diferente. Se só me faltassem os outros, vá um homem consola-se mais ou menos das pessoas que perde; mais falto eu mesmo, e esta lacuna é tudo. O que aqui está é, mal comparando, semelhante à pintura que se põe na barba e nos cabelos, e que apenas conserva o hábito externo, como se diz nas autópsias; o interno não agüenta tinta. Uma certidão que me desse vinte anos de idade poderia enganar os estranhos, como todos os documentos falsos, mas não a mim. Os amigos que me restam são de data recente; todos os antigos foram estudar a geologia dos campos-santos. Quanto às amigas, algumas datam de quinze anos, outras de menos, e quase todas crêem na mocidade. Duas ou três fariam crer nela aos outros, mas a língua que falam obriga muita vez a consultar os dicionários, e tal freqüência é cansativa.

Entretanto, vida diferente não quer dizer vida pior, é outra cousa a certos respeitos, aquela vida antiga aparece-me despida de muitos encantos que lhe achei; mas é também exato que perdeu muito espinho que a fez molesta, e, de memória, conservo alguma recordação doce e feiticeira. Em verdade, pouco apareço e menos falo. Distrações raras. O mais do tempo é gasto em hortar, jardinar e ler; como bem e não durmo mal.

Ora, como tudo cansa, esta monotonia acabou por exaurir-me também. Quis variar, e lembrou-me escrever um livro. Jurisprudência. filosofia e política acudiram-me, mas não me acudiram as forças necessárias. Depois, pensei em fazer uma "História dos Subúrbios" menos seca que as memórias do Padre Luís Gonçalves dos Santos relativas à cidade; era obra modesta, mas exigia documentos e datas como preliminares, tudo árido e longo. Foi então que os bustos pintados nas paredes entraram a falar-me e a dizer-me que, uma vez que eles não alcançavam reconstituir-me os tempos idos, pegasse da pena e contasse alguns. Talvez a narração me desse a ilusão, e as sombras viessem perpassar ligeiras, como ao poeta, não o do trem, mas o do Fausto: Aí vindes outra vez, inquietas sombras?...

Fiquei tão alegre com esta idéia, que ainda agora me treme a pena na mão. Sim, Nero, Augusto, Massinissa, e tu, grande César, que me incitas a fazer os meus comentários, agradeço-vos o conselho, e vou deitar ao papel as reminiscências que me vierem vindo. Deste modo, viverei o que vivi, e assentarei a mão para alguma obra de maior tomo. Eia, comecemos a evocação por uma célebre tarde de novembro, que nunca me esqueceu. Tive outras muitas, melhores, e piores, mas aquela nunca se me apagou do espírito. É o que vais entender, lendo.

CAPÍTULO III/ A DENÚNCIA

Ia entrar na sala de visitas, quando ouvi proferir o meu nome e escondi-me atrás da porta. A casa era a da Rua de Mata-cavalos, o mês novembro, o ano é que é um tanto remoto, mas eu não hei de trocar as datas à minha vida só para agradar às pessoas que não amam histórias velhas; o ano era de 1857.

--D. Glória, a senhora persiste na idéia de meter o nosso Bentinho no seminário? É mais que tempo, e já agora pode haver uma dificuldade.

--Que dificuldade?

--Uma grande dificuldade.

Minha mãe quis saber o que era. José Dias, depois de alguns instantes de concentração, veio ver se havia alguém no corredor; não deu por mim, voltou e, abafando a voz, disse que a dificuldade estava na casa ao pé, a gente do Pádua.

--A gente do Pádua?

--Há algum tempo estou para lhe dizer isto, mas não me atrevia. Não me parece bonito que o nosso Bentinho ande metido nos cantos com a filha do Tartaruga, e esta é a dificuldade, porque se eles pegam de namoro, a senhora terá muito que lutar para separá-los.

--Não acho. Metidos nos cantos?

--É um modo de falar. Em segredinhos, sempre juntos. Bentinho quase que não sai de lá. A pequena

already exist large datasets with action labels from various online platforms that researchers have used for imitation learning.[9,10] When large labeled datasets do not exist, the canonical strategy for training capable agents is reinforcement learning (RL),[11] which can be sample inefficient and expensive for hard-exploration problems.[12–18] Many virtual tasks, e.g. navigating websites, using Photoshop, booking flights, etc., can be very hard to learn with RL and do not have large, commonly available sources of labeled data.[19,20] In this paper, we seek to extend the paradigm of training large, general-purpose foundation models to sequential decision domains by utilizing freely available internet-scale unlabeled video datasets with a simple semi-supervised imitation learning method. We call this method Video PreTraining (VPT) and demonstrate its efficacy in the domain of Minecraft.

Existing semi-supervised imitation learning methods aim to learn with few or no explicit action labels; however, they generally rely on the policy's ability to explore the environment throughout training, making them susceptible to exploration bottlenecks.[21–25] Furthermore, most prior semi-supervised imitation learning work was tested in the relatively low data regime; because we experiment with *far* more data ($\sim$70k hours of unlabeled video), we hypothesize that we can achieve good performance with a much simpler method, a trend that has proven true for pretraining in other modalities such as text.[1] In particular, given a large but unlabeled dataset, we propose generating pseudo-labels by gathering a small amount of labeled data to train an inverse dynamics model (IDM) that predicts the action taken at each timestep in a video. Behavioral cloning (BC) can require a large amount of data because the model must learn to infer intent and the distribution over future behaviors from only past observations. In contrast, the inverse dynamics modeling task is simpler because it is *non-causal*, meaning it can look at both past and future frames to infer actions. In most settings, environment mechanics are far simpler than the breadth of human behavior that can take place within the environment, suggesting that non-causal IDMs could require far less data to train than causal BC models. Using pseudo-labels generated from the IDM, we then train a model to mimic the distribution of behavior in the previously unlabeled dataset with standard behavioral cloning at scale, which does not require any model rollouts and thus does not suffer from any potential exploration bottlenecks in the environment. Finally, we show we can fine-tune this model to downstream tasks with either behavioral cloning or reinforcement learning.

Figure 1: Example Minecraft crafting GUI. Agents use the mouse and keyboard to navigate menus and drag and drop items.

We chose to test our method in Minecraft because (a) it is one of the most actively played games in the world[26] and thus has a wealth of commonly available video data online, (b) it is a fairly open-ended sandbox game with an extremely wide variety of potential things to do, build, and collect, making our results more applicable to real-world applications such as computer usage, which also tends to be varied and open-ended, and (c) it has already garnered interest by the RL community as a research domain due to its complexity and correspondingly difficult exploration challenges.[27–31] In this work we use the native human interface for Minecraft so that we can (1) most accurately model the human behavior distribution and reduce domain shift between video data and the environment, (2) make data collection easier by allowing our human contractors to play the game without modification, and (3) eliminate the need to hand-engineer a custom interface for models to interact with the environment. This choice means that our models play at 20 frames per second and must use a mouse and keyboard interface to interact with human GUIs for crafting, smelting, trading, etc., including dragging items to specific slots or navigating the recipe book with the mouse cursor (Fig. 1). Compared to prior work in Minecraft that uses a lower frame rate and constructs crafting and attacking macros,[30,32–34] using the native human interface drastically increases the environment's exploration difficulty, making most simple tasks near impossible with RL from scratch. Even the simple task of gathering a single wooden log while already facing a tree takes 60 consecutive attack actions with the human interface, meaning the chance for a naive random policy to succeed is $\frac{1}{2}^{60}$. While this paper shows results in Minecraft only, the VPT method is general and could be applied to any domain.

In Section 4 we show that the VPT foundation model has nontrivial zero-shot performance, accomplishing tasks impossible to learn with RL alone, such as crafting planks and crafting tables (tasks requiring a human proficient in Minecraft a median of 50 seconds or $\sim$970 consecutive actions). Through fine-tuning with behavioral cloning to smaller datasets that target more specific behavior distributions, our agent is able to push even further into the technology tree, crafting stone tools

CAPÍTULO V / O AGREGADO

Nem sempre ia naquele passo vagaroso e rígido. Também se descompunha em acionados, era muita vez rápido e lépido nos movimentos, tão natural nesta como naquela maneira. Outrossim, ria largo, se era preciso, de um grande riso sem vontade, mas comunicativo, a tal ponto ás bochechas, os dentes, os olhos, toda a cara, toda a pessoa, todo o mundo pareciam rir nele. Nos lances graves, gravíssimo.

Era nosso agregado desde muitos anos; meu pai ainda estava na antiga fazenda de Itaguaí, e eu acabava de nascer. Um dia apareceu ali vendendo-se por médico homeopata; levava um Manual e uma botica. Havia então um andaço de febres; José Dias curou o feitor e uma escrava, e não quis receber nenhuma remuneração. Então meu pai propôs-lhe ficar ali vivendo, com pequeno ordenado. José Dias recusou, dizendo que era justo levar a saúde à casa de sapé do pobre.

--Quem lhe impede que vá a outras partes? Vá aonde quiser, mas fique morando conosco.

--Voltarei daqui a três meses.

Voltou dali a duas semanas, aceitou casa e comida sem outro estipêndio, salvo o que quisessem dar por festas. Quando meu pai foi eleito deputado e veio para o Rio de Janeiro com a família, ele veio também, e teve o seu quarto ao fundo da chácara. Um dia, reinando outra vez febres em Itaguaí, disse-lhe meu pai que fosse ver a nossa escravatura. José Dias deixou-se estar calado, suspirou e acabou confessando que não era médico. Tomara este título para ajudar a propaganda da nova escola, e não o fez sem estudar muito e muito; mas a consciência não lhe permitia aceitar mais doentes.

--Mas, você curou das outras vezes.

--Creio que sim; o mais acertado, porém, é dizer que foram os remédios indicados nos livros. Eles, sim, eles, abaixo de Deus. Eu era um charlatão... Não negue; os motivos do meu procedimento podiam ser e eram dignos; a homeopatia é a verdade, e, para servir à verdade, menti; mas é tempo de restabelecer tudo.

Não foi despedido, como pedia então; meu pai já não podia dispensá-lo. Tinha o dom de se fazer aceito e necessário; dava-se por falta dele, como de pessoa da família. Quando meu pai morreu, a dor que o pungiu foi enorme, disseram-me; não me lembra. Minha mãe ficou-lhe muito grata, e não consentiu que ele deixasse o quarto da chácara; ao sétimo dia. depois da missa, ele foi despedir-se dela.

--Fique, José Dias.

--Obedeço, minha senhora.

Teve um pequeno legado no testamento, uma apólice e quatro palavras de louvor. Copiou as palavras, encaixilhou-as e pendurou-as no quarto, por cima da cama. "Esta é a melhor apólice", dizia ele muita vez. Com o tempo, adquiriu certa autoridade na família, certa audiência, ao menos; não abusava, e sabia opinar obedecendo. Ao cabo, era amigo, não direi ótimo, mas nem tudo é ótimo neste mundo. E não lhe suponhas alma subalterna; as cortesias que fizesse vinham antes do cálculo que da índole. A roupa durava-lhe muito; ao contrário das pessoas que enxovalham depressa o vestido novo, ele trazia o velho escovado e liso, cerzido, abotoado, de uma elegância pobre e modesta. Era lido, posto que de atropelo, o bastante para divertir ao serão e à sobremesa, ou explicar algum fenômeno, falar dos efeitos do calor e do frio, dos pólos e de Robespierre. Contava muita vez uma viagem que fizera à Europa, e confessava que a não sermos nós, já teria voltado para lá; tinha amigos em Lisboa, mas a nossa família, dizia ele, abaixo de Deus, era tudo.

--Abaixo ou acima? perguntou-lhe tio Cosme um dia.

--Abaixo, repetiu José Dias cheio de veneração.

E minha mãe, que era religiosa, gostou de ver que ele punha Deus no devido lugar, e sorriu

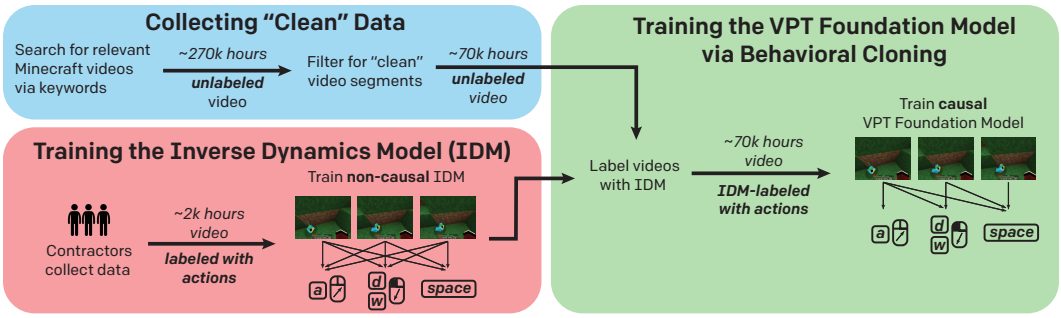

Figure 2: Video Pretraining (VPT) Method Overview.

3 Methods

Inverse Dynamics Models (IDM) VPT, illustrated in Figure 2, requires we first collect a small amount of labeled contractor data with which to train an inverse dynamics model $p_{\text{IDM}}(a_t|o_{1\ldots T})$, which seeks to minimize the negative log-likelihood of an action at timestep t given a trajectory of T observations $o_t \ : \ t \in [1...T]$. In contrast to an imitation learning policy, the IDM can be non-causal, meaning its prediction for a_t can be a function of both past and *future events*, i.e. $o_{t'>t}$. Compared to the behavioral cloning objective of modeling the distribution of human intent given past frames only, we hypothesize that inverting environment dynamics is easier and more data efficient to learn. Indeed, Sec. 4.1 will show that the IDM objective is much easier to learn, and furthermore Sec. 4.6 will show that with very little labeled data (as few as 100 hours) we can train a fairly accurate IDM. This IDM can be used to label online videos, providing the large amount of data required for the harder task of behavioral cloning. See appendices D and B for IDM training and data collection details.

Data Filtering We gather a large dataset of Minecraft videos by searching the web for related keywords (Appendix A). Online videos often (1) include overlaid artifacts, such as a video feed of the player's face, channel logos, watermarks, etc., (2) are collected from platforms other than a computer with different gameplay, or (3) are from different game modes, e.g. in Minecraft we only want "survival mode" where players start from scratch and must gather or craft all their items. We call data "clean" if it does not contain visual artifacts and is from survival mode, and call all other data "unclean." With enough data, a large enough model, and enough training compute, a BC model trained on both unclean and clean videos would likely still perform well in a clean Minecraft environment. However, for simplicity and training compute efficiency, we choose to filter out unclean segments of video (note that a video may contain both clean and unclean segments). We do this by training a model to filter out unclean segments using a small dataset (8800) of images sampled from online videos labeled by contractors as clean or unclean (Appendix A.2).

VPT Foundation Model We train a foundation model with standard behavioral cloning, i.e. minimizing the negative log-likelihood of actions predicted by the IDM on clean data. For a particular trajectory of length T we minimize

$$\min_{\theta} \sum_{t \in [1...T]} -\log \pi_\theta(a_t|o_1, \ldots, o_t), \text{ where } \ a_t \sim p_{\text{IDM}}(a_t|o_1, \ldots, o_t, \ldots, o_T) \tag{1}$$

As we will see in the following sections, this model exhibits nontrivial zero-shot behavior and can be fine-tuned with both imitation learning and RL to perform even more complex skills.

4 Results

4.1 Performance of the Inverse Dynamics Model

The IDM architecture is comprised primarily of a temporal convolution layer, a ResNet[62] image processing stack, and residual unmasked attention layers, from which the IDM simultaneously predicts keypresses and mouse movements (see Appendix D for IDM architecture and training details). A key hypothesis behind our work is that IDMs can be trained with a relatively small amount of labeled data. While more data improves both mouse movement and keypress predictions, our best

Parei na varanda; ia tonto, atordoado, as pernas bambas, o coração parecendo querer sair-me pela boca fora. Não me atrevia a descer à chácara, e passar ao quintal vizinho. Comecei a andar de um lado para outro, estacando para amparar-me, e andava outra vez e estacava. Vozes confusas repetiam o discurso do José Dias:

"Sempre juntos..."

"Em segredinhos..."

"Se eles pegam de namoro..."

Tijolos que pisei e repisei naquela tarde, colunas amareladas que me passastes à direita ou à esquerda, segundo eu ia ou vinha, em vós me ficou a melhor parte da crise, a sensação de um gozo novo, que me envolvia em mim mesmo, e logo me dispersava, e me trazia arrepios, e me derramava não sei que bálsamo interior. Às vezes dava por mim, sorrindo, um ar de riso de satisfação, que desmentia a abominação do meu pecado. E as vozes repetiam-se confusas;

"Em segredinhos..."

"Sempre juntos..."

"Se eles pegam de namoro..."

Um coqueiro, vendo-me inquieto e adivinhando a causa, murmurou de cima de si que não era feio que os meninos de quinze anos andassem nos cantos com as meninas de quatorze, ao contrário, os adolescentes daquela idade não tinham outro ofício, nem os cantos outra utilidade. Era um coqueiro velho, e eu cria nos coqueiros velhos, mais ainda que nos velhos livros. Pássaros, borboletas, uma cigarra que ensaiava o estilo, toda a gente viva do ar era da mesma opinião.

Com que então eu amava Capitu, e Capitu a mim? Realmente, andava cosido às saias dela, mas não me ocorria nada entre nós que fosse deveras secreto. Antes dela ir para o colégio, eram tudo travessuras de criança; depois que saiu do colégio, é certo que não estabelecemos logo a antiga intimidade, mas esta voltou pouco a pouco, e no último ano era completa. Entretanto, a matéria das nossas conversações era a de sempre. Capitu chamava-me às vezes bonito, mocetão, uma flor - outras pegava-me nas mãos para contar-me os dedos. E comecei a recordar esses e outros gestos e palavras, o prazer que sentia quando ela me passava a mão pelos cabelos, dizendo que os achava lindíssimos. Eu, sem fazer o mesmo aos dela, dizia que os dela eram muito mais lindos que os meus. Então Capitu abanava a cabeça com uma grande expressão de desengano e melancolia, tanto mais de espantar quanto que tinha os cabelos realmente admiráveis - mas eu retorquia chamando-lhe maluca. Quando me perguntava se sonhara com ela na véspera, e eu dizia que não, ouvia-lhe contar que sonhara comigo, e eram aventuras extraordinárias, que subíamos ao Corcovado pelo ar, que dançávamos na lua, ou então que os anjos vinham perguntar-nos pelos nomes, a fim de os dar a outros anjos que acabavam de nascer. Em todos esses sonhos andávamos unidinhos. Os que eu tinha com ela não eram assim, apenas reproduziam a nossa familiaridade, e muita vez não passavam da simples repetição do dia. alguma frase, algum gesto. Também eu os contava. Capitu um dia notou a diferença, dizendo que os dela eram mais bonitos que os meus, eu, depois de certa hesitação, disse-lhe que eram como a pessoa que sonhava... Fez-se cor de pitanga.

Pois, francamente, só agora entendia a comoção que me davam essas e outras confidências. A emoção era doce e nova, mas a causa dela fugia-me, sem que eu a buscasse nem suspeitasse. Os silêncios dos últimos dias, que me não descobriam nada, agora os sentia como sinais de alguma cousa, e assim as meias palavras, as perguntas curiosas, as respostas vagas, os cuidados, o gosto de recordar a infância. Também adverti que era fenômeno recente acordar com o pensamento em Capitu, e escutá-la de memória, e estremecer quando lhe ouvia os passos. Se se falava nela, em minha casa, prestava mais atenção que dantes, e, segundo era louvor ou crítica, assim me trazia gosto ou desgosto mais intensos que outrora, quando éramos somente companheiros de travessuras. Cheguei a pensar nela durante as missas daquele mês, com intervalos, é verdade, mas com exclusivismo também.

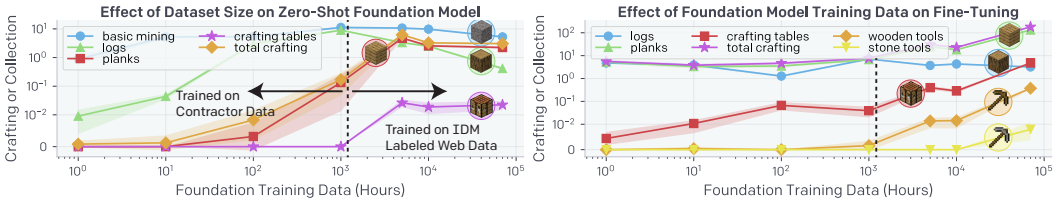

Figure 8: **(Left)** Zero-shot rollout performance of foundation models trained on varying amounts of data. Models to the left of the dashed black line (points $\leq$1k hours) were trained on contractor data (ground-truth labels), and models to the right were trained on IDM pseudo-labeled subsets of `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.

contractor data, and those trained on 5k hours and above are trained on subsets of `web_clean`, which does not contain any IDM contractor data. Scaling training data increases log collection, mining, and crafting capabilities. The zero-shot model only begins to start crafting crafting tables at over 5000 hours of training data. When fine-tuning each foundation model to `contractor_house`, we see that crafting rates for crafting tables and wooden tools increase by orders of magnitude when using the entire $\sim$70k hour `web_clean` dataset. We furthermore only see the emergence of crafting stone tools at the largest data scale.

4.6 Effect of Inverse Dynamics Model Quality on Behavioral Cloning

This section investigates how downstream BC performance is affected by IDM quality. We train IDMs on increasingly larger datasets and use each to independently label the `earlygame_keyword` dataset (this smaller dataset was chosen due to a limited compute budget). We then train a BC model from scratch on each dataset and report game statistics for each model as a function of IDM contractor dataset size (Fig. 9).

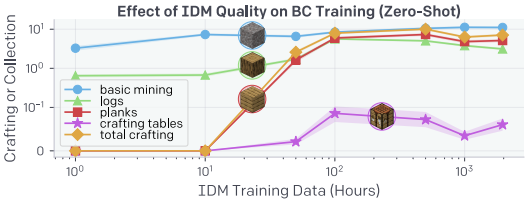

Figure 9: Zero-shot performance of BC models trained from scratch on the `earlygame_keyword` dataset labeled with IDMs that were trained on increasing amounts of contractor data.

IDMs trained on at least 10 hours of data are required for any crafting, and the crafting rate increases quickly up until 100 hours of data, after which there are few to no gains and differences are likely due to noise. Similarly, crafting tables are only crafted after 50 or more hours of IDM data, and again gains plateau after 100 hours. While in all previous experiments we use our best IDM trained on 1962 hours of data, these results suggest we could reduce that number to as low as 100 hours.

5 Discussion and Conclusion

The results presented in this paper help pave the path to utilizing the wealth of unlabeled data on the web for sequential decision domains. Compared to generative video modeling or contrastive methods that would only yield representational priors, VPT offers the exciting possibility of directly *learning to act* during pretraining and using these learned behavioral priors as extremely effective exploration priors for RL. VPT could even be a better general representation learning method even when the downstream task is not learning to act in that domain—for example, fine-tuning to explain what is happening in a video—because arguably the most important information in any given scene would be present in features trained to correctly predict the distribution over future human actions. We leave this intriguing direction to future work.

Future work could improve results with more data (we estimate we could collect >1M hours) and larger, better-tuned models. Furthermore, all the models in this work condition on past observations only; we cannot ask the model to perform specific tasks. Appendix I presents preliminary experiments on conditioning our models on closed captions (text transcripts of speech in videos), showing they