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

Speaker: Pirogov Slava
Reviewer-researcher: Vinogradova Uliana
Coder: Spirin Ivan

Minecraft? Who?

Wooden Log
280 actions
14 seconds

→

Wooden Plank
860 actions
43 seconds

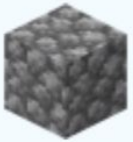
→

Crafting Table
960 actions
48 seconds

→

Wooden Pickaxe
1,390 actions
1.2 minutes

→



Cobblestone
2,050 actions
1.7 minutes

→

Stone Pickaxe
2,790 actions
2.3 minutes

Method Overview

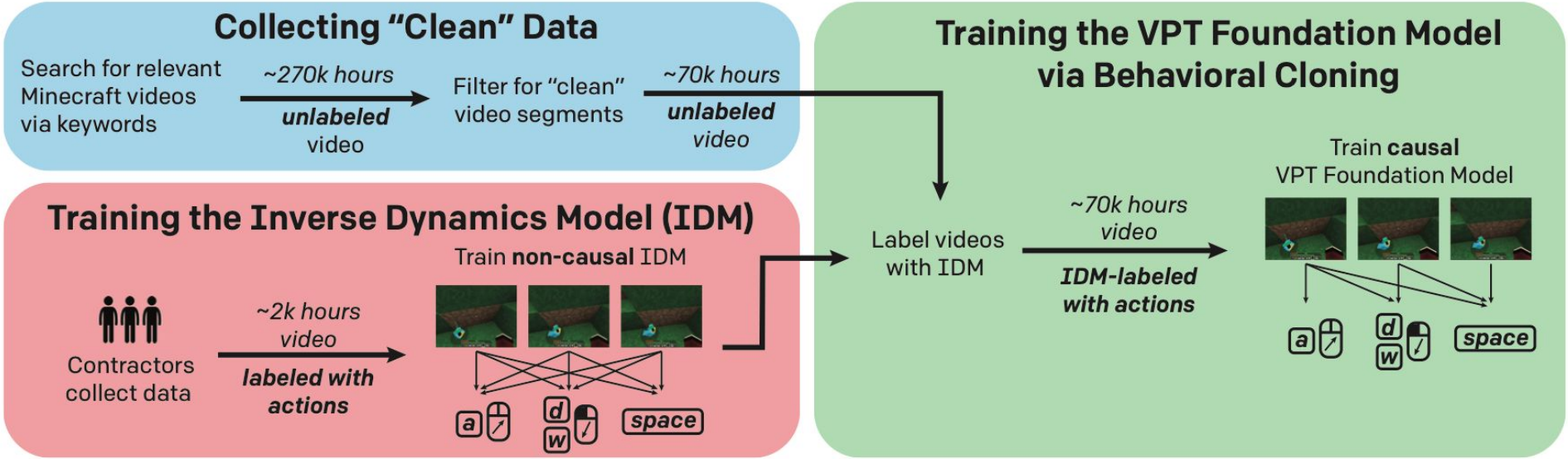


Figure 2: Video Pretraining (VPT) Method Overview.

Motivation

1. Large general foundation models are cool
2. Decision domains are complex
3. Most actively played game in the world
4. Wide action space
5. A lot of works by RL community

Method Overview

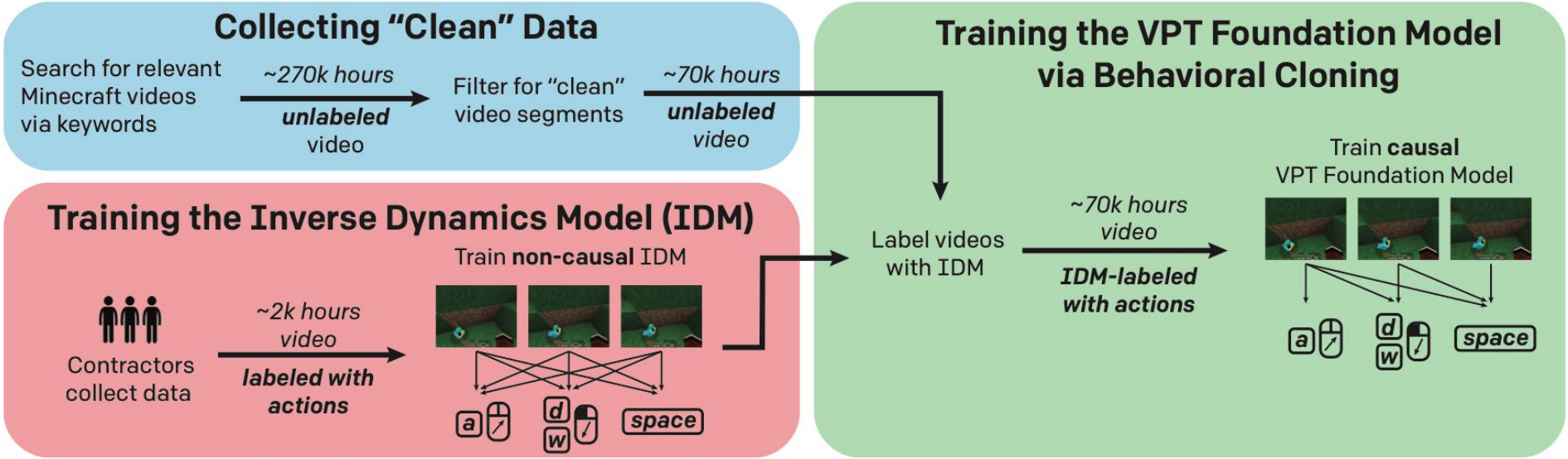


Figure 2: Video Pretraining (VPT) Method Overview.

Datasets. Unclean

keywords

minecraft survival longplay
minecraft gameplay no webcam
minecraft gameplay survival mode
minecraft survival tutorial
minecraft survival guide
minecraft survival let's play
minecraft survival for beginners
minecraft beginners guide
ultimate minecraft starter guide
minecraft survival guide 1.16
minecraft how to start a new survival world
minecraft survival fresh start
minecraft survival let's play episode 1
let's play minecraft episode 1
minecraft survival 101
minecraft survival learning to play
how to play minecraft survival
how to play minecraft
minecraft survival basic
minecraft survival for noobs
minecraft survival for dummies
how to play minecraft for beginners
minecraft survival tutorial series
minecraft survival new world
minecraft survival a new beginning
minecraft survival episodio 1
minecraft survival епизод 1
minecraft survival 1. bölüm
i made a new minecraft survival world

Blacklist keywords

{ps3, ps4, ps5, xbox 360, playstation, timelapse, multiplayer, minecraft pe, pocket edition, skyblock, realistic minecraft, how to install, how to download, realmcraft, animation}

270k hours total

Datasets. Cleanest

1. Contractors labeled a set of random video frames (images) from Minecraft videos (N=8800) for 3 classes

2. On “No Artifacts” we using ResNet-based CLIP model to get embeddings

3. Train SVM. 3 fps, if > 80% - clean, and more than 5 seconds - we take it in dataset

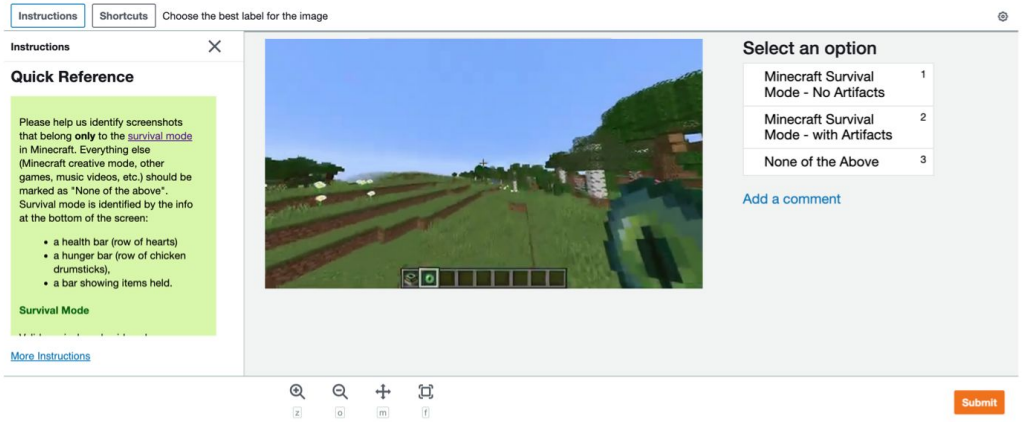


Figure 10: Amazon Mechanical Turk worker interface showing an example labeling task

Figure 11: **(Left)** Sample image for Class 1: `Minecraft Survival Mode - No Artifacts`. **(Middle)** Sample image for Class 2: `Minecraft Survival Mode - with Artifacts` – Image contains annotations and picture-in-picture of the narrator. **(Right)** Sample image for Class 3: `None of the Above` – Image is missing the hotbar as well as health and armor bars, indicating that it was not captured during survival mode gameplay

Method Overview

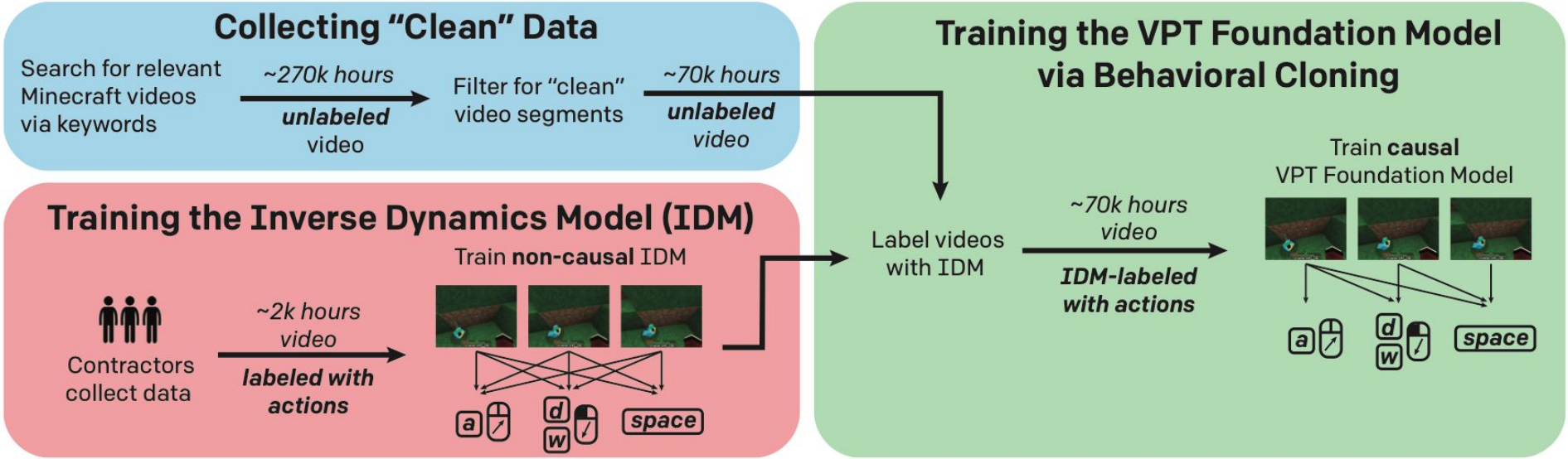


Figure 2: Video Pretraining (VPT) Method Overview.

Datasets. Contractor

“We are collecting data for training AI models in Minecraft. You’ll need to install java, download the modified version of Minecraft (that collects and uploads your play data), and play Minecraft survival mode! Paid per hour of gameplay. Prior experience in Minecraft not necessary. We do not collect any data that is unrelated to Minecraft from your computer.”

Paid 20$ per hour on UpWork freelancing platform (minus Upwork platform fees and applicable taxes)

Minecraft environment details

Image from 640x360 to 128x128

FPS from 60-100 to 20 for client and server

Similar settings like brightness and other graphics settings

Agent can die

Seems to be not determined seed (not sure)

Action space

Action	Human action	Description
`forward`	W key	Move forward.
`back`	S key	Move backward.
`left`	A key	Strafe left.
`right`	D key	Strafe right.
`jump`	space key	Jump.
`inventory`	E key	Open or close inventory and the 2x2 crafting grid.
`sneak`	shift key	Move carefully in current direction of motion. In the GUI it acts as a modifier key: when used with `attack` it moves item from/to the inventory to/from the hotbar, and when used with `craft` it crafts the maximum number of items possible instead of just 1.
`sprint`	ctrl key	Move fast in the current direction of motion.
`attack`	left mouse button	Attack; In GUI, pick up the stack of items or place the stack of items in a GUI cell; when used as a double click (attack - no attack - attack sequence), collect all items of the same kind present in inventory as a single stack.
`use`	right mouse button	Place the item currently held or use the block the player is looking at. In GUI, pick up the stack of items or place a single item from a stack held by mouse.
`drop`	Q key	Drop a single item from the stack of items the player is currently holding. If the player presses ctrl-Q then it drops the entire stack. In the GUI, the same thing happens except to the item the mouse is hovering over.
`hotbar.[1-9]`	keys 1 – 9	Switch active item to the one in a given hotbar cell.

Only disallow typing arbitrary letters (useful for search bar in craft book)

Discretize mouse position by 11 bins for each axis (X and Y)

Method Overview

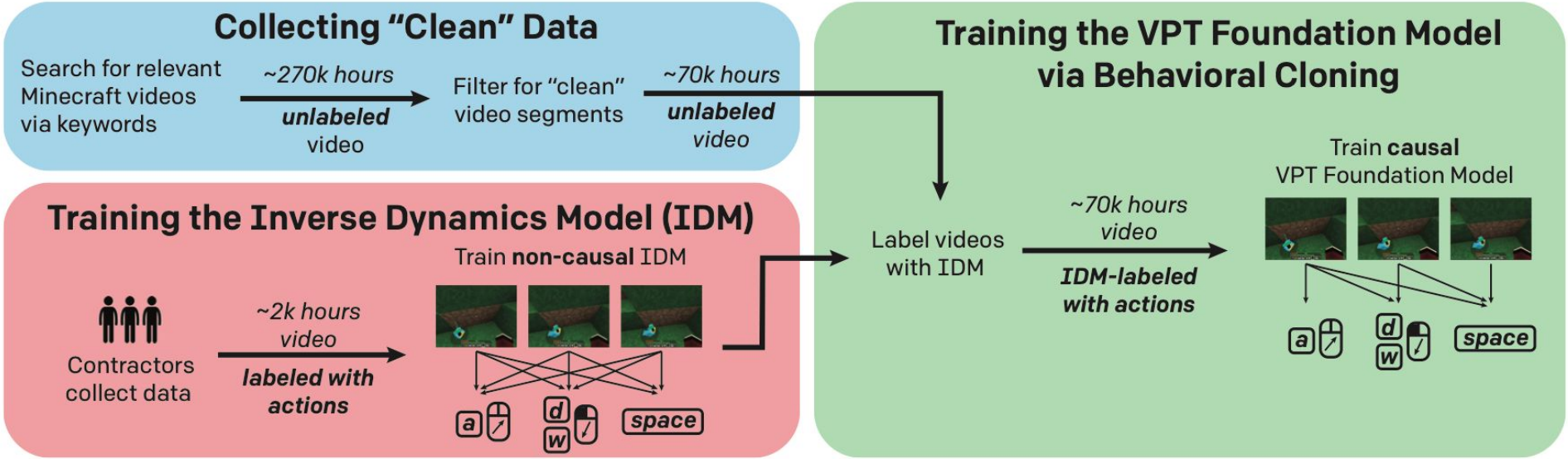
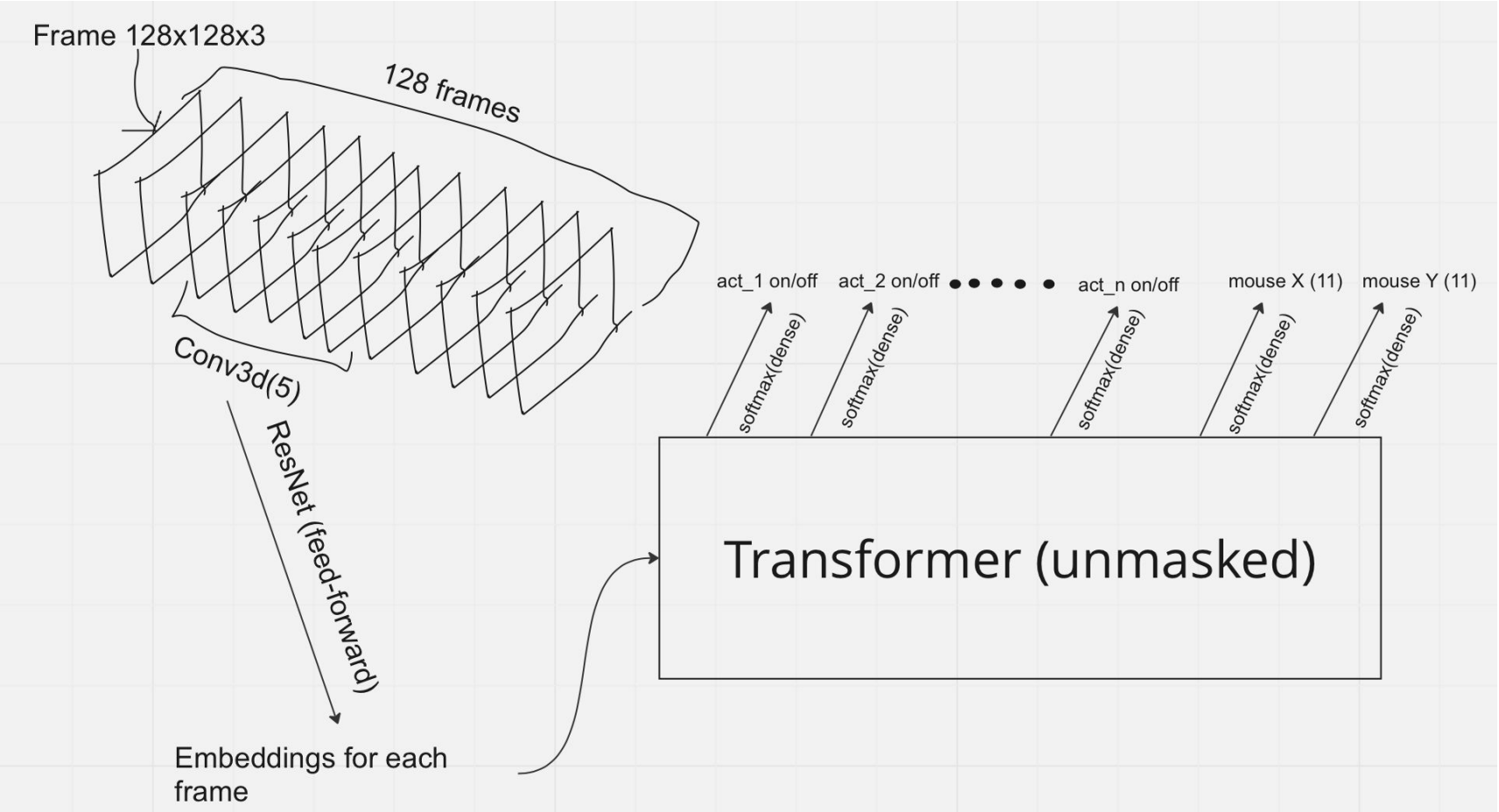


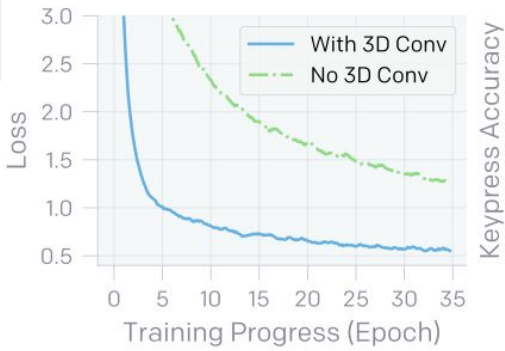
Figure 2: Video Pretraining (VPT) Method Overview.

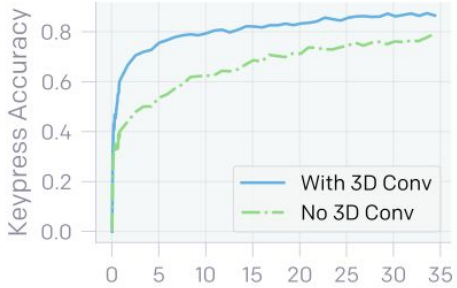
Inverse Dynamics Model

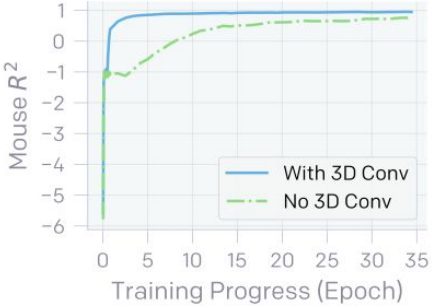


Inverse Dynamics Model

1. Looking in the future
2. 0.5 billion trainable weights
3. 3D convolutional is really important
4. Non-casual. At time index ***t*** we are looking at ***t-2, t-1, t, t+1, t+2***
5. ResNet
6. Transformer
7. LayerNorm + ReLU. Fan-In initialization, biases are zero
8. 4 days on 32 A100 GPUs







Cleaning Dataset with Inverse Dynamics

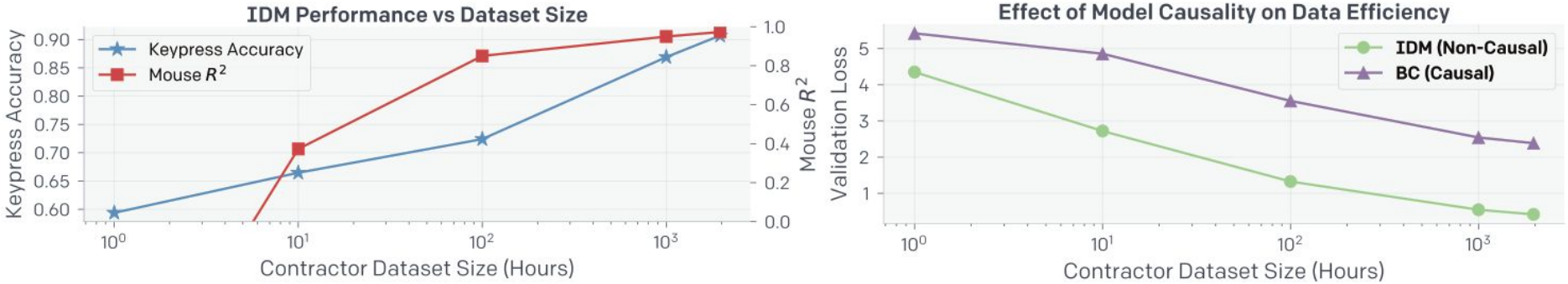


Figure 3: **(Left)** IDM keypress accuracy and mouse movement R^2 (explained variance[61]) as a function of dataset size. **(Right)** IDM vs. behavioral cloning data efficiency.

Method Overview

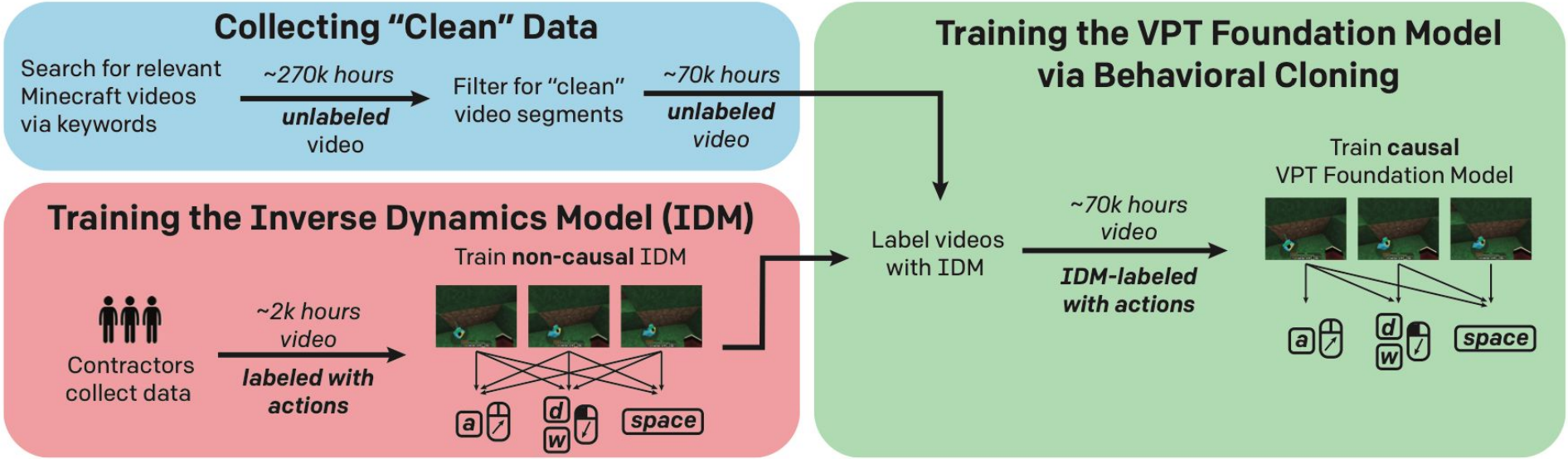
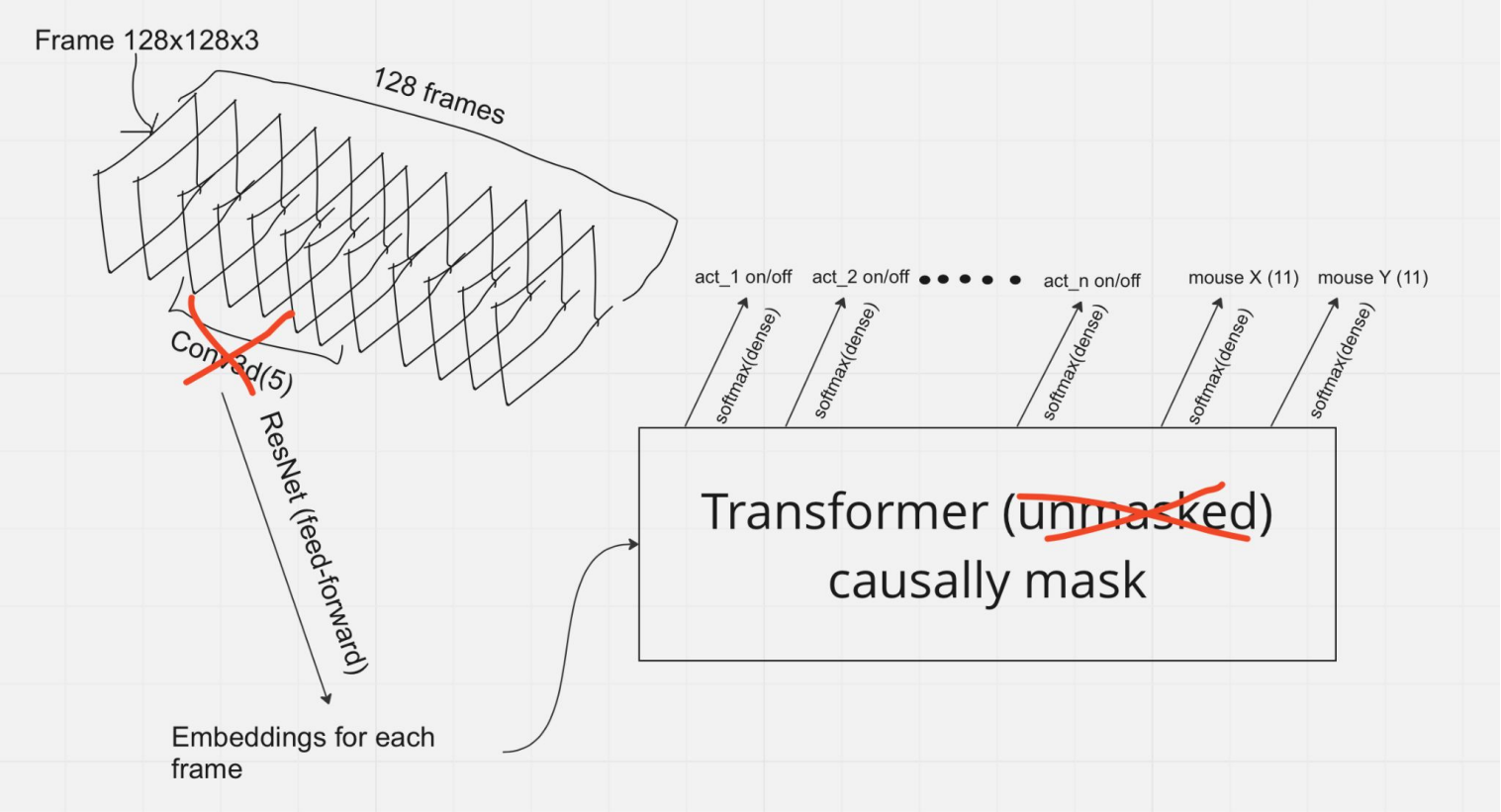


Figure 2: Video Pretraining (VPT) Method Overview.

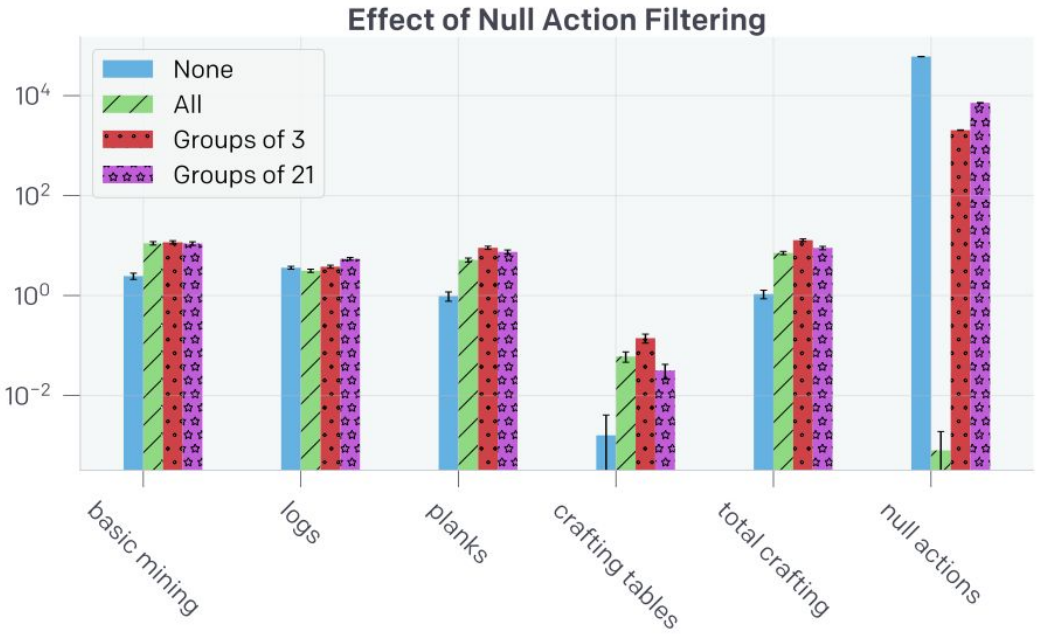
Foundation Model Behaviour Cloning



Transformer-XL style

Foundation Model Behaviour Cloning. Troubles

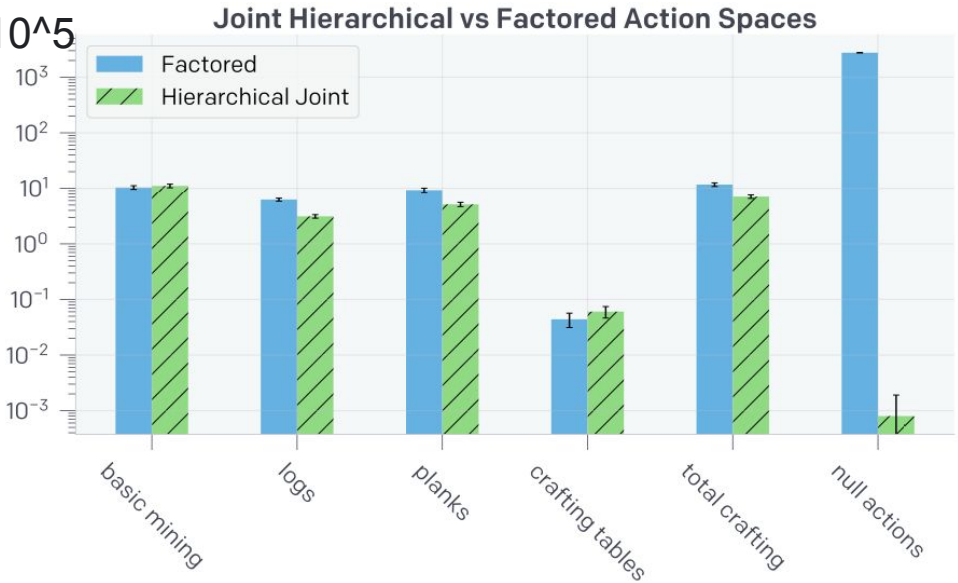
- Humans have 35% null actions - BC model 95%
- Remove null if 1/3/21 null frames



Foundation Model
Behaviour Cloning. Troubles

Behaviour distribution

1. Factored action space - bad approach
2. Full joint distribution - 2^20 x 11^2 ≈ 1.2 * 10^8
3. Ban mutually exclusive actions ≈ 5.2 * 10^5
4. Hierarchical binary action for camera - 2 head with 121 and 8461 dimension



Method Overview

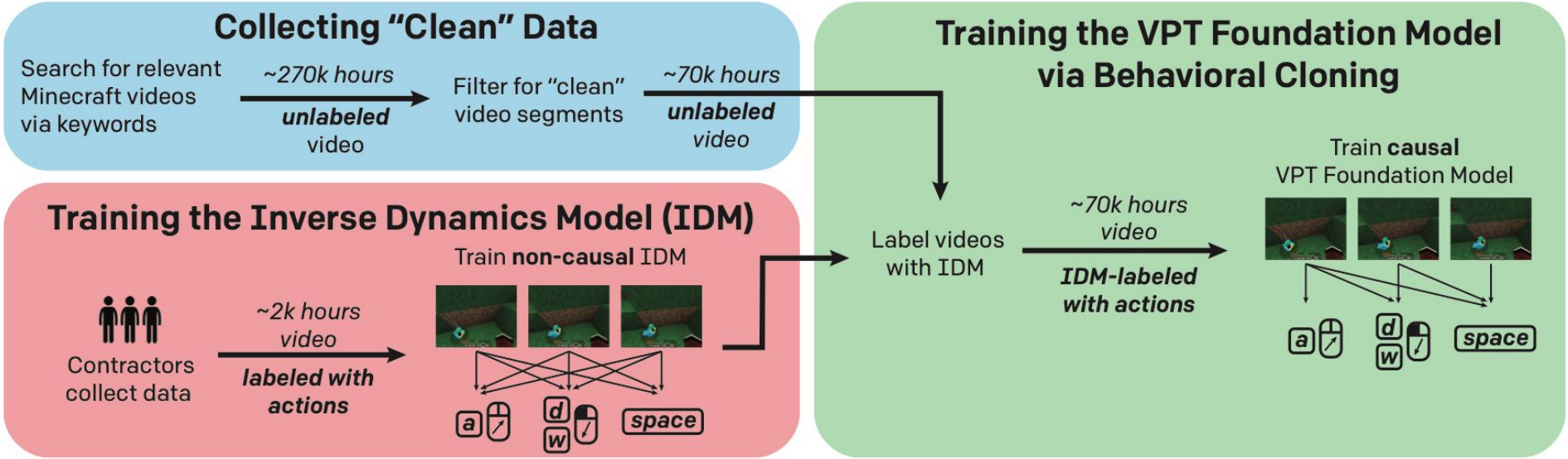
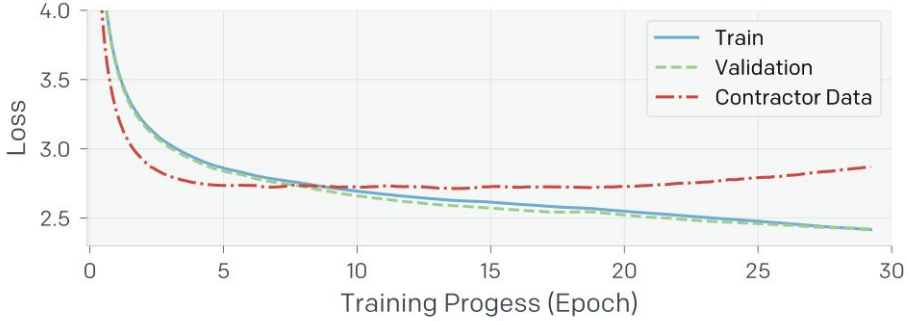


Figure 2: Video Pretraining (VPT) Method Overview.

VPT zero-shot

https://www.youtube.com/playlist?list=PLNAOIb_agjf3U3rSvG_BCWqJ869NdBhcP

https://openai.com/blog/vpt/



Behaviour Cloning Fine-Tune

Similar as casual BC

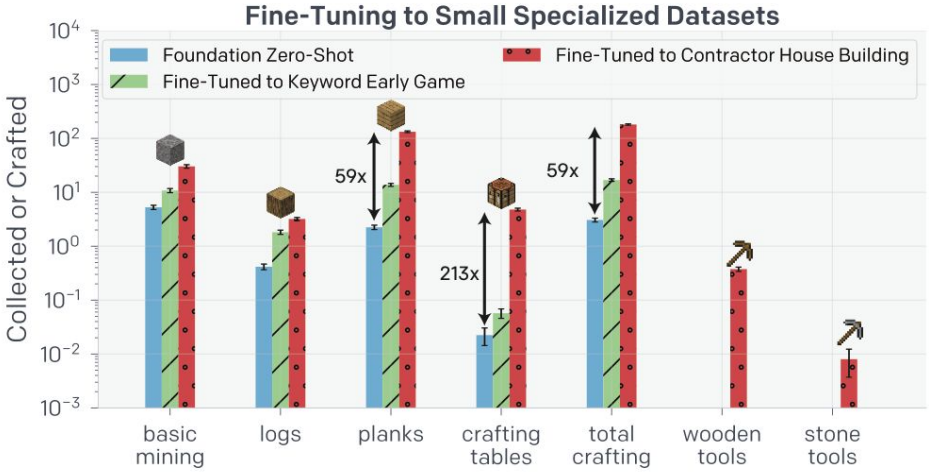
early_game dataset - 16xA100 GPUs for 6 hours

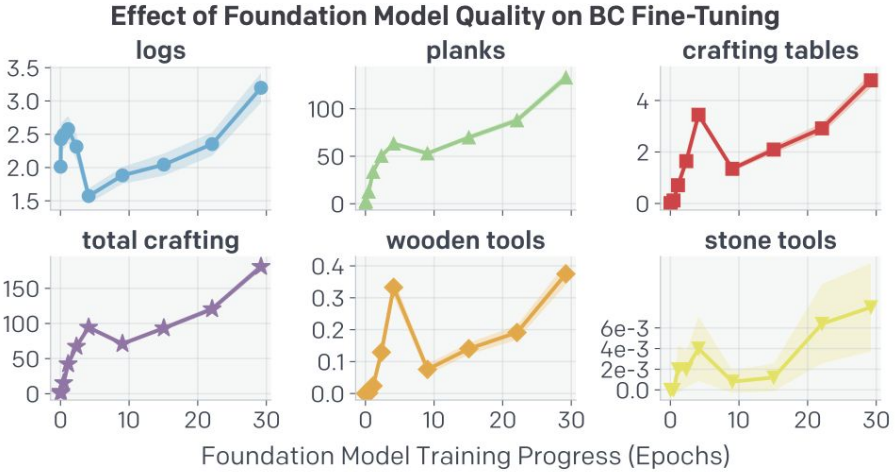
or contractor_house dataset - 16xA100 for 2 days

Behaviour Cloning Fine-Tune

https://www.youtube.com/playlist?list=PLNAOIb_agjf2yDSs4AqcoyPv4z_eWUiKm

https://openai.com/blog/vpt/





Reinforcement Learning Fine-Tune

Phasic Policy Gradient (PPG) - based on Proximal Policy Optimization (PPO)

Generalized Advantage Estimation (GAE)

!Applying KL loss between RL model and frozen pretrained model

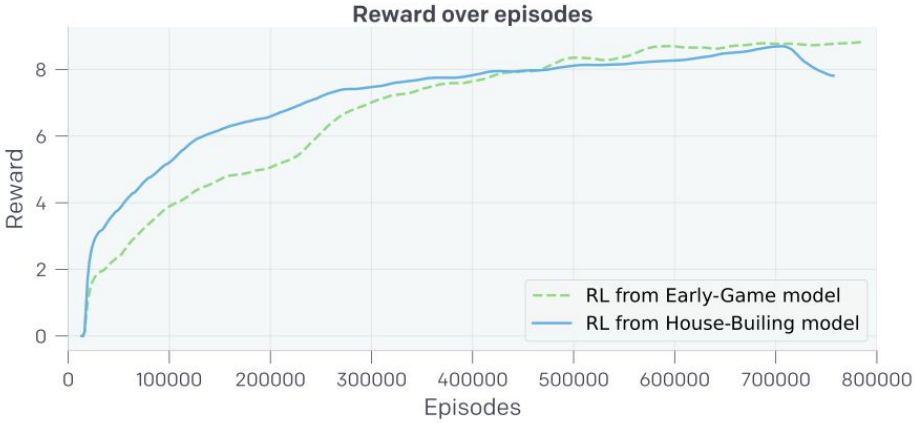
144 hours on 80 GPUs and 56,719 CPUs

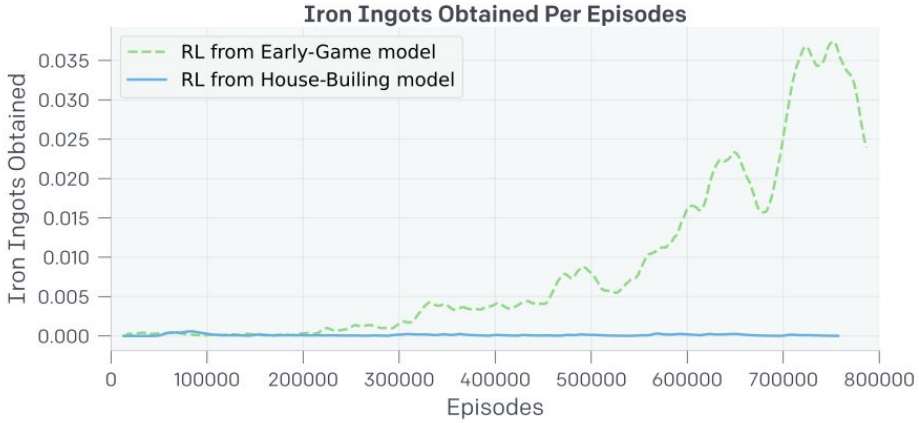
Reinforcement Learning Fine-Tune

Item	Quantity rewarded	Reward per item
Log	8	1/8
Planks	20	1/20
Stick	16	1/16
Crafting table	1	1
Wooden pickaxe	1	1
Cobblestone	11	1/11
Stone pickaxe	1	1
Furnace	1	1
Coal	5	2/5
Torch	16	1/8
Iron ore	3	4/3
Iron ingot	3	4/3
Iron pickaxe	1	4
Diamond	inf	8/3
Diamond pickaxe	inf	8

Table 7: Reward per item and total quantity rewarded.

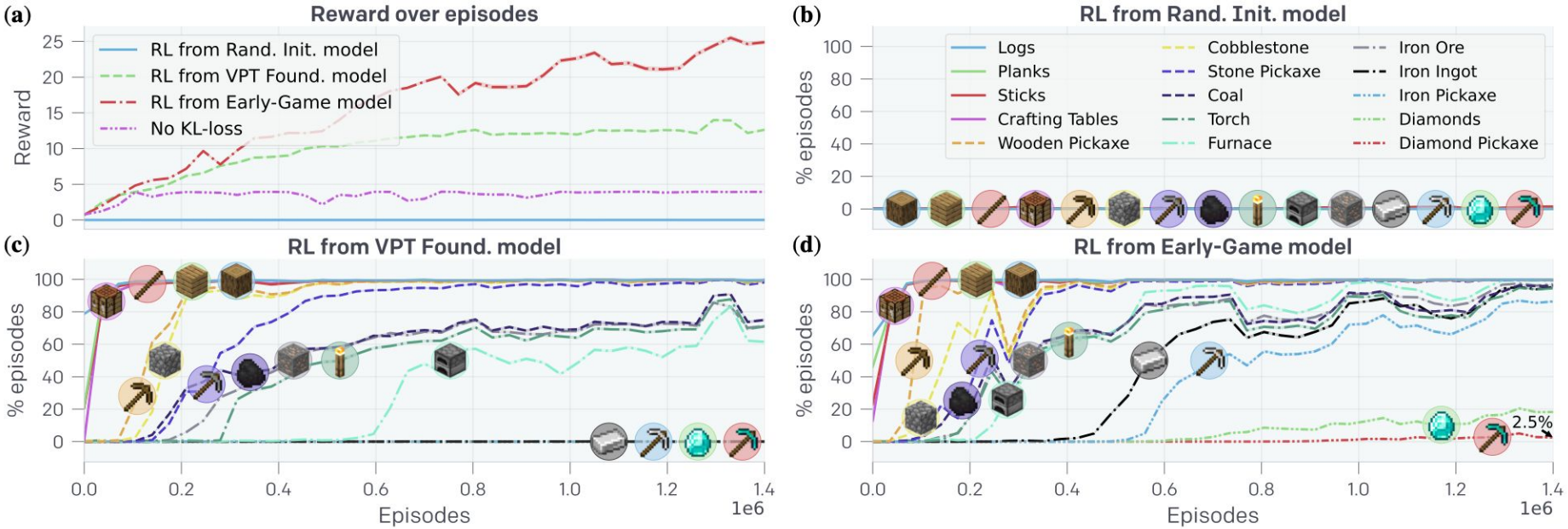
Reinforcement Learning Fine-Tune



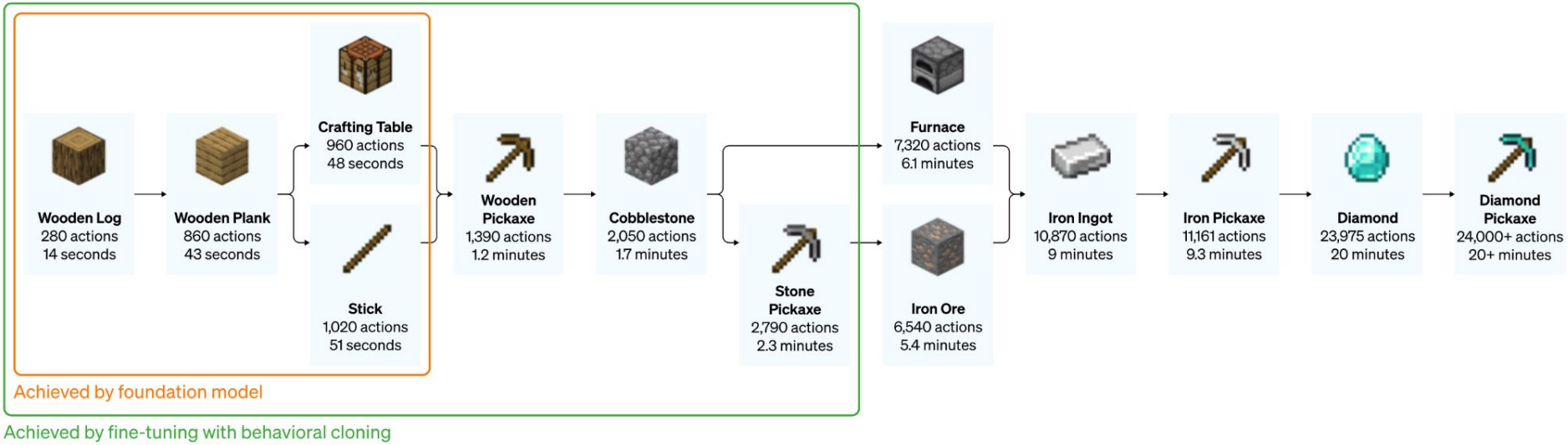


Reinforcement Learning Fine-Tune

https://www.youtube.com/playlist?list=PLNAOIb_agjf3e_UKweM5pQUSfTw8r-Wfc

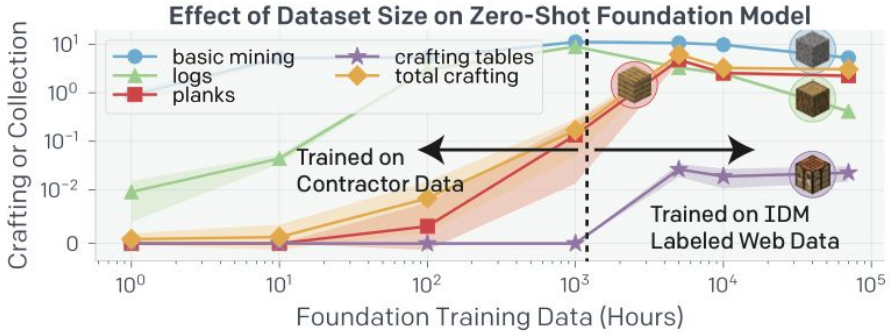


Reinforcement Learning Fine-Tune



Achieved by fine-tuning with behavioral cloning

Data Scaling Properties of the Foundation Model



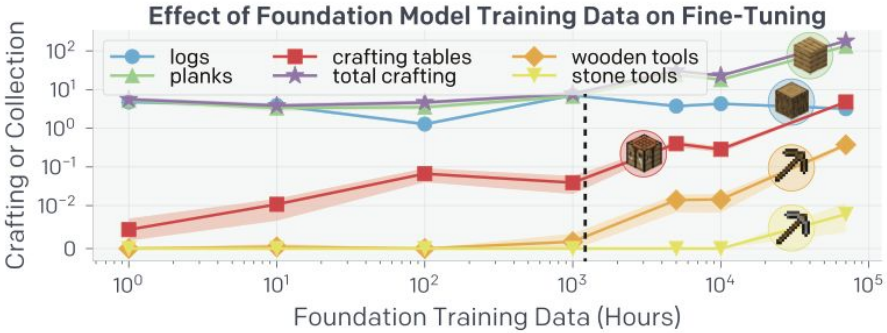
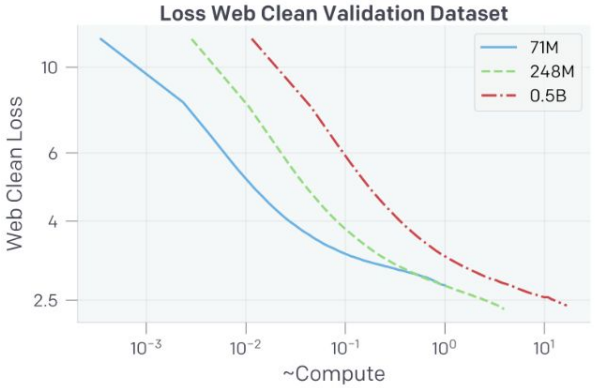
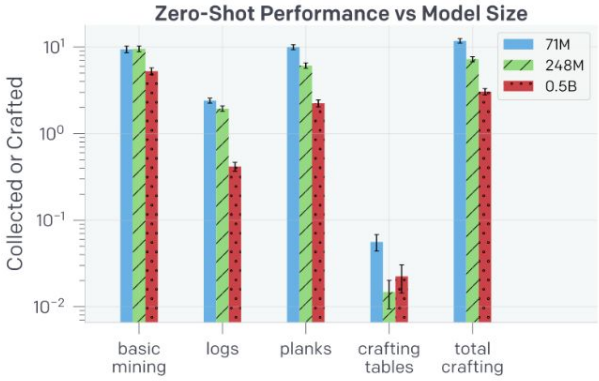
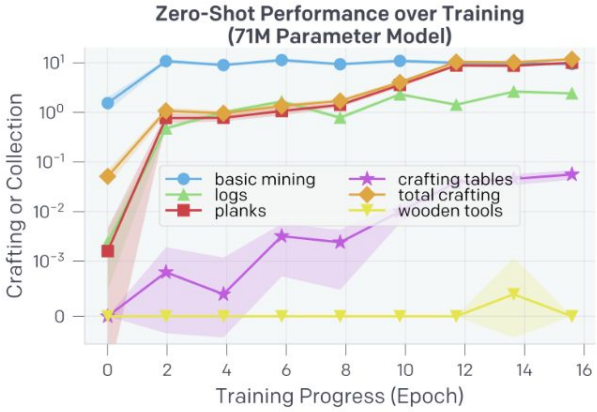


Figure 8: **(Left)** Zero-shot rollout performance of foundation models trained on varying amounts of data. Models to the left of the dashed black line (points $\leq$1k hours) were trained on contractor data (ground-truth labels), and models to the right were trained on IDM pseudo-labeled subsets of `web_clean`. Due to compute limitations, this analysis was performed with smaller (71 million parameter) models except for the final point, which is the 0.5 billion parameter VPT foundation model. **(Right)** The corresponding performance of each model *after* BC fine-tuning each model to the `contractor_house` dataset.

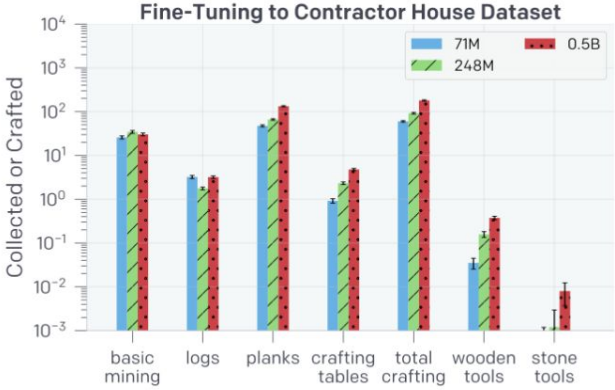
Foundation Model scaling







Foundation Model scaling



Results

Now we can label videos in decision domain and then training some models

Navigating websites, Photoshop, booking flights, …

Money for IDM labeling - 2000 (100) hours * 20$ + $12k = $40k + $12k (datashpere prices) - 70k hours of data

Text conditioning