

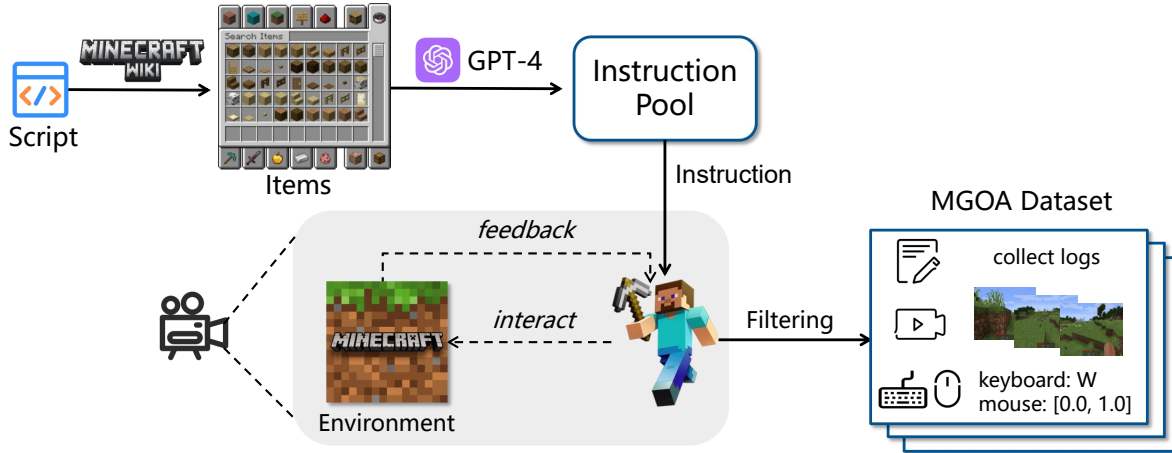


Figure 3. The pipeline for generating the MGOA dataset. First, we extracted item names from the Minecraft Wiki and employed GPT-4 to generate corresponding instructions. These instructions were then provided as input to STEVE-1, enabling it to interact with the environment to accomplish the tasks. During task execution, each observation was paired with its corresponding action, resulting in the creation of goal-observation-action pairs.

using or placing items. In Minecraft, precise mouse movements are important when completing complex tasks that need open inventory or crafting table. In order to achieve both the same action space with MineDojo [4], we abstract the craft and the smelt action into action space. The detailed action space is described in Table 1.

C. MGOA Dataset

In Minecraft, there is still a lack of sufficient high-quality goal-observation-action pairs to support the training of Optimus-2. To address this, we propose an automated dataset construction process aimed at creating high-quality Minecraft Goal-Observation-Action (MGOA) datasets. Through this method, MGOA contains 25,000 videos, providing about 30M goal-observation-action pairs. It contains 8 *Atomic Tasks* across 5 tech levels: ‘Log 🌳’, ‘Seed 🌱’, ‘Dirt 🟤’, ‘Stone 🟠’, ‘Iron ⚙️’, ‘Gold 🟡’, ‘Diamond 💎’, ‘Redstone 🔴’. Note that the *Atomic Tasks* in MGOA require minimal steps and can typically be completed within 2 ~ 3 minutes. For instance, the task ‘Iron ⚙️’ involves mining iron with a stone pickaxe, without the need to gather raw materials to craft the stone pickaxe. The statistics for the MGOA dataset is shown in Figure 2. We provide several examples of the dataset in the `MGOA_samples` folder within the supplementary materials. We will release this dataset to contribute to the development of open-world agents within the community.

C.1. Dataset Construction

Pipeline. Inspired by Li et al. [8], we employed a prior policy (STEVE-1 [9] in our work) to perform specific tasks

in Minecraft, and recorded the corresponding videos and actions to generate goal-observation-action pairs. As illustrated in Figure 3, we employed a custom script to extract item names from the Minecraft Wiki². Using these item names, we queried GPT-4³ with a predefined prompt template to generate task instructions, thereby constructing an Instruction Pool. The task instructions from the Instruction Pool serve as input to STEVE-1 [9], enabling it to interact with the environment to complete the tasks. During task execution, each frame and corresponding action were recorded and stored. To expedite data generation, we instantiated multiple policies and used parallelization to quickly produce large amounts of data.

Data Filtering. We judged task success based on environmental feedback. For example, feedback like “obtained new item, diamond axe” indicated that the task “craft a diamond axe” was successfully completed. During the dataset generation process, we observed a significant amount of low-quality video data due to limitations in the policy’s ability to follow instructions. Examples of low-quality data included task failures or task completion timeouts. To address this issue, we established two filtering criteria to ensure data quality: (1) only retaining data from successfully completed tasks, and (2) removing data for tasks that lasted longer than 2 minutes. These criteria allowed us to automatically filter out low-quality data, significantly reducing the cost of constructing the dataset. As a result, we obtained a high-quality MGOA dataset consisting of 25,000 samples.

²<https://minecraft.wiki/>

³<https://openai.com/index/gpt-4-research/>

C.2. Comparison with Existing Datasets

Previous gameplay videos were primarily obtained through two methods below.

Video Platform: For example, MineDojo [4] collected game videos uploaded by human players on platforms such as YouTube and Twitter, combining the video content with corresponding titles or subtitles to form video-text pairs. However, this dataset lacked recorded actions. To address this, VPT [1] used an Inverse Dynamics Model (IDM) to generate action sequences from the videos. However, the actions predicted by the IDM model are only approximations, which introduces a potential risk of misalignment between the frames and the corresponding actions.

Human Contractors: VPT [1] hired human players to freely explore Minecraft and used the frames and actions to construct a video-action dataset. However, this dataset lacked explicit natural language instructions. To create goal-observation-action pairs, STEVE-1 [9] used GPT-3.5 to generate specific task descriptions based on the gameplay, thereby integrating natural language instructions into the dataset. However, they provide only approximately 32k aligned goal-observation-action pairs, which remains a relatively scarce amount of data.

In addition, some work [11, 14] have utilized GPT-4V to generate image captions, task planning, and reflections, thereby creating image-text pairs that form instruction-following datasets.

Distinct from the aforementioned datasets, the MGOA dataset directly captures agents performing specific tasks, offering clear natural language instructions with a one-to-one correspondence between observations and actions. Furthermore, through rigorous data filtering, redundant action sequences that do not contribute to task completion are excluded from MGOA. In addition, compared to the small-scale goal-observation-action datasets currently available, MGOA offers 25,000 videos, encompassing approximately 30 million goal-observation-action pairs. This dataset is not only significantly larger but also highly scalable in an automated manner.

D. Training Details

D.1. Training Pipeline

One of the key factors in implementing our proposed method lies in the efficient alignment of language with the observation-action sequence, and subsequently translating language space into the action space. To tackle this problem, we adopt a two-phase training approach. First, we align language with the observation-action sequence via behavior pre-training. Then, we transform the language space into the action space through action fine-tuning.

Behavior Pre-training: During the pre-training phase, we integrated the Vision-guided Behavior Encoder into the

Table 2. Hyperparameter setting for pre-training and finetuning.

Hyperparameter	Pre-training	Fine-tuning
Optimizer	AdamW	AdamW
Learning Rate	0.0001	0.00004
Warmup Steps	0	0
Epochs	5	10
Batch Size	32	2048
Num. Frames	5M	80M
LoRA_r	64	64
LoRA_alpha	128	128

model. We used OpenAI Contractor Dataset [1] and a subset of MGOA as training data, which comprised approximately 5,000 videos. To balance efficiency and effectiveness, we freeze the visual encoder, then tune the Vision-guided Behavior Encoder along with a large language model (LoRA [6]). During pre-training, we set the learning rate to 0.0001 and trained for 5 epochs. The hyperparameter settings are shown in Table 2.

Action Fine-tuning: During the fine-tuning phase, we adapted the general MLLM DeepSeek-VL-1.3B [10] to the Minecraft environment, transitioning the model’s output space from language to low-level actions. We fine-tuned it using OpenAI Contractor Dataset [1] and MGOA, which comprises approximately 20,000 videos. In this phase, we freeze the Vision-guided Behavior Encoder, visual encoder, and large language model (LoRA), and only fine-tuned the action head. During fine-tuning, we set the learning rate to 0.00004 and train for 10 epochs. The hyperparameter settings are shown in Table 2.

D.2. Implementation Details

For the planner, we follow Li et al. [8], employing Multimodal Hybrid Memory empowered GPT-4V for planning and reflection. For the policy, we train the GOAP through the above pipeline. All experiments were conducted on 8x NVIDIA L40 GPUs. For the MGOA dataset, data collection and filtering were conducted in parallel, taking approximately 7 days. Training required around 2 days, while inference and evaluation on atomic tasks, long-horizon tasks, and open-ended instruction tasks took approximately 4 days.

E. Benchmark

E.1. Evaluation Tasks

The evaluation tasks are divided into three categories: *Atomic Tasks*, *Long-horizon Tasks*, and *Open-ended Instruction Tasks*. For each task, the agent’s environment is



Figure 4. Examples of *Atomic Task*. The agent must follow the instructions to collect resources. These four tasks represent the basic capabilities of the agent. The more resources are collected, the stronger the basic capabilities of the agent will be.

randomly initialized each time, and every task is executed at least 30 times. For *Atomic Tasks*, we follow the setting of prior work [9, 14], which requires the agent to execute the task within 2 minutes. We then report the average reward for the task, defined as the number of items obtained. For *Open-ended Instruction Tasks* and *Long-horizon Tasks*, we report the average success rate (SR) for each task.

Atomic Tasks. As shown in Figure 4, Atomic Tasks are short-term skills in Minecraft, such as “chop a tree to get logs”, “mine dirt”, “collect seeds”, and “dig down to mine stone”, etc.

Long-horizon Tasks. As shown in Figure 5, Long-Horizon Tasks are a sequence of *Atomic Tasks*. For example, “craft an iron sword from scratch” requires completing the atomic tasks of “chop 7 logs”, “craft 21 planks”, “craft 5 sticks”, “craft 1 crafting table”, and so on. These *Atomic Tasks* are interdependent, meaning that the failure of any single atomic task will result in the failure of the entire *Long-horizon Task*.

Open-ended Instruction Tasks. Open-Ended Instruction Tasks are not limited to predefined text formats; rather, they involve flexible language directives that prompt the agent to accomplish long-horizon tasks. These tasks evaluate the

agent’s capacity to interpret and execute instructions expressed in open-ended natural language. We selected Torch 🔥, Rail 🚊, Golden Shovel 🛟, Diamond Pickaxe ⚒️, and Compass 🧭 as evaluation tasks. Instruction for each task are shown in Table 3, Table 4, Table 5, Table 6 and Table 7.

E.2. Baselines

In this section, we provide a brief overview of existing Minecraft agents and compare them with our proposed Optimus-2. Current agents can be broadly categorized into two types: policy-based agents and planner-policy agents.

Policy-based Agents. Policy-based agents [1–4, 9] refer to those trained through reinforcement learning or imitation learning, capable of completing atomic tasks within Minecraft. However, due to limitations in instruction understanding and reasoning abilities, they struggle to accomplish long-horizon tasks.

Planner-Policy Agents. Planner-policy agents [7, 8, 11–14] refer to non-end-to-end architectures that utilize a MLLM (Multi-Layered Language Model) as a planner to decompose complex instructions into a sequence of sub-goals executable by a policy. While significant progress has been made, the current performance bottleneck stems from

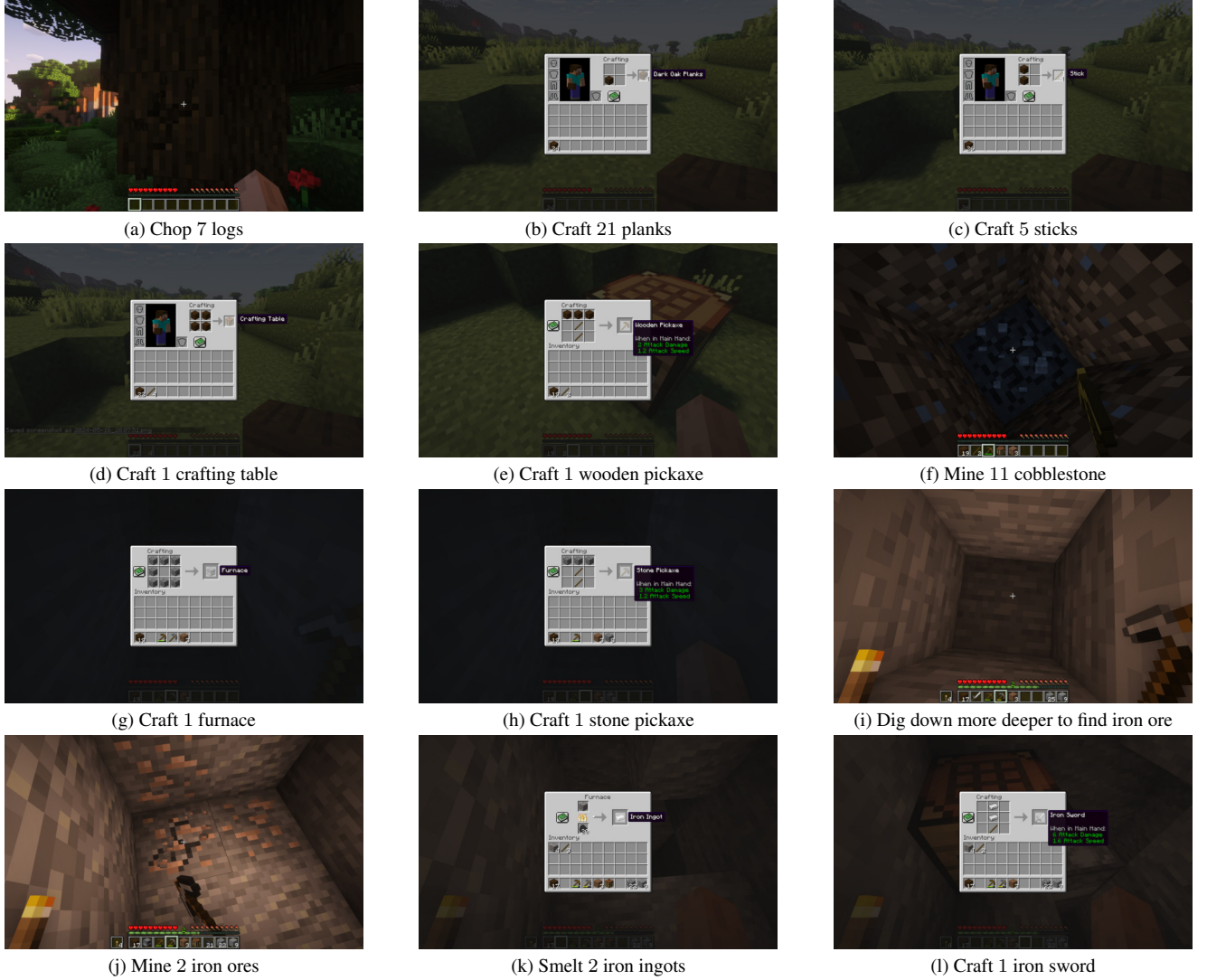


Figure 5. An example of *long-horizon task* “crafting an iron sword”. The agent must sequentially complete each atomic task in order to successfully craft the iron sword. Failure in any of the atomic tasks will result in the failure of the entire long-horizon task.

the policy’s ability to effectively understand and execute the sub-goals generated by the planner.

Comparison with Existing Agents. As a core contribution of this work, we propose a novel Goal-Observation-Action Conditioned Policy, GOAP. It integrates two key components: an Action-Guided Behavior Encoder for modeling observation-action sequences, and an MLLM for aligning sub-goals with these sequences. Leveraging the MLLM’s advanced understanding of open-ended instructions, GOAP demonstrates superior instruction-following capabilities compared to existing policies. On top of GOAP, the proposed agent, Optimus-2, exhibits superior performance in long-horizon tasks, outperforming the current state-of-the-art across all seven task groups.

F. Experimental Results

In this section, we report the experimental results of Optimus-2 on each *Long-horizon task*.

F.1. Results on Long-horizon Task

In this section, we report the results of Optimus-2 on each *Long-horizon Task*, with details including task name, numbers of sub-goals, success rate (SR), and eval times. As shown in Tables 8 and 9, Optimus-2 demonstrates superior performance across all 67 *Long-horizon Tasks*. Since Optimus-2 is randomly initialized in arbitrary environments for each task execution, the experimental results also highlight its generalization capability across diverse environments.

G. Case Study

In this section, we provide additional cases to illustrate the differences in the ability of VPT (text) [1], STEVE-1 [9], and Optimus-2 to perform *Open-ended Instruction Tasks*. We provide different open-ended instructions requiring the agent to perform tasks across various biomes. As shown in Figure 6, Figure 7, and Figure 8, Optimus-2 effectively completes all tasks, while VPT (text) and STEVE-1 fail due to limitations in language understanding and multimodal perception capabilities. Moreover, we provide several demo videos of Optimus-2 performing long-horizon tasks in the `Optimus2_videos` folder within the supplementary materials.

Table 3. Open-ended instruction examples of “Craft a torch”

Planner	Instruction
GPT4V	<p>Plan 1 1. First, I need to locate and chop down a couple of trees to collect 4 logs. 2. Next, I should craft these logs into 13 planks. 3. Then, I need to use some of the planks to craft 3 sticks. 4. I will use some of the planks to craft a crafting table. 5. Next, I should craft a wooden pickaxe using some of the planks and sticks. 6. With the wooden pickaxe, I need to find and mine at least 1 coal from a nearby coal deposit or cave. 7. Finally, I can use one stick and one coal to craft a torch.</p>
	<p>Plan 2 1. First, I need to find a tree to gather 4 logs. 2. Next, I have to craft the logs into planks. 3. From the planks, I will craft 3 sticks. 4. I need to craft a crafting table using some of the planks. 5. I should then craft a wooden pickaxe to mine coal. 6. I will locate and mine 1 piece of coal using the wooden pickaxe. 7. Finally, I will use the coal and sticks to craft a torch.</p>
	<p>Plan 3 1. I have to find a tree to get 4 logs. 2. Next, I have to craft planks from the logs. 3. Get sticks from some of the planks. 4. I need a crafting table, which requires planks to craft. 5. I should craft a wooden pickaxe using the crafting table. 6. I need to find and mine some coal using the wooden pickaxe. 7. Finally, with the stick and coal, I can craft a torch.</p>
GLM4V	<p>Plan 1 1. Locate and chop down a tree to gather 4 logs. 2. Convert the logs into 13 planks using the crafting menu. 3. Use the planks to craft a crafting table. 4. Craft 3 sticks from the remaining planks. 5. With the crafting table and sticks, craft a wooden pickaxe if one is not already available. 6. Use the wooden pickaxe to dig for coal or kill some Creepers or zombies for coal drops. 7. Once 1 coal is obtained, use the crafting table to combine it with a stick to craft 1 torch.</p>
	<p>Plan 2 1. Locate and chop down a tree to gather the 4 logs required. 2. Convert the logs into 13 planks using the crafting menu, as this will be the material needed for crafting the crafting table and the wooden pickaxe. 3. Use some of the planks to craft a crafting table, which is necessary for advanced crafting recipes. 4. Craft a wooden pickaxe using some of the planks and sticks. The wooden pickaxe will be used to mine coal. 5. Gather 3 sticks, which can be obtained from the logs or by crafting them from planks. 6. Use the wooden pickaxe to mine coal. Only 1 coal is needed to craft a torch. 7. Once all materials are gathered, open the crafting table and combine 1 coal and 1 stick to craft a torch.</p>
	<p>Plan 3 1. Locate and gather 4 logs from the nearest trees. 2. Convert the logs into 13 planks using the crafting menu. 3. Use some planks to craft a crafting table for more complex recipes. 4. Craft a wooden pickaxe using some planks and sticks to gather coal and other resources. 5. Convert some planks into 3 sticks required for crafting torches and the wooden pickaxe. 6. Use the wooden pickaxe to mine coal ore, ensuring to gather at least 1 coal. 7. With all materials gathered, open the crafting table and craft 1 torch using 1 coal and 1 stick.</p>

Table 4. Open-ended instruction examples of “Craft a rail”

Planner	Instruction
GPT4V	<p>Plan 1 1. First, locate and punch down a tree to acquire 7 logs. 2. Convert these logs into planks to obtain a total of 25 planks. 3. Use some of these planks to craft 5 sticks. 4. Create a crafting table using 4 planks. 5. Craft a wooden pickaxe using 3 planks and 2 sticks. 6. Use the wooden pickaxe to mine 11 cobblestones. 7. Craft a furnace using 8 cobblestones. 8. Craft a stone pickaxe using 3 cobblestones and 2 sticks. 9. Use the stone pickaxe to mine 6 iron ores. 10. Smelt the iron ores in the furnace to create 6 iron ingots. 11. Finally, craft a rail using the iron ingots and sticks.</p> <p>Plan 2 1. I have to find a tree to get 7 logs. 2. Next, I have to craft the logs into 25 planks. 3. Craft 5 sticks from the planks. 4. I need to create a crafting table. 5. Use the crafting table to craft a wooden pickaxe. 6. Use the wooden pickaxe to mine 11 pieces of cobblestone. 7. I’ll then craft a furnace using the cobblestone. 8. Craft a stone pickaxe with the cobblestone and sticks. 9. Use the stone pickaxe to mine 6 pieces of iron ore. 10. Smelt the iron ore in the furnace to obtain 6 iron ingots. 11. Finally, use the crafting table to craft a rail using the iron ingots and sticks.</p> <p>Plan 3 1. I have to find a tree to get 7 logs. 2. Next, I need to craft planks from the logs. 3. Get sticks from the planks. 4. Craft a crafting table using some of the planks. 5. Craft a wooden pickaxe with sticks and planks to mine stone. 6. Mine 11 cobblestone with the wooden pickaxe. 7. Craft a furnace using the cobblestone. 8. Craft a stone pickaxe using cobblestone and sticks. 9. Use the stone pickaxe to mine 6 iron ore. 10. Smelt the iron ore in the furnace to get 6 iron ingots. 11. Finally, use the iron ingots and sticks to craft a rail.</p>
	<p>Plan 1 1. Locate and gather 7 logs from nearby trees. 2. Use the crafting menu to convert the logs into 25 planks. 3. Craft 5 sticks from the planks. 4. Place down a crafting table to access more crafting recipes. 5. Use the crafting table to craft a wooden pickaxe. 6. With the wooden pickaxe, mine 11 cobblestones from the ground or from stone blocks. 7. Craft a furnace using the cobblestones. 8. Upgrade the wooden pickaxe to a stone pickaxe by crafting it at the crafting table with the cobblestones. 9. Use the stone pickaxe to mine 6 iron ore from the appropriate rock layers. 10. Smelt the 6 iron ore in the furnace to produce 6 iron ingots. 11. Finally, use the crafting table to craft 1 rail using the 6 iron ingots and 1 stick.</p> <p>Plan 2 1. Locate and gather 7 logs from nearby trees. 2. Use the crafting menu to convert the logs into 25 planks. 3. Craft 5 sticks using the planks in the crafting menu. 4. Place a crafting table to enable the crafting of more complex items. 5. Craft a wooden pickaxe using some of the planks and sticks. 6. Use the wooden pickaxe to mine 11 cobblestones from stone outcrops. 7. Construct a furnace using the cobblestones. 8. Upgrade to a stone pickaxe by crafting it with the cobblestones and sticks. 9. Use the stone pickaxe to mine 6 iron ore from iron veins. 10. Smelt the 6 iron ore in the furnace to produce 6 iron ingots. 11. Finally, use the crafting table to craft 1 rail using the iron ingots and sticks.</p> <p>Plan 3 1. Find and chop down a tree to gather 7 logs. 2. Use the crafting menu to convert the logs into 25 planks. 3. Craft 5 sticks from the planks. 4. Place down a crafting table to access more crafting options. 5. Craft a wooden pickaxe using some of the planks and sticks. 6. Use the wooden pickaxe to mine 11 cobblestones from stone. 7. Construct a furnace using some of the cobblestones. 8. Upgrade the wooden pickaxe to a stone pickaxe by crafting it with the cobblestones. 9. Search for and mine 6 iron ore blocks with the stone pickaxe. 10. Smelt the iron ore in the furnace to produce 6 iron ingots. 11. Finally, use the crafting table to craft 1 rail using the iron ingots and sticks.</p>
GLM4V	

Table 5. Open-ended instruction examples of “Craft a golden shovel”

Planner	Instruction
GPT4V	<p>Plan 1 1. Start by finding and chopping down trees to collect 9 logs. 2. Convert the logs into 35 planks using the crafting interface. 3. Use some of the planks to craft 8 sticks. 4. Create a crafting table using some of the planks. 5. Craft a wooden pickaxe. 6. Mine 19 cobblestone. 7. Craft a furnace using 8 of the cobblestones. 8. Craft a stone pickaxe. 9. Mine 3 iron ores using the stone pickaxe. 10. Smelt the iron ore in the furnace. 11. Craft an iron pickaxe using sticks and iron ingots. 12. Locate and mine a gold block using the iron pickaxe. 13. Smelt the gold block in the furnace to get 1 gold ingot. 14. Craft a golden shovel using the gold ingot and sticks.</p> <p>Plan 2 1. Start by finding and chopping down trees to collect 9 logs. 2. Use the logs to craft planks. 3. Craft sticks. 4. Use some of the planks to craft a crafting table. 5. Use the crafting table to craft a wooden pickaxe. 6. Use the wooden pickaxe to mine 19 pieces of cobblestone. 7. Use some of the cobblestone to craft a furnace. 8. Use the remaining cobblestone to craft a stone pickaxe. 9. Mine iron ore with the stone pickaxe. 10. Smelt the iron ore in the furnace to obtain 3 iron ingots. 11. Use 2 of the iron ingots and sticks to craft an iron pickaxe. 12. Use the iron pickaxe to mine gold ore or find a gold block; you need at least 1 gold ingot. 13. If mining gold ore, smelt it in the furnace to obtain gold ingots. 14. Use the gold ingot and stick to craft a golden shovel.</p> <p>Plan 3 1. Find and cut down trees to collect 9 logs. 2. Use the logs to craft 35 planks. 3. Craft 8 sticks using the planks. 4. Use some of the planks to craft a crafting table. 5. Craft a wooden pickaxe using planks and sticks. 6. Mine cobblestone until you have 19 pieces. 7. Use some cobblestone and the crafting table to craft a furnace. 8. Craft a stone pickaxe using cobblestone and remaining sticks. 9. Use the stone pickaxe to mine 3 pieces of iron ore. 10. Smelt the iron ore in the furnace to get 3 iron ingots. 11. Craft an iron pickaxe. 12. Use the iron pickaxe to mine gold ore to obtain gold blocks. 13. Use 1 gold ingot to craft a golden shovel. 14. use a gold ingot and sticks to craft a golden shovel using the crafting table.</p>
	<p>Plan 1 1. Find and chop down trees to gather 9 logs. 2. Convert the logs into 35 planks using the crafting menu. 3. Craft 8 sticks from the planks. 4. Build a crafting table to expand the crafting options. 5. Craft a wooden pickaxe with the planks and sticks. 6. Mine cobblestone until you have a total of 19. 7. Craft a furnace using 8 cobblestone. 8. Use the furnace to smelt the cobblestone into 19 stone, then use the crafting table to craft a stone pickaxe with the stone and sticks. 9. Use the stone pickaxe to mine 3 iron ore. 10. Smelt the iron ore in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe using the iron ingots and sticks at the crafting table. 12. Use the iron pickaxe to mine a gold block. 13. Smelt the gold block in the furnace to get 1 gold ingot. 14. Finally, use the crafting table to craft the golden shovel using 1 gold ingot and 2 sticks.</p> <p>Plan 2 1. Find and gather 9 logs from nearby trees. 2. Convert the logs into 35 planks using a crafting table. 3. Craft 8 sticks from the planks. 4. Build a crafting table. 5. Use the crafting table to craft a wooden pickaxe with the planks and sticks. 6. Use the wooden pickaxe to mine 19 cobblestones from stone. 7. Construct a furnace to smelt materials. 8. Craft a stone pickaxe using the cobblestones and sticks. 9. Use the stone pickaxe to mine 3 iron ore from iron deposits. 10. Smelt the 3 iron ore in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe with the iron ingots and sticks. 12. Use the iron pickaxe to mine a gold block. 13. Smelt the gold block in the furnace to get 1 gold ingot. 14. Finally, craft the golden shovel using the gold ingot and stick in the crafting table.</p> <p>Plan 3 1. Find and gather 9 logs from nearby trees. 2. Use the crafting menu to convert the logs into 35 planks. 3. Craft 8 sticks from the planks. 4. Craft a crafting table using some of the planks. 5. Use the crafting table to craft a wooden pickaxe with some of the sticks and planks. 6. Use the wooden pickaxe to mine 19 cobblestones from stone. 7. Craft a furnace using some of the cobblestones. 8. Use the furnace to smelt the cobblestones into 19 stone, then craft a stone pickaxe with the sticks and stone. 9. Mine 3 iron ore from the appropriate rock faces. 10. Smelt the iron ore in the furnace to produce 3 iron ingots. 11. Use the crafting table to craft an iron pickaxe with the iron ingots and sticks. 12. Use the iron pickaxe to mine a gold block. 13. Smelt the gold block in the furnace to produce 1 gold ingot. 14. Finally, use the crafting table to craft the golden shovel with the gold ingot and 2 sticks.</p>
GLM4V	

Table 6. Open-ended instruction examples of “Craft a diamond pickaxe”

Planner	Instruction
GPT4V	<p>Plan 1 1. First, locate and chop down trees to gather 9 logs. 2. Convert the logs into 35 planks. 3. Use some of the planks to craft 8 sticks. 4. Create a crafting table using 4 planks. 5. With the crafting table, craft a wooden pickaxe using planks and sticks. 6. Use the wooden pickaxe to collect at least 11 cobblestones from stone. 7. With the cobblestones, craft a furnace. 8. Use more cobblestones to craft a stone pickaxe. 9. Use the stone pickaxe to mine 3 iron ores. 10. Smelt the iron ores into 3 iron ingots using the furnace. 11. Craft an iron pickaxe with iron ingots and sticks. 12. Use the iron pickaxe to mine 3 diamonds. 13. Finally, craft the diamond pickaxe using the diamonds and sticks.</p> <p>Plan 2 1. Begin by locating a tree to gather 9 logs. 2. Convert the logs into planks. 3. Use some of the planks to craft sticks. 4. Craft a crafting table to facilitate further crafting. 5. Create a wooden pickaxe to start mining stone. 6. Mine 11 cobblestones. 7. Use the cobblestones to craft a furnace. 8. Craft a stone pickaxe to advance to mining iron ore. 9. Locate and mine 3 iron ores using the stone pickaxe. 10. Smelt the iron ores in the furnace to produce 3 iron ingots. 11. Use two of the iron ingots and sticks to craft an iron pickaxe. 12. Locate a diamond ore vein and mine 3 diamonds using the iron pickaxe. 13. Finally, craft the diamond pickaxe with the 3 diamonds and sticks.</p> <p>Plan 3 1. I need to find a tree to gather 9 logs. 2. Convert the logs into 35 planks. 3. Use some planks to craft 8 sticks. 4. Craft a crafting table using some of the planks. 5. With the crafting table, make a wooden pickaxe to mine stone. 6. Mine 11 cobblestones using the wooden pickaxe. 7. Craft a furnace using the collected cobblestones. 8. Make a stone pickaxe to efficiently mine iron ore. 9. Find and mine 3 iron ores using the stone pickaxe. 10. Smelt the iron ores in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe using some planks, sticks, and the iron ingots. 12. Locate a source of diamonds and use the iron pickaxe to mine 3 diamonds. 13. Finally, use the crafting table along with the sticks and diamonds to craft a diamond pickaxe.</p>
	<p>Plan 1 1. Start by finding trees to gather 9 logs. 2. Convert the logs into 35 planks. 3. Craft 8 sticks from the planks. 4. Build a crafting table to expand crafting options. 5. Use the crafting table to make a wooden pickaxe. 6. With the wooden pickaxe, mine 11 cobblestones. 7. Craft a stone pickaxe using the cobblestones. 8. Construct a furnace to smelt materials. 9. Use the stone pickaxe to mine 3 iron ores. 10. Smelt the iron ores in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe with the iron ingots. 12. Locate and mine 3 diamonds, which may require exploring caves or mining deep underground with the iron pickaxe. 13. Finally, use the crafting table to combine the diamonds and the sticks to craft a diamond pickaxe.</p> <p>Plan 2 1. Locate and chop down trees to gather a total of 9 logs. 2. Use the logs to craft 35 planks, as this will be the base material for several other items. 3. Craft 8 sticks from the planks, which will be needed to create various tools. 4. Craft a crafting table using some of the planks, as this is essential for crafting more complex items. 5. Use the crafting table to craft a wooden pickaxe with the planks and sticks, which will be used to mine cobblestone. 6. Mine 11 cobblestone using the wooden pickaxe. 7. Craft a stone pickaxe with the cobblestone and sticks, as this is more durable and efficient for mining. 8. Construct a furnace using the cobblestone, which will be required to smelt ores into ingots. 9. Mine 3 iron ore with the stone pickaxe. 10. Smelt the iron ore in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe using the iron ingots and sticks, as this is necessary to mine diamonds. 12. Find and mine diamond ore with the iron pickaxe to gather 3 diamonds. 13. Finally, use the crafting table to craft a diamond pickaxe with the diamonds and sticks, completing the task.</p> <p>Plan 3 1. Begin by locating trees to gather 9 logs. 2. Convert the logs into 35 planks. 3. Craft 8 sticks from the planks. 4. Build a crafting table to expand crafting options. 5. Use the crafting table to craft a wooden pickaxe. 6. With the wooden pickaxe, mine 11 cobblestones. 7. Craft a stone pickaxe using the cobblestones. 8. Construct a furnace to smelt materials. 9. Use the stone pickaxe to mine 3 iron ore. 10. Smelt the iron ore in the furnace to produce 3 iron ingots. 11. Craft an iron pickaxe with the iron ingots. 12. Locate and mine diamond ore with the iron pickaxe to obtain 3 diamonds. 13. Finally, use the crafting table to craft a diamond pickaxe with the diamonds.</p>
GLM4V	

Table 7. Open-ended instruction examples of “Craft a compass”

Planner	Instruction
GPT4V	<p>Plan 1 1. Locate a tree and gather 10 logs. 2. Convert the logs into 38 planks. 3. Craft 8 sticks from the planks. 4. Use some planks to craft a crafting table. 5. With the crafting table, craft a wooden pickaxe. 6. Mine cobblestone until you have 12 pieces. 7. Upgrade your tools by crafting a stone pickaxe. 8. Construct a furnace using some of the cobblestone. 9. Use the stone pickaxe to mine 7 iron ore. 10. Smelt the iron ore in the furnace to produce 7 iron ingots. 11. Craft an iron pickaxe using the iron ingots. 12. Find and mine redstone using the iron pickaxe to obtain 1 redstone dust. 13. Finally, use the remaining iron ingots and redstone dust at the crafting table to craft a compass.</p> <p>Plan 2 1. Start by finding and chopping down trees to collect 10 logs. 2. Use the logs to craft 38 planks. 3. Craft sticks using the planks to get 8 sticks. 4. Craft a crafting table from the planks. 5. Use the crafting table to create a wooden pickaxe. 6. Locate and mine 12 pieces of cobblestone using the wooden pickaxe. 7. Craft a stone pickaxe from cobblestone and sticks. 8. Use the cobblestone to craft a furnace. 9. Mine 7 iron ores using the stone pickaxe. 10. Smelt the iron ore in the furnace to get 7 iron ingots. 11. Craft an iron pickaxe using iron ingots and sticks. 12. Search for and mine redstone using the iron pickaxe. 13. Finally, use the iron ingots and redstone to craft a compass.</p> <p>Plan 3 1. First, I need to find a tree and collect 10 logs. 2. Convert the logs into planks, ensuring I get at least 38 planks. 3. Use some of the planks to craft sticks, collecting at least 8 sticks. 4. Use some of the planks to create a crafting table. 5. Craft a wooden pickaxe. 6. Locate cobblestone and mine at least 12 pieces using the wooden pickaxe. 7. Use some cobblestone to craft a furnace. 8. Craft a stone pickaxe for improved mining. 9. Search for iron ore and mine at least 7 using the stone pickaxe. 10. Smelt the iron ore in the furnace to produce 7 iron ingots. 11. Craft an iron pickaxe using some of the iron ingots and sticks. 12. Locate a redstone ore and mine it using the iron pickaxe to obtain redstone. 13. Finally, use 4 of the iron ingots and the redstone to craft a compass at the crafting table.</p>
	<p>Plan 1 1. Locate and gather 10 logs from nearby trees. 2. Convert the logs into 38 planks using the crafting menu. 3. Craft 8 sticks from the planks. 4. Build a crafting table to access more crafting options. 5. Use the crafting table to make a wooden pickaxe. 6. With the wooden pickaxe, mine 12 cobblestones from stone. 7. Craft a stone pickaxe using the cobblestones and the crafting table. 8. Construct a furnace to smelt materials. 9. Use the stone pickaxe to mine 7 iron ores from the appropriate rock formations. 10. Smelt the 7 iron ores in the furnace to produce 7 iron ingots. 11. Craft an iron pickaxe with the iron ingots at the crafting table. 12. Mine for redstone, which can sometimes be found in the ground. 13. Use the crafting table to craft a compass with the iron ingot and redstone.</p> <p>Plan 2 1. Locate and chop down trees to gather the required 10 logs. 2. Convert the logs into 38 planks using the crafting menu. 3. Craft 8 sticks from the planks. 4. Use some planks to craft a crafting table. 5. With the crafting table and the sticks and planks crafted, make a wooden pickaxe. 6. Use the wooden pickaxe to mine 12 cobblestone. 7. Craft a stone pickaxe using the cobblestone. 8. Construct a furnace with the cobblestone. 9. Use the stone pickaxe to mine 7 iron ore. 10. Smelt the iron ore in the furnace to produce 7 iron ingots. 11. Craft an iron pickaxe using the iron ingots and the crafting table. 12. Search for redstone; gather at least 1 redstone. 13. Finally, use the crafting table to combine the redstone with the iron ingots to craft a compass.</p> <p>Plan 3 1. Find and chop down trees to gather 10 logs. 2. Convert the logs into 38 planks. 3. Craft 8 sticks from the planks. 4. Craft a crafting table using some of the planks. 5. Use the crafting table to craft a wooden pickaxe. 6. With the wooden pickaxe, mine cobblestone until you have at least 12. 7. Upgrade the wooden pickaxe to a stone pickaxe using the cobblestone. 8. Construct a furnace out of the cobblestone. 9. Use the stone pickaxe to mine 7 iron ore. 10. Smelt the 7 iron ore in the furnace to produce 7 iron ingots. 11. Craft an iron pickaxe using the iron ingots and sticks at the crafting table. 12. Search for redstone; you will need at least 1 piece. 13. Finally, use the crafting table to combine the redstone with the iron ingots to craft a compass.</p>
GLM4V	

Table 8. The results of Optimus-2 on the Wood Group, Stone Group, and Iron Group. SR denotes success rate.

Group	Task	Sub-Goal Num.	SR	Eval Times
Wood	Craft a wooden shovel	6	100.00	40
	Craft a wooden pickaxe	5	100.00	30
	Craft a wooden axe	5	97.37	38
	Craft a wooden hoe	5	100.00	30
	Craft a stick	4	100	30
	Craft a crafting table	3	93.02	43
	Craft a wooden sword	5	100.00	30
	Craft a chest	4	100.00	30
	Craft a bowl	4	100.00	30
	Craft a ladder	4	100.00	30
Stone	Craft a stone shovel	8	89.47	57
	Craft a stone pickaxe	10	98.00	50
	Craft a stone axe	10	94.44	54
	Craft a stone hoe	8	95.74	47
	Craft a charcoal	9	85.71	42
	Craft a smoker	9	90.00	40
	Craft a stone sword	8	95.45	44
	Craft a furnace	9	94.44	36
	Craft a torch	8	89.36	47
Iron	Craft an iron shovel	13	52.08	48
	Craft an iron pickaxe	13	56.00	50
	Craft an iron axe	13	48.15	54
	Craft an iron hoe	13	56.60	53
	Craft a bucket	13	45.10	51
	Craft a hopper	14	54.90	51
	Craft a rail	13	51.02	49
	Craft an iron sword	12	56.52	46
	Craft a shears	12	48.28	58
	Craft a smithing table	12	53.33	45
	Craft a tripwire hook	13	55.56	45
	Craft a chain	13	52.17	46
	Craft an iron bars	12	51.06	47
	Craft an iron nugget	12	54.55	44
	Craft a blast furnace	14	52.27	44
	Craft a stonecutter	13	52.27	44

Table 9. The results of Optimus-2 on the Gold group, Diamond Group, Redstone Group, and Armor Group. SR denotes success rate.

Group	Task	Sub Goal Num.	SR	Eval Times
Gold	Craft a golden shovel	16	8.93	56
	Craft a golden pickaxe	16	11.29	62
	Craft a golden axe	16	8.93	56
	Craft a golden hoe	16	8.96	67
	Craft a golden sword	16	8.20	61
	Smelt and craft an golden ingot	15	9.68	62
Diamond	Craft a diamond shovel	15	15.91	44
	Craft a diamond pickaxe	15	11.76	34
	Craft a diamond axe	16	11.00	36
	Craft a diamond hoe	15	15.91	44
	Craft a diamond sword	15	11.11	36
	Dig down and mine a diamond	15	11.42	35
	Craft a jukebox	15	13.15	38
Redstone	Craft a piston	16	28.33	60
	Craft a redstone torch	16	27.69	65
	Craft an activator rail	18	25.81	62
	Craft a compass	23	28.36	67
	Craft a dropper	16	30.30	66
	Craft a note block	16	25.40	63
Armor	Craft shield	14	45.16	62
	Craft iron chestplate	14	43.86	57
	Craft iron boots	14	40.35	57
	Craft iron leggings	14	8.57	35
	Craft iron helmet	14	47.46	56
	Craft diamond helmet	17	9.09	33
	Craft diamond chestplate	17	7.89	38
	Craft diamond leggings	17	5.41	37
	Craft diamond boots	17	12.50	40
	Craft golden helmet	17	13.89	36
	Craft golden leggings	17	12.20	41
	Craft golden boots	17	10.26	39
	Craft golden chestplate	17	10.00	40

Agent	Instruction: <i>I want to get some logs to craft wooden sword, what should I do first?</i>							Success
OpenAI								✗
STEVE-1								✗
Optimus-2								✓

Figure 6. An illustration of VPT (text) [1], STEVE-1 [9], and Optimus-2 executing the open-ended instruction, “I want to get some logs to craft wooden sword, what should I do first?”. Existing policies are limited by their instruction comprehension abilities and thus fail to complete the task, whereas GOAP leverages the language understanding capabilities of the MLLM, enabling it to accomplish the task.

Agent	Instruction: <i>I need coal for heating. What should I do?</i>							Success
OpenAI								✗
STEVE-1								✗
Optimus-2								✓

Figure 7. An illustration of VPT (text) [1], STEVE-1 [9], and Optimus-2 executing the open-ended instruction, “I need coal for heating. What should I do?”. Existing policies are limited by their instruction comprehension abilities and thus fail to complete the task, whereas GOAP leverages the language understanding capabilities of the MLLM, enabling it to accomplish the task.

Agent	Instruction: <i>I want to collect some seeds, Can you help me?</i>							Success
OpenAI								✗
STEVE-1								✗
Optimus-2								✓

Figure 8. An illustration of VPT (text) [1], STEVE-1 [9], and Optimus-2 executing the open-ended instruction, “I want to collect some seeds, Can you help me?”. Existing policies are limited by their instruction comprehension abilities and thus fail to complete the task, whereas GOAP leverages the language understanding capabilities of the MLLM, enabling it to accomplish the task.

References

- [1] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampe-dro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022. [4](#), [5](#), [7](#), [15](#)
- [2] Shaofei Cai, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. Open-world multi-task control through goal-aware representation learning and adaptive horizon prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13734–13744, 2023.
- [3] Shaofei Cai, Bowei Zhang, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. Groot: Learning to follow instructions by watching gameplay videos. In *The Twelfth International Conference on Learning Representations*, 2023.
- [4] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35: 18343–18362, 2022. [3](#), [4](#), [5](#)
- [5] William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019. [2](#)
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. [4](#)
- [7] Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16426–16435, 2024. [5](#)
- [8] Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *arXiv preprint arXiv:2408.03615*, 2024. [3](#), [4](#), [5](#)
- [9] Shalev Lifshitz, Keiran Paster, Harris Chan, Jimmy Ba, and Sheila McIlraith. Steve-1: A generative model for text-to-behavior in minecraft. *Advances in Neural Information Processing Systems*, 2023. [3](#), [4](#), [5](#), [7](#), [15](#)
- [10] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. [4](#)
- [11] Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. *arXiv preprint arXiv:2312.07472*, 2023. [4](#), [5](#)
- [12] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023.
- [13] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*, 2023.
- [14] Zihao Wang, Shaofei Cai, Zhancun Mu, Haowei Lin, Ceyao Zhang, Xuejie Liu, Qing Li, Anji Liu, Xiaojian Ma, and Yitao Liang. Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents. *arXiv preprint arXiv:2407.00114*, 2024. [4](#), [5](#)