RoboPlan: A Multimodal Long-Horizon Planning Robotics Benchmark

Anonymous Author(s)
Affiliation
Address
email

Abstract: We present RoboPlan, a large-scale multimodal and cross-embodiment dataset and benchmark for long-horizon planning. We use scalable ways for acquiring real-world data and language labels with high throughput and high diversity: crowd-sourcing tasks from users and operators, collecting long continuous episodes, using crowd-sourced labeling and automatically generating tasks from it, using human embodiment data along with robotic embodiment data. We analyze the effects of mixing cross-embodiment data as well as multi-task data and find that it generally increases performance, broadens capabilities and increases collection throughput. Examples in RoboPlan are formatted as video-text pairs, where videos of robot or human actions are annotated with texts of multi-turn visual question answering (VQA), detailing the intermediate thoughts and actions required to achieve the session goals. This dataset covers a total of 300 hours of videos, over 1 million conversations and 100k instructions. We then define a novel benchmark based on this dataset, using an intervention rate metric to assess a robot's level of autonomy in accomplishing predefined goals without human interference. We validate our approach by running several state-of-the-art visual language models (VLM) over the dataset, and demonstrate the feasibility of long-horizon planning in the real-world with low intervention rate. Moreover, we find that video language models in general work better than image language models, indicating the necessity of modeling visual temporal dynamics in long-horizon planning tasks. Finally, our results corroborate our assumption that models trained on human data can effectively transfer to robot setup, revealing the potential of our approach in facilitating large-scale, cost-effective training for robot planning problems. We commit to releasing the data at a later date.

Keywords: Long-horizon planning, Benchmark, Multimodality

1 Introduction

Long-horizon planning is a challenging task for robotics, as it requires agents to reason on how to accomplish high level objectives, and predict and plan accordingly. However, there are few videotext datasets or benchmarks designed specifically for robotic long-horizon planning. To address this problem, we introduce RoboPlan, a large-scale multimodal dataset and benchmark for long-horizon planning on robots. RoboPlan has several important features which we believe can make it a valuable contribution to the research community.

First, the RoboPlan dataset is formulated as aligned video-language pairs. Given a full episode of a robot completing a high-level goal, the episode is decomposed into a multi-turn conversation, with each turn consisting of a human asking the robot a question about its planning, affordances, and actions on a smaller task. Each question is paired with a first-person video view of the environment, and the changes that the robot is making within it. This framework enforces the robot to make its decisions grounded on its visual input, as well as introduce a structured reasoning process that allows for transparency and interpretability in its trajectory. To the best of our knowledge, this is a novel format for a robotics dataset.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

Second, RoboPlan is cross-embodiment, and is collected as a mixture of robot and human sessions. Given the same high-level objective, RoboPlan includes data of humans completing the same objective. This was motivated by the cost of robot data collection; human data is much cheaper and faster to collect by comparison, making it much more feasible for large-scale data preparation. Moreover, we hypothesize that given the same data collection budget, models can obtain similar or better planning abilities when trained on a combination of robot and human data, as demonstrated in Section 3.3.

Finally, to ensure that we faithfully measure the generalization ability of planning models, the RoboPlan dataset contains a diverse variety of tasks. These tasks are collected with a bottom-up strategy in which the tasks are crowd-sourced by users and tele-operators. Data collection occurred across multiple buildings and environments, totaling 1.9k unique long-horizon tasks and 29k medium-horizon tasks. The richness of the tasks makes RoboPlan a diverse testbed for improving and evaluating model generalization ability.

The associated RoboPlan benchmark provides the ability to evaluate the level of autonomy of a robot in accomplishing tasks without human interference. We validate our dataset and benchmark by testing multiple state-of-the-art visual language models (VLM) on our benchmark. Due to the importance of multimodal understanding for successful planning and task completion for robots, we experiment with both image-language and video-language understanding models. Our results show that in the offline evaluation setting, VLMs are able to achieve a relatively low (i.e. good) intervention rate, but the rate increases significantly in the online setting. Moreover, we observe better quality yielded by video-language models than image-language models, indicating that modeling long-term temporal dynamics of visual inputs is essential to the improvement of planning accuracy. Our results also support our hypothesis that models trained on the hybrid of robot and human data transfers effectively to increasing generalization performance for the robot tasks, justifying our cross-embodiment approach in data collection.

Our contributions can be summarized as follows:

1. We propose RoboPlan, a large-scale multimodal cross-embodied benchmark for training and evaluating robotic long-horizon planning.
2. We demonstrate the feasibility of achieving low intervention rate using state-of-the-art VLMs, and argue for the importance of video sequence modeling and multi-task training for the improvement of planning accuracy.
3. We show that the cross-embodiment data collection procedure can facilitate large-scale robotics model training, due to the effective transfer learning between human and robot operation data.

2 The RoboPlan Benchmark

Fig. 1 shows an episode from the RoboPlan benchmark. Each episode is decomposed into a sequence of tasks, each consisting of a text question and a video segment. The following eight tasks are defined:

1. **Planning** Given the high-level goal, determine the immediate next step required to accomplish it, or all steps required to accomplish to goal. For example, `current goal is: Please get a water bottle and put it on Tomas's desk. Q: immediate next step? A: Open the fridge`
2. **Planning with Context** An extension of **Planning** that includes contextual information of the steps that have already occurred. Example: `current goal is: Please get a water bottle and put it on Tomas's desk. steps so far: 1- Open the fridge 2- Put water bottle on the table Q: immediate next step? A: Close the fridge`
3. **Planning Remaining Steps** A further extension of **Planning with Context** that requires the robot to answer with all steps remaining until the goal is completed. `current goal is: Please get a water bottle and put it on Tomas's desk. steps so far: 1- Open the fridge 2- Put water bottle on the table Q: immediate next step? A: 3- Close the fridge 4- Bring water bottle to Tomas's desk ...`
4. **Discriminative Affordance** Given a step, ask whether this is possible with yes or no. Example: `put the apple on the counter Q: possible right now? A: yes`
5. **Generative Affordance** Ask for a step that can be taken at the current time. Example: `Q: what action is possible right now? A: stack the glasses`

6. **Success** Given a step, ask if it has been executed successfully. Example: `pick up the pen Q: satisfied? A: no`
7. **Future Prediction** Ask for a likely future step. Example: `Q: what is likely to happen next? A: put the orange in the bowl`
8. **Past Prediction** Ask for the step that has just occurred. `Q: what just happened? A: Put the memory card packet on the stack of memory card packet`

The first task of an episode is **Planning Remaining Steps**, asking the robot to formulate a long-horizon plan to execute to accomplish the goal. As the robot executes each step of this plan, it is prompted with multiple tasks to determine **Affordance** (what actions are possible), **Success** (which actions have succeeded), and **Prediction** (which actions need to be done afterwards).

Intervention Since the high-level goal is decomposed into a sequence of tasks, there is a consistent and regularized framework for determining whether or not the robot is proceeding towards the goal correctly. These questions can be viewed as an interactive conversation between a questioning human and an answering robot. In particular, the RoboPlan benchmark allows for human *intervention* within the robot's trajectory — for example, if the robot responds incorrectly to a particular task, a human can intervene to overwrite its prediction with the correct answer, allowing it to continue the execution of subsequent tasks.

Chain-of-Thought in Natural Language 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-of-thought for language model prompting [1]. We also note concurrent work [2] which demonstrates that mimicking step-by-step human thought improves planning accuracy.

2.1 Dataset

As part of the benchmark, we collect and publish the RoboPlan dataset with both training and evaluation splits. As shown in Fig. 1, we first asked human users to provide a list of common tasks that they would like to see a robot butler perform for them in office or kitchen environments. We then record first-person videos executing these tasks in two different embodiments: (1) using a tele-operated robot with a single arm, and (2) with a human using a single arm, and holding a camera in their other hand. After videos were collected, we crowdsourced hindsight relabels for video segments, in which workers answered several questions on planning, success, future prediction, etc. From this data, the tasks described in Section 2 are generated automatically with heuristics, for example the future prediction task can be constructed by first extracting the video before a segment, then combining the question "what is likely to happen next?" with the instruction found in the segment. All tasks are constructed using different videos before, during or after a segment.

Task length We focus on tasks that require long-horizon planning. Therefore the collected long-horizon episodes last on average 1 minute and 42 seconds. The medium-horizon tasks segments labeled in hindsight last on average 13 seconds.

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 from a robotics perspective. Unlike existing robotics works [3] 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. The sessions were across 3 office buildings, covering 1,939 unique long-horizon tasks and 29,367 unique medium-horizon tasks.

Dataset Statistics The dataset contains 312.6 hours of videos or 13 days, collected across 3 office buildings. These videos correspond to 2859 robot embodiment episodes and 2672 human embodiment episodes. There is a total of 5531 long horizon instructions with an average execution length of 102 seconds with 1939 unique values among them. The dataset also has 111,046 medium horizon instructions, with an average execution time of 13 seconds with 29,367 unique samples among them. From these instructions, we construct a visual question-answering (VQA) dataset of 1+ million (video, VQA conversation) pairs. Because evaluation of freeform text answers are performed by humans, 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.

3 Experiments

3.1 Visual Language Models and Planning Method Baselines

To accomplish the long-horizon planning tasks, robots need to be controlled by strong VLMs for visual understanding and linguistic reasoning. We therefore consider the following state-of-the-art VLMs as our baseline models for experiment.

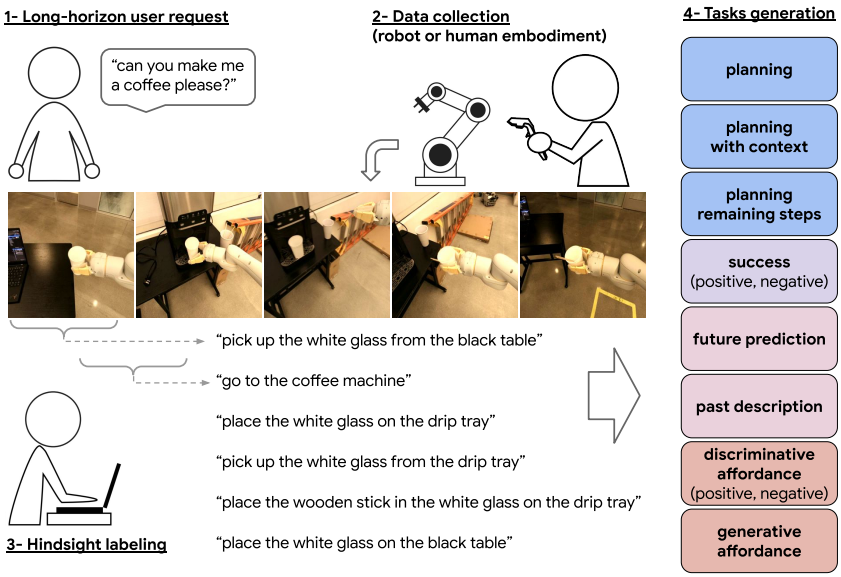

Figure 1: Data collection, instructions labeling and tasks generation.

VideoCoCa [4] is a video language model extending CoCa [5]. It uses an encoder-decoder architecture combining contrastive pretraining (like CLIP [6]) as well as generative pretraining (like SimVLM [7]) 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 RoboPlan video-text datasets.

PaLM-E [8] is a visual language model built from pretrained ViT [9] and PaLM [10] LLM models, which projects images into the token embedding space of the pretrained LLM. We use the PaLM-E-562B model, trained only on single-image examples, despite testing it in 2-image tasks. In our experiments we test PaLM-E-562B *zero-shot*, without training on RoboPlan dataset. This establishes baseline performance for a SoTA visual language model tested on the dataset without training.

Planning Methods. We experiment with four baseline planning methods: two of which use VideoCoCa and PaLM-E (zero-shot), as end-to-end planning models. As two other baselines, we adapt the methods of **SayCan** [3] and **Grounded Decoding** [11], which use a text-only LLM (for which we use PaLM [10]) in either phrase-level or token-level decoding guided by a visual affordance function. In these experiments for SayCan and Grounded Decoding, we provide a VideoCoCa model trained to perform affordances as a strong visual affordance function.

Baseline Results for the methods from Sec. 3.1 are shown in Tab. 1. The VideoCoca model trained on the RoboPlan dataset demonstrates an intervention rate much lower than the other methods tested. In particular, the SayCan and Grounded Decoding methods, although they use VideoCoca for te affordance function, have particularly high rates of interventions, suggesting the advantage of performing end-to-end planning. The PaLM-E method is only tested in the zero-shot setting, not trained on RoboPlan, and so accordingly does not perform as well as the finetuned method, but establishes a strong zero-shot baseline for future work.

3.2 Evaluation Method

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.

3.3 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

Planning Model	Visual Affordance Function	# calls to VLM for 30k instr.	Inference Time*	Model Size	Intervention Rate
SayCan	VideoCoca *(F-t)*	30k	>20s per step	540B	95.8%
Grounded Decoding	VideoCoca *(F-t)*	10	†	540B	97.6%
PaLM-E *(Zero-Shot)*	–	1	∼30s	562B	57.9%
VideoCoca *(Fine-tuned)*	–	1	∼1s	383M	28.8%

Table 1: Comparison of baseline methods for the RoboPlan benchmark tasks. † Grounded decoding inference time depends on beam size. * Inference time depends on inference compute – we note the models as used.

Figure 2: Examples of 3 embodiments in the dataset: robot, human (single) arm, human using a grasping tool.

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 Figure 3 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 3-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.

3.4 Tasks Transfer via Cross-Embodiment Data

In Fig. 4, we compare error rates on the test split using VideoCoCa-RoboPlan 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 [3] 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 RoboPlan as affordance function to SayCan for planning.

Similarly, we evaluate on the joint human and robot test split in 7.1 Fig. 6. 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. We also explore in Fig. 7 the effect of training on multiple tasks versus training specialized models on reduced sets of tasks. 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.

3.5 End-to-end Long-horizon Inference Evaluation

In Fig. 5, we present the answers of our model trained on both robot and human embodiment data for a full episode when queried for the immediate next step given a long-horizon instruction. We use the temporal segments provided in hindsight by human annotators on an existing test episode. For each segment, we retrieve a short video right before the segment starts and ask the model what should be done next using the following

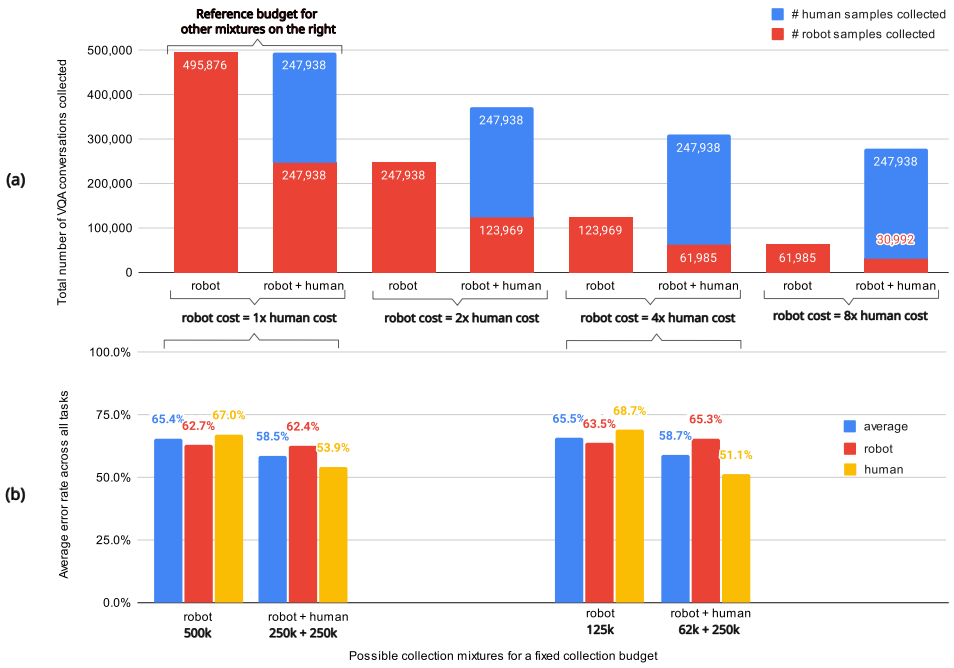

Figure 3: **Possible embodiment mixtures for a fixed collection budget.** This graph illustrates the possible trade-offs in total amounts of VQA samples collected for a fixed collecting budget and depending on the collection cost ratios between robot and human embodiments. In (a) we simulate different 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.

prompt: "current goal is: [long-horizon instruction] Q: immediate next step? A: ". We then submit each answer for human review and infer the intervention rate given how many steps had incorrect answers. We present more qualitative runs in Section 7.2.

3.6 Observations

From the experimental results above, we make the following observations:

Feasibility of low-intervention rate The intervention rate depends critically on the performance of VLMs. Using a strong VLM, it is possible for the robot to do long-horizon planning with relatively low intervention rate. This indicates that the contemporary multimodal models are strong enough to help robots understand visual scenes and reason over the action steps in controlled environments, and it is critical to build stronger VLMs to achieve higher levels of robotic autonomy.

Importance of multi-task training Multitask training has been demonstrated to be effective in facilitating transfer learning, improving models' generalization ability and versatility []. Similar observations hold in our experiments. We find in Fig. 7 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. However the performance difference is small, and a robotics setup benefits more largely from broad and general answering capabilities.

Importance of video modeling In order to perform tasks accurately, visual grounding over time horizon is important. We verify this assumption by comparing VideoCoCa trained with different number of frames (1, 2, 4, 8, 16). The results are presented in Table 13 in Appendix 7.5. As expected, modeling with more frames yields better results, as it captures longer temporal dynamics for more accurate visual grounding.

4 Related Work

Vision-Language Models. Recently many methods [6, 12, 13, 5, 7, 14, 9] 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 [4] on conversation data grounded in long-horizon videos. The advantage of this VLM is

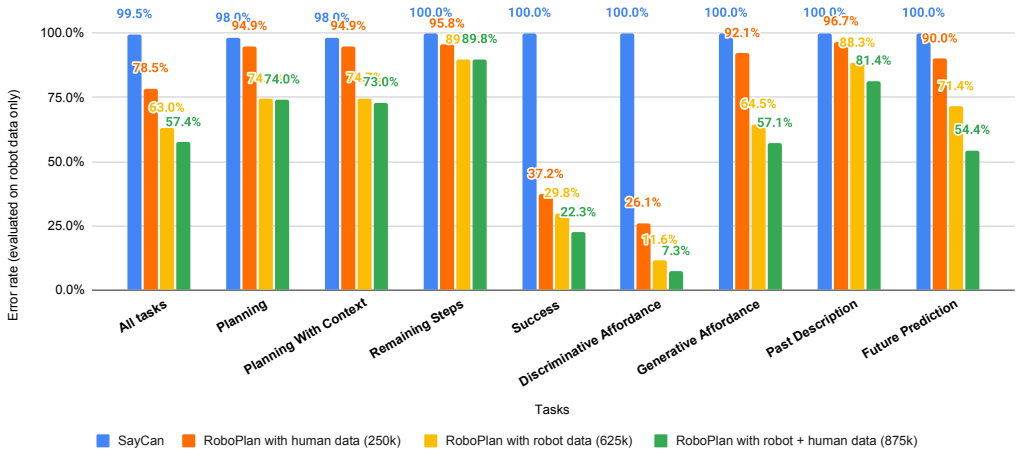

Figure 4: **Error rates on robot-only test set**, comparing models trained on robot only, human only or both embodiments. We observed that while it is not trained on robot data, the model trained on human data still performs with less than 100% error. We also find that the cross-embodiment training is beneficial even when evaluated on robot data only.

that it is the encoder can consume full videos which helps in fine-grained temporal reasoning required to solve the tasks introduced in the RoboPlan benchmark.

Video Captioning. Our task is closely related to the task of video captioning [15, 16, 17, 18, 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 [20, 21, 22, 23, 24, 25, 26, 27] that include videos of humans performing tasks with text narrations or question-answer annotations. Ego4D is the most similar dataset to the RoboPlan 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. Like RoboPlan, TEACh[28] 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. [29] 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 [3, 8, 30, 31, 32]. One advantage of using LLMs to plan is that the output of these models can be used as input to language-conditioned policies [33, 34, 35] that may have been trained independently.

Intervention Rate. Intervention Rate is a commonly used evaluation metric [36, 37, 38] in robotics and self-driving car literature for measuring the performance of policies. In this work, we use it as a metric for evaluating video question answering performance when a model is deployed in unseen test environments. Since, robots encounter new scenes and objects as they explore new scenes, we find intervention metric to be a better indicator of the capabilities of the model rather than a metric calculated on an offline dataset.

Chain of Thought Prompting. [39, 40, 1] 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.

5 Limitations

While we successfully collected a larger number of video-text examples over a wide range of tasks, our approach is limited in the following ways. First, all tasks were accomplished in unimanual manner. To further

Figure 5: **Full episodes runs with Intervention** given a long-horizon instruction and an existing video. We show the human labels on the left in blue, correct answer green and incorrect in red. For each picture (we show pictures here for simplicity but the model is fed the last few seconds before this picture as input), we give the long-horizon in a prompt and ask what the immediate next step should be. The human evaluator rates each answer and provides correction if needed. We report the rate of intervention at the end of the run.

expand the variety of tasks, we will consider introducing bimanual operations in future work. Secondly, the ways robots and humans perform tasks may differ, potentially impeding transfer learning between human and robot data. Thirdly, human intervention and oversight of robot task execution is time consuming, making this evaluation procedure difficult to be deployed at large scale. Lastly, we have not compared the effectiveness of the proposed human-and-robot dataset/benchmark with human-only dataset/benchmarks like Ego4D [27], EpicKitchens [41] etc., which merit careful study in our future work.

We also acknowledge that in this work, we did not conduct evaluations on a combined planning and mobile-manipulation setting. Rather, we opted to focus on high-level planning only. This is well motivated as decoupling planning and manipulation allows us to study the full breadth of possible tasks. We will explore combining high level with low-level policies in future works.

6 Conclusion

In conclusion, we hope that this dataset will serve as a benchmark for the robotics community working on solving grounded multimodal reasoning in complex real world settings. We also hope that cross embodiment transfer from human and robot data will usher in a greater possibility of accelerating robot learning by use of larger human task execution datasets. We show that video sequence models and multi task training improve planning performance over comparable methods.

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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*, pages 8748–8763. PMLR, 2021.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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*, pages 4904–4916. PMLR, 2021.

[13] 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*, pages 12888–12900. PMLR, 2022.

[14] 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*, pages 16399–16409, 2022.

[15] 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*, pages 4213–4222, 2018.

[16] 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*, 19(9):2045–2055, 2017.

[17] 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*, pages 6504–6512, 2017.

[18] 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.

[19] 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*, pages 17949–17958, 2022.

[20] 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)*, pages 720–736, 2018.

[21] 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.

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

[23] 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*, pages 9127–9134, 2019.

[24] 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.

[25] 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*, pages 1686–1697, 2021.

[26] 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)*, pages 9777–9786, June 2021.

[27] 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*, pages 18995–19012, 2022.

[28] 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*, volume 36, pages 2017–2025, 2022.

[29] W. Huang, P. Abbeel, D. Pathak, and I. 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.

[30] 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.

[31] 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. URL https://openreview.net/forum?id=1QMMUB4zfl.

[32] 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.

[33] 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. URL https://openreview.net/forum?id=8kbp23tSGYv.

[34] 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.

[35] 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.

[36] 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*, pages 33–40, 2006.

[37] 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)*, pages 197–198. IEEE, 2013.

[38] D. Riedelbauch, N. Höllerich, and D. Henrich. Benchmarking teamwork of humans and cobots–an overview of metrics, strategies, and tasks. *IEEE Access*, 2023.

[39] 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)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi:10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

[40] 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.

[41] 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.