Interactive Language: Talking to Robots in Real Time

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding
James Betker, Robert Baruch, Travis Armstrong, Pete Florence

Robotics at Google

Abstract

We present a framework for building interactive, real-time, natural language-instructable robots in the real world, and we open source related assets (dataset, environment, benchmark, and policies). Trained with behavioral cloning on a dataset of hundreds of thousands of language-annotated trajectories, a produced policy can proficiently execute an order of magnitude more commands than previous works: specifically, we estimate a 93.5% success rate on a set of 87,000 unique natural language strings specifying raw end-to-end visuo-linguo-motor skills in the real world. We find that the same policy is capable of being guided by a human via real-time language to address a wide range of precise long-horizon rearrangement goals, e.g. "*make a smiley face out of blocks*". The dataset we release comprises nearly 600,000 language-labeled trajectories, an order of magnitude larger than prior available datasets. We hope the demonstrated results and associated assets enable further advancement of helpful, capable, natural-language-interactable robots. See videos at https://interactive-language.github.io.

1 Introduction

The goal of building a robot that can follow a diverse array of natural language instructions has been a longstanding goal of AI research, since at least the SHRDLU [1] experiments starting in the late 1960s. While recent research on this topic has been abundant [2–9], few efforts have actually produced a robot that (i) exists in the real world, and (ii) can capably respond to a large number of rich, diverse language commands. We expect that future research will continue to produce larger and more diverse sets of behaviors, either by sequencing raw skills together [10] or growing the number of raw skills themselves [11]. However, we are also interested in (iii), the capacity to follow *interactive* language commands, by which we mean that the robot reacts capably and in-the-moment to new natural language instructions provided during ongoing task execution. Although we might expect such a robot to be possible given current methods, natural language-interactable robots are frequently slow in practice, and often use blocking parameterized skills [7, 10] or simplifying self-resetting behaviors [9, 12] that prohibit this kind of live, real-time interaction.

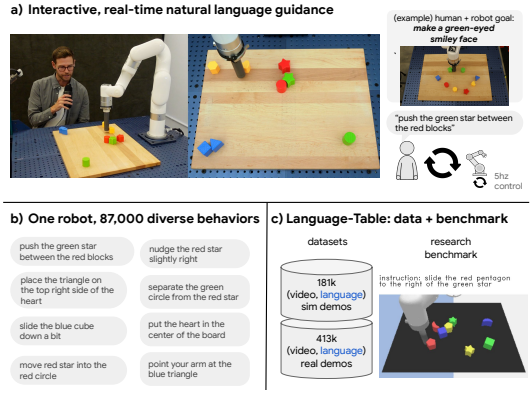

Figure 1: Real-time language, diverse robot behaviors. a) Over the course of 5 minutes, a human guides a robot to precisely rearrange objects a table into a desired shape, with real-time natural language as the only mechanism for specifying behaviors. b) We demonstrate a single robot that can capably address 87,000 behaviors specified entirely in natural language. c) We release Language-Table, a suite of human-collected datasets and benchmark for learning continuous visuolinguomotor control.

NeurIPS 2022 Workshop on Robot Learning: Trustworthy Robotics, Virtual, Virtual

In this paper, we (i) contribute and demonstrate a large scale imitation learning framework, *Interactive Language*, for producing real-world, real-time-interactable[1], natural-language-instructable robots (Fig. 1, a). In terms of scale, the produced robot policies can address 87,000 unique commands at an estimated 93.5% success rate (Fig. 1, b), with continuous 5Hz visuo-linguo-motor control. To accelerate further research in this setting, we (ii) we release *Language-Table* (Fig. 1, c), a dataset and simulated multitask imitation learning benchmark. With nearly 600,000 diverse demonstrations across simulation and the real world, Language-Table is, to our knowledge, the largest natural language conditioned imitation learning dataset of its kind by an order of magnitude (Table 1). We additionally (iii) show that through *interactive language guidance* in the form of occasional human natural-language feedback, the robot can accomplish a wide array of complex long-horizon rearrangements such as *"put the blocks into a smiley face with green eyes"* that require multiple minutes of precise coordinated control (Figure 2, left). We also (iv) find that real-time language competency unlocks new capabilities like simultaneous, multi-robot instruction (Figure 2, right).

2 Related work

Imitation learning (see review [14]), the perspective we adopt in this work, provides a simple and stable way for robots to acquire behaviors from human expert demonstrations. While historically imitation learning has been applied to individual tasks from instrumented state [15–18], the desire for more general purpose robots has motivated study into policies capable of learning multiple skills at once from on-board sensory observations like RGB pixels [19–21]. To condition multiple behaviors, prior setups have relied on discrete task identifiers [22], which can be difficult to scale to many tasks, or goal images [23–25], which can be impractical to provide in real world scenarios. Alternatively, a long history of prior work in broader AI research [1–6, 11] has investigated natural language as a more convenient task specification format (survey [26]), with some results on physical robots [7–9, 12]. However, instruction-following robots rarely leverage the full capabilities of continuous control continuous control, instead using simpler, parameterized action spaces [6, 7, 27, 28]. Furthermore, once provided, language conditioning is typically presumed fixed over robot execution [8–10, 12], with little opportunity for subsequent interaction by the instructor. Our work exists in a larger setting of humans modifying or correcting the online behavior of autonomous agents [29], historically addressed in forms like teleoperation [30–32], kinesthetic teaching [33], or sparse human preference feedback [34]. Certain works have studied language as a means of correction, but typically do so under simplifying assumptions that we relax in the current work. For example, [35], [36], [37], and [38] study language corrections, but under the respective simplifying assumptions of hand-defined optimization for grounding, undivided operator attention, paired iterative corrections at training time, and presumed access to motion planners and task cost functions. Additionally, to the best of our knowledge, none of these works support multiple-Hz iterative specification over the course of execution. Closest to our approach is [11] and [28], which study imitation learned language-interactive agents, but entirely in simulation and under varying degrees of actuation realism. Our work, in contrast, studies the first combination, to our knowledge, of real-time natural language guidance of a physical robot engaged in continuous visuomotor manipulation.

3 Interactive Language: Data Collection and Imitation Learning Framework

We introduce *Interactive Language*, summarized in Figure 3, an imitation learning framework for training real-time natural-language-interactable robots. Interactive Language combines a scalable method for collecting varied real world language-conditioned demonstration datasets with straightforward language conditioned behavioral cloning (LCBC).

Data Collection. In our framework, operators continuously teleoperate a variety of long horizon behaviors, without low-level task definition, segmentation, or episodic resets. Each collect episode lasts $\sim$10 minutes before a break, and is guided by multiple randomly chosen long horizon prompts $p \in \mathcal{P}$ (e.g. "*make a square shape out of the blocks*"), drawn from the set of target long horizon goals, which teleoperators are free to follow or ignore. We do not assume all of the data collected for each prompt p is optimal (each p is discarded after collecting). This strategy shares assumptions with "play" collection [24], but additionally guides collect towards temporally extended low-entropy states like lines,

[1]For the scope of this paper, by real-time we mean new language conditioning can occur in the "blink of an eye", i.e. approximately 3 Hz [13] or greater.

shapes, and complex arrangements. This collect procedure yields a *semi-structured, optimality-agnostic* collection $\mathcal{D}_{\text{collect}} = \{\tau_i\}_{i=0}^{\mathcal{D}_{\text{collect}}}$, which we convert into natural language conditioned demonstrations $\mathcal{D}_{\text{training}} = \{(\tau, l)_i\}_{i=0}^{\mathcal{D}_{\text{training}}}$, using a new variant of hindsight language relabeling [11] we call "Event-Selectable Hindsight Relabeling" (Fig.3, left). Here, we ask crowdsourced annotators watch a full collection episode, then find K coherent behaviors ($K = 24$ in our case), using marking the start and end frame of each behavior, and describing it as an open vocabulary natural language command reached in hindsight. See Table 5 for a comparison to prior "random window" relabeling techniques.

Policy Learning. In Figure 4, we describe our transformer-based [39] neural network policy architecture, mapping from video and text to continuous actions, which we refer to as LAVA ("Language Attends to Vision to Act"). Each training example consists of $(s, a, l)_i \sim \mathcal{D}_{\text{training}}$, where $s \in \mathbb{R}^{\text{seqlen} \times 640 \times 320 \times 3}$ is RGB observation history. We pass each frame in the video s through a Imagenet-pretrained ResNet [40, 41] and embed instruction l using a pretrained (and in-domain finetuned) CLIP text encoder [42], then fuse multi-scale vision and language embeddings using a "Language-Attends-to-Vision" cross-attention transformer block. The sequence output is fed to a temporal prenorm [43] transformer, which is then average pooled over time and fed to an MLP which outputs next action a. All the policies we present are deterministic and simply trained with mean squared error behavioral cloning [44]: $\min_\theta \sum_{(s,a,l)\sim\mathcal{D}_{\text{training}}} ||a - \pi_\theta(s, l)||_2^2$, e.g. as in [9, 45].

4 Language-Table: Datasets and Environment

To facilitate further research in language-conditioned visuomotor learning, we release *Language-Table*, which consists of (i) a suite of datasets and (ii) a simulated multi-task environment and benchmark. Language-Table provides our human-relabeled $\mathcal{D}_{\text{training}}$ and the underlying human-teleoperated $\mathcal{D}_{\text{collect}}$, both in simulation and the real world. The $\mathcal{D}_{\text{training}}$ real and sim datasets are highlighted in Table 1 – an order of magnitude larger than comparable, previously-available datasets.

Language-Table's simulated environment resembles our real-world tabletop manipulation scenario, which consists of an xArm6 robot, constrained to move in a 2D plane with a cylindrical end-effector as in [50], in front of a smooth wooden board with a fixed set of 8 plastic blocks, comprising 4 colors and 6 shapes (Fig. 2). In both simulation and real collection, we use high-rate human teleoperation with a 3rd person view (line-of-sight in real). Actions are 2D delta Cartesian setpoints, from the previous setpoint to the new one. We batch collected training and inference data to 5hz observations and actions. The Language-Table benchmark computes automated metrics for 5 task families, with 696 unique task variations. We note that policy hyperparameters ordered by success in Language-Table have thus far been ordered similarly in real-world performance, providing a degree of validation for the simulated benchmark's relevancy to real world robotics.

Dataset	# Traj. (k)	# Unique (k)	Physical Actions	Real	Available
Episodic Demonstrations					
BC-Z [9]	25	0.1	✓	✓	✓
SayCan [10]	68	0.5	✓	✓	✗
Playhouse [28]	1,097	779	✗	✗	✗
Hindsight Language Labeling					
BLOCKS [46, 47]	30	n/r	✗	✗	✓
LangLFP [11]	10	n/r	✓	✗	✗
LOREL [8, 48]	6	1.7	✓	✓	✓
CALVIN [49]	20	0.4	✓	✗	✓
Language-Table	**594**	**198**	✓	✓	✓
(*real*+*sim*)	(*413*+*181*)	(*119*+*79*)			

Table 1: Comparison of human-guided, language-labeled trajectory datasets.

5 Policy Results, Discussion, and Conclusion

We present experiments aimed at answering the following questions: Q1: How capably can the system follow a wide variety of short-horizon natural language conditioned commands? Q2: How capably can these skills be composed through interactive language guiding to accomplish a wide variety of multi-step long-horizon compositional rearrangements, and what is the benefit of this guidance? Q3: Can one operator simultaneously guide several robots equipped with our policy?

Q1: Diverse short-horizon language conditionable skills. To study Q1, we estimate a 95% confidence interval on average success over the 87,588 natural language instructions collected via crowdsourcing (20 randomly selected instructions, 10 trials each) (Table 4), with results reported in Table 2. We see that Interactive Language obtains a 93.5% expected average success rate over the instruction set, 95% CI [90.08%, 96.92%]. See examples of diverse learned behaviors in Figure 5. This is the

largest set of language conditioned behaviors, to our knowledge, a real-world policy has been shown to capably address.

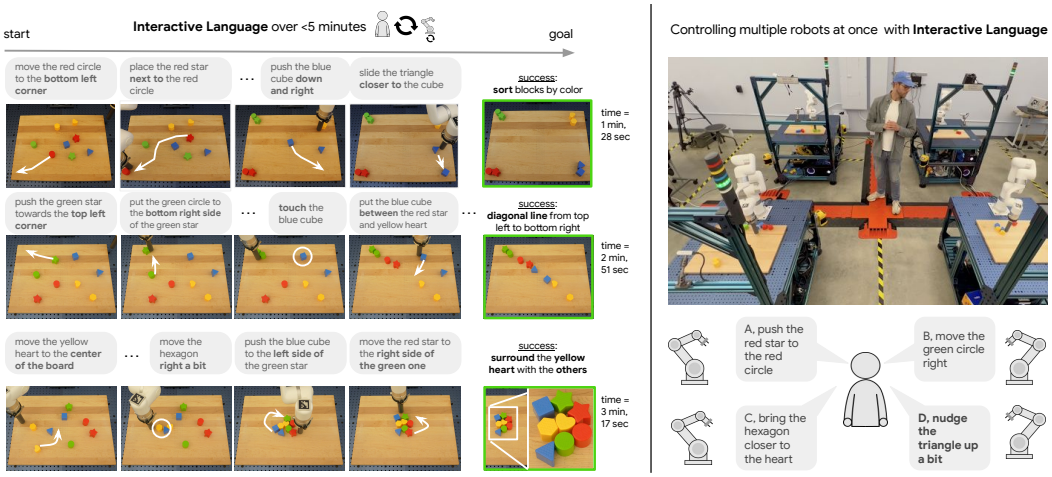

Figure 2: Capabilities explored with Interactive Language. Left: **Long horizon language guidance** allows a human to guide a single policy to achieve a wide variety of long horizon precise rearrangement goals. Right: **Simultaneous multi-robot control**. Real time language allows a single human operator to guide multiple robots at once through the same long horizon task, without requiring undivided attention to any one robot.

Instruction	Success %
(87k more...)	...
nudge the yellow heart a bit right	80%
place the red star above the blue cube	90%
push the group of blocks left a bit	100%
Average over 87k, CI 95%	**93.50%** $\pm$ **3.42%**

Table 2: Real world: Evaluating a wide variety of short horizon language conditionable skills.

Interaction style	Avg. # instructions given	Success %
Open-loop	6.5	25.0% +- 18.98%
Real-time (ours)	15	**85.0%** **+- 15.65%**

Table 3: Real world: long horizon goal reaching via real-time human language guidance.

Q2, Q3: Long-horizon real-time language guidance. To answer Q2, we define over 100,000 language-distinct compositional goal states on our tabletop from 11 high level families (e.g. make high-level shapes, sort by color, place all blocks in specific locations, arrange into lines, etc.), then sample 20 uniformly from all 11. See Figure 2 for examples of different goal states. We evaluate each long horizon goal 3 times from randomly reset board states, yielding 60 total evaluations of a single policy. Results show (Table 3, see video) our policy obtains an 85.0% expected average success rate on this diverse set of goals, $95\%\,CI\,[69.35\%, 100.00\%]$. Even though our policies do not do so fully autonomously, we believe the fact that a real robot can address such a large and varied set of goals with real-time language feedback suggests a synergistic mode of teaching robots to be even more capable: robots learn a set of general-purpose low-level skills, and humans put them together in a familiar way using natural language, interrupting at any time to offer situation-specific corrections. To quantify the benefit, we perform the same evaluation as in the previous section, but the human operator commits up front to the set and order of commands they will provide. We present results for this ablation in Table 3, finding that performance deteriorates from 85% to 25% when real-time language is removed, highlighting the dependence on sufficient feedback, not only for the low-level poicy but for the agent providing it instructions. Finally, in Figure 2 (see video as well), we find affirmative evidence for Q3. We see that four robots equipped with Interactive Language policies can be guided at the same time by one operator. This language guided multi-robot control is, as far as we know, a capability not yet demonstrated in the literature. See the appendix for additional experiments comparing our transformer-based policy architecture compare to an existing visuo-linguo-motor baseline and studying how our presented approach scales with varying amounts of data.

Conclusion We have presented and analyzed the Interactive Language framework and we provide a number of associated assets, notably the Language-Table dataset and environment. We believe the scale of the dataset assets, the recipe used to produce them, the scale of the demonstrated policy diversity, and the exploration of new capabilities, each offer benefit to the research community in further advancing capable, realtime-conditionable visuo-linguo-motor robots.

References

[1] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[2] S. R. Branavan, H. Chen, L. Zettlemoyer, and R. Barzilay, "Reinforcement learning for mapping instructions to actions," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 82–90.

[3] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 1507–1514.

[4] D. Chen and R. Mooney, "Learning to interpret natural language navigation instructions from observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 859–865.

[5] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Experimental robotics*. Springer, 2013, pp. 403–415.

[6] F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro, "Emergent systematic generalization in a situated agent," *arXiv preprint arXiv:1910.00571*, 2019.

[7] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 894–906.

[8] S. Nair, E. Mitchell, K. Chen, S. Savarese, C. Finn, *et al.*, "Learning language-conditioned robot behavior from offline data and crowd-sourced annotation," in *Conference on Robot Learning*. PMLR, 2022, pp. 1303–1315.

[9] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[10] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

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

[12] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor, "Language-conditioned imitation learning for robot manipulation tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 139–13 150, 2020.

[13] K.-A. Kwon, R. J. Shipley, M. Edirisinghe, D. G. Ezra, G. Rose, S. M. Best, and R. E. Cameron, "High-speed camera characterization of voluntary eye blinking kinematics," *Journal of the Royal Society Interface*, vol. 10, no. 85, p. 20130227, 2013.

[14] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[15] D. A. Pomerleau, "Alvinn: An Autonomous Land Vehicle in a Neural Network," Carnegie-Mellon University, Tech. Rep., 1989.

[16] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1431, pp. 537–547, 2003.

[17] S. M. Khansari-Zadeh and A. Billard, "Learning stable nonlinear dynamical systems with gaussian mixture models," *IEEE Transactions on Robotics*, vol. 27, no. 5, pp. 943–957, 2011.

[18] S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert, "Learning movement primitives," in *Robotics research. the eleventh international symposium*. Springer, 2005, pp. 561–572.

[19] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[20] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.

[21] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5628–5635.

[22] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[23] Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp, "Goal-conditioned imitation learning," *Advances in neural information processing systems*, vol. 32, 2019.

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

[25] Y. Chebotar, K. Hausman, Y. Lu, T. Xiao, D. Kalashnikov, J. Varley, A. Irpan, B. Eysenbach, R. Julian, C. Finn, *et al.*, "Actionable models: Unsupervised offline reinforcement learning of robotic skills," *arXiv preprint arXiv:2104.07749*, 2021.

[26] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel, "A survey of reinforcement learning informed by natural language," *arXiv preprint arXiv:1906.03926*, 2019.

[27] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6629–6638.

[28] D. I. A. Team, J. Abramson, A. Ahuja, A. Brussee, F. Carnevale, M. Cassin, F. Fischer, P. Georgiev, A. Goldin, T. Harley, *et al.*, "Creating multimodal interactive agents with imitation and self-supervised learning," *arXiv preprint arXiv:2112.03763*, 2021.

[29] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics," 2014.

[30] D. Rakita, B. Mutlu, and M. Gleicher, "An autonomous dynamic camera method for effective remote teleoperation," in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2018, pp. 325–333.

[31] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "Hg-dagger: Interactive imitation learning with human experts," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8077–8083.

[32] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa, "Learning from interventions: Human-robot interaction as both explicit and implicit feedback," in *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[33] P. Kormushev, S. Calinon, and D. G. Caldwell, "Imitation learning of positional and force skills demonstrated via kinesthetic teaching and haptic input," *Advanced Robotics*, vol. 25, no. 5, pp. 581–603, 2011.

[34] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[35] A. Broad, J. Arkin, N. Ratliff, T. Howard, B. Argall, and D. C. Graph, "Towards real-time natural language corrections for assistive robots," in *RSS Workshop on Model Learning for Human-Robot Communication*, 2016.

[36] S. Karamcheti, M. Srivastava, P. Liang, and D. Sadigh, "Lila: Language-informed latent actions," in *Conference on Robot Learning*. PMLR, 2022, pp. 1379–1390.

[37] J. D. Co-Reyes, A. Gupta, S. Sanjeev, N. Altieri, J. Andreas, J. DeNero, P. Abbeel, and S. Levine, "Guiding policies with language via meta-learning," *arXiv preprint arXiv:1811.07882*, 2018.

[38] P. Sharma, B. Sundaralingam, V. Blukis, C. Paxton, T. Hermans, A. Torralba, J. Andreas, and D. Fox, "Correcting robot plans with natural language feedback," *arXiv preprint arXiv:2204.05186*, 2022.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[43] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.

[44] D. A. Pomerleau, "Efficient Training of Artificial Neural Networks for Autonomous Navigation," *Neural Comput.*, vol. 3, 1991.

[45] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[46] Y. Bisk, D. Yuret, and D. Marcu, "Natural language communication with robots," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 751–761.

[47] Y. Bisk, K. J. Shih, Y. Choi, and D. Marcu, "Learning interpretable spatial operations in a rich 3d blocks world," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[48] B. Wu, S. Nair, F.-F. Li, and C. Finn, "Example-driven model-based reinforcement learning for solving long-horizon visuomotor tasks," in *5th Annual Conference on Robot Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=_daq0uh6yXr

[49] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, 2022.

[50] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

Appendix for *Interactive Language*

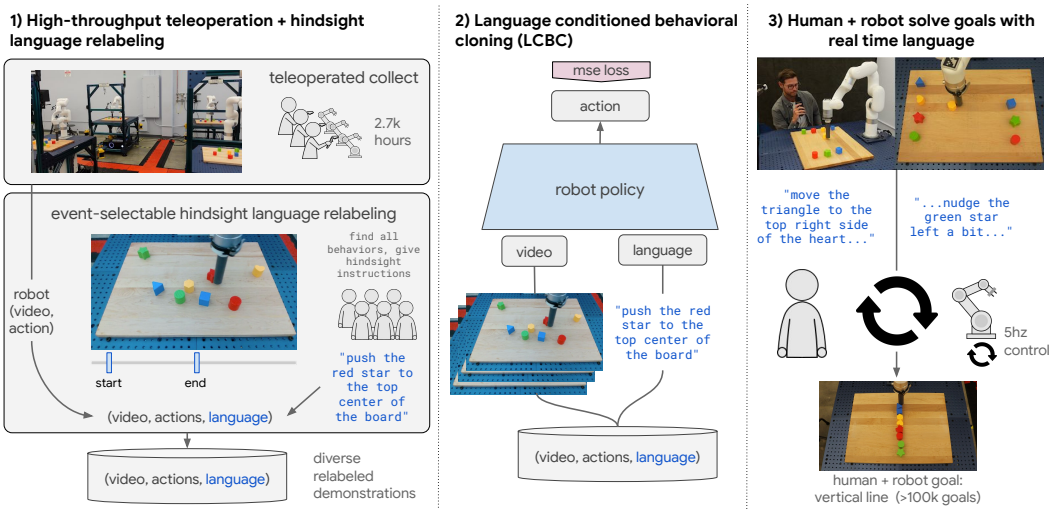

Figure 3: Interactive Language: a large scale robot imitation learning framework for real-time language. Stage 1: First, high throughput robot data collection with multiple operators. Post-collection, relabel robot video and actions into language conditioned demonstrations using event-selectable hindsight relabeling. Stage 2: do simple language conditioned behavioral cloning. Stage 3: Human guides a single learned policy in real-time using natural language to accomplish hundreds of thousands of goals.

A Data Collection

See Table 4 for statistics on our dataset collection. We find, perhaps surprisingly, that the main bottleneck in our data operation is *not* robot teleoperation but rather the crowdsourced language annotation that follows, with 18.06% of the raw data having undergone annotation prior to model training (5.5x as much unlabeled collected data as annotated data). This is true even though there are 16x as many hindsight annotators as robots. Bottlenecks like this may be addressed by exploiting language-free co-training [11], or by simply continuing to horizontally scale crowdsourced annotators. A simple way to address this bottleneck may be to continue to horizontally scale more crowdsourced annotators, or to make use of goal conditioned co-training capable of exploiting language-free robot data [11].

Real-World Data Collection	
Total robots	4
Total teleoperators	10
Total episodes	16.4k
Average episode length (minutes)	9.9
Total hours of collect time	2.7k
Hindsight Relabeling	
Total crowdsourced annotators	64
Total relabeled demonstrations obtained	299k
Total unique relabeled instructions	87k
Average relabeled demonstration length (seconds)	5.8
Total number of hours of relabeled demonstrations obtained	488
Total instruction hours / Collect hours	18.06%

Table 4: Statistics: real-world collection and relabeling. This data snapshot went into training and is a subset of the full Language-Table data.

	Has contact	Object/location -directed instructions	Compound instructions
Random window [8, 11]	86%	47%	16%
Event-selectable (ours)	91%	83%	$< 1\%$
Real test instructions	89%	84%	$< 1\%$

Table 5: Which relabeling strategy aligns best with test-time language?

B Event-selectable Hindsight Relabeling

A drawback of prior "random window" relabeling systems [8, 11] is that random windows are not guaranteed to contain "usefully describable" actions. We instead ask annotators to watch the full collect video, then find K coherent behaviors ($K = 24$ in our case), using temporal segmentation tools to mark the start and end frame of each behavior, and phrasing their descriptions as natural language commands. We instead ask annotators to watch the full collect video, then find K coherent behaviors ($K = 24$ in our case). Annotators have the ability to mark the start and end frame of each behavior, and are asked to phrase their text descriptions as natural language commands. In Table 5, we compare event-selectable relabeling to prior "random window" relabeling on a subset of our training data, finding that while both strategies tend to describe contact-rich behaviors, our analysis suggests event-selectable relabeling yields more well-matched data: fewer complex compound instructions, and more compositionally directed instructions.

C Architecture Details

In Figure 4, we describe our transformer-based [39] neural network policy architecture, mapping from video and text to continuous actions, which we refer to as LAVA ("Language Attends to Vision to Act"). Each training example consists of $(s,a,l)_i \sim \mathcal{D}_{\text{training}}$, where $s \in \mathbb{R}^{\text{seqlen} \times 640 \times 320 \times 3}$ is RGB observation history. We pass each frame in the video s through a convnet to obtain multi-scale visual feature descriptors (features at multiple layers). Our convnet consists of two Imagenet-pretrained ResNet [40,41] layers and two additional learned convolutional layers. l is embedded using a pretrained CLIP text encoder [42], which is finetuned on our in-domain data, but remains fixed during policy training. We fuse visual and lingual information using a "Language-Attends-to-Vision" transformer block, which performs cross-attention with the sentence token acting as query, and flattened multi-scale visual tokens acting as keys and values. This operation is applied to each image, and the sequence output is fed to a temporal prenorm [43] transformer, which is average pooled and fed to a deep residual multi-layer perceptron (MLP), outputting the predicted next action a.

D Short horizon behaviors

In Figure 5 see Interactive Language rollouts on a sample of the >87,000 crowdsourced natural language instructions we evaluate.

E Ablations

In Figure 6, we present results in simulation ablating (i) our transformer-based policy architecture LAVA against the FiLM-conditioned ResNet architecture in [9] and (ii) the amount of data provided to policy training. We report average success and SPL over the multi-task benchmark in Language-Table (see "Environment and Benchmark" in Section 4), and all numbers are reported with confidence intervals over three seeded training runs. We see the presented architecture is indeed responsible for significant gains over prior work in SPL, a path-length-aware success metric we find correlates best with real world quality in our setup. When sweeping the amount of training data, we find that policy performance is seeing diminishing returns, but not yet plateauing across each doubling of data. While perhaps surprising given the scale of our collect, we believe that this result highlights the environment's complexity as well as the difficulty of open vocabulary visuomotor learning.

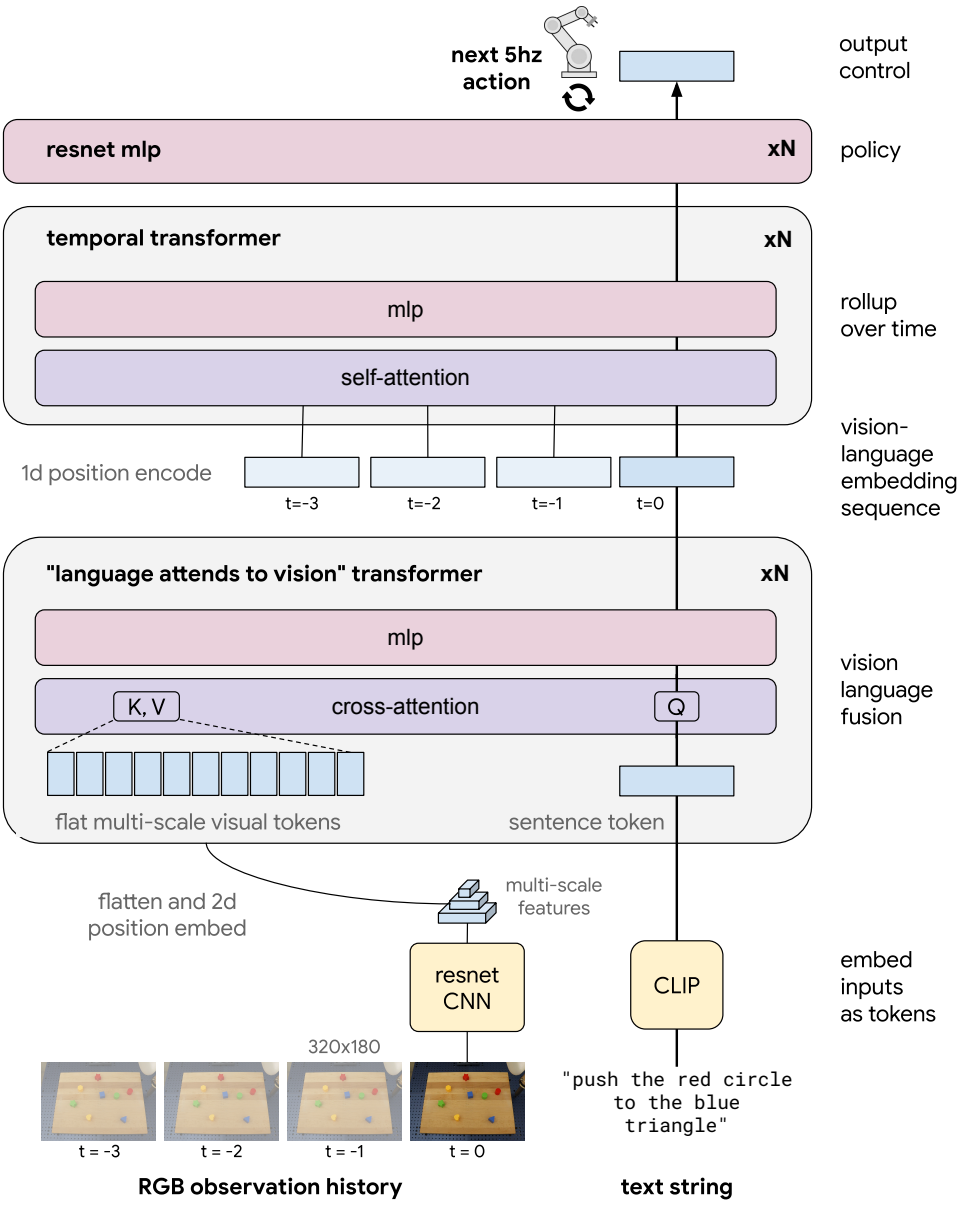

Figure 4: LAVA: our transformer-based architecture for language conditioned visuomotor control.

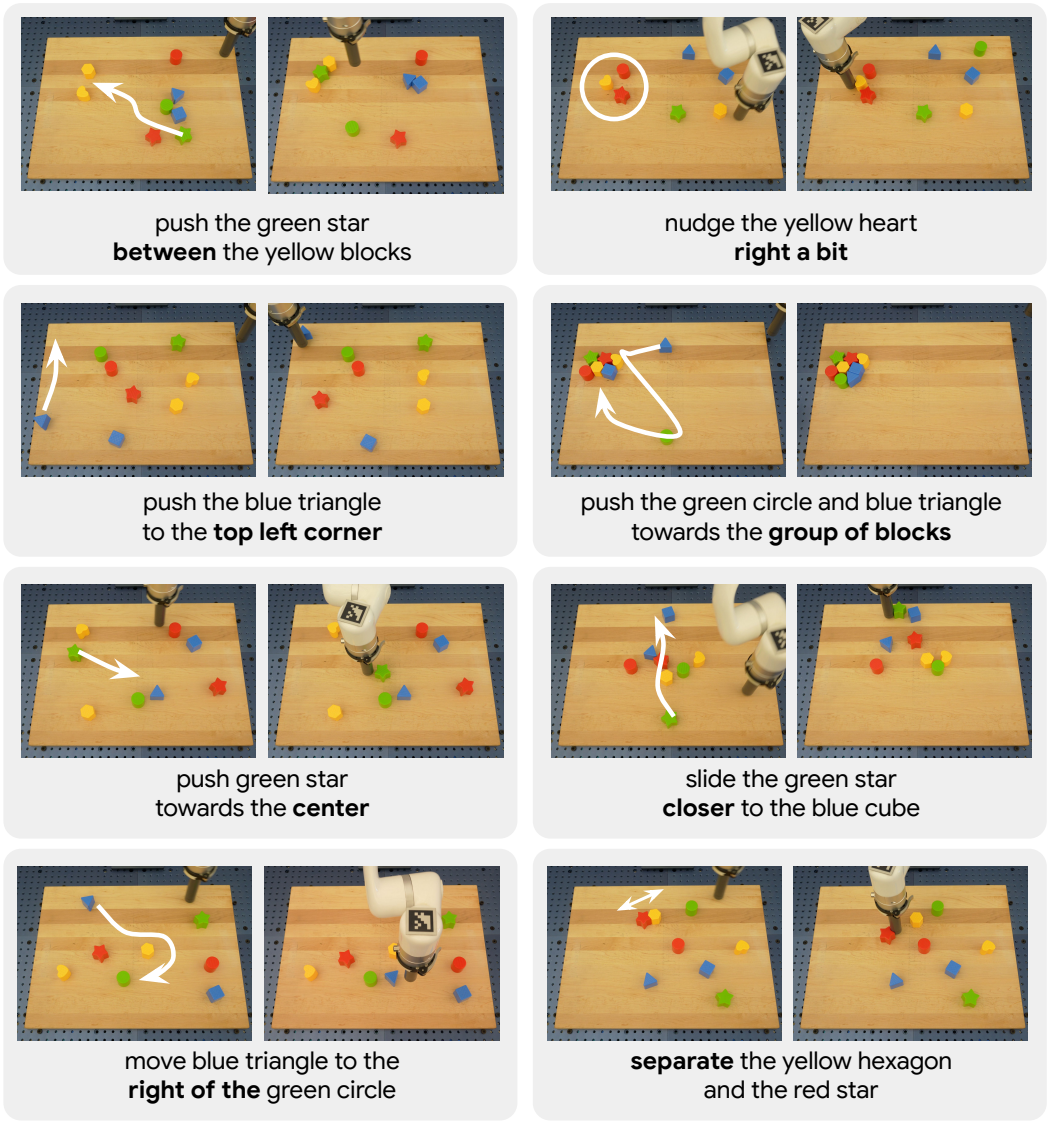

Figure 5: Learning a wide variety of short-horizon open vocabulary behaviors. Interactive Language rollouts on a sample of the >87,000 crowdsourced natural language instructions we evaluate.

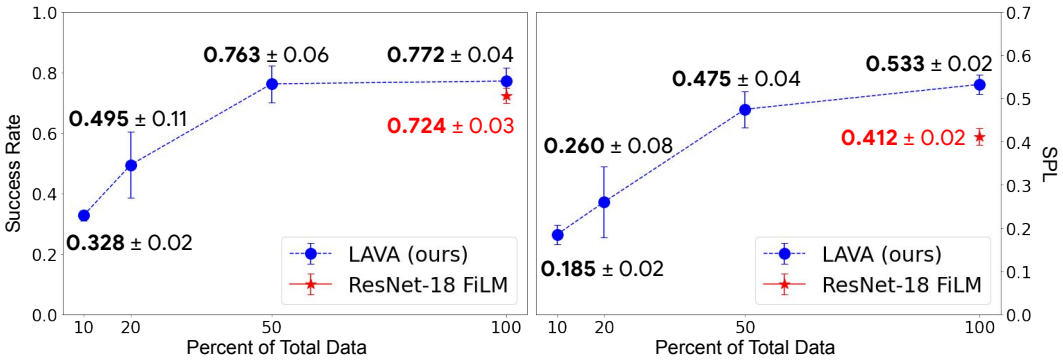

Figure 6: Ablations in simulation. We compare our LAVA transformer architecture to a baseline ResNet-18 FiLM model from [9], as well as ablate the amount of data provided to training. We find the average success-weighted path length (SPL) to be a better indicator of qualitative performance than (unweighted) average success.