Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos

Arzhantsev Andrey, 202

Sources

- github: https://github.com/openai/Video-Pre-Training/tree/main
- MineRL package: https://github.com/minerllabs/minerl
- competition: https://www.aicrowd.com/challenges/neurips-2022-minerl-basalt-competition
- blogpost: https://openai.com/research/vpt

Fine-Tune

- .mp4 video

https://drive.google.com/file/d/13frzJVAy4CjvcpEi7TLUPtWszIGvvgtc/view?usp=sharing

- .jsonl actions file

https://drive.google.com/file/d/1Wx47fllzua1Ztny4t65KfxXDap9T7Wod/view?usp=sharing

2022 minerl basalt competition

VPT to large to fine-tune

FindCave

MakeWaterfall

MakeVillageAnimalPen

BuildVillageHouse

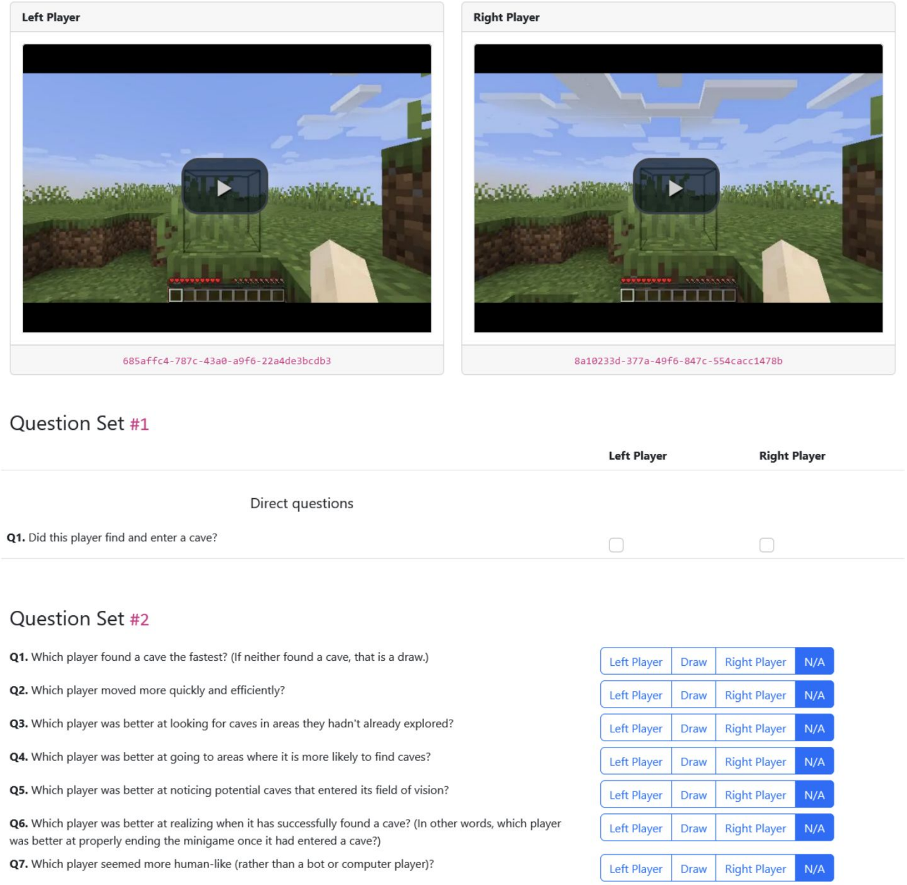
BEDD: The MineRL BASALT Evaluation and Demonstrations Dataset for Training and Benchmarking Agents that Solve Fuzzy Tasks

https://arxiv.org/pdf/2312.02405.pdf

Evaluation

BASALT competition environments do not include reward functions

Human-eval: Labeling who is better from pairs of gameplays



685affc4-787c-43a0-a9f6-22a4de3bcdb3

8a10233d-377a-49f6-847c-554cacc1478b

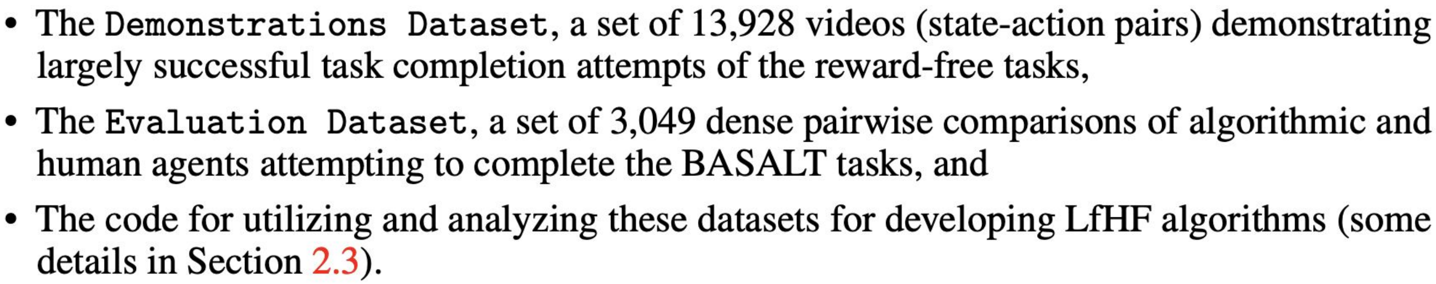
Question Set #1

	Left Player	Right Player
Direct questions		
Q1. Did this player find and enter a cave?	□	□

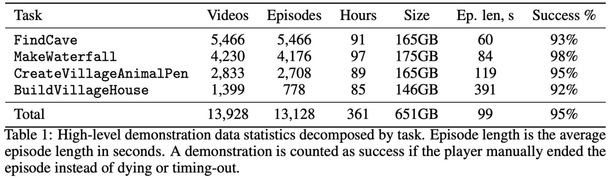
Question Set #2

Q1. Which player found a cave the fastest? (If neither found a cave, that is a draw.)	Left Player	Draw	Right Player	N/A
Q2. Which player moved more quickly and efficiently?	Left Player	Draw	Right Player	N/A
Q3. Which player was better at looking for caves in areas they hadn't already explored?	Left Player	Draw	Right Player	N/A
Q4. Which player was better at going to areas where it is more likely to find caves?	Left Player	Draw	Right Player	N/A
Q5. Which player was better at noticing potential caves that entered its field of vision?	Left Player	Draw	Right Player	N/A
Q6. Which player was better at realizing when it has successfully found a cave? (In other words, which player was better at properly ending the minigame once it had entered a cave?)	Left Player	Draw	Right Player	N/A
Q7. Which player seemed more human-like (rather than a bot or computer player)?	Left Player	Draw	Right Player	N/A

BEDD

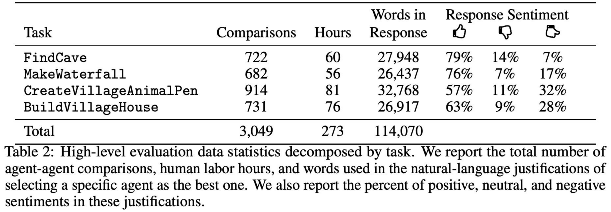


- The Demonstrations Dataset, a set of 13,928 videos (state-action pairs) demonstrating largely successful task completion attempts of the reward-free tasks,
- The Evaluation Dataset, a set of 3,049 dense pairwise comparisons of algorithmic and human agents attempting to complete the BASALT tasks, and
- The code for utilizing and analyzing these datasets for developing LfHF algorithms (some details in Section 2.3).



Task	Videos	Episodes	Hours	Size	Ep. len, s	Success %
FindCave	5,466	5,466	91	165GB	60	93%
MakeWaterfall	4,230	4,176	97	175GB	84	98%
CreateVillageAnimalPen	2,833	2,708	89	165GB	119	95%
BuildVillageHouse	1,399	778	85	146GB	391	92%
Total	13,928	13,128	361	651GB	99	95%

Table 1: High-level demonstration data statistics decomposed by task. Episode length is the average episode length in seconds. A demonstration is counted as success if the player manually ended the episode instead of dying or timing-out.



			Words in	Response Sentiment		
Task	Comparisons	Hours	Response	👍	👎	👉
FindCave	722	60	27,948	79%	14%	7%
MakeWaterfall	682	56	26,437	76%	7%	17%
CreateVillageAnimalPen	914	81	32,768	57%	11%	32%
BuildVillageHouse	731	76	26,917	63%	9%	28%
Total	3,049	273	114,070			

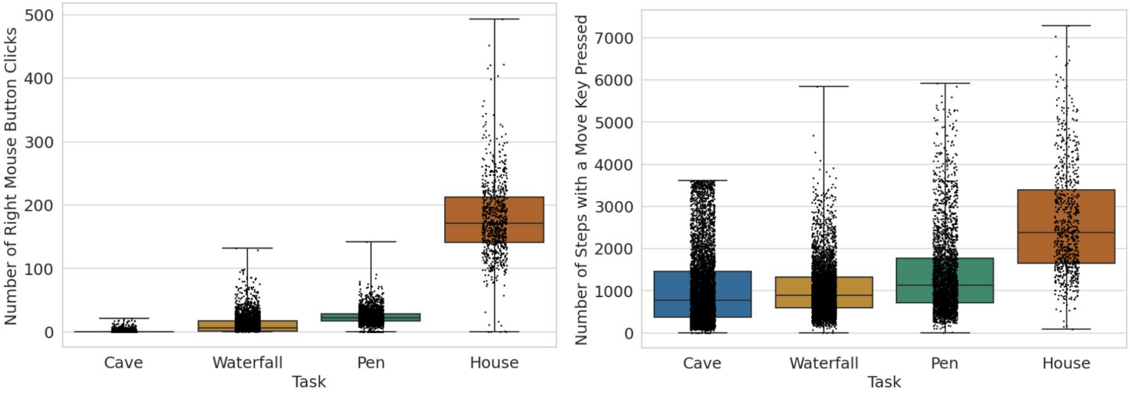
Table 2: High-level evaluation data statistics decomposed by task. We report the total number of agent-agent comparisons, human labor hours, and words used in the natural-language justifications of selecting a specific agent as the best one. We also report the percent of positive, neutral, and negative sentiments in these justifications.

Analysis (dataset)

general goal - define proxy metrics

difficulty == length of the demonstration

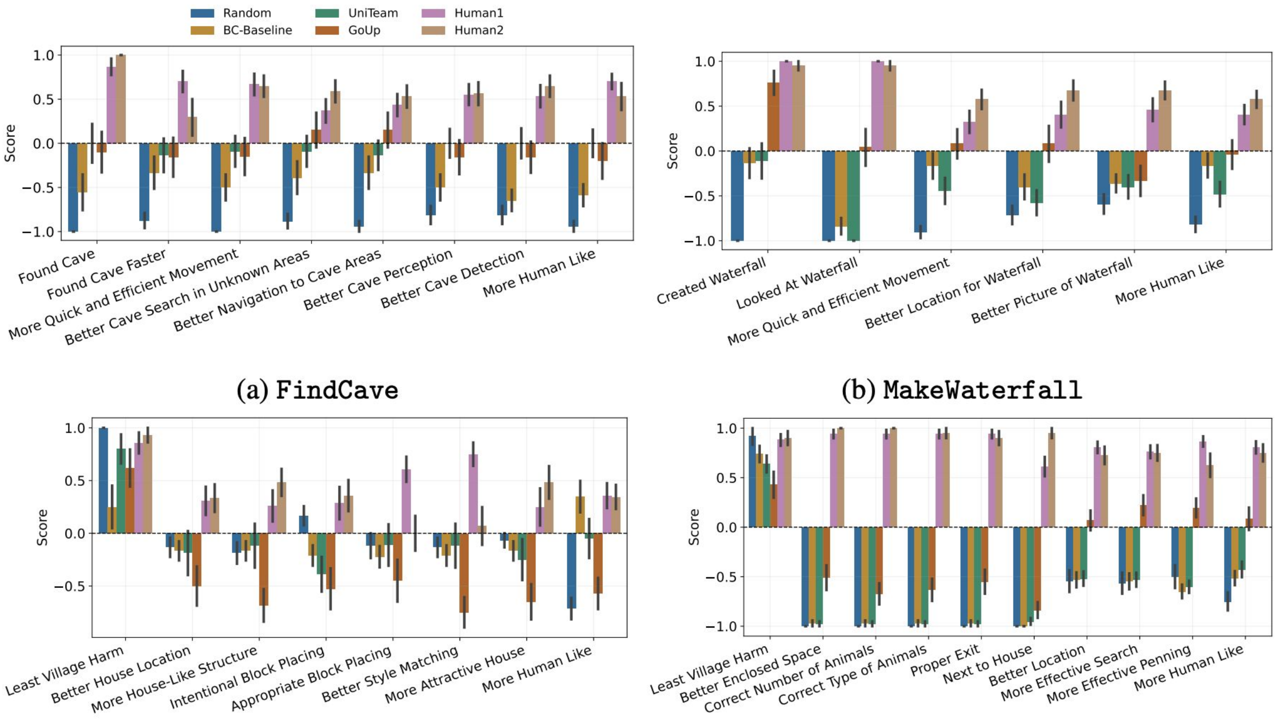
right mouse button clicks == the number of blocks placed



(a) Right mouse button clicks

(b) Movement key presses

A Retrospective of the MineRL BASALT 2022 Competition



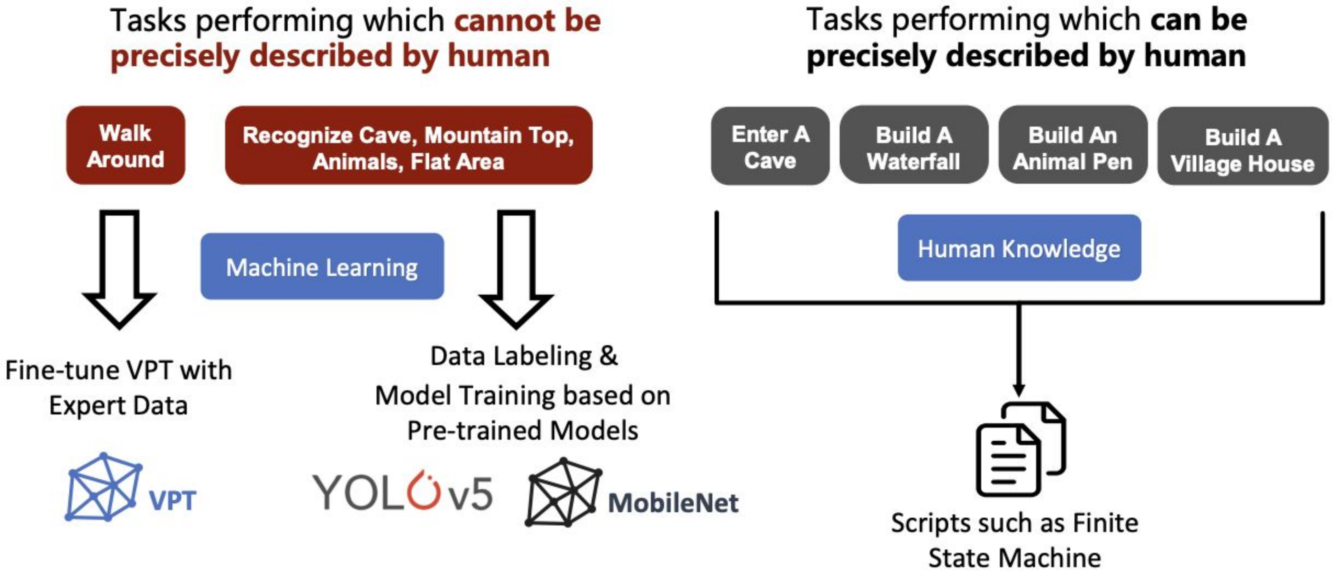
(a) `FindCave`

(b) `MakeWaterfall`

(c) `BuildVillageHouse`

(d) `AnimalPen`

GoUp



Tasks performing which **cannot be precisely described by human**

Walk Around

Recognize Cave, Mountain Top, Animals, Flat Area

Machine Learning

Fine-tune VPT with Expert Data

VPT

Data Labeling & Model Training based on Pre-trained Models

YOLOv5

MobileNet

Tasks performing which **can be precisely described by human**

Enter A Cave

Build A Waterfall

Build An Animal Pen

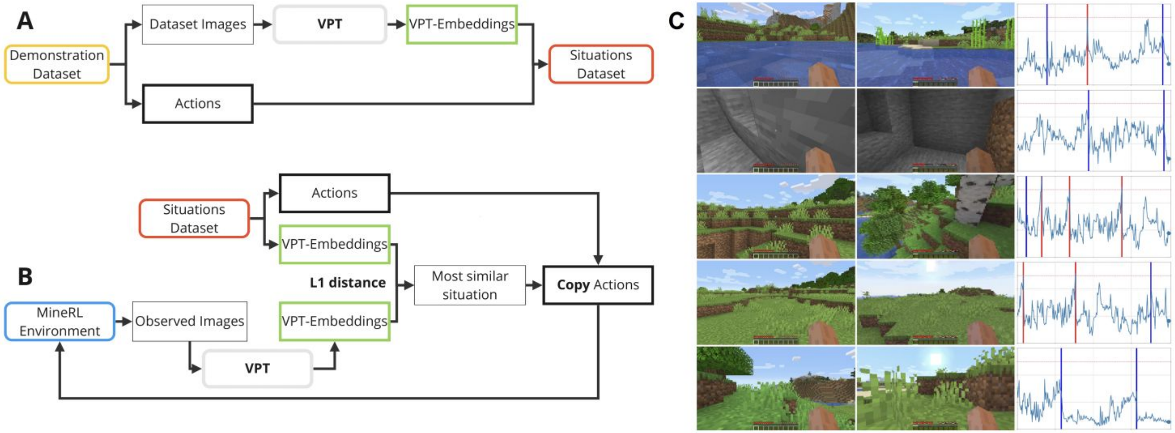
Build A Village House

Human Knowledge

Scripts such as Finite State Machine

UniTeam

L1 distance between their embedded current situation and the embedded situations from the expert’s dataset -> copy nearest action


A
Demonstration Dataset
Dataset Images
VPT
VPT-Embeddings
Actions
Situations Dataset
B
Situations Dataset
Actions
VPT-Embeddings
L1 distance
Most similar situation
Copy Actions
MineRL Environment
Observed Images
VPT-Embeddings
VPT
C