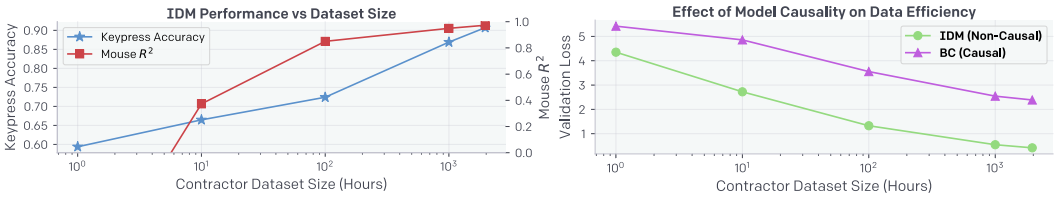

Figure 3: **(Left)** IDM keypress accuracy and mouse movement R^2 (explained variance[61]) as a function of dataset size. **(Right)** IDM vs. behavioral cloning data efficiency.

IDM trains on only 1962 hours of data (compared to the ∼70k hours of clean data we collected from the internet) and achieves 90.6% keypress accuracy and a 0.97 R^2 for mouse movements evaluated on a held-out validation set of contractor-labeled data (Figure 3 left).

Figure 3 (right) validates our hypothesis that IDMs are far more data efficient than BC models, likely because inverting environment mechanics is far easier than modeling the entire distribution of human behavior. The IDM is two orders of magnitude more data efficient than a BC model trained on the same data and improves more quickly with more data. This evidence supports the hypothesis that it is more effective to use contractor data within the VPT pipeline by training an IDM than it is to train a foundation model from contractor data directly (Sections 4.5 and 4.6 provide additional evidence).

4.2 VPT Foundation Model Training and Zero-Shot Performance

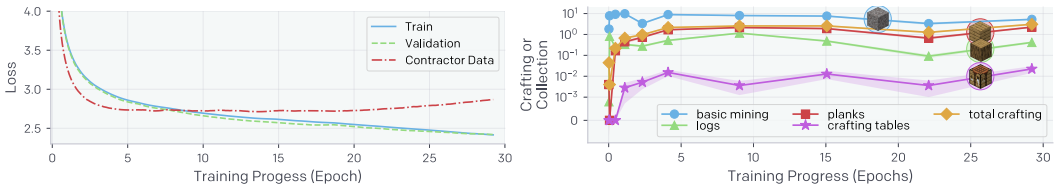

Figure 4: **(Left)** Training and validation loss on the `web_clean` internet dataset with IDM pseudo-labels, and loss on the main IDM contractor dataset, which has ground-truth labels but is out-of-distribution (see text). **(Right)** Amount a given item was collected per episode averaged over 2500 60-minute survival episodes as a function of training epoch, shaded with the standard error of the mean. Basic mining refers to collection of dirt, gravel, or sand (all materials that can be gathered without tools). Logs are obtained by repeatedly hitting trees for three seconds, a difficult feat for an RL agent to achieve as we show in Sec. 4.4. Planks can be crafted from logs, and crafting tables crafted from planks. Crafting requires using in-game crafting GUIs, and proficient humans take a median of 50 seconds (970 consecutive actions) to make a crafting table.

We now explore the emergent behavior learned by a behavioral cloning policy trained on an extremely large, but noisy, internet dataset labeled with our IDM. To collect the unlabeled internet dataset, we searched for publicly available videos of Minecraft play with search terms such as "minecraft survival for beginners." These searches resulted in ∼270k hours of video, which we filtered down to "clean" video segments yielding an *unlabeled* dataset of ∼70k hours, which we refer to as `web_clean` (Appendix A has further details on data scraping and filtering). We then generated pseudo-labels for `web_clean` with our best IDM (Section 3) and then trained the VPT foundation model with behavioral cloning. Preliminary experiments suggested that our model could benefit from 30 epochs of training and that a 0.5 billion parameter model was required to stay in the efficient learning regime[63] for that training duration (Appendix H), which took ∼9 days on 720 V100 GPUs.

We evaluate our models by measuring validation loss (Fig. 4, left) and rolling them out in the Minecraft environment. Unless otherwise noted, in all environment evaluations we spawn agents in a standard survival mode game where they play for 60 minutes, i.e. 72000 consecutive actions, and we plot the mean and shade the standard error of the mean for various game statistics such as crafting and collection rates (Fig. 4, right). The VPT foundation model quickly learns to chop down trees to collect logs, a task we found near impossible for an RL agent to achieve with the native human interface (Sec. 4.4). It also learns to craft those logs into wooden planks and then use those planks