GROOT-1.5: Learning to Follow Multi-Modal Instructions from Weak Supervision

Shaofei Cai [*1] **Bowei Zhang** [*1] **Zihao Wang** [1] **Xiaojian Ma** [2] **Anji Liu** [3] **Yitao Liang** [1]

Abstract

This paper studies the problem of learning an agent policy that can follow various forms of instructions. Specifically, we focus on *multi-modal* instructions: the policy is expected to accomplish tasks specified in 1) a reference video, a.k.a. one-shot demonstration; 2) a textual instruction; 3) an expected return. Canonical goal-conditioned imitation learning pipelines require strong supervision (labeled data) in the form of $\langle \tau, c \rangle$ (τ denotes a trajectory $(s_1, a_1, \dots)$ and c denotes an instruction) from *all* modalities, which can be hard to obtain. To this end, we propose GROOT-1.5 to learn from mostly unlabeled data τ plus a relatively small amount of data with strong supervision $\langle \tau, c \rangle$. The key idea is a novel algorithm to learn a shared intention space from the trajectories τ themselves and labels c, *i.e.*, *semi-supervised learning*. We evaluate GROOT-1.5 on various benchmarks including open-world Minecraft, Atari games, and robotic manipulation and it has demonstrated strong steerability and performance on these tasks.

1. Introduction

Developing policies that can follow multi-modal instructions to solve open-ended tasks in open-world environments is a long-standing challenge both in robotics and AI research. With the development of large-scale pre-training (Brown et al., 2020; Baker et al., 2022; Brohan et al., 2022), the research paradigm for instruction-following policies has shifted from reinforcement learning to supervised learning. As a major approach within supervised learning, researchers collect extensive demonstration data and annotate each demonstration with multi-modal instructions, such as videos (Duan et al., 2017; Jang et al., 2022) and texts (Padalkar et al., 2023; Lynch et al., 2023), through hindsight relabeling. The performance of such policies can increase with the size of the dataset. However, annotating demonstration data with high-quality, diverse instructions is prohibitively expensive, making it difficult to scale up these methods. In contrast, another line of work (Lynch et al., 2020b; Ajay et al., 2020; Cai et al., 2023b) avoids collecting any additional annotated data and instead learns from the demonstration data in a self-supervised manner, such as by jointly learning an encoder and an instruction-following policy through a (variational) auto-encoding framework (Kingma & Welling, 2013). The learned policy typically conditions on the target image (Lynch et al., 2020b) or the reference video (Cai et al., 2023b). On the surface, a reference video can represent any task, but due to the ambiguity of the video, the learned latent space may collapse into a specific meaning. For example, the encoder model might capture the dynamics between adjacent frames in a video, thus learning a latent representation of the action sequence, a process we refer to as "mechanical imitation." While this can significantly reduce the final action reconstruction objective, such a latent space is not desired. Some efforts mitigate this issue by imposing constraints on the latent space using techniques such as discretization (Van Den Oord et al., 2017) and KL divergence to a prior distribution. However, these still cannot guarantee consistency of human and agent's video interpretations, which may hurt the agent's steerability.

In this paper, we advocate for learning of a latent intention space that aligns with human priors by utilizing both instruction-labeled and pure demonstration data. Intuitively, observing the instruction labels yields a deterministic intention distribution rather than merely observing demonstration data. Hence, we propose an uncertainty-driven intention learning framework that aligns instruction data with demonstration data within the intention space, thereby reducing the degrees of freedom in learning the latent intention space. Furthermore, this framework also supports learning behaviors from extensive unlabeled demonstration data. We have tested this methodology across three representative environments, including Atari Games, Language Table, and Minecraft, demonstrating its robust capability to follow

[*]Equal contribution [1]Peking University, Beijing, China [2]Beijing Institute of Technology, Beijing, China [3]University of California, Los Angeles, USA. Correspondence to: Yitao Liang <yitaol@pku.edu.cn>.

MFM-EAI Workshop on Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

multi-modal instructions. Our experiments further show that extensive unsupervised demonstration data can significantly enhance the agent's steerability and performance.

2. Preliminaries and Problem Formulation

We explore the development of instructable agents capable of following open-ended instructions to interact with environments and accomplish diverse tasks. An instruction may take various forms, such as text, image, video, audio, or expected returns, and is intended to convey behavioral intentions or specify tasks within the environment. Following common practice, we model the environment as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{I}, \mathcal{R}, \gamma \rangle$. Formally, the objective is to learn a policy $\pi : \mathcal{S} \times \mathcal{I} \times \mathcal{A} \to \mathbb{R}^+$ that maximizes the expected cumulative reward $\mathbb{E}_{i\sim\mathcal{I}}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$, where r_t denotes the reward obtained at each time step t. Given the complexity of designing effective reward functions, deriving such a policy from play data presents a more scalable approach.

We can collect two types of data from the web: a large set of unlabeled demonstrations $\mathcal{D}_{\text{dem}} = \{\tau_i\}_M$ and a relatively small set of annotated demonstrations $\mathcal{D}_{\text{ins}} = \{(\tau_i, c_i)\}_N$, where $N \ll M$. An annotated demonstration refers to one that is accompanied by a label, such as text or cumulative episode rewards, which explains the behavior or outcome of the demonstration. Our goal is to learn a shared latent intention space $\mathcal{Z}$, a multimodal instruction encoder $e(z|c)$, and an intention-conditioned policy $\pi(a|s, z)$. Note that, c can also be a video instruction. We define intention as a minimum succinct representation that can guide the policy to reconstruct the next action given past states. And an instruction c of a demonstration can contain multiple intentions. Prior works have demonstrated the feasibility of learning these components for image and video instructions from $\mathcal{D}_{\text{dem}}$ using self-supervised techniques such as latent variable generative models (Cai et al., 2023b) and the hindsight relabeling trick (Lifshitz et al., 2023). However, the resulting intention space often lacks sufficient constraints, offering no semantic guarantees. To address this challenge, it is advantageous to introduce auxiliary knowledge about human biases, which can help structure the intention space more effectively. This paper explores the problem of jointly learning a semantically rich intention space from both annotation-free and annotated demonstrations.

3. Uncertainty-driven Intention Learning Framework

We begin with an illustrative example on the language table platform (Lynch et al., 2023), depicted in Figure 1. Here, a demonstration is annotated with the instruction `slide the red pentagon below the blue moon`. From the initial state of the demonstration, multiple potential intentions can be inferred. Because it is possible to manipulate each pair of objects. Observing the full state sequence significantly reduces the uncertainty in the intention distribution, although ambiguities remain. However, any possible intention, whether object-centric or position-centric, could independently reconstruct the action sequence. When the textual annotation is also considered, the intention distribution is expected to converge sharply, aligning closely with the specific semantic intention we seek to capture. Based on these insights, we propose a novel uncertainty-driven intention learning framework. This framework is predicated on two main principles: (1) *a more deterministic intention distribution should influence the shaping of a more uncertain one*; (2) *the intention distribution characterized by greater uncertainty should, in turn, constrain the learning of the more deterministic one*. We find that the Kullback–Leibler (KL) divergence metric aptly fulfills these criteria. Further details on the dataset and training pipeline will be discussed subsequently.

We seek to build a dataset $\mathcal{D}_{\text{unc}} = \{(\tau, c_a \prec c_b)\}$ where each data point contains two different views c_a and c_b of the demonstration τ, $c_a \prec c_b$ indicates that c_a is a subset of c_b in terms of the induced underlying intentions. We show that such data points can be generated from both the demonstration-only data and the annotated demonstration data. As for each demonstration τ from $\mathcal{D}_{\text{dem}}$ or $\mathcal{D}_{\text{ins}}$, we can always create a training sample $(\tau, s_{1:|\tau|} \prec s_1)$. This is because one can infer more deterministic intentions when observing the whole states $s_{1:|\tau|}$ (video) compared with only observing the start state s_1. As for the annotated demonstration (τ, c) from $\mathcal{D}_{\text{ins}}$, we can additionally create a sample $(\tau, c \prec s_{1:|\tau|})$ for training, where c can either be a language instruction or the cumulative rewards of $\sum_{t=0}^{|\tau|} r_t$. Generally, a demonstration may owe many reasonable explanations while the annotated instruction is only one of them, which makes $c \prec s_{1:|\tau|}$ satisfied.

Our framework comprises two learnable modules: an encoder $e_\phi(z|c)$ that maps the instruction to an intention distribution, and an intention-conditioned policy $\pi_\theta(a|s, z)$ that interacts with the environment, with ϕ and θ denoting the parameters. We optimize the modules by targeting the constraint behavior cloning objective in an end-to-end manner:

$$\min \mathbb{E}_{(\tau, c_a \prec c_b)\sim\mathcal{D}_{\text{unc}}} \left[\mathcal{L}_{\text{BC}} + \beta\mathcal{L}_{\text{KL}}\right], \tag{1}$$

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{z\sim e_\phi(\cdot|c_a)} \left[\sum_{t=1}^{|\tau|} -\log \pi_\theta(a_t|s_{1:t}, z)\right], \tag{2}$$

$$\mathcal{L}_{\text{KL}} = D_{KL}\left(e_\phi(z|c_a) \parallel e_\phi(z|c_b)\right) \tag{3}$$

where β is the balancing coefficient. This formulation links the behavior, intention, and instruction through the gradient from the action reconstruction loss. The KL diver-

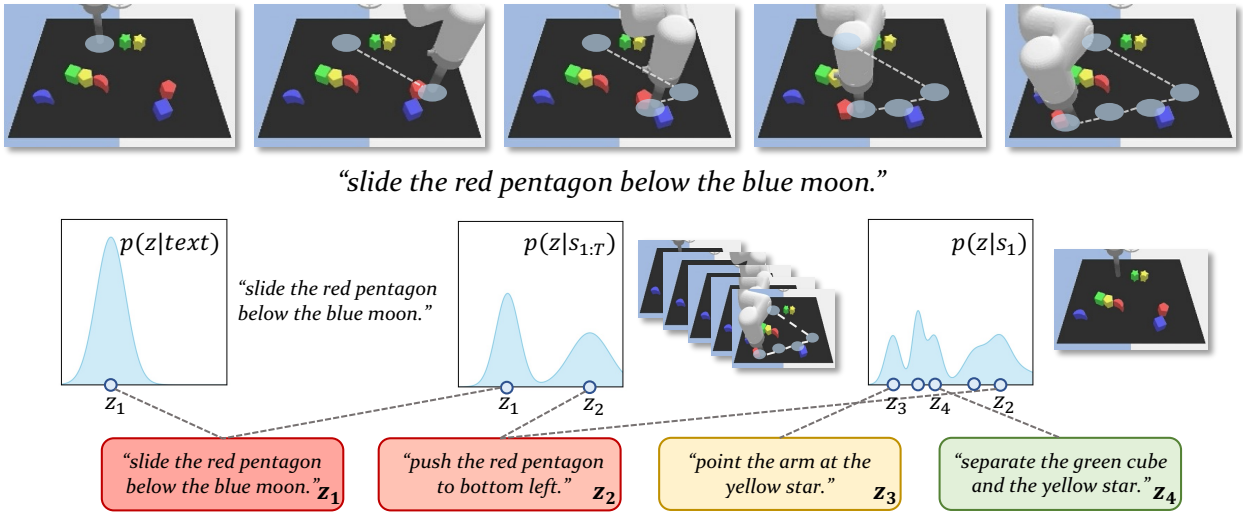

Figure 1: **An example to illustrate the uncertainty of intention space given different conditions.** We present an example of a text-annotated demonstration in the language table. From this example, we have three types of instructions: text, video, and the initial state. When observing the text and video, one can infer the intentions z_1 and $\{z_1, z_2\}$, respectively. Given only the initial state, a broader range of potential intentions, such as z_3 and z_4, can be inferred. For clarity, we use text to approximately describe an intention z in the language table, although it is important to note that the intention itself is not textual.

gence ensures that knowledge is transferred from $e_\phi(z|c_a)$ to $e_\phi(z|c_b)$, enhancing the generalization and improving the shared intention space. It is important to note that our objective, while similar, is inherently different from the frameworks of VAE and CVAE. In the traditional VAE framework, the latent variable z is always sampled from the posterior conditioned on the complete state sequence $s_{1:|\tau|}$. In contrast, our framework allows for sampling from a distribution conditioned on text if the text specifies more deterministic intentions than the states of the corresponding demonstration. In such cases, the encoding distributions conditioned on text and states are represented in the KL divergence as $D_{KL}(e_\phi(z|\text{text}) \parallel e_\phi(z|s_{1:|\tau|}))$. Details of model architecture can be found in Appendix B.

4. Capabilities and Analysis

Environment and Benchmarks. We conduct experiments across three types of representative environments: classical 2D game-playing benchmarks on Atari (Bellemare et al., 2013), 3D open-world game-playing benchmarks on Minecraft (Johnson et al., 2016; Lin et al., 2023), and Robotics benchmarks on language table simulators (Lynch et al., 2023). These three simulators are used to assess whether GROOT-1.5 can be effectively steered by returns (Chen et al., 2021; Mnih et al., 2015), reference videos, and textual instructions, respectively. Note that a key challenge in the language table environment is its significant ambiguity in the intention that may arise from the given demonstration, shown in Figure 1;

Training Datasets. We leverage existing datasets from Agarwal et al. (2020) in the d4rl (Fu et al., 2020) format for **Atari** games containing approximately 10M frames per Atari game. We further normalize the returns across the whole datasets with $\mu = 0, \sigma = 1$. For the **language table**, we utilize a dataset (Lynch et al., 2023) comprising 181k trajectories across six different types of tasks. In **Minecraft**, we employ the contractor dataset collected by Baker et al. (2022) consisting of about 160M frames, albeit without any annotations.

Table 1: **Evaluation results** of GROOT-1.5 on Robotics language table tasks (Lynch et al., 2023) and open-world Minecraft tasks (Guss et al., 2019) with instructions on different modalities.

Group	Type	Task Description	Success Rate (%)
Language Table	text	block to block	44
		block to absolute location	42
		block to block relative location	40
		block to relative location	46
		separate	92
		overall	**53**
Open-world Minecraft	video	Chop tree	100
		Mine stone	100
		Collect seeds	100
		Build pillar	100
		Hoe and plant wheat	100
		Hunt animals	90
		Cross the river by boat	90
		Dig down three fill one up	60
		Craft furnace with crafting_table	60
		overall	**89**

Evaluation Methods. We will test two evaluation methods: (1) conditioning the intention distribution on a reference video, and (2) conditioning the intention distribution on instructions from other modalities, such as text and returns.

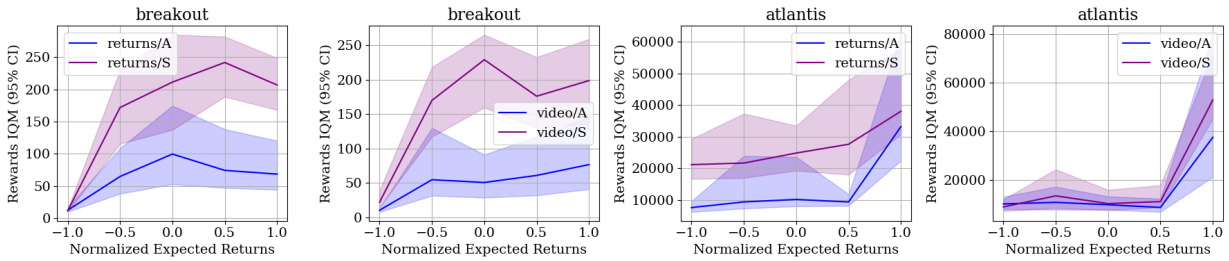

Figure 2: **Ablation study on the training dataset in Atari games.** "A", an abbreviation for "annotation", refers to a dataset containing 30% demonstrations, each associated with a returns label. The dataset "S", standing for "semi-supervised", extends "A" by including a large number of **label-free** demonstrations. We evaluate the performance of each trained policy when conditioned on both returns and reference videos. When conditioning on returns, the normalized returns are directly input into the encoder. When conditioning on video, we first retrieve a demonstration labeled with similar returns (error < 0.1) and then input the state sequence.

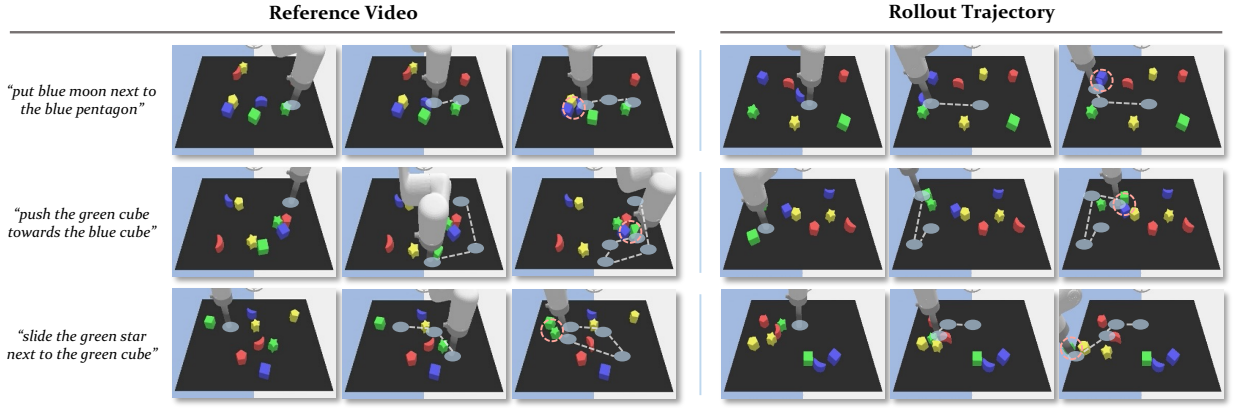

Figure 3: **GROOT-1.5 can infer the intention behind the reference video and follow it to complete tasks.** The **left** visualizes three reference videos along with their textual descriptions. The right figure displays the policy's rollout trajectories when conditioned on the reference videos. The white dashed line represents the arm's movement trajectory, the red dashed circle highlights the arm's final position.

The first method is applicable across all environments. Additionally, in the language table environment, we can sample intentions from a text-conditioned distribution, and in Atari games, from a returns-conditioned distribution.

Experimental Results. The steerability of GROOT-1.5 in different contexts is shown in Figure 2, Figure 3 and Table 1.

On **Atari** Atlantis and Breakout games, we first use the labeled dataset (have returns, with about 30% of the full demonstrations) to train GROOT-1.5-A. We further use the left unlabeled 70% demonstration to joint train GROOT-1.5-S. As depicted in Figure 2, GROOT-1.5-S significantly outperforms GROOT-1.5-A, which indicates that incorporating additional label-free demonstrations can enhance both the performance and the steerability of GROOT-1.5.

GROOT-1.5 is capable of inferring the underlying intention from a reference video and using it to complete tasks. As depicted in Figure 3, GROOT-1.5 with different reference videos (annotated with different texts) in **language table** accurately infers intentions and successfully executes tasks to *"put the blue moon to the blue pentagon"* and *"push the green cube towards the blue cube"*. Upon viewing the third reference video, GROOT-1.5 seems like inferring the intention *"move the green star to the middle left"*, diverging from the original explanation of the reference video. This discrepancy, although somewhat expected, highlights the inherent ambiguity within the inferred intention space.

Experimental results in Table 1 also show GROOT-1.5 can achieve lots of basic skills in 3D open-world **Minecraft**. We also list the GROOT-1.5's capability to follow language instructions on robotics tasks. GROOT-1.5 can follow instructions with different modality and show great generalization performance on different types of tasks and environments.

5. Conclusions

In this paper, we investigate how to jointly learn a latent intention space and a multi-modal instruction-following policy under a weak supervision setting. We propose an uncertainty-driven intention learning framework, implemented using the Transformer architecture for GROOT-1.5. We demonstrate the ability of GROOT-1.5 to comprehend three modalities—text, video, and returns—across three representative environments.

References

Abramson, J., Ahuja, A., Barr, I., Brussee, A., Carnevale, F., Cassin, M., Chhaparia, R., Clark, S., Damoc, B., Dudzik, A., et al. Imitating interactive intelligence. *arXiv preprint arXiv:2012.05672*, 2020.

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Ajay, A., Kumar, A., Agrawal, P., Levine, S., and Nachum, O. Opal: Offline primitive discovery for accelerating offline reinforcement learning. *arXiv preprint arXiv:2010.13611*, 2020.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. URL https://api.semanticscholar.org/CorpusID:248476411.

Andrychowicz, M., Crow, D., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. *ArXiv*, abs/1707.01495, 2017. URL https://api.semanticscholar.org/CorpusID:3532908.

Badia, A. P., Piot, B., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, Z. D., and Blundell, C. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, pp. 507–517. PMLR, 2020.

Baker, B., Akkaya, I., Zhokhov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *ArXiv*, abs/2206.11795, 2022. URL https://api.semanticscholar.org/CorpusID:249953673.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R. C., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M. S., Salazar, G., Sanketi, P. R., Sayed, K., Singh, J., Sontakke, S. A., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q. H., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-1: Robotics transformer for real-world control at scale. *ArXiv*, abs/2212.06817, 2022. URL https://api.semanticscholar.org/CorpusID:254591260.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Cai, S., Wang, Z., Ma, X., Liu, A., and Liang, Y. Open-world multi-task control through goal-aware representation learning and adaptive horizon prediction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13734–13744, 2023a. URL https://api.semanticscholar.org/CorpusID:256194112.

Cai, S., Zhang, B., Wang, Z., Ma, X., Liu, A., and Liang, Y. Groot: Learning to follow instructions by watching gameplay videos. In *The Twelfth International Conference on Learning Representations*, 2023b.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In *Neural Information Processing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:235294299.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jan 2019. doi: 10.18653/v1/p19-1285. URL http://dx.doi.org/10.18653/v1/p19-1285.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805,

2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Duan, Y., Andrychowicz, M., Stadie, B., Jonathan Ho, O., Schneider, J., Sutskever, I., Abbeel, P., and Zaremba, W. One-shot imitation learning. *Advances in neural information processing systems*, 30, 2017.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. *Advances in neural information processing systems*, 31, 2018.

Fu, J., Kumar, A., Nachum, O., Tucker, G., and Levine, S. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Guss, W. H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M. M., and Salakhutdinov, R. Minerl: A large-scale dataset of minecraft demonstrations. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:199000710.

Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., and Gould, S. Vln-bert: A recurrent vision-and-language bert for navigation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1643–1653, 2020. URL https://api.semanticscholar.org/CorpusID:227228335.

Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.-C., Jia, B., and Huang, S. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

Jang, E., Irpan, A., Khansari, M., Kappler, D., Ebert, F., Lynch, C., Levine, S., and Finn, C. Bc-z: Zero-shot task generalization with robotic imitation learning. *ArXiv*, abs/2202.02005, 2022. URL https://api.semanticscholar.org/CorpusID:237257594.

Johnson, M., Hofmann, K., Hutton, T., and Bignell, D. The malmo platform for artificial intelligence experimentation. In *International Joint Conference on Artificial Intelligence*, 2016. URL https://api.semanticscholar.org/CorpusID:9953039.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Lee, K.-H., Nachum, O., Yang, M. S., Lee, L., Freeman, D., Guadarrama, S., Fischer, I., Xu, W., Jang, E., Michalewski, H., et al. Multi-game decision transformers. *Advances in Neural Information Processing Systems*, 35: 27921–27936, 2022.

Lifshitz, S., Paster, K., Chan, H., Ba, J., and McIlraith, S. A. Steve-1: A generative model for text-to-behavior in minecraft. *ArXiv*, abs/2306.00937, 2023. URL https://api.semanticscholar.org/CorpusID:258999563.

Lin, H., Wang, Z., Ma, J., and Liang, Y. Mcu: A task-centric framework for open-ended agent evaluation in minecraft. *arXiv preprint arXiv:2310.08367*, 2023.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

Lynch, C. and Sermanet, P. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., and Sermanet, P. Learning latent plans from play. In *Conference on robot learning*, pp. 1113–1132. PMLR, 2020a.

Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., and Sermanet, P. Learning latent plans from play. In *Conference on robot learning*, pp. 1113–1132. PMLR, 2020b.

Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T., and Florence, P. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

Majumdar, A., Aggarwal, G., Devnani, B., Hoffman, J., and Batra, D. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *ArXiv*, abs/2206.12403, 2022. URL https://api.semanticscholar.org/CorpusID:250048645.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. URL https://api.semanticscholar.org/CorpusID:205242740.

Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Pashevich, A., Schmid, C., and Sun, C. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15942–15952, 2021.

Raad, M. A., Ahuja, A., Barros, C., Besse, F., Bolt, A., Bolton, A., Brownfield, B., Buttimore, G., Cant, M., Chakera, S., et al. Scaling instructable agents across many simulated worlds. *arXiv preprint arXiv:2404.10179*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J. T., et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Schmidhuber, J. Reinforcement learning upside down: Don't predict rewards - just map them to actions. *ArXiv*, abs/1912.02875, 2019. URL https://api.semanticscholar.org/CorpusID:208857600.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X. S., and Liang, Y. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. *Advances in Neural Information Processing Systems*, 36, 2023a.

Wang, Z., Cai, S., Liu, A., Jin, Y., Hou, J., Zhang, B., Lin, H., He, Z., Zheng, Z., Yang, Y., et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*, 2023b.

Wang, Z., Liu, A., Lin, H., Li, J., Ma, X., and Liang, Y. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024.

Yu, T., Quillen, D., He, Z., Julian, R. C., Hausman, K., Finn, C., and Levine, S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *ArXiv*, abs/1910.10897, 2019. URL https://api.semanticscholar.org/CorpusID:204852201.

Zhang, Y. and Chai, J. Hierarchical task learning from language instructions with unified transformers and self-monitoring. *arXiv preprint arXiv:2106.03427*, 2021.

A. Related Works

Learning Policy in Diverse Domains. Learning policies to address sequential control problems in both real and virtual environments presents a significant challenge. Researchers have devised numerous algorithms across varied domains such as robotic manipulation (Yu et al., 2019; Lynch et al., 2023), video games (Bellemare et al., 2013; Guss et al., 2019), and embodied navigation (Hong et al., 2020; Savva et al., 2019; Huang et al., 2023). These algorithms are typically categorized based on their reliance on the reward function, falling into two main categories: reinforcement learning (RL) and imitation learning (IL). In the case of video games offering dense rewards, such as those on the ALE platform (Bellemare et al., 2013), practitioners often deploy online RL-based algorithms capable of surpassing human performance (Mnih et al., 2015; Badia et al., 2020). Nonetheless, the primary limitations of this approach include low training efficiency, risky environmental interactions, and limited generalization to new tasks. These issues make it challenging to apply such methods in physical (Padalkar et al., 2023) or embodied environments (Guss et al., 2019), where both a reliable rewards function and inexpensive environment interactions are absent. Under these constraints, imitation learning becomes the predominant research approach. As a form of supervised learning, IL benefits from the efficiency of batch processing and the capability to scale up with large datasets, leveraging the computational power of Transformers (Zhang & Chai, 2021; Pashevich et al., 2021; Jang et al., 2022). The RT-X series (Brohan et al., 2022; 2023; Padalkar et al., 2023) has made strides in addressing robotic manipulation tasks by fitting a Transformer to a vast corpus of expert demonstrations through imitation learning, showcasing remarkable zero-shot generalization. (Baker et al., 2022) has developed a Transformer-based policy in Minecraft using internet-scale gameplay videos and refined it to tackle the diamond challenge successfully. Riding on this momentum, (Schmidhuber, 2019) suggests framing the reinforcement learning problem within a supervised learning context. Furthermore, (Chen et al., 2021; Lee et al., 2022) have introduced the "decision transformer", designed to model the joint distribution of rewards, states, and actions derived from offline experiences, underscoring the potential for unified policy learning within the Transformer architecture.

Learning Policy to Follow Instructions. Equipping a policy with the capability to follow instructions is crucial for developing a generally capable agent. A common approach involves collecting language annotations from offline demonstrations and training a language-conditioned policy (Abramson et al., 2020; Brohan et al., 2022; 2023; Padalkar et al., 2023; Lynch et al., 2023; Reed et al., 2022; Cai et al., 2023a; Huang et al., 2023; Raad et al., 2024; Wang et al., 2023a;b; 2024). Given the compositional nature of natural language, this strategy enables the policy to generalize to some unseen tasks. However, acquiring and processing high-quality language annotations can be prohibitively expensive. Alternatively, some researchers advocate using anticipated future outcomes as instructions to guide the policy. (Majumdar et al., 2022) proposed learning an image-goal conditioned navigation policy using the hindsight relabeling trick (HER) (Andrychowicz et al., 2017), and subsequently aligning the goal space with the textual modality. (Lifshitz et al., 2023) employed a similar strategy in Minecraft environments to train a policy that addresses open-ended tasks. Diverging from the use of the HER trick, (Lynch et al., 2020a; Ajay et al., 2020) utilize generative latent variable models to process label-free demonstrations, thereby enabling a plan-conditioned policy. (Cai et al., 2023b) extends this approach within the Minecraft setting by directly employing a posterior encoder to interpret reference videos during inference. Compared with these methodologies, fewer studies have investigated policy learning under weak supervision. (Lynch & Sermanet, 2020) suggests learning a shared latent space, conditioning the decoder on both languages and goal images, with the latter being generated using the HER trick. In contrast, (Jang et al., 2022) substitutes the goal image with a video label under a fully supervised learning framework. Our work focuses on developing a universally instructable policy learning framework under weak supervision settings.

B. Model Architecture

This section outlines the architectural design choices employed in our approach. GROOT-1.5 utilizes a Transformer encoder-decoder architecture, augmented with a probabilistic latent space. We detail the components of the model in a structured sequence: *extract representations*, *encode instructions*, and *decode actions*. More details on model selection are provided in the Appendix.

Extract Representations. This paragraph elaborates on the backbone networks used to extract representations from various data modalities. We denote the modalities of image observation, language instruction, and expected returns as o^{i}, o^{w}, and o^{r}, respectively. For vision inputs, we utilize a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2020) initialized with CLIP (Radford et al., 2021) weights. Specifically, the $t-$step image observation o^{i}_t is resized to 224×224 and processed to extract 7×7 patch embeddings $x^{\mathrm{i}}_t = \left\langle x^{\mathrm{i}}_{t,[1]}, \cdots, x^{\mathrm{i}}_{t,[49]} \right\rangle$. The video representation x^{v} is then composed of the averages of

these embeddings across the video frames, denoted as $x^{\mathrm{v}} = \langle \mathrm{avg}(x^{\mathrm{i}}_1), \cdots, \mathrm{avg}(x^{\mathrm{i}}_T) \rangle$, where $\mathrm{avg}(\cdot)$ refers to spatial average pooling to minimize computational overhead and T represents the video length. Textual inputs are processed using the BERT encoder (Devlin et al., 2019) of the CLIP model. Rather than utilizing the [CLS] token as the final representation, we retain all word embeddings generated by BERT as $x^{\mathrm{w}} = \left\langle x^{\mathrm{w}}_{[1]}, \cdots \right\rangle$. The BERT model parameters are kept frozen during training. For the scalar-form modality of expected returns, we employ a simple Multi-Layer Perceptron (MLP) to process these values, represented as $x^{\mathrm{r}} \leftarrow \mathrm{MLP}(o^{\mathrm{r}})$. These embeddings are then forwarded to subsequent modules.

Encode Multimodal Instructions with Non-Causal Transformer. Recent works (Reed et al., 2022; Lu et al., 2023; Team et al., 2023) have demonstrated the Transformer's effectiveness in capturing both intra-modal and inter-modal relationships, which inspires us to adopt a unified Transformer encoder for encoding multi-modal instructions. This approach offers two significant advantages: (1) It eliminates the need for designing separate architectures and tuning hyperparameters for each modality. (2) It promotes the sharing of underlying representations across different modalities. Instructions are represented as a sequence of embeddings. Prior to encoding, each embedding is augmented with a modality-specific marker. For instance, video instructions are represented as $\langle x^{\mathrm{v}}_1 + [\mathrm{VID}], \cdots, x^{\mathrm{v}}_T + [\mathrm{VID}] \rangle$, and textual instructions as $\langle x^{\mathrm{w}}_1 + [\mathrm{TXT}], \cdots \rangle$. Both [VID] and [TXT] are distinct learned embeddings. The encoder outputs a set of parameters defining an intention distribution, denoted as $\Phi \leftarrow e_\phi(c)$. Given that the desired intention distribution may be multimodal, we utilize a Gaussian mixture model for the encoder's output, represented as $\Phi = \{\pi, \mu, \sigma\}$, where π refers to a categorical distribution, and μ and σ are the parameters for Gaussian components. We apply the Monte-Carlo sampling method with the implicit reparameterization trick (Figurnov et al., 2018) to calculate the behavior cloning loss and KL loss in Equation 1 for stable gradients.

Decode Action with Causal Transformer. Given an intention $z \sim p(z|\Phi)$ and a temporal sequence of perceptual observations $o^{\mathrm{i}}_{1:t}$, the policy aims to decode the next action a_t. Following prior works (Baker et al., 2022; Cai et al., 2023b; Raad et al., 2024), we employ the Transformer-XL model (Dai et al., 2019) in our policy network, which enables causal attention to past memory states and facilitates smooth predictions. Additionally, we utilize the shared vision backbone to extract vision representations, thereby representing perceptual inputs as $x^{\mathrm{i}}_{1:t}$. A significant challenge with this approach is low efficiency: each new observation x^{i}_t adds up to 49 tokens to the input sequence, substantially increasing memory and computational demands. To address this issue, we introduce a *pre-fusion mechanism* inspired by (Abramson et al., 2020; Lynch et al., 2023; Alayrac et al., 2022). Specifically, we deploy a lightweight cross-attention module $\texttt{XATTN}(\mathrm{q} = \cdot; \mathrm{kv} = \cdot)$ to perform spatial pooling on x^{i}_t, using z as the *query* and $\left\langle x^{\mathrm{i}}_{t,[1]}, \cdots, x^{\mathrm{i}}_{t,[49]} \right\rangle$ as the *keys* and *values*:

$$x^z_t \leftarrow \texttt{XATTN}(\mathrm{q} = z; \mathrm{kv} = x^{\mathrm{i}}_{t,[1]}, \cdots, x^{\mathrm{i}}_{t,[49]}). \tag{4}$$

This *pre-fusion mechanism* not only reduces the sequence length but also enhances the integration of perceptual and intention representations. Utilizing the intention-fused representations $x^z_{1:t}$ as the input sequence, we articulate the action decoding process in an *autoregressive* manner:

$$a_t \leftarrow \texttt{TransformerXL}(x^z_1, \cdots, x^z_t). \tag{5}$$