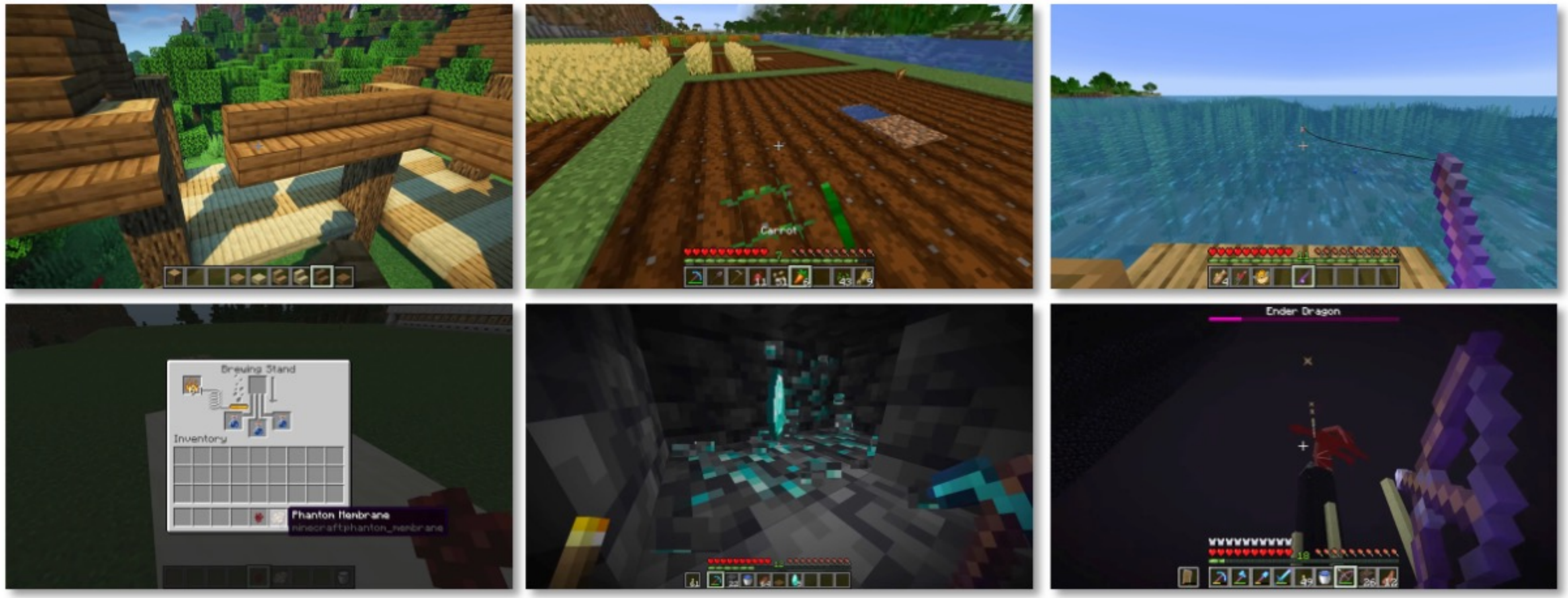
GROOT: Learning to Follow Instructions by Watching Gameplay Videos

Shaofei Cai, Bowei Zhang, Zihao Wang, Xiaojian Ma, Anji Liu, Yitao Liang

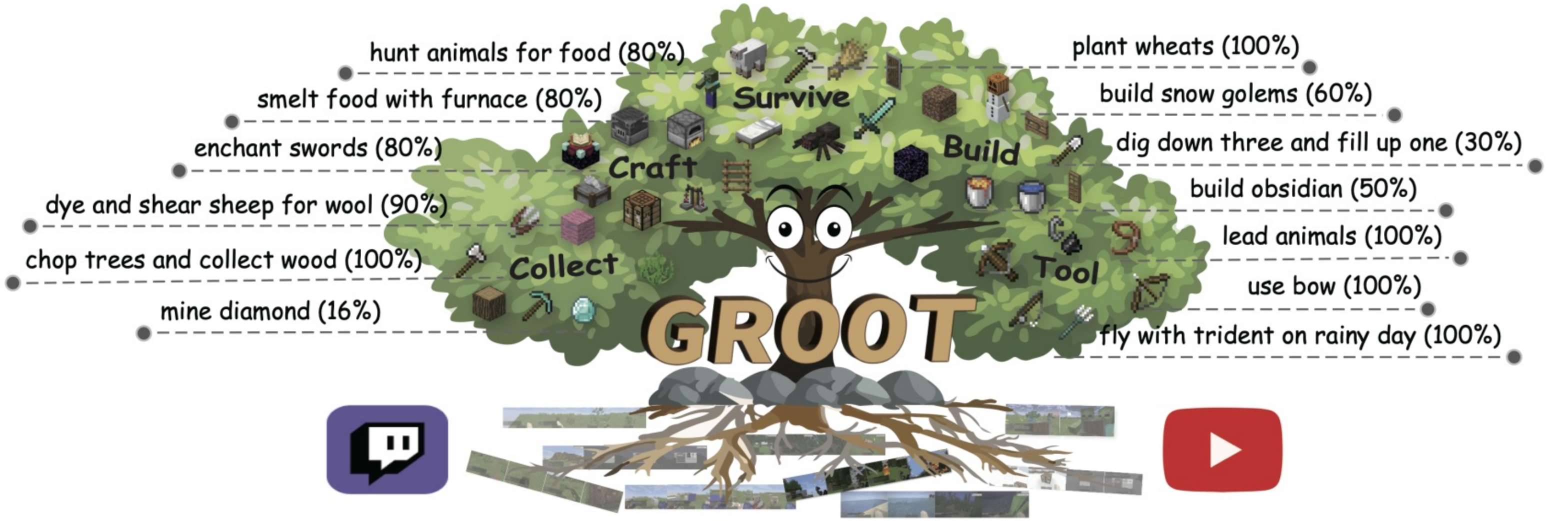
Problem



Some example tasks in Minecraft

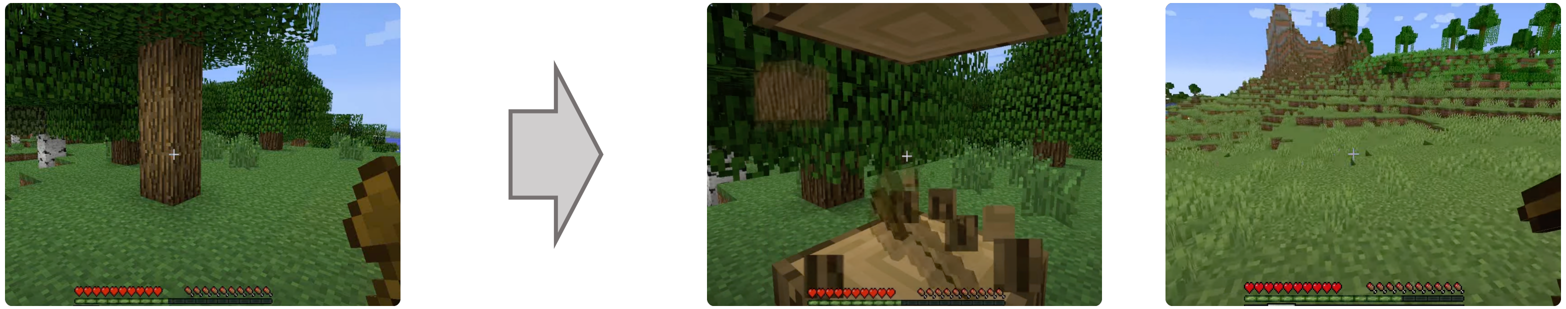
Developing a **policy** that can follow **open-ended instructions** and complete multiple tasks in open-world environments such as Minecraft is challenging and important.

Motivation



- The self-supervised pre-training paradigm can promote large-scale task learning.
- The reference video as instruction interface is expressive while the training data is easy to collect.

Goal Space Discovery via Future State Prediction

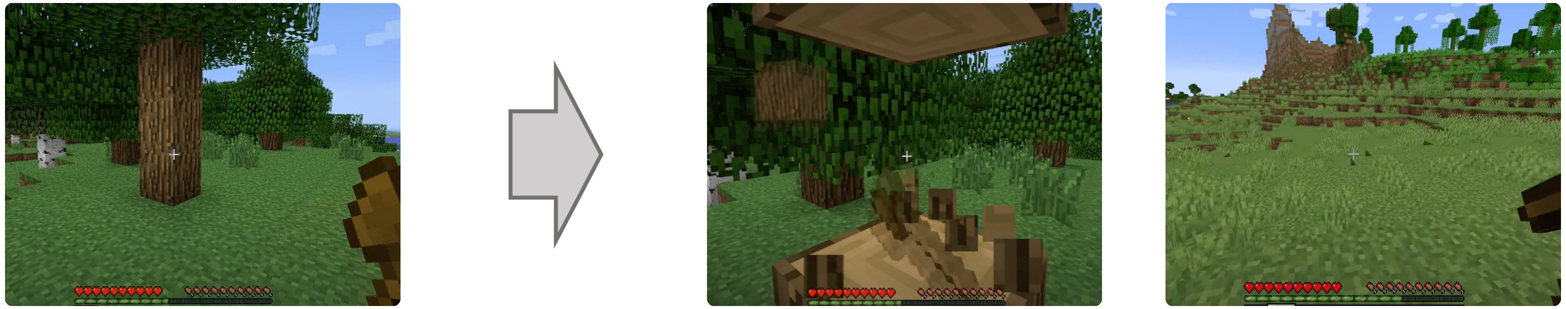


standing in plains *chop the tree or by pass the tree ?*

Q: How can we want to induce a goal space from a given gameplay dataset $\mathcal{D} = \{(s_{1:T})\}_M$?

Imagine you are standing in front of a tree. The next states you will see depend on what you want to do (goal), chop the tree of by pass the tree.

Goal Space Discovery via Future State Prediction



standing in plains *chop the tree or by pass the tree ?*

A: We create a generative pre-training task called future state prediction $p(s_{t+1:T}|s_{1:t})$. This process can be modeled using the variational autoencoder framework:

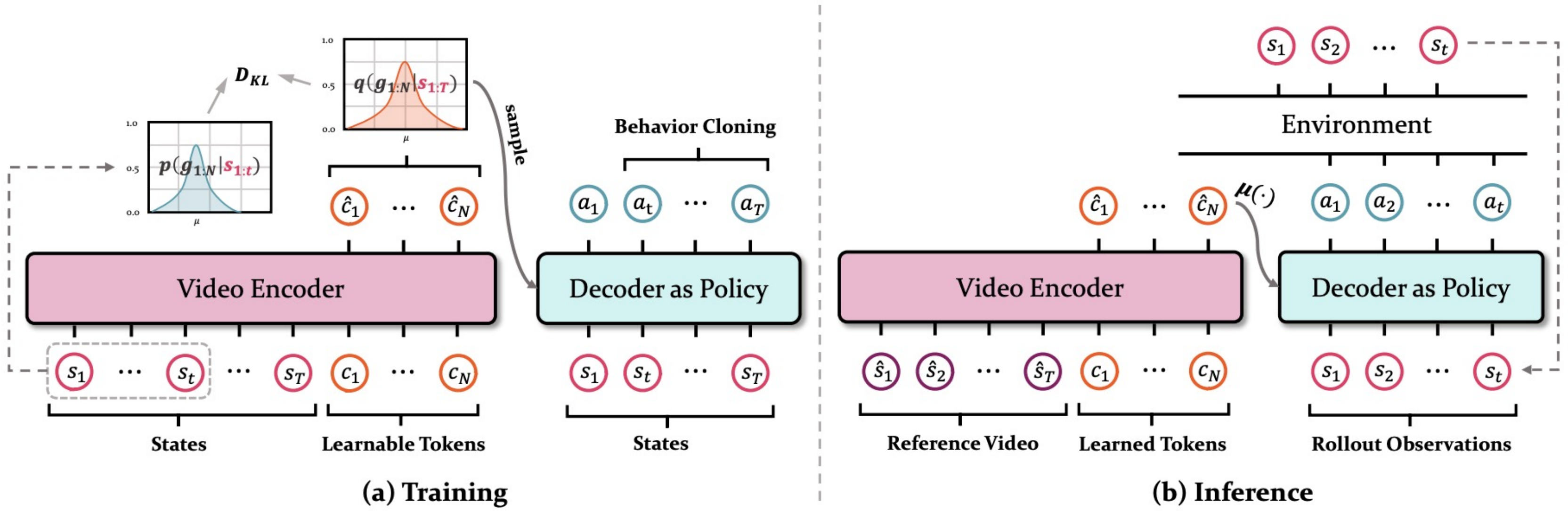
$$\log p_\theta(s_{t+1:T}|s_{1:t}) = \log \sum_g p_\theta(s_{t+1:T}, g|s_{1:t})$$
$$\geq \mathbb{E}_{g\sim q_\phi(\cdot|s_{1:T})} \left[\log p_\theta(s_{t+1:T}|s_{1:t}, g)\right] - D_{\mathrm{KL}}\left(q_\phi(g|s_{1:T}) \parallel p_\theta(g|s_{1:t})\right)$$

Goal Space Discovery via Future State Prediction

Since we want to learn a **policy** instead of a **video generator**, we breakdown $p_\theta(s_{t+1:T}|s_{1:t}, g)$ into components contributed by a **goal-conditioned policy** $\pi_\theta(a_\tau|s_{1:\tau}, g)$ and an **inverse dynamic model** $p_\theta(a_\tau|s_{1:\tau+1})$.

$$\log p(s_{t+1:T}|s_{1:t}) \geq \underbrace{\sum_{\tau=t}^{T-1} \mathbb{E}_{g\sim q_\phi(\cdot|s_{1:T}), a_\tau\sim p_\theta(\cdot|s_{1:\tau+1})} \left[\log \pi_\theta(a_\tau|s_{1:\tau}, g)\right]}_{\text{behaviour cloning}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(g|s_{1:T}) \parallel p_\theta(g|s_{1:t})\right)}_{\text{goal space constraint (KL regularization)}}$$

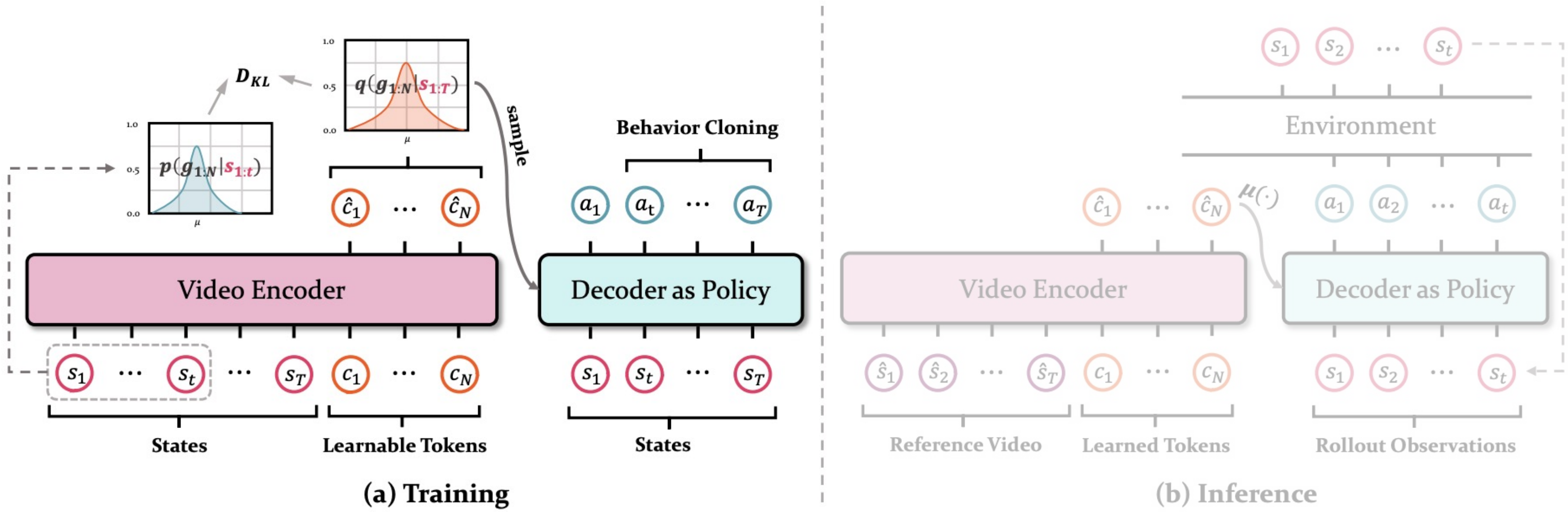
Architecture Design



A video encoder (non-causal transformer) learns to extract the semantic meaning and transfer the video into the goal embedding space.

A goal-conditioned policy (causal transformer) is learned to predict actions following the given instructions.

Architecture Design

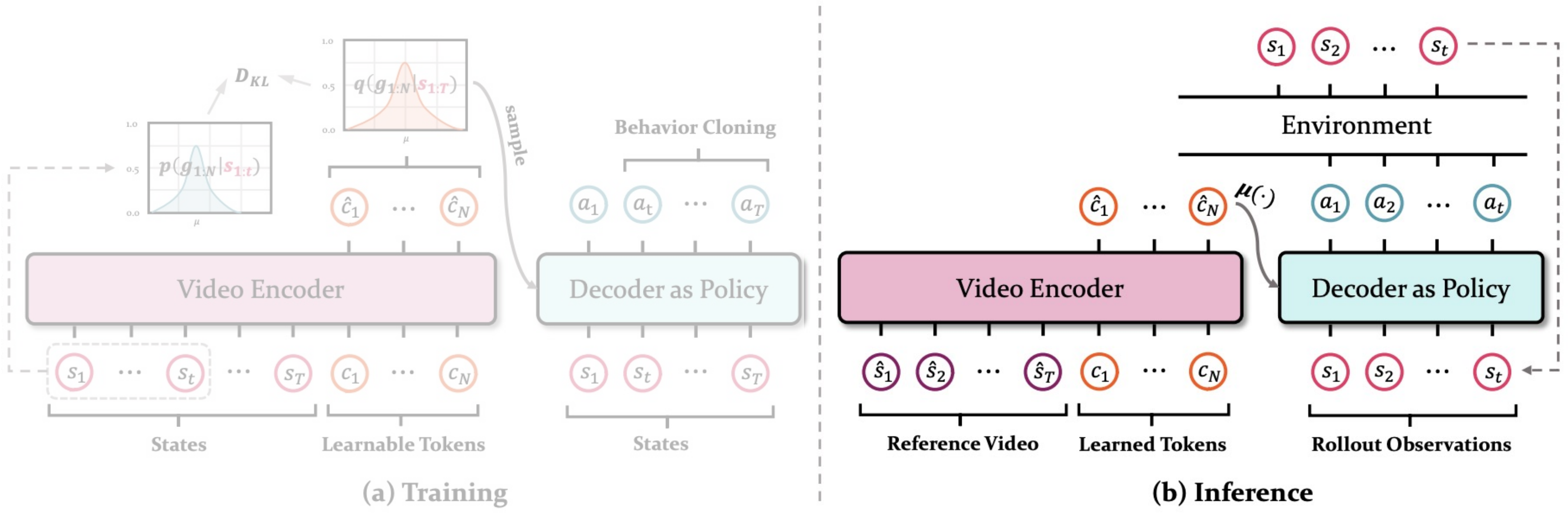


Training:

Given a gameplay video ($s_{1:T}$), we label them with inverse dynamic model and obtain ($s_{1:T}, a_{1:T}$). The objective function is

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\substack{(s_{1:T}, a_{1:T}) \sim \mathcal{D} \\ g \sim q_\phi(\cdot|s_{1:T})}} \left[\sum_{\tau=t}^{T-1} - \log \pi_\theta(a_\tau | s_{1:\tau}, g) + \lambda_{KL} D_{KL} \left(q_\phi(g|s_{1:T}) \,\|\, p_\theta(g|s_{1:t}) \right) \right]$$

Architecture Design



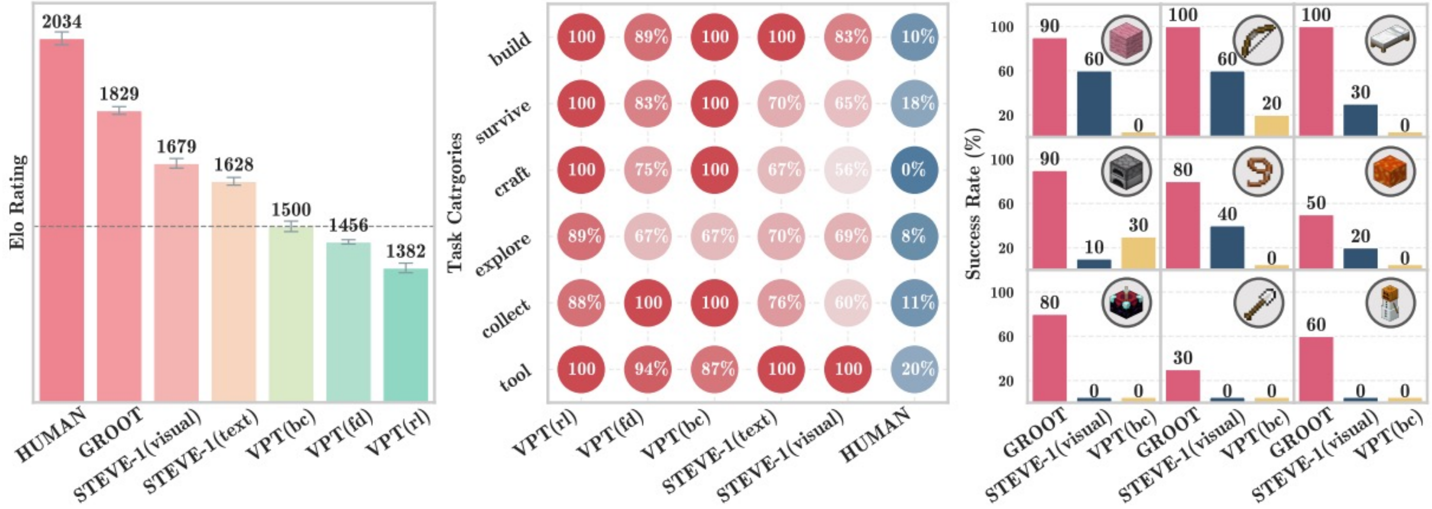
Inference:

Any reference video is passed into the video encoder to obtain goal embeddings that drive the policy to interact with the Minecraft environment.

Minecraft SkillForge Benchmark

We create a diverse benchmark called **Minecraft SkillForge**. It covers **30** tasks from **6** major categories of representative skills in Minecraft, including <u>*collect*</u>, <u>*explore*</u>, <u>*craft items*</u>, <u>*tool use*</u>, <u>*survive*</u>, and <u>*build*</u>.

Human Evaluation on Elo Rating System



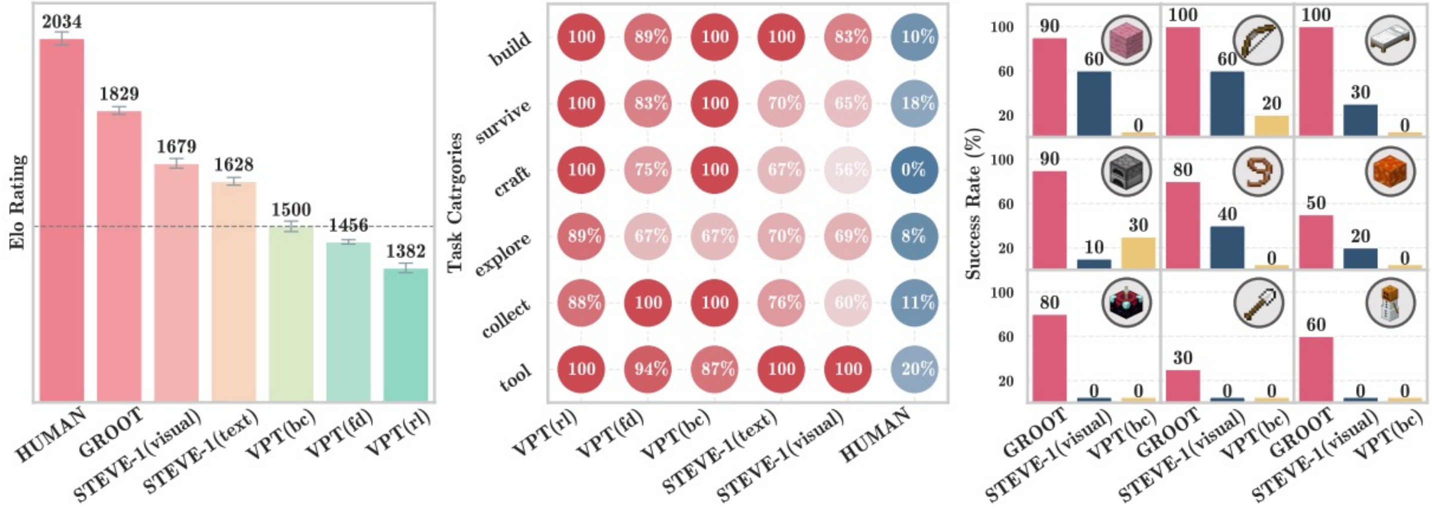
(a) Elo Rating Comparison

(b) Winning Rate of GROOT vs. Baselines

(c) Success Rate Comparison

Although there is a large performance gap compared with human players (2034), GROOT (1829) has significantly surpassed the current state-of-the-art STEVE-1 series (1679) and condition-free VPT series (1500).

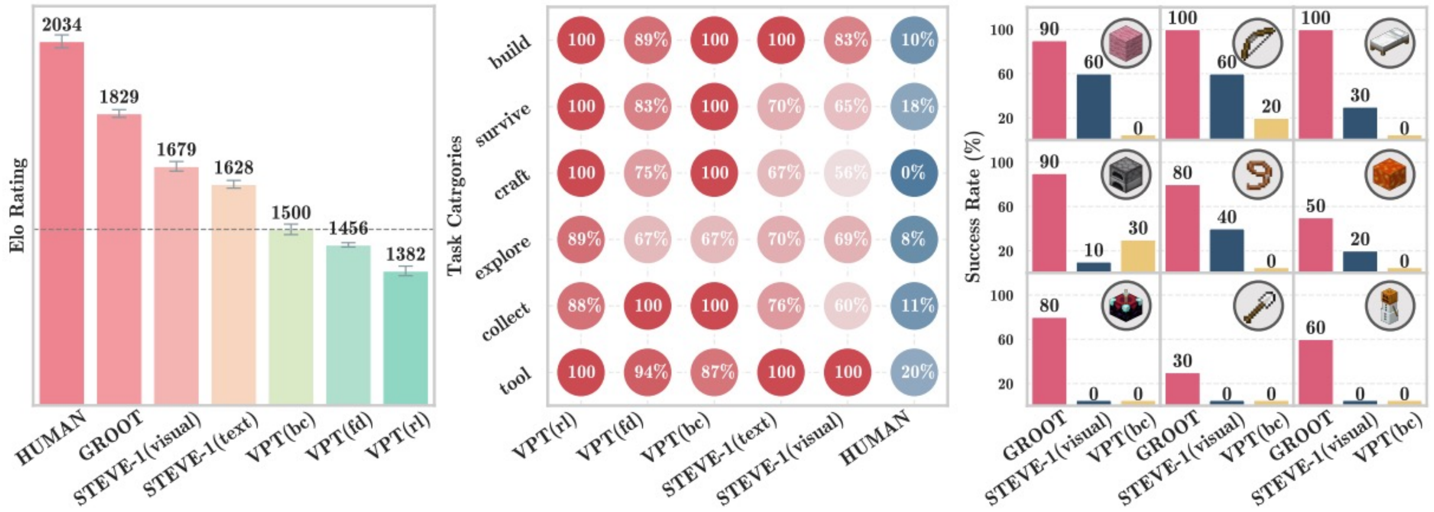
Human Evaluation on Elo Rating System



(a) Elo Rating Comparison (b) Winning Rate of GROOT vs. Baselines (c) Success Rate Comparison

On all the tasks, GROOT achieves over 50% winning rate against current SOTA baselines, especially on less common tasks “build” and “tool”.

Human Evaluation on Elo Rating System



(a) Elo Rating Comparison (b) Winning Rate of GROOT vs. Baselines (c) Success Rate Comparison

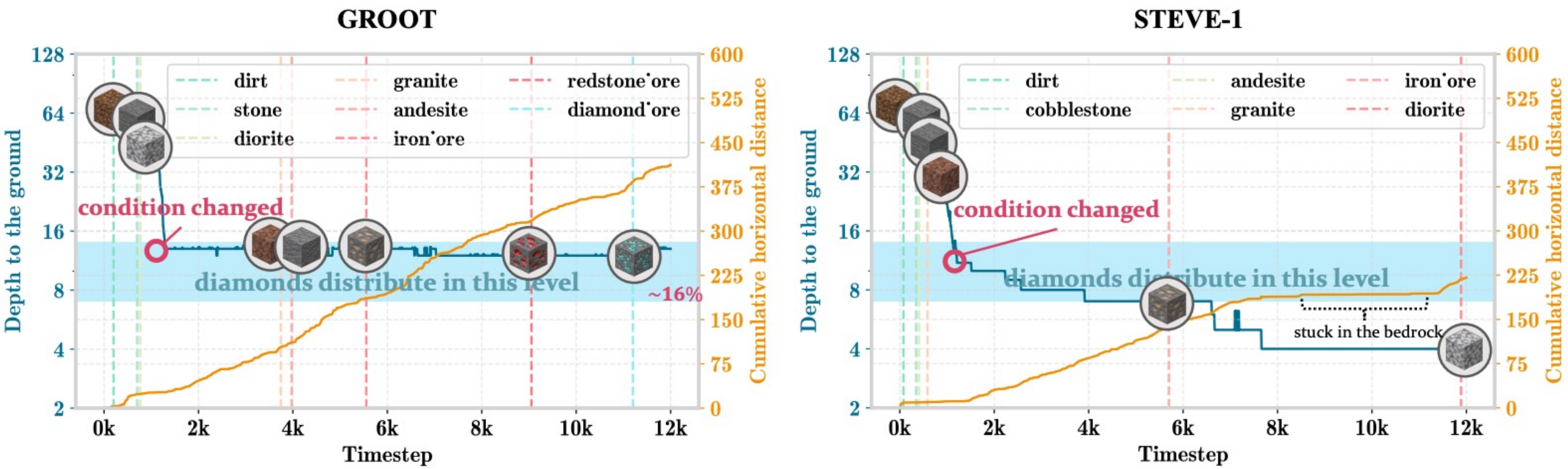
GROOT is the only that achieves non-zero success rate on challenging "enchantment", "dig 3 down fill 1 up", and "build snow golems" tasks.

Visualization of Goal Space

(a) Random Initialized **(b) GROOT w/o KL** **(c) GROOT w/ KL**

After being trained via self-supervised learning, the encoded video with the similar semantics are clustered together.

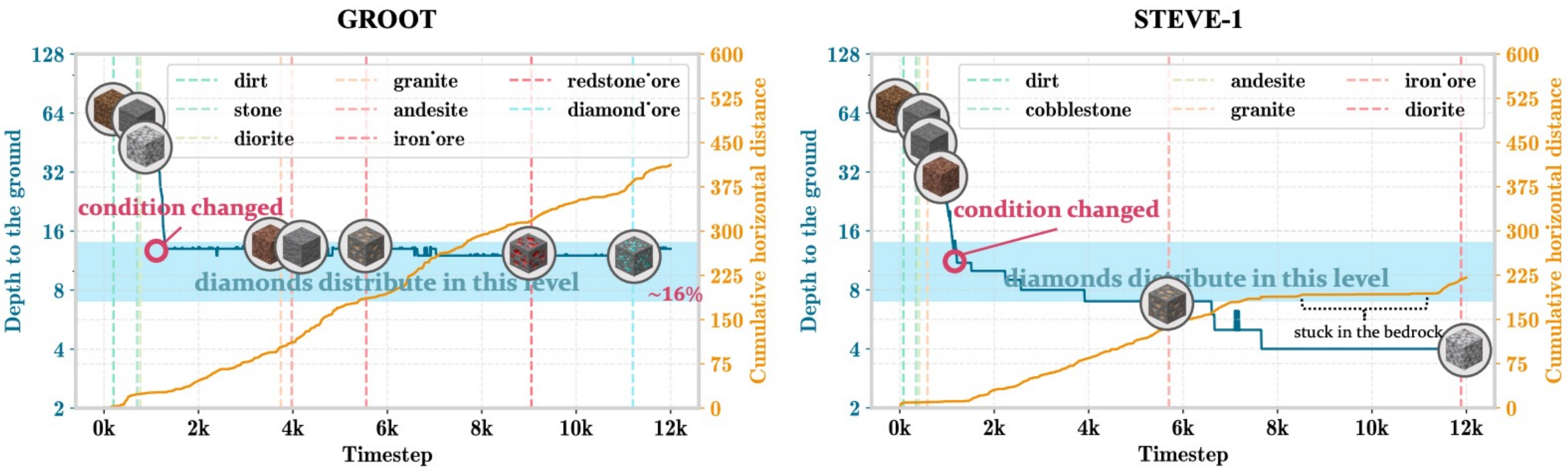
Chain of Instructions



By chaining “dig down” and “mine horizontally” instructions, GROOT achieves 16% diamond obtaining success rate with 10 minutes.

STEVE-1 struggle to obtain diamond because of inability of expressing mining horizontally.

Chain of Instructions



GROOT can be integrated with the LLM planner to solve complex and long-horizon tasks.

Thanks