RoboVQA: Multimodal Long-Horizon Reasoning for Robotics

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a scalable, bottom-up and intrinsically diverse data collection scheme that can be used for high-level reasoning with long and medium horizons and that has 2.2x higher throughput compared to traditional narrow top-down step-by-step collection. We collect realistic data by performing any user requests within the entirety of 3 office buildings and using multiple embodiments (robot, human, human with grasping tool). With this data, we show that models trained on all embodiments perform better than ones trained on the robot data only, even when evaluated solely on robot episodes. We explore the economics of collection costs and find that for a fixed budget it is beneficial to take advantage of the cheaper human collection along with robot collection. We release a large and highly diverse (29,520 unique instructions) dataset dubbed RoboVQA containing 829,502 (video, text) pairs for robotics-focused visual question answering. We also demonstrate how evaluating real robot experiments with an intervention mechanism enables performing tasks to completion, making it deployable with human oversight even if imperfect while also providing a single performance metric. We demonstrate a single video-conditioned model named RoboVQA-VideoCoCa trained on our dataset that is capable of performing a variety of grounded high-level reasoning tasks in broad realistic settings with a cognitive intervention rate 46% lower than the zero-shot state of the art visual language model (VLM) baseline and is able to guide real robots through long-horizon tasks. The performance gap with zero-shot state-of-the-art models indicates that a lot of grounded data remains to be collected for real-world deployment, emphasizing the critical need for scalable data collection approaches. Finally, we show that video VLMs significantly outperform single-image VLMs with an average error rate reduction of 19% across all VQA tasks. Thanks to video conditioning and dataset diversity, the model can be used as general video value functions (e.g. success and affordance) in situations where actions needs to be recognized rather than states, expanding capabilities and environment understanding for robots. Data and videos are available at anonymous-robovqa.github.io

1 Introduction

The field of textual high-level reasoning has seen major breakthroughs recently with large language models (LLMs) [28, 4], while progress has also been made in visual language models (VLMs) [8], high-level reasoning that is grounded in the real world remains a challenging task and critical for robotics. Can the state-of-the-art VLMs trained on available multimodal datasets perform grounded tasks with high accuracy in the real-world? We aim to answer the question by showing that new large scale data collection are still needed to achieve lower error rates outside of lab environments. A

Submitted to NeurIPS 2023 6th Robot Learning Workshop: Pretraining, Fine-Tuning, and Generalization with Large Scale Models. Do not distribute.

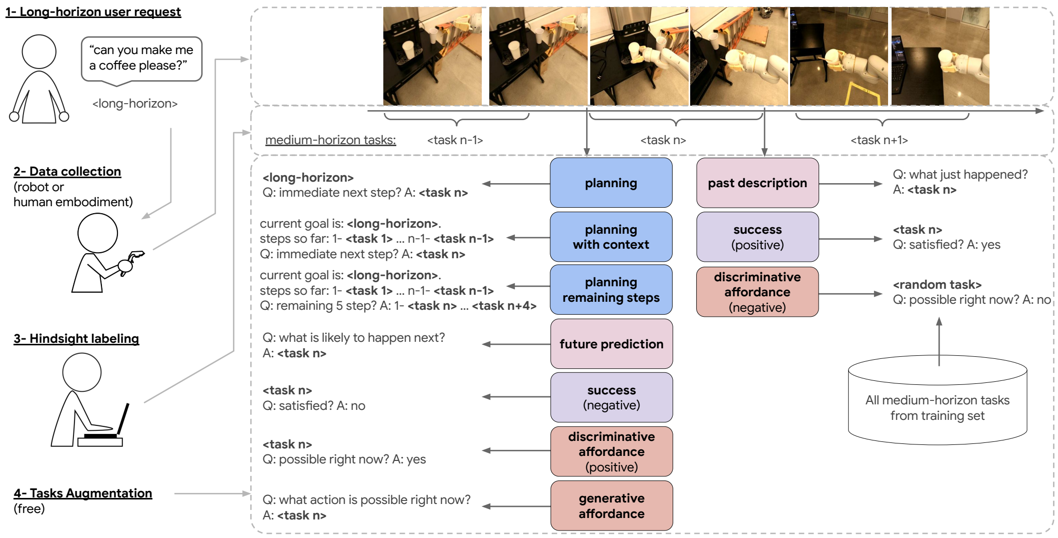

Figure 1: Data collection procedure: Given long-horizon user requests, a human operator teleoperates a robot to fulfill the task. Medium-horizon tasks are then labeled in hindsight via crowd-sourcing, with temporal segmentation and task instruction for each segment. Finally, from a sequence of labeled segments, we automatically generate 10 types of question/answer pairs.

major difficulty for VLMs stems from the high-dimensionality of the real world which, accordingly requiring large amounts of multimodal data (video, language, actions) for training. Hence a major contribution of our work is to validate more efficient data collection approaches than the traditional top-down step-by-step collection [2], by reducing overheads such as resets and scene preparations and leveraging the low costs of human embodiment collection. With a crowd-sourced bottom-up approach where long-horizon tasks are decided by real users the resulting medium-horizon steps are naturally highly diverse, relevant and on-distribution for users. Not only it is a more efficient way to collect medium-horizon steps, we also get long-horizon coherent sequences which can train models to perform planning tasks. With a 2.2x throughput increase compared to the traditional method, it is preferable to collect data this way even if long-horizon tasks are not needed. While we do collect robot actions in this dataset, the focus of this paper is on high-level reasoning tasks, we can hence train on embodiments which do not come with motor commands and observe transfer of knowledge between embodiments. We find in Sec. 9.3 that for a fixed collection budget, it is beneficial for high-level reasoning to jointly with cheaper human embodiment even when evaluating on the robot embodiment only.

Our contributions can be summarized as follows:

1. We demonstrate a scalable, bottom-up and intrinsically diverse data collection scheme that can be used for high-level reasoning with long and medium horizons and that has 2.2x higher throughput compared to traditional narrow top-down step-by-step collection and show additional cheap human embodiment data improves performance.

2. We release a large and diverse cross-embodiment dataset of 829,502 (video, text) pairs for robotics-focused visual question answering.

3. We demonstrate a single video-conditioned model trained on the dataset that is capable of performing a variety of tasks with higher accuracy than baselines and is able to guide real robots through long-horizon tasks.

4. We establish a robotics VQA benchmark and long-horizon planning benchmark with an intervention mechanism on real robots providing a single performance metric and enabling performing tasks to completion, making it deployable with human oversight even when imperfect.

2 Data

Collection & Dataset: In Fig. 1 we describe the collection process, from user request to VQA tasks generation. We collect episodes from any long-horizon tasks within the entirety of 3 office buildings and with 3 embodiments (Fig. 3), resulting in 238 hours of video (10 days), 5,246 long-horizon episodes and 92,948 medium-horizon episodes. The average long-horizon episode lasts 102 seconds, the medium-horizon average is 14s. Because evaluation of freeform text answers are performed by humans in our experiments, we keep the validation and test sets small on purpose with approximately 1,000 VQA entries for each (coming from 50 episodes each). While there can be overlap in scenes between training and val/test, there is no overlap in episodes. For more statistics, see Sec. 9.2.

Task diversity: To ensure that our dataset and benchmark do not overfit to a specific environment, domain or task, we collect examples over a wide range of tasks compared to more traditional collections [1] where a fixed and small list of tasks is decided in advance by researchers and engineers in a top-down fashion. We opt for a bottom-up approach where a large number of tasks are crowd-sourced by users and tele-operators. This favors breadth and a better alignment with a distribution of requests coming from real users. This results in high tasks diversity (26,798 unique medium-horizon instructions, 2,722 unique long-horizon instructions).

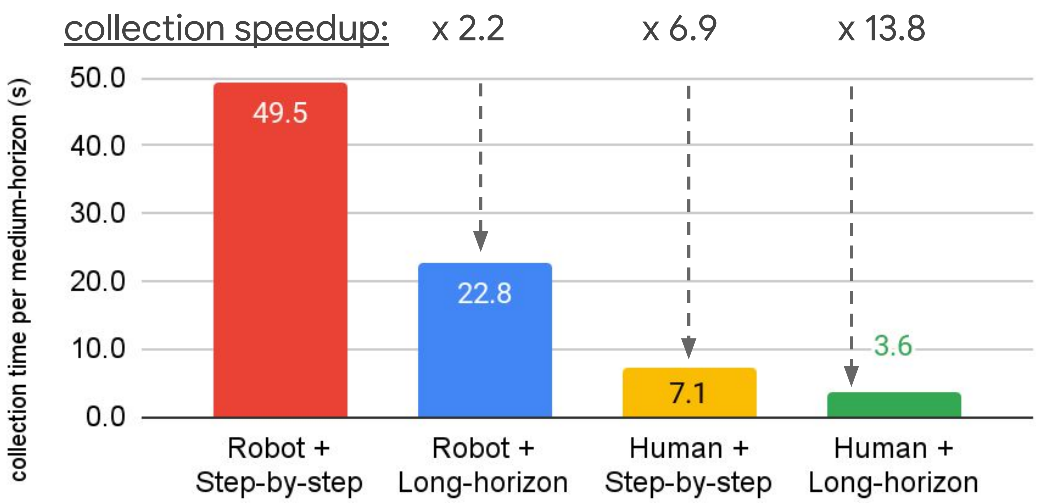

Figure 2: Throughput gains compared to the traditional top-down step-by-step collection approach. The throughput of our long-horizon collection is 2.2x higher for robot collection and 13.8x higher with human bodies (compared to the robot used in our experiments).

Throughput and costs: Much of the throughput gains reported in Fig. 2 come from collecting medium-horizon episodes in a continuous fashion without needing to reset the scene or the robot. Note that the hindsight labeling process can be parallelized via crowd-sourcing and does not impact the throughput if performed in parallel, however it remains a cost in the collection budget. The VQA tasks however are generated for free by taking advantage of the known sequence of past and future tasks and positioning the questions in time with respect to different known semantic points (e.g. before or after a medium-horizon task was performed).

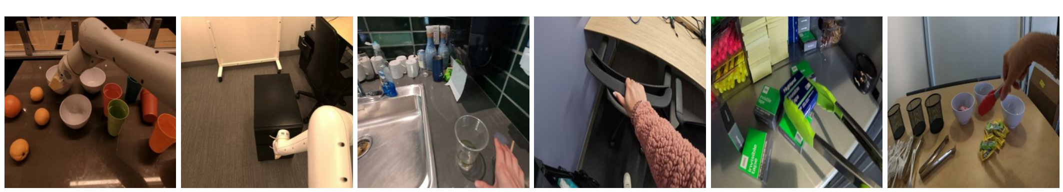

Figure 3: Examples of 3 embodiments in the dataset: robot, human (single) arm, human using a grasping tool.

Chain-of-Thought: Decomposing high-level goals into the defined tasks allows for robots to manifest its thinking process when carrying out long-horizon plans. Moreover, these tasks are provided as

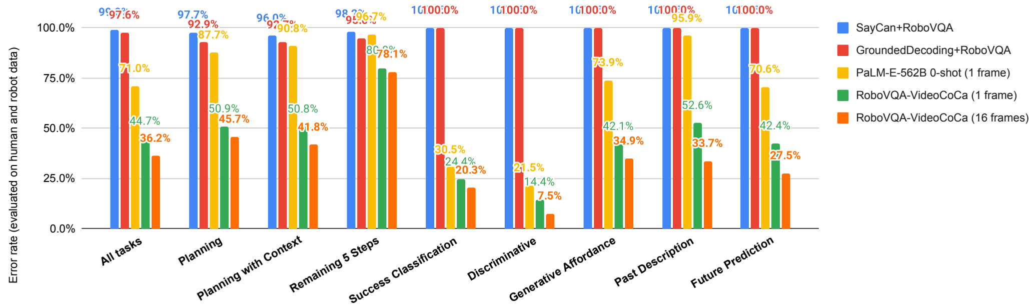

Figure 4: VQA Error rates: we evaluate all models on the test set using human raters. We observe that state-of-the-art methods do not perform well in realistic settings in zero-shot, thus motivating the need for further scalable data collections. We also observe substantial gains when using video (16 frames) vs image conditioning.

natural language questions and answers, and can be viewed as a series of Visual Question Answering (VQA) steps. This formulation is similar to chain-of-thought for language model prompting [39]. We also note concurrent work [12] which demonstrates that mimicking step-by-step human thought improves planning accuracy.

3 Models

3.1 RoboVQA-VideoCoCa

We train a new model called RoboVQA-VideoCoCa derived from the **VideoCoCa** model [41], which is a video language model extending CoCa [43]. It uses an encoder-decoder architecture combining contrastive pretraining (like CLIP [31]) as well as generative pretraining (like SimVLM [38]) between video and text modalities. Unless otherwise stated, we use a VideoCoCa base model of 383M parameters with the initial checkpoint trained on image-captioning tasks as the original paper did, and fine-tune the model on the RoboVQA video-text datasets. We choose a video-conditioned model to explore the importance of video in answering the visual questions in our dataset and find substantial benefits to video conditioning (see Fig. 17 and 16).

3.2 Baselines

To compare with our finetuned model, we consider the following state-of-the-art baselines which have similar capabilities in visual question answering and planning for robotics.

PaLM-E [8] is a visual language model built from pretrained ViT [3] and PaLM [4] LLM models, which projects images into the token embedding space of the pretrained LLM. In our experiments we test PaLM-E-562B *zero-shot*, without training on RoboVQA dataset. While not finetuning is not a head to head comparison of models, the point of this comparison is establish how well state-of-the-art models trained on prior datasets can perform in the real world, and motivate further scalable data collection efforts to address the remaining performance gap.

Planning Methods. We experiment with four baseline planning methods: two of which use RoboVQA-VideoCoCa and PaLM-E (zero-shot), as end-to-end planning models. As two other baselines, we adapt the methods of **SayCan** [1] and **Grounded Decoding** [14], which use a text-only LLM (PaLM [4]) in either phrase-level or token-level decoding guided by a visual affordance function (using RoboVQA-VideoCoCa as a video value function for affordance).

4 Benchmarks

4.1 VQA Benchmark

We first evaluate the model performance on individual tasks, where each task consists of a video segment and a question. The inference result is compared using exact match against prior human

Model	Cognitive Model: Training procedure	Cognitive Model: Size	Cognitive Model: Inference time	Cognitive Model: # frames	Physical Model (policy)	Multi-turn Long-Horizon Planning: Total # tasks	Multi-turn Long-Horizon Planning: Total # steps	Multi-turn Long-Horizon Planning: domain	Multi-turn Long-Horizon Planning: bodies	Intervention Rate (per episode average): cognitive	Intervention Rate (per episode average): physical	Intervention Rate (per episode average): average
Evaluation #1: 100 long-horizon multi-turn planning tasks on **pre-recorded** videos (robot and human embodiments)												
SayCan / PaLM	Language pretraining only & RoboVQA Affordance Model	540B	150h+ (30k affordances)	1	Pre-recorded video	100	854	Broad	Robot & Human (50/50%)	98.8%	100% (teleop.)	99.4%
Grounded Decoding / PaLM			~10s (8 affordances)	1						95.5%		97.8%
PaLM-E	(Zero-Shot) Finetuned on SayCan/ Fractal	12B	1s	1						81.4%		90.7%
RoboVQA-VideoCoCa (ours)	Finetuned on RoboVQA	383M	1s	16						**44.0%**		72.0%
Evaluation #2: 10 long-horizon multi-turn planning tasks in a **live real-world** setting, with human teleoperation as policy												
PaLM-E	(Zero-Shot) Finetuned on SayCan/ Fractal	12B	1s	1	Live human teleop.	10	~60	Broad	Robot	78.2% ± 7.6%	100% (teleop.)	92.8%
RoboVQA-VideoCoCa (ours)	Finetuned on RoboVQA	383M	1s	16						**47.67%** ± 9.1%		73.8%
Evaluation #3: 1 long-horizon multi-turn planning tasks in a live real-world setting with a policy X for control (**fully autonomous**)												
RoboVQA-VideoCoCa (ours)	Finetuned on RoboVQA	383M	1s	16	policy X	1	5	Narrow / Easy	Robot	40.0%	0% (easy tasks)	20.0%

Figure 5: Planning benchmarks with Intervention: evaluation #1 evaluates 854 planning steps on long-horizon episodes from RoboVQA dataset, evaluation #2 is performed live on a robot teleoperated by a human, while evaluation #3 is controlled end-to-end by our model and a policy. Note that thanks to human intervention in the loop, all tasks are performed to completion even when the model makes mistakes.

evaluation results stored in a central database as correct/incorrect for the video-question pair. The inference results for which no match is found are then collected for human raters to evaluate. During evaluation, a human rater is presented with the exact video segment and question as presented to the model. The rater is asked to either mark the model-generated answer as correct or incorrect, in which case the rater can propose a correct answer. All answers are added to the database, with the correctness of each answer marked accordingly.

We report the error rate for all models in Fig. 4 and find that there remains a substantial gap in performance for zero-shot state-of-the-art models compared to the finetuned model. While this is not too surprising, it is a valid question to ask when seeing good qualitative results by recent VLMs. Here we quantitatively prove that further scalable data collection efforts are required when deploying in the real world. In this graph we also make the case for video conditioning over image conditioning by presenting substantial gains with the former.

4.2 Planning Benchmark with Intervention

Intervention: In Fig. 5, we propose 3 different evaluations of long-horizon planning. Each evaluation is measured by intervention rate, which we further decompose into *cognitive* for the high-level text domain and *physical* for the low-level motor command domain. However all progress can be measured with the single intervention rate which averages the cognitive and physical rates. This distinction is useful when physical actions are teleoperated (100% physical intervention) to decouple high-level evaluations from low-level ones. Because the RoboVQA dataset is very broad and diverse, we need an evaluation procedure that can test that entire breadth. Current low-level policies however tend to only perform in very narrow domains, this decoupling thus allows us to test the full breadth of tasks in evaluations #1 and #2. See Fig. 6 for an example of cognitive intervention in the chat window between the user, the model and the intervention operator.

Offline Video Results: In evaluation #1, we run models on 100 long-horizon episodes (robot and human embodiments) from the RoboVQA dataset which amounts to 854 planning steps in total. Models are given the long-horizon instruction and need to output medium-horizon plans, which are graded by humans. Note that the SayCan and Grounded Decoding baselines have slow inference time which makes them impractical to run in a live settings (hence not showing in other evaluations). Similarly, the inference time of the PaLM-E 562B model is too slow for real time (30s), so we use a smaller version here. Note that despite being is 30x smaller, our model outperforms the state-of-the-art model by 46%.

Live Real-world Results: In evaluation #2, the high-level models are given a long-horizon instruction and provide medium-horizon plans in real time to a real robot teleoperated by a human. In evaluation #3, a policy is deployed instead of a human teleoperator, but the domain is a lot narrower given the limited abilities of the policy. See videos of these evaluations at anonymous-robovqa.github.io. While with evaluation #3 we can obtain a much lower intervention rate thanks to the policy deployment, the domain is a lot narrower and emphasizes the need for a decoupled evaluation for high-level reasoning in broad domains.

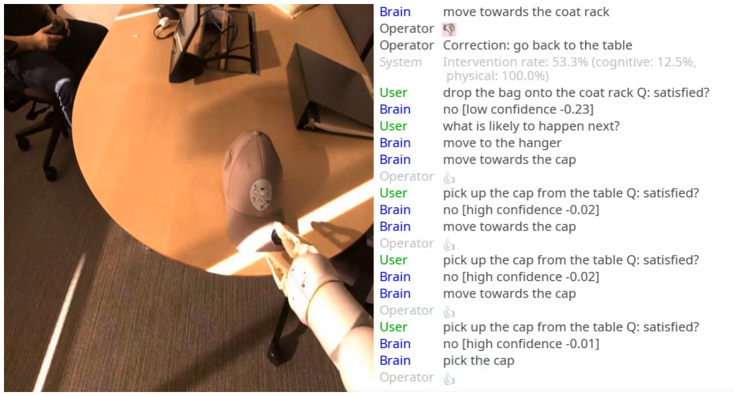

Figure 6: Example of grounded chat with cognitive intervention. Our model "Brain" is tasked with the following task at the beginning of the chat: "take the bag and cap on the desk and hang them on the coat rack" in this case. The bottom of the chat shows the most recent messages. The model is ran on an existing long-horizon video from the RoboVQA dataset and produces medium-horizon plans to fulfill the long-horizon request. An operator is in the chatroom and validates each plan or provides a correction if incorrect. The user is also able to ask questions at any point in time. Here we see that the operator intervened and the system reported a cognitive intervention rate of 12.5% at this point of the episode.

5 Analysis

5.1 Task Augmentation Matters

In Fig. 7 we trained models on different following set of tasks: planning only, context-planning only, planning + success + affordance, context-planning + success + affordance, or all tasks. Note that when comparing planning vs. all tasks, the model trained on planning only sees 38M examples of planning task, while the one trained on all tasks sees roughly 1/8 the number of samples for the planning task. We find that the model trained on all tasks is often better of comparable than the models dedicated to a subset of tasks, with the exception of the success task. For example training on all tasks leads to better planning (70.9% error) compared to training on planning only (77.2% error). From a collection cost perspective, it is interesting to note that despite coming from the exact same set of instructions, the free tasks augmentation yields better results at no extra cost, hence task augmentation matters for performance and collection scalability.

5.2 Tasks Transfer via Cross-Embodiment Data

In Fig. 14, we compare error rates on the test split using RoboVQA-VideoCoCa trained on robot embodiment only, human embodiment only, and their combination. The test set contains only robot embodiment data. Despite cross-embodiment, we find that errors are below 100% for all tasks when training on human data only, indicating human data by itself is useful to acquire a grounded understanding of videos with robot embodiment. Furthermore, training on both embodiments performs better than training on robot data only, indicating that extra data with human embodiment does not hurt performance when evaluating on the robot embodiment. We use [1] as a baseline, which uses a small, fixed list of 60 tasks and can only be evaluated on the planning task. We also provide the affordance answers from RoboVQA as affordance function to SayCan for planning. Similarly, we evaluate on the joint human and robot test split in Fig. 15. While it is not surprising that training on both embodiments performs best on the robot+human test set, we also shows it is the most general model as it performs better in all situations. More analysis is available in Sec. 9.3.

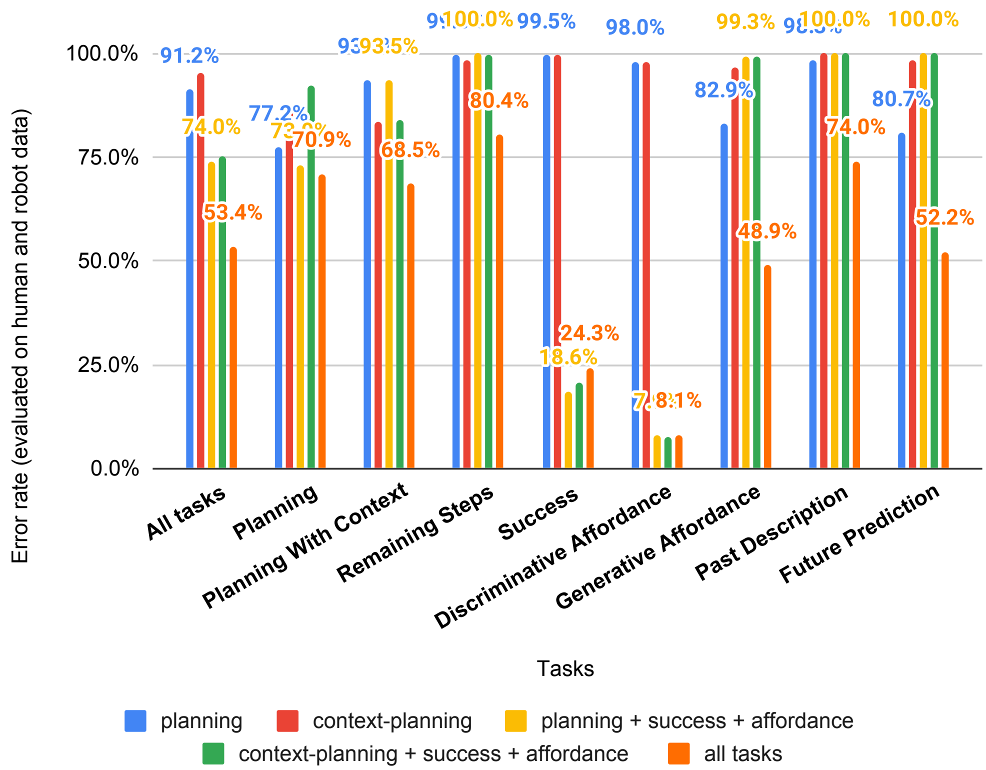

Figure 7: Error rates for models trained with different sets of tasks. Each model is trained and evaluated on the (robot + human) dataset, but using different subsets of tasks. We find that training on all tasks leads to better planning (70.9% error) compared to training on planning only (77.2% error).

5.3 Importance of Video modeling

We investigate performance gains from video by training our model with (1, 2, 4, 8, 16) frames in 16 and find substantial error reductions in Fig. 17 between 1 and 16 frames. As expected, modeling with more frames yields better results, as it captures longer temporal dynamics for more accurate visual grounding.

5.4 Video Value-Functions

We evaluate our model as a general grounded value-function from video and observe that it can provide stable binary detections as shown in Fig. 8. Moreover, when filtering by the confidence of the yes/no tokens, we can further improve the accuracy of the success detection. These value functions can be used for closed-loop planning to know when a step is performed. Additionally, thanks to the dataset breadth and to video conditioning, the value functions can give richer understanding than traditional image-based success or affordance detectors.

6 Related Work

Vision-Language Models. Recently many methods [31, 16, 18, 43, 38, 11, 3] have been proposed that aim to train vision-language models (VLMs) on large-scale image-text pair datasets. We find the features learned by these methods generalize to robotic datasets. In this work, we also fine-tune a pre-trained vision language model called VideoCoCa [41] on conversation data grounded in long-horizon videos. The advantage of this VLM is that it is the encoder can consume full videos which helps in fine-grained temporal reasoning required to solve the tasks introduced in the RoboVQA benchmark.

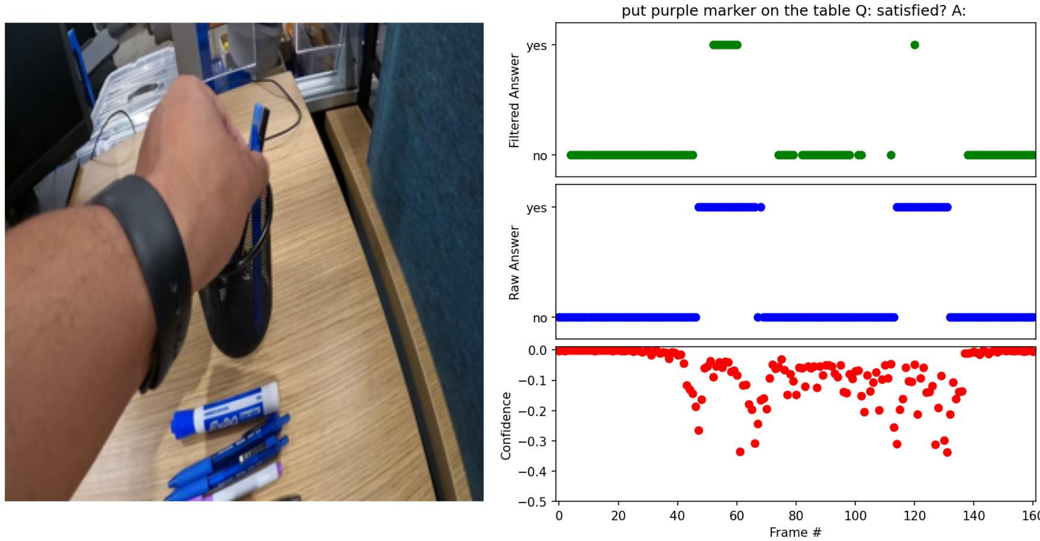

Figure 8: RoboVQA-VideoCoCa used for video success detection. In blue are the raw answers to the question "put purple marker on the table Q: satisfied? A:", the confidence is shown in red and the answer filted by confidence is shown in green.

Video Captioning. Our task is closely related to the task of video captioning [37, 9, 30, 22, 19] which is a well studied problem in computer vision. In fact, we fine-tune a pre-trained video-captioning model VideoCoCa on these long-horizon videos. Different from the video captioning problem, all the videos in our fine-tuning dataset are egocentric. Also, we collect segment labels for a long-horizon task executed by either a robot or human. Furthermore, we augment these segments with a variety of question-answer pairs that add more supervision to the model so that an agent can execute long-horizon tasks.

Video Datasets with Text Annotations. Recently many large-scale video datasets have been introduced [7, 33, 17, 44, 26, 42, 40, 10] that include videos of humans performing tasks with text narrations or question-answer annotations. Ego4D is the most similar dataset to the RoboVQA dataset because Ego4D also has egocentric view of daily human activities annotated with dense narrations. However, our dataset differs in two key aspects. First, we collect human and robot interactions in the same environment. Second, our focus is on tasks that a robot is capable of doing. We hope that by lowering the domain gap between the human and robot videos we can achieve more transfer from human videos (which are faster to collect) to robot videos. [25] also explores scalable ways to collect language data with unstructured play [23], however they rely on an LLM requiring a prompt with a scene description that matches the environment's state and is limited to 25 medium-horizon instructions. Like RoboVQA, TEACh[29] is another dataset that also contains interactive dialogues required to solve household tasks. However, TEACh consists of data in simulated environments while our dataset is collected in real kitchen and office environments with both humans and robots.

Language Models for Planning. [13] used a large language model (LLM) to produce plans for robotic tasks. This has been followed up by many works that also use LLMs to produce feasible next steps for a robot [1, 8, 35, 34, 21]. One advantage of using LLMs to plan is that the output of these models can be used as input to language-conditioned policies [15, 2, 24] that may have been trained independently.

Intervention Rate. Intervention Rate is a commonly used evaluation metric [36, 27, 32] in robotics and self-driving car literature for measuring the performance of policies. In this work, we use it as a metric and as a mean to perform all tasks to completion, a necessary condition for real-world deployment.

Chain of Thought Prompting. [20, 5, 39] use the idea of prompting a language model with the process or steps to perform a reasoning task. The authors observe that prompting allows the model to improve performance on symbolic reasoning tasks like algebraic problems. Inspired by those results, we also provide rationale or thought supervision to the model by providing the sub-tasks as hindsight labels for successfully achieving the long-horizon task.

7 Limitations

Some long-horizon episodes may be too repetitive and easy, thus we have filtered out episodes with more than 5 identical medium-horizon steps. Subsequently we observed gains in generalization. Additionally we have not compared the effectiveness of the proposed human-and-robot dataset/benchmark with human-only dataset/benchmarks like Ego4D [10], EpicKitchens [6] etc., which merit careful study in our future work.

8 Conclusion

We have shown a long-horizon collection approach with higher throughput and high diversity and breadth and released the resulting dataset for the benefit of the robotics community. We have demonstrated on real robots a number of capabilities learned with this dataset and established planning benchmarks with intervention as a metric and as a means for deployment.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022.

[3] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[5] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[8] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.

[9] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9): 2045–2055, 2017.

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[11] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16399–16409, 2022.

[12] Shengran Hu and Jeff Clune. Thought cloning: Learning to think while acting by imitating human thinking. 2023.

[13] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *CoRR*, abs/2201.07207, 2022. URL `https://arxiv.org/abs/2201.07207`.

[14] Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, et al. Grounded decoding: Guiding text generation with grounded models for robot control. *arXiv preprint arXiv:2303.00855*, 2023.

[15] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-z: Zero-shot task generalization with robotic imitation learning. In *5th Annual Conference on Robot Learning*, 2021. URL `https://openreview.net/forum?id=8kbp23tSGYv`.

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[19] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022.

[20] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL `https://aclanthology.org/P17-1015`.

[21] Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023.

[22] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[23] Corey Lynch and Pierre Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020. URL `https://arxiv.org/abs/2005.07648`.

[24] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *arXiv preprint arXiv:2210.06407*, 2022.

[25] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[26] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[27] Robin R Murphy and Debra Schreckenghost. Survey of metrics for human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 197–198. IEEE, 2013.

[28] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[29] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-driven Embodied Agents that Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[30] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Dominik Riedelbauch, Nico Höllerich, and Dominik Henrich. Benchmarking teamwork of humans and cobots–an overview of metrics, strategies, and tasks. *IEEE Access*, 2023.

[33] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[34] Tom Silver, Varun Hariprasad, Reece S Shuttleworth, Nishanth Kumar, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. PDDL planning with pretrained large language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. URL `https://openreview.net/forum?id=1QMMUB4zfl`.

[35] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *arXiv preprint arXiv:2212.04088*, 2022.

[36] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 33–40, 2006.

[37] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4213–4222, 2018.

[38] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022.

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[40] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021.

[41] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners, 2023.

[42] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022.

[44] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019.

9 Appendix

9.1 Random frames from training set

Figure 9: Random frames from training set.

9.2 Dataset Statistics

As reported in Fig. 10, the entire dataset is a collection of 5246 long-horizon episodes (5046 for training and 100 for validation). Each episode has 1 long-horizon instruction and a varying number of medium horizon instructions that are temporally segmented. There are 2638 unique long-horizon instructions in the training set. Each unique long-horizon instruction has an average of 2.01 episodes collected, median is 1 and maximum is 90. See Fig. 11 for the number of training episodes per long-horizon instruction. In Fig. 12 we show the number of training episodes that have the same long-horizon instruction as a test episode. We find that 46% of the test episodes do not have a long-horizon match in the training set. We show random frames from the training set in Fig. 9 and random long and short horizon instructions from the training set in 9.4. We also provide extensive analysis of the language found in the training set in 9.5 by automatically breaking down short-horizon instructions by categories (objects, actions, locations and attributes) using an LLM. This analysis found 2862 objects (e.g. ”tissue box”, ”purple color plate”), 680 skills or verbs (e.g. ”add something into something” or ”go out of a room”), 3322 locations or spatial relations (e.g. ”in the green plate”, ”left trash can”) and 901 attributes (e.g. shapes, color). Note that these numbers are only indicative as some objects can be redundantly described for example, see 9.5 for more details.

9.3 Comparing Embodiment Mixtures

Robot collection throughput will often be a factor of the cost including time, money, tele-operator training and availability, hardware maintenance etc., while humans are already expert of their own embodiment, collecting data with much less cost and cycle than robots. When factoring in all of these parameters into a collection budget, we can see that robot-to-human collection cost ratios and throughputs can vary wildly depending on all of these parameters. It is hence a critical question while scaling up data collection to know which data mixture for a given budget leads to the lowest error rates.

We explore this question in Fig. 13 by looking at the data yields for a fixed collection budget of 500,000 VQA conversations, and report the performance for different configurations in Figure 13-b to analyze the trade-offs between different mixtures. We find that even if the robot-human ratio is 1.0 and only evaluating on the robot test set, the error rate is comparable when training on the equal robot250k-human250k mixture (62.4%) compared to the full 500k robot dataset (62.7%), while also being significantly lower on the human test set (53.9% vs 67.0%). Not only there is no downside for the robot performance to mix human data, it also makes the model more general and usable for other applications that require human embodiment understanding.

Similarly we find that when the robot-human cost ratio is 4.0, the performance of the mixed dataset (robot-62k + human-250k) on the robot test set is similar to the robot-only 125k dataset (65.3% vs 63.5%) while also being significantly lower on the human test set (51.1% vs 68.7%). We also observe that the performance gains seem rather small when training on 500k robot samples vs 125k, and that performance on human data degrades slightly when increasing robot data from 62k to 250k. We conclude that this analysis validates the common intuition that human data collection is an efficient way to scale up data collection for robots, despite the embodiment differences.

9.4 Instructions Samples

We print 50 random instructions from the training set for both long-horizon and short-horizon below to get a sense of what the data looks like.

50 long-horizon instructions:

- please place all of the highlighters into the pen holder
- please clean up the spill and put cup back on mouse pad
- Please flip the bowls and pickup the yellow, pink and green candies from the floor and place them in bowls. Then restock the chips into the bin.
- please grab a small bin from the cart, place it on the table, put the red pens on the table in it, then put it back on the supply cart
- empty the chips onto the counter

	Entire dataset	% of data	Training set	Validation set
VQA tasks (8 types)				
# (video, text) pairs	829,502	-	798,429	18,248
Long-horizon instructions				
# instructions	5,246	-	5,046	100
# unique instructions	2,722	-	2,638	94
average length	163.4s (2m 7s)	-	163.6s	161.0s
Medium-horizon instructions				
# instructions	92,948	-	89,227	1,850
# unique instructions	26,798	-	25,880	885
average length	14.2s	-	14.2s	13.5s
Episodes				
# episodes	5,246	100.0%	5,046	100
# robot episodes	2,350	44.8%	2,274	41
# human episodes	2,896	55.2%	2,772	59
total duration	238.0 hours (~10 days)	-	229.3 hours (~10 days)	4.5 hours
average # medium-horizon steps per episode with low overlap (<.5)	9.5	-	9.5	10.0
Locations (# long-horizon episodes)				
Building 1	3,190	60.8%	3,078	58
Building 2	1,507	28.7%	1,442	32
Building 3	485	9.2%	464	10
Unkown building	64	1.2%	62	0
Language analysis (approximate)				
# unique objects	2862	-	2773	254
# unique verbs	680	-	671	115
# unique locations	3322	-	3199	220
# unique attributes	901	-	861	108
Robot data				
# long-horizon instructions	2350	-	2274	41
# medium-horizon instructions	61153	-	58916	1140
# unique long-horizon instructions	1214	-	1181	37
# unique medium-horizon instructions	19448	-	18772	597
total duration	185.3 hours			
Human data				
# long-horizon instructions	2896	-	2772	59
# medium-horizon instructions	31795	-	30311	710
# unique long-horizon instructions	1551	-	1499	57
# unique medium-horizon instructions	8786	-	8499	300
total duration	52.7 hours			

Figure 10: Dataset statistics.

- Please flip the bowls and pickup the yellow, pink and green candies from the floor and place them in bowls. Then place the tongs into the bins.

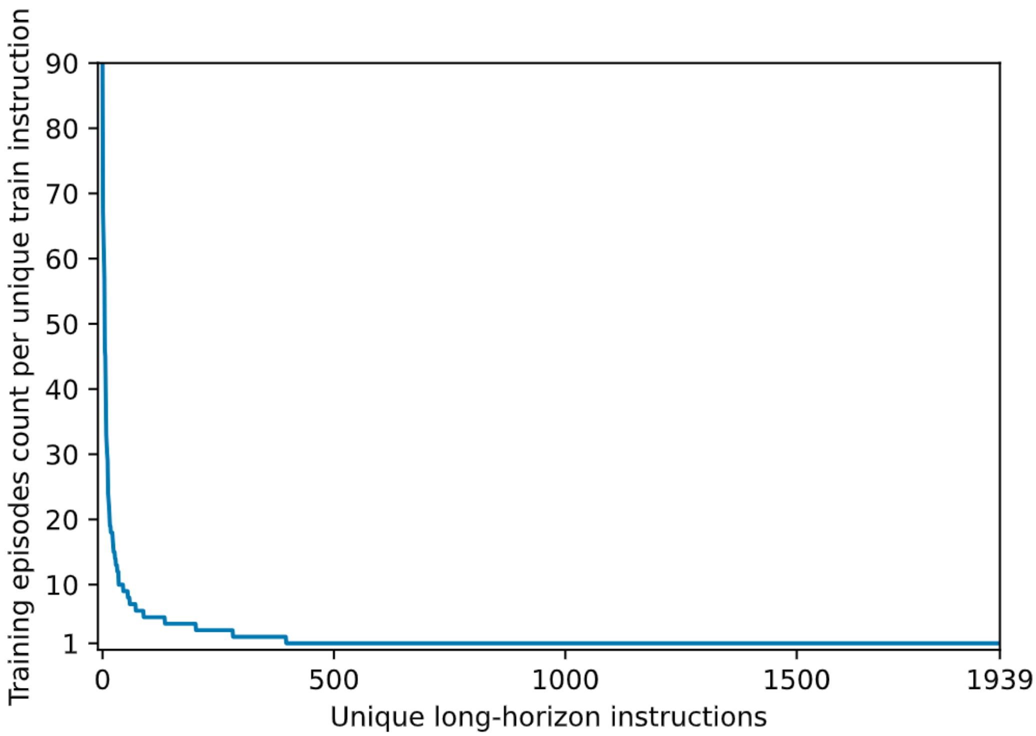

Figure 11: Number of training episodes per unique instruction: the maximum number of episodes for a unique long-horizon instruction is 90, the average 2.01 and the median is 1. There are 3894 training episodes which yield 1939 unique long-horizon instructions.

- Please flip the bowls and pickup the yellow, pink and green candies from the floor and place them in bowls. Then pick up the tongs from floor and place in bins.
- please clean up the pistachios spill on desk
- I am feeling a little sick, can you please get me a covid test in the cabinet at the end of the building, as well as return it back onto my desk.
- put fruit on the bookshelf
- fill the bowl with apples
- prepare a cup of coffee with the espresso machine.
- place candies into middle bowl and blue chip bag in left bowl
- place items from counter to bin
- I don't want the water anymore. Can you pour the water into the sink and then throw the cup away
- move items from table to cart
- can you take the wireless mouse box out of the filing cabinet and put it on top of the table for me
- I am done using the room can you turn off all the lamps.
- Tidy up the mk table by straightening out the fruit labels, lining up the utensil holders and straightening the honey bottle platform
- there is rubbish on the table, please throw them away into the correct places in the disposal bins on the floor by the door
- i'm done writing in my notebook, please close it up and return the pen to the pen holder
- please bring my green shopping bag from the coat rack to the table
- separate the toys and microfiber cloths into different baskets.

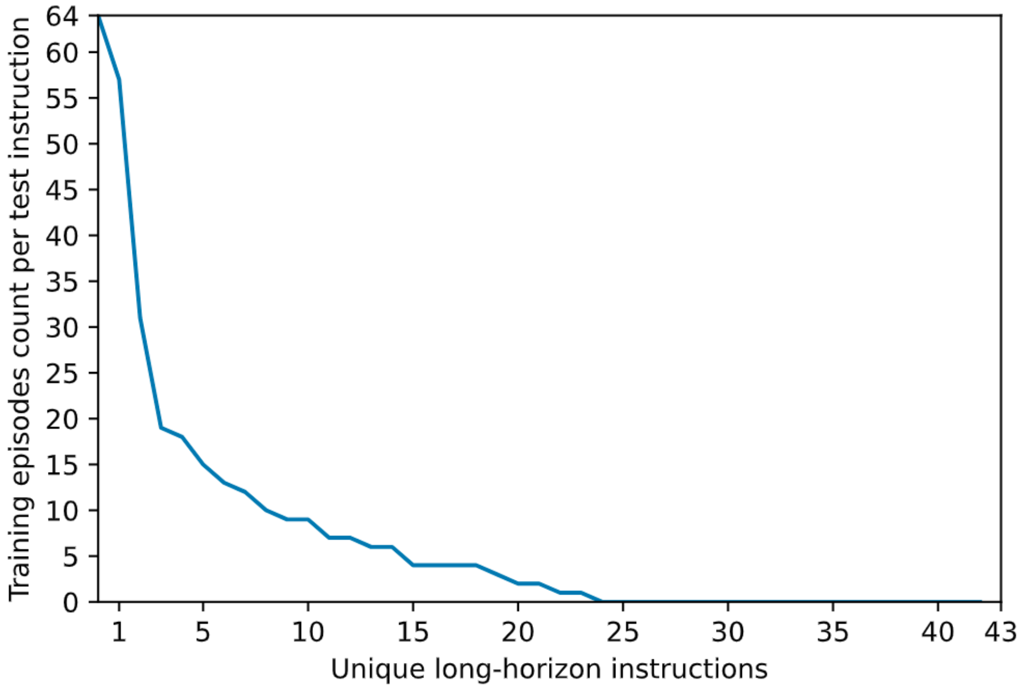

Figure 12: Number of training episodes that have the same long-horizon instruction as a test episode. Test episodes were sampled randomly and hence follow a similar distribution as observed in Fig. 11. Among the 43 episodes in the test set, we find that 23 of them have at least one episode with the same long-horizon instruction in the training set. For 20 of them (46% of the test set), the long-horizon instruction is not present in the training set.

- please remove the chips from the bowl and place them in the top draw.
- I am done drinking the coffee can you throw it in a trash can and get me some laffy taffy from MK kitchen to my desk.
- please put the sugar packets in the tray
- Can you refill my water cup and replace the cap and straw?
- Restock the Numi tea boxes into the correct places
- put the chips in the bin.
- put all the snacks in the tray.
- move the mouse box from the Whitney conference room to the dining booth
- Please place the cookie squares into the tray.
- please stock caddy for phone room
- pick the apple out of the jar and take it to phone room 2a3
- place only the green pears in the bowl
- Restock the ice packs and bandage rolls
- put all the screwdrivers in the cup
- please get the colored plastic cups from the top drawer and put them on the countertop
- empty bin onto the table
- open locker 17. then bring bag of chips from desk 2p2a to locker. close locker 17.
- throw away the cocunut water
- Put the red pens in the cup and bring them to a table in the mk, then bring the large postit notes to the table also

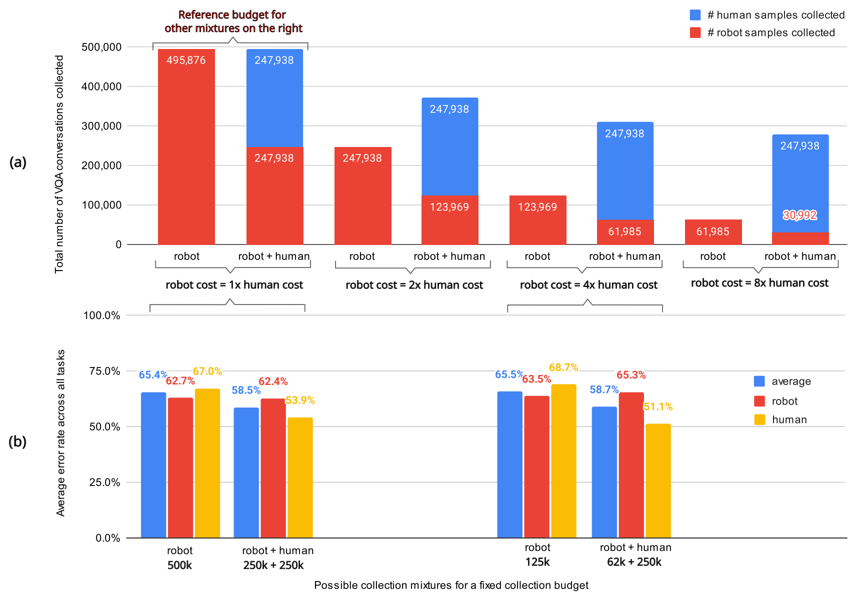

Figure 13: Possible embodiment mixtures for a fixed collection budget. This graph illustrates the possible trade-offs in total amounts of VQA samples collected for a fixed collecting budget and depending on the collection cost ratios between robot and human embodiments. In (a) we simulate different cost ratios by reducing the dataset size of the robot-embodiment dataset while keeping an equal budget for each embodiment. We calibrate this graph with a reference fixed budget that can produce approximately 500,000 VQA conversations at human collection cost. In (b) we report the error rates of each mixture (average error rate over all tasks). We find that mixing embodiments is overall beneficial even when the collection costs are the same and even when evaluating on the robot embodiment data only.

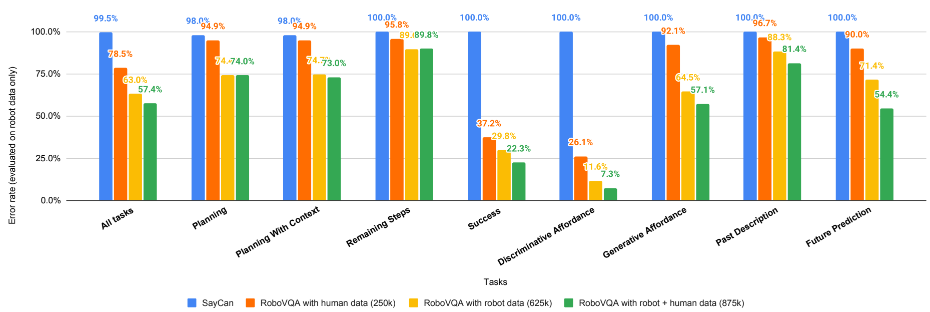

Figure 14: Error rates on robot-only test set, comparing models trained on robot only, human only or both embodiments. We observed that while it is not trained on robot data, the model trained on human data still performs with less than 100% error. We also find that the cross-embodiment training is beneficial even when evaluated on robot data only.

- make a virtal line of the plants and sort them by hight
- please pick up the trash on the table and throw it away into the compost
- bring a usb c charger from the bookshelf to the desk in the whitney room
- take out duck from plate on counter in a group
- put duck into the basket
- i'm finished with this hint water, please go recycle it in the micro kitchen for me and then bring me back a bag of lesser evil popcorn, cheese flavor
- Please flips the bowls then seperate the green, yellow and pink candy. Then remove the tongs and the forks from bins and place them on table.
- put the fruits in the basket

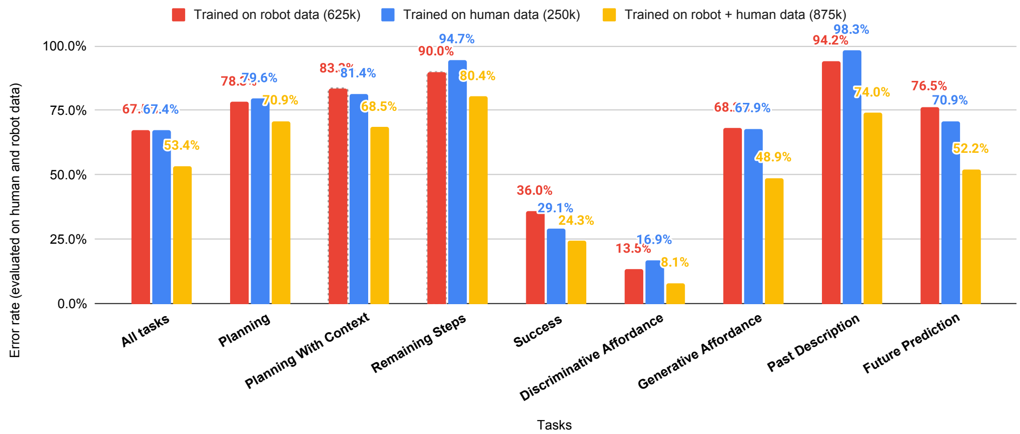

Figure 15: Error rates on robot+human test set. While it is expected that the model trained on both embodiments performs best, this graph illustrates that this model has the most breadth in capabilities and embodiments.

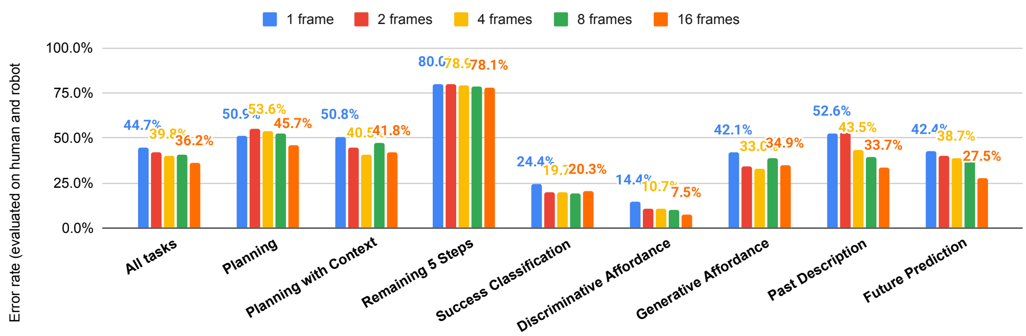

Figure 16: Error rates for video model trained with different number of frames. The model is trained on 875k samples (robot + human) and evaluated on the (robot + human) test set. We find that 16 frames yields the best results.

50 medium-horizon instructions:

- Touch the green bag
- go away from the table
- Grab the tissue
- place the banana into the small bowl
- drop the cups on the table
- place strawberry hint water bottle in the tray
- place green marker in the cup
- Drop the green candy packet in the container
- Place the black book on the table
- Pick the bag on the table
- Arrange the white packet in tray
- open the cap of jar
- place the yellow packet in glass
- Put the tilted cup up right on the table
- Release the orange marker into the left holder

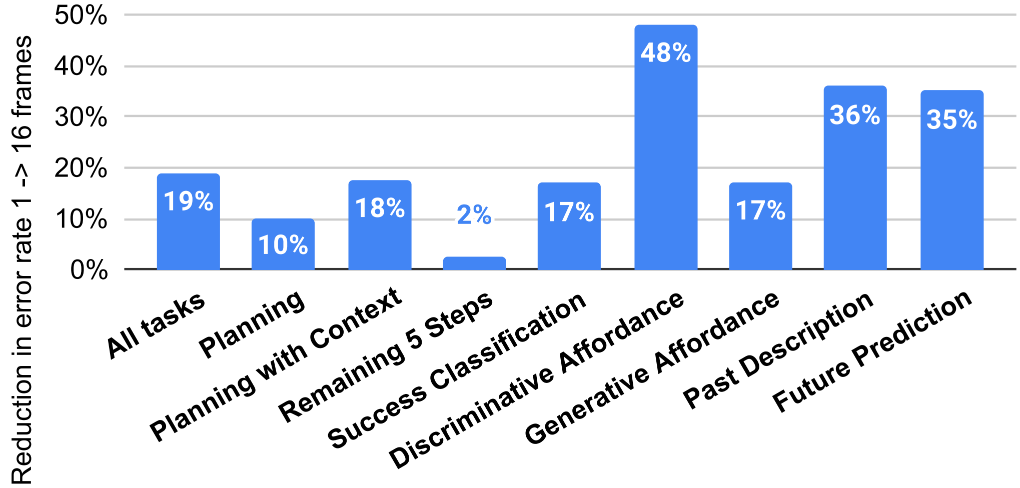

Figure 17: Rate of error reductions when training a model with 16 frames as input versus 1

- Turn to the right
- drop yellow candy into the left bowl
- place the cup backward
- drop the blue pen on a table
- open the white box
- Put orange bowl in the box
- place tissue in the tray
- Put the banana on the white table
- move away from the rack
- place 2 pistachio in the vessel
- move away from the hanger
- Place the square symbol in the baby pink box
- Move your arm towards the right chair
- place the lead on the glass
- Put the paper bag in the black container
- put paper clip in the rectangular stand
- move to the orange packet
- throw the tissue paper in dustbin
- Place the red pen on the table
- move towards the apple
- Move away from the hint bottle
- Go to the right side chair
- Place the left indoor plant on the table
- draw R on board
- put sugar packets in the container
- Place the 2 red packets on the table
- move to the orange cable on the table
- Drop the white pebble in the transparent glass

- drop the black container in the box
- Draw a diagonal line from left
- place the black cart to the corner
- Put blue cup on the table
- drop the apple on the floor
- Place the red can in fridge
- pick the sanitizer

9.5 Dataset Language Statistics Analysis by LLM

We use an LLM to extract different attributes from each short-horizon instruction from the training set and find:

- 1795 objects, e.g. ”tissue box”, ”purple color plate”.
- 494 actions, e.g. ”add something into something”, ”go out of a room”.
- 2064 locations, e.g. ”in the green plate”, ”left trash can”.
- 462 attributes, e.g. shapes, color.

Note that no clustering is performed and these lists contain redundant descriptions for each categories, the counts above are not meant to represent unique instances. In subsequent sections we display the full lists for each category above along with their parent categories inferred by the LLM.