2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

RoboVQA: Multimodal Long-Horizon Reasoning for Robotics

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, Pete Florence, Wei Han, Robert Baruch, Yao Lu, Suvir Mirchandani, Peng Xu, Pannag Sanketi, Karol Hausman, Izhak Shafran, Brian Ichter, Yuan Cao
Google DeepMind

***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 robovqa.github.io**

I. Introduction

The field of textual high-level reasoning has seen major breakthroughs recently with large language models (LLMs) [1, 2], while progress has also been made in visual language models (VLMs) [3], 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 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 [4], 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. IX-C of [5] 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,

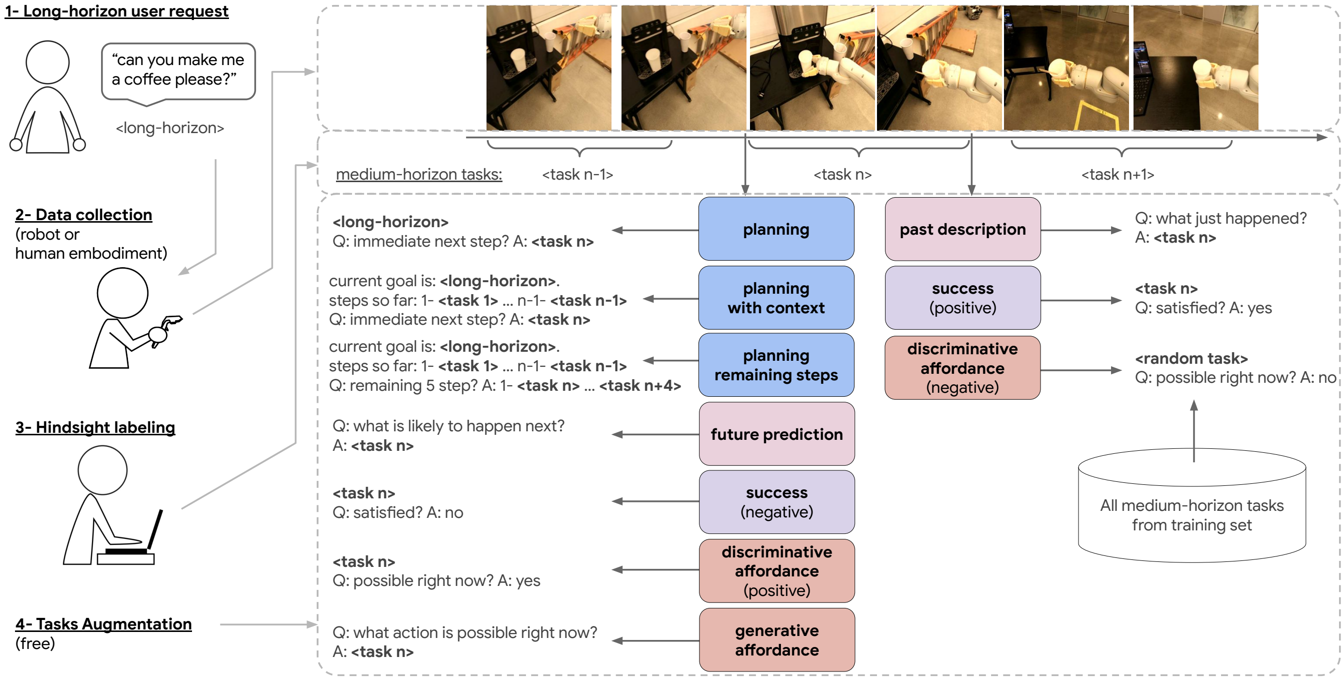

Fig. 1: Data collection procedure: Given long-horizon user requests, a human operator teleoperates a robot to fulfill the task. Medium-horizon tasks are then labeled in hindsight via crowd-sourcing, with temporal segmentation and task instruction for each segment. Finally, from a sequence of labeled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

making it deployable with human oversight even when imperfect.

II. 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. IX-B of [5].

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 [6] 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).

Throughput and costs: Much of the throughput gains reported in Fig. 2 come from collecting medium-horizon

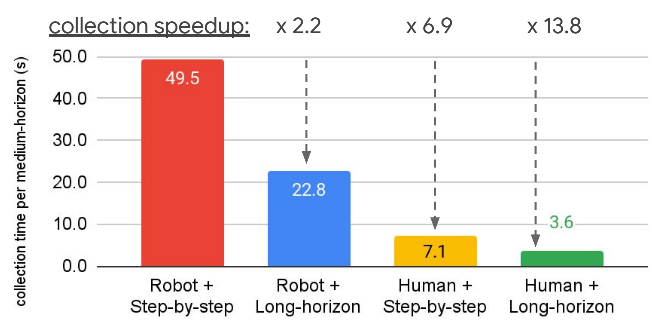

Fig. 2: Throughput gains compared to the traditional top-down step-by-step collection approach. The throughput of our long-horizon collection is 2.2x higher for robot collection and 13.8x higher with human bodies (compared to the robot used in our experiments).

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).

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 natural language questions and answers, and can be viewed as a series of Visual Question Answering (VQA) steps. This formulation is similar to chain-

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

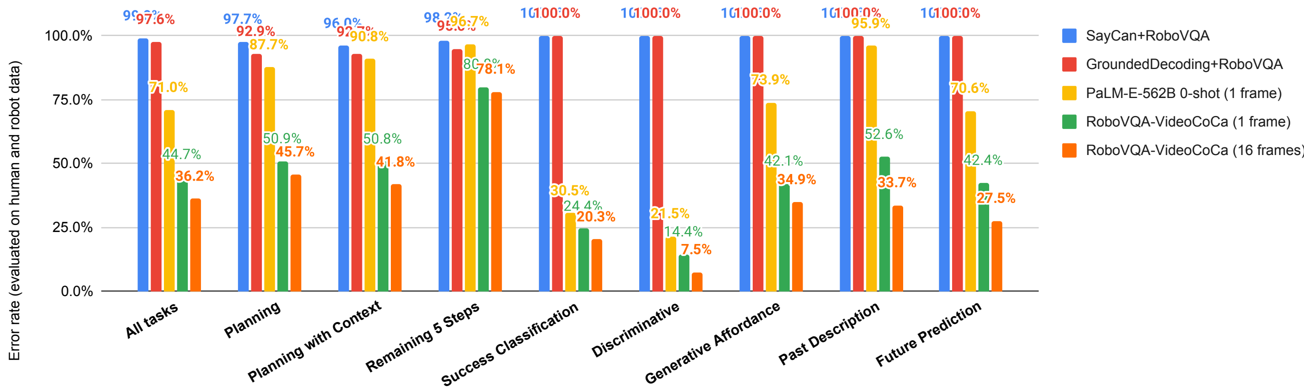

Fig. 4: VQA Error rates: we evaluate all models on the test set using human raters. We observe that state-of-the-art methods do not perform well in realistic settings in zero-shot, thus motivating the need for further scalable data collections. We also observe substantial gains when using video (16 frames)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

of-thought for language model prompting [7]. We also note concurrent work [8] which demonstrates that mimicking step-by-step human thought improves planning accuracy.

III. Models

A. RoboVQA-VideoCoCa

We train a new model called RoboVQA-VideoCoCa derived from the **VideoCoCa** model [9], which is a video language model extending CoCa [10]. It uses an encoder-decoder architecture combining contrastive pretraining (like CLIP [11]) as well as generative pretraining (like SimVLM [12]) 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. 16 and 17 in [5]).

B. 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 [3] is a visual language model built from pretrained ViT [13] and PaLM [2] 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** [6] and **Grounded Decoding** [14], which use a text-only LLM (PaLM [2]) in either phrase-level or token-level decoding guided by a visual affordance function (using RoboVQA-VideoCoCa as a video value function for affordance).

IV. Benchmarks

A. 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 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.

	Cognitive Model				Physical Model	Multi-turn Long-Horizon Planning				Intervention Rate		
	Training		Inference		(policy)	Total				(per episode average)		
Model	procedure	Size	time	# frames		# tasks	# steps	domain	bodies	cognitive	physical	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%

Fig. 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.

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.

B. 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 ($\sim$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 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.

V. Analysis

A. 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

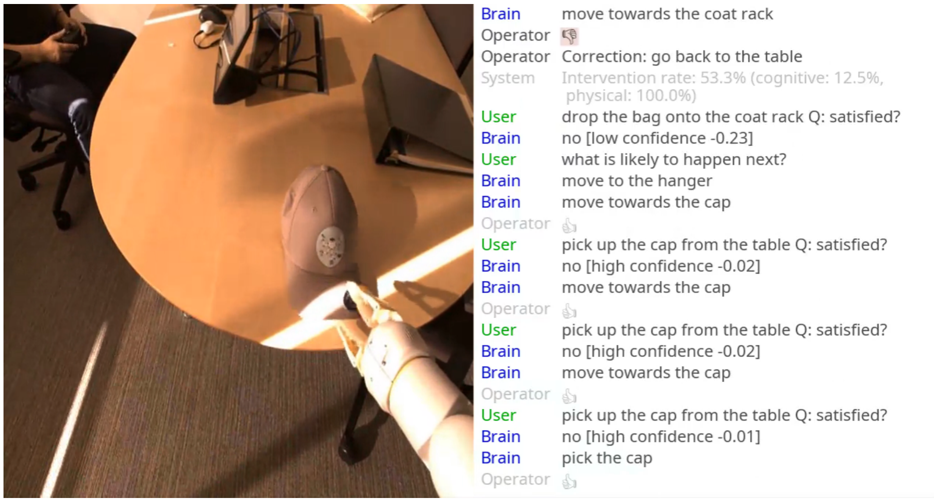

Fig. 6: Example of grounded chat with cognitive intervention. Our model ”Brain” is tasked with the following task at the beginning of the chat: ”take the bag and cap on the desk and hang them on the coat rack” in this case. The bottom of the chat shows the most recent messages. The model is ran on an 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.

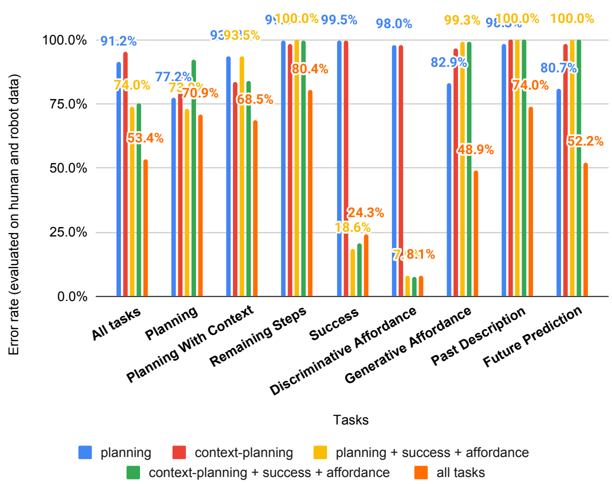

Fig. 7: Error rates for models trained with different sets of tasks. Each model is trained and evaluated on the (robot + human) dataset, but using different subsets of tasks. We find that training on all tasks leads to better planning (70.9% error) compared to training on planning only (77.2% error).

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.

B. Tasks Transfer via Cross-Embodiment Data

In Fig. 14 in [5], 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 [6] 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 in [5]. 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. IX-C of [5].

C. Importance of Video modeling

We investigate performance gains from video by training our model with (1, 2, 4, 8, 16) frames in Fig. 16 in [5] and find substantial error reductions in Fig. 17 in [5] 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.

D. 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

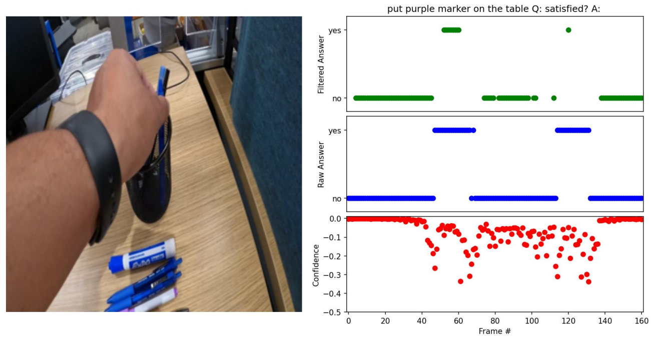

Fig. 8: RoboVQA-VideoCoCa used for video success detection. In blue are the raw answers to the question "put purple marker on the table Q: satisfied? A:", the confidence is shown in red and the answer filted by confidence is shown in green.

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.

VI. Related Work

Vision-Language Models. Recently many methods [11, 15, 16, 10, 12, 17, 13] 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 [9] 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.

Video Captioning. Our task is closely related to the task of video captioning [18, 19, 20, 21, 22] 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 [23, 24, 25, 26, 27, 28, 29, 30] 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. [31] also explores scalable ways to collect language data with unstructured play [32], 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[33] 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. [34] 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 [6, 3, 35, 36, 37]. One advantage of using LLMs to plan is that the output of these models can be used as input to language-conditioned policies [38, 4, 39] that may have been trained independently.

Intervention Rate. Intervention Rate is a commonly used evaluation metric [40, 41, 42] 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. [43, 44, 7] 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.

VII. 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 [30], EpicKitchens [45] etc., which merit careful study in our future work.

VIII. 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.

Acknowledgment

We thank Tran Pham, Elio Prado, Gavin Gonzalez, Jodilyn Peralta, Joseph Dabis, Alex Luong, Jodexty Therlonge, Rochelle Dela Cruz, Brianna Zitkovich, Emily Perez, Eric Tran, Huong T Tran, Dee M, Utsav Malla, Samuel Wan, Justice Carbajal and Scott Lehrer, Joel Magpantay, Sarah Nguyen, Steve Vega and Jonathan Vela for their contributions to data collection.

REFERENCES

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

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[3] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, "Palm-e: An embodied multimodal language model," in *arXiv preprint arXiv:2303.03378*, 2023.

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

[5] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi, P. Florence, W. Han, R. Baruch, Y. Lu, S. Mirchandani, P. Xu, P. Sanketi, K. Hausman, I. Shafran, B. Ichter, and Y. Cao, "Robovqa: Multimodal long-horizon reasoning for robotics," in *arXiv preprint arXiv:2311.00899*, 2023.

[6] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[8] S. Hu and J. Clune, "Thought cloning: Learning to think while acting by imitating human thinking," 2023.

[9] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Videococa: Video-text modeling with zero-shot transfer from contrastive captioners," 2023.

[10] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022.

[11] 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.

[12] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," 2022.

[13] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, "Pali: A jointly-scaled multilingual language-image model," 2023.

[14] W. Huang, F. Xia, D. Shah, D. Driess, A. Zeng, Y. Lu, P. Florence, I. Mordatch, S. Levine, K. Hausman *et al.*, "Grounded decoding: Guiding text generation with grounded models for robot control," *arXiv preprint arXiv:2303.00855*, 2023.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[17] T. Gupta, A. Kamath, A. Kembhavi, and D. 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*, 2022, pp. 16 399–16 409.

[18] X. Wang, W. Chen, J. Wu, Y.-F. Wang, and W. Y. Wang, "Video captioning via hierarchical reinforcement learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4213–4222.

[19] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[20] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6504–6512.

[21] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[22] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. 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*, 2022, pp. 17 949–17 958.

[23] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[24] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[25] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," in *EMNLP*, 2018.

[26] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019, pp. 9127–9134.

[27] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *ICCV*, 2019.

[28] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[29] J. Xiao, X. Shang, A. Yao, and T.-S. 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)*, June 2021, pp. 9777–

9786.
[30] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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*, 2022, pp. 18 995–19 012.
[31] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.
[32] C. Lynch and P. Sermanet, "Grounding language in play," *arXiv preprint arXiv:2005.07648*, 2020. [Online]. Available: https://arxiv.org/abs/2005.07648
[33] A. Padmakumar, J. Thomason, A. Shrivastava, P. Lange, A. Narayan-Chen, S. Gella, R. Piramuthu, G. Tur, and D. Hakkani-Tur, "TEACh: Task-driven Embodied Agents that Chat," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2017–2025.
[34] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," *CoRR*, vol. abs/2201.07207, 2022. [Online]. Available: https://arxiv.org/abs/2201.07207
[35] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," *arXiv preprint arXiv:2212.04088*, 2022.
[36] T. Silver, V. Hariprasad, R. S. Shuttleworth, N. Kumar, T. Lozano-Pérez, and L. P. Kaelbling, "PDDL planning with pretrained large language models," in *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. [Online]. Available: https://openreview.net/forum?id=1QMMUB4zfl
[37] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, "Llm+ p: Empowering large language models with optimal planning proficiency," *arXiv preprint arXiv:2304.11477*, 2023.
[38] 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 *5th Annual Conference on Robot Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=8kbp23tSGYv
[39] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *arXiv preprint arXiv:2210.06407*, 2022.
[40] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common metrics for human-robot interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 33–40.
[41] R. R. Murphy and D. Schreckenghost, "Survey of metrics for human-robot interaction," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 197–198.
[42] D. Riedelbauch, N. Höllerich, and D. Henrich, "Benchmarking teamwork of humans and cobots–an overview of metrics, strategies, and tasks," *IEEE Access*, 2023.
[43] W. Ling, D. Yogatama, C. Dyer, and P. 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)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 158–167. [Online]. Available: https://aclanthology.org/P17-1015
[44] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.
[45] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.