Don't Copy the Teacher: Data and Model Challenges in Embodied Dialogue

So Yeon Min[1] **Hao Zhu**[2] **Ruslan Salakhutdinov**[1] **Yonatan Bisk**[2]
Machine Learning[1] and Language Technologies[2] at Carnegie Mellon University
{soyeonm,hzhu2,rsalakhu,ybisk}@andrew.cmu.edu

Abstract

Embodied dialogue instruction following requires an agent to complete a complex sequence of tasks from a natural language exchange. The recent introduction of benchmarks (Padmakumar et al., 2022) raises the question of how best to train and evaluate models for this multi-turn, multi-agent, long-horizon task. This paper contributes to that conversation, by arguing that imitation learning (IL) and related low-level metrics are actually misleading and do not align with the goals of embodied dialogue research and may hinder progress.

We provide empirical comparisons of metrics, analysis of three models, and make suggestions for how the field might best progress. First, we observe that models trained with IL take spurious actions during evaluation. Second, we find that existing models fail to ground *query* utterances, which are essential for task completion. Third, we argue evaluation should focus on higher-level semantic goals. [1]

1 Introduction

Dialogue is key to how humans collaborate; through dialogue, we query information, confirm our understanding, or banter in a friendly manner. Since communication helps us work more efficiently and successfully, it is only natural to imbue for collaborative agents with this same ability. Most work has focused on grounded dialogues for embodied navigation (Thomason et al., 2020; Chi et al., 2019; Roman et al., 2020) or limited interaction (Suhr et al., 2019), which are narrower domains than the larger instruction following literature (Tellex et al., 2011, 2020; Shridhar et al., 2020; Blukis et al., 2018, 2021; Min et al., 2021).

The first step towards engaging in a dialogue, is being able to understand and learn from it. Picture a child watching their parents with the goal to learn by imitation. They witness instructions, clarifications, mistakes, and banter. Begging the question: *What should one learn from noisy natural dialogues?*

Unlike in alinguistic tasks where modeling humans has recently proved helpful for search strategies (Deitke et al., 2022), we focus on language based tasks that require learning lexical-visual-action correspondences. We discuss and compare three paradigms: Instruction Following (IF), actions from Entire Dialogue History (EDH) and Trajectory from Dialogue (TfD). The novel TEACh dataset (Padmakumar et al., 2021) proposes EDH as the primary metric and uses the Episodic Transformer (ET) (Pashevich et al., 2021) trained with behavior cloning as their baseline. We also include comparisons to the EDH competitive Symbiote[2] system and we adapt FILM (Min et al., 2021), a recent method for general IF, to dialog instruction following (DIF) on TEACH. FILM and Symbiote belong to a different family of models, focusing on abstract planning trained at a higher semantic level than behavior cloning. This approach appears crucial for generalization and TfD evaluations.

Most importantly, we analyze the human behaviors in TEACH and the corresponding effect on ET, Symbiote, and FILM, as representatives of existing model classes. From our findings, we suggest there are three major challenges the community must tackle to move forward in the nascent field of Dialogue based Instruction Following:

Recognizing mistakes Behavior cloning encourages replication of low-level errors, but not high-level intentions. Agents should learn to construe high-level intentions of demonstrations and to deviate from demonstration errors.

Grounding queries No approaches correctly ground *"queries"* requesting information.

[1] Code to be released at https://github.com/soyeonm/TEACh_FILM

[2] Model outputs provided by correspondence with the team.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9361 - 9368
December 7-11, 2022 ©2022 Association for Computational Linguistics

Evaluation Agent evaluation should focus on achieving goals rather than immitating procedures.

2 Related Work

Instruction Following A plethora of works have been introduced for instruction following without dialogue (Chen and Mooney, 2011; Matuszek et al., 2012); an agent is expected to perform a task given a language instruction at the beginning and visual inputs at every time step. Representative tasks are Visual Language Navigation (Anderson et al., 2018; Fried et al., 2018; Zhu et al., 2020) and instruction following (IF) (Shridhar et al., 2020; Singh et al., 2020), which demands both navigation and manipulation. Popular methods rely on imitation learning (Pashevich et al., 2021; Singh et al., 2020) and modularly trained components (Blukis et al., 2021; Min et al., 2021) (e.g. for mapping and depth).

Dialogue Instruction Following Instruction Following with Dialogue (She et al., 2014) has mostly addressed navigation. Thomason et al. (2020); Suhr et al. (2019) built navigation agents that ground human-human dialogues, while Chi et al. (2019); Nguyen and Daumé III (2019) showed that obtaining clarification via simulated interactions can improve navigation. Manipulation introduces grounding query utterances that involve more complex reasoning than in navigation-only scenarios (Tellex et al., 2013); for example, the agent may hear that the object of interest (e.g. “apple”) is inside “the third cabinet to the right of the fridge.”

Imitation Learning vs Higher semantics While behavior cloning (BC) is a popular method used to train IF agents, it assumes that expert demonstration is optimal (Zhang et al., 2021; Wu et al., 2019). TEACh demonstrations are more “ecologically valid" (de Vries et al., 2020) but correspondingly suboptimal, frequently containing mistakes and unnecessary actions. Popular methods that deal with suboptimal demonstrations involve annotated scoring labels or rankings for the quality of demonstrations (Wu et al., 2019; Brown et al., 2019). Such additional annotations are not available in existing IF and DIF benchmarks. In this work, we empirically demonstrate the effect of noisy demonstrations on an episodic trained with BC for DIF.

3 Tasks

TEACh focuses on two tasks: Entire Dialogue History and Trajectory from Dialogue. Despite what the name implies, EDH is an evaluation over partial dialogues (e.g. from state S_t begin execution to S_T). TfD starts an agent at S_0 and asks for a complete task completion provided the full dialogue.

In both settings, the agent (driver) completes household tasks conditioned on text, egocentric RGB observations, and the current view. An instance of a dialogue will take the form of a command: *Prepare coffee in a clean mug. Mugs are in the microwave.*, the agent response *How many do I need?*, and commander's answer: *One*, together with a sequence of RGB frames and actions that the agent performed during the dialogue. As in this example, the agent has to achieve multiple subtasks (e.g. find mug in the microwave, clean mug in the sink, turn on the coffee machine, etc) to succeed.

In TfD, the full dialogue history is given, and the agents succeeds if it completes the full task itself (e.g. make coffee). In EDH, the dialog history is partitioned into “sessions” (e.g. Fig. 1) with the corresponding action/vision/dialogue history until the first utterance of the commander (*Prepare* $\sim$ *microwave.*) being the first session and those after it being the second. In EDH evaluation, the agent takes one session as input and predicts actions until the next session. An agent succeeds if it realizes all state changes (e.g. *Mug: picked up*) that the human annotator performed. Succinctly, TfD measures the full dialogue while EDH evaluates subsequences.

4 Models

TEACh is an important new task for the community. We analyze the provided baseline (ET), retrofit the ALFRED FILM model, and requested outputs from the authors of Symbioteon the EDH leaderboard.

ET is a transformer for direct sequence imitation approach, that produces low-level actions conditioned on the accumulated visual and linguistic contexts. In contrast, FILM consists of four submodules - semantic mapping, language processing, semantic policy, and deterministic policy modules. For the adaptation, we refactored the original code of FILM to the TEACH API, retrained the learned components of the semantic mapping module for the change in height and camera horizon, and retrained/rewrote the language processing module to take a dialogue history as input. The language processing (LP) module of FILM maps an instruction to a *task type* and instruction-specific *arguments*. For TfD this maps a dialogue to a sequence of tasks, while for EDH only the subsequence is mapped to

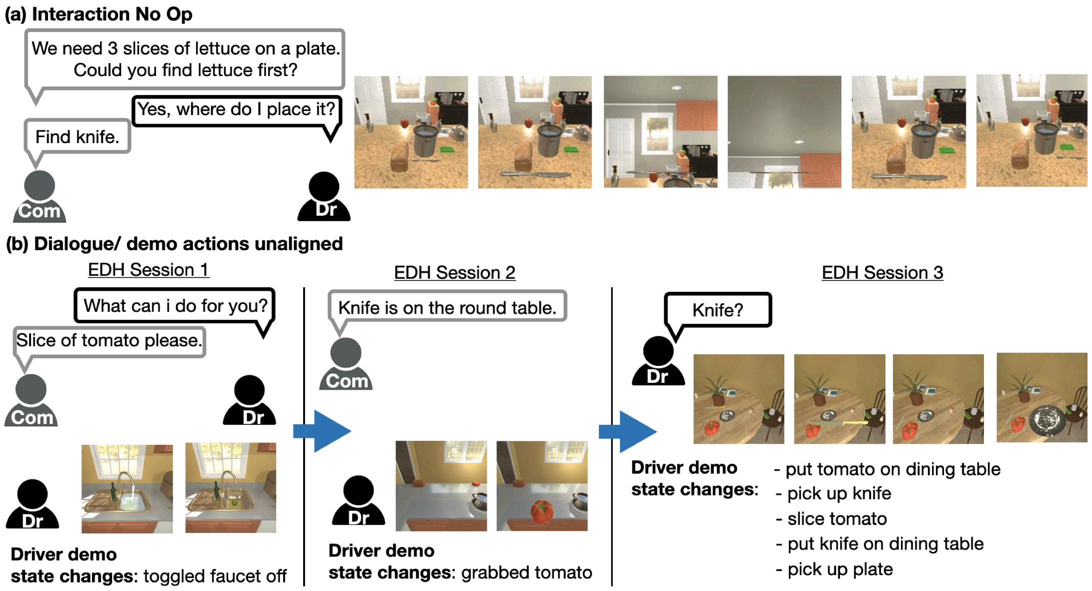
(a) Interaction No Op
We need 3 slices of lettuce on a plate. Could you find lettuce first?
Yes, where do I place it?
Find knife.
Com
Dr
(b) Dialogue/ demo actions unaligned
EDH Session 1
What can i do for you?
Slice of tomato please.
Com
Dr
Dr
Driver demo
state changes: toggled faucet off
EDH Session 2
Knife is on the round table.
Com
Dr
Driver demo
state changes: grabbed tomato
EDH Session 3
Knife?
Dr
Driver demo
state changes:
- put tomato on dining table
- pick up knife
- slice tomato
- put knife on dining table
- pick up plate

Figure 1: Examples of suboptimal demonstrations that can be harmful for training and evaluation. (a: **no-op**) The driver grabs a knife, looks up and down, and put its down, although nowhere in the dialogue indicates to do these actions, nor do they facilitate the high-level goal. (b: **unaligned** intent) In EDH sessions 1 and 2, the commander asks for an item (a slice of tomato) and provides the location of the knife, but the driver performs unaligned actions. In session 3, the driver suddenly asks "knife?", but performs a long sequence of implied but not stated actions.

an immediate action. Symbiote is a competitive modular method for EDH whose language understanding component is designed for dialogues (§A).

5 Challenges of Human Traces

First we present how TEACh and, by extension, future embodied dialogue settings present novel training and evaluation challenges as the data, by virtue of its authenticity, includes substantial noise in both the training and evaluation (despite filtering by the authors §C). See §B for how statistics were computed, for those not explained in this section.

5.1 Explanation of Metrics

Evaluation for both EDH and TFD is done by SR (success rate), GC (goal condition success rate), and their path-length-weighted versions. Success Rate (SR) is a binary indicator of whether all subtasks were completed. The definition of "subtasks" is different for EDH and TfD; for the former, they are all tasks required to realize state changes done by the **human demonstration** that are relevant to the ultimate task (e.g. The demo state changes in each session of Fig.1 (b)). Thus, the state changes brought by the human is considered ground truth in EDH evaluation; this brings multiple challenges further discussed in §5.2. On the other hand, for TfD, the subtasks are independent of what was done in the demo; for example, as long as an agent "slices the tomato" correctly for the task of Fig.1 (b), its SR will be 1 for this task. [3]

The goal-condition success (GC) of a model is the ratio of goal-conditions completed at the end of an episode. Both SR and GC can be weighted by (path length of the expert trajectory)/ (path length taken by the agent); these are called path length weighted SR (PLWSR) and path length weighted GC (PLWGC). Higher is better for all metrics.

5.2 Challenges in Evaluation

Irrelevant Actions Humans often explore the environment, or simply play around in the middle of a task. This means they may flip a switch completely unrelated to the goal. Table 1 are representative state changes that do *not* have direct correspondence with the dialogue, and the percentage of human demonstrations that contain these actions.

It is not always clear if this behavior is because of misunderstandings, boredom, or curiosity. For example, we can classify a large number of navigation and interaction "No Op"s, or action sequences that return to the original state (e.g. turning around

[3] While the github repository https://github.com/alexa/teach#downloading-the-dataset mentions that the EDH tasks were filtered so that "the state changes checked for to evaluate success are only those that contribute towards task success in the main task of the gameplay session the EDH instance is created from", we find that even after this filtering, there exist many EDH tasks with subotpimal demonstrations as in Fig.1.

Unnecessary State Changes	Val Seen	Val Unseen
Coffee Machine on/off	47.73	47.54
Picked up and not placed	25.49	23.41
Faucet on/off	12.68	10.59
Stove/ Microwave on/off	35.61	28.31
⇒ Total	38.98	35.60

Table 1: Representative state changes that do not have direct correspondence with the dialogue, and the percentage of human demonstrations that contain these actions. The action types listed here bring "state changes" that are counted during EDH evaluation. For example, an agent would "fail" an EDH task if the human annotator of the task left coffee machine off at the end, although the task (e.g. "Make coffee") or dialogue itself does not mention that it be left on.

in place). In principle, these might be information seeking, to build a better map of the environment, but in practice, many of the demonstrations do not seem to exhibit those properties, particularly in extreme cases like repeatedly picking up and putting down the same object. The percentages of prevalence of these unnecessary actions in both training and validation are shown in Table 2.

Suboptimal Actions	Train	V. Seen	V. Un
Navigation No Op			
Turn Left/ Right x 4	3.07	2.30	2.33
Forward + Backward	4.80	7.57	4.23
Pan Right + Pan Left	5.83	4.77	8.10
Turn Right + Turn Left	13.13	13.49	10.89
⇒ Total	22.41	23.03	21.36
Interaction No Op			
Toggle off + on same obj	1.39	1.81	1.58
Open + Close same obj	1.46	1.80	1.30
Place + Pick up same obj	25.06	27.80	28.34
⇒ Total	26.76	30.10	30.57
Interaction w. unrelated obj.	14.10	16.61	13.59
Demo unaligned w. dialog	25.25	22.49	23.40

Table 2: Representative unnecessary action types that do not have associations with the high level goal or the dialogue, and the percentage of demonstrations that contain these action types in train/ valid seen/ valid unseen splits.

The prevalence of these actions can be viewed as a positive for realism and even helpful if teaching how to search, but pose a challenge for evaluation.

Penalizing Agents for Accuracy Using a human's action trace as the ground truth, means agents are penalized for skipping erroneous actions. This leads to a misleading mismatch in performance between EDH and TfD. Additionally, EDH inflates model performance as it includes subsequences which are nearly deterministic (e.g. all but the last "placing" action). Table 3 contains EDH scores for our three comparison models and TfD for ET/FILM. As suggested by authors of related papers, we treat Unseen Success Rate as the most important metric (seen in **blue**).

	Valid Seen		Valid Unseen	
	GC	SR	GC	**SR**
Entire Dialogue History (EDH)				
E.T.	15.7 [4.1]	10.2 [1.7]	9.1[1.7]	7.8[0.9]
SYMBIOTE	25.9 [5.3]	**16.1 [2.6]**	17.2 [2.9]	10.1 [1.2]
FILM	**26.4 [5.6]**	14.3 [2.1]	**18.3 [2.7]**	**10.2 [1.0]**
Trajectory from Dialogue (TfD)				
E.T.	1.4 [4.8]	1.0 [0.2]	0.4 [0.6]	0.5 [0.1]
FILM	**5.8 [11.6]**	**5.5 [2.6]**	**6.1 [2.5]**	**2.9 [1.0]**

Table 3: EDH and TfD performances of E.T., Symbiote, and FILM. While the SR on TfD is very low for all models, E.T.'s performance on TfD drops significnatly due to replication of errors and lack of grounding of high-level semantics.

Note, that an ideal evaluation would capture both "actions in context" and "task success." In the following section breakdown the overall numbers presented here to understand if models more carefully.

5.3 Challenges in Training

Behavior Cloning with Suboptimal Demonstrations We find that ET trained with behavior cloning repeats the same mistakes in novel scenes that are frequent in demonstrations. We examine two kinds of mistakes in demonstrations - (1) No Op interactions, in which consecutive interactions produces futile state changes (e.g. Placing and immediately picking up the same object) and (2) Interactions with unrelated objects (e.g. picking up "saltshaker" while making coffee). In Table 4 we compare what percent of model predictions in seen and unseen scenes replicate the no-op behavior.

Suboptimal Actions	ET		Symbiote		FILM	
	S	U	S	U	S	U
No Op (same obj)						
Toggle off + on	0.0	0.1	0.2	0.1	0.0	0.0
Open + Close	2.5	1.5	0.0	0.2	0.0	0.0
Place + Pick up	45.1	47.1	4.9	2.5	0.0	0.0
⇒ Total	46.2	48.1	5.1	2.8	0.0	0.0
Unrelated obj.	24.0	20.3	27.5	30.6	15.9	12.10

Table 4: Percentage of tasks in which a model exhibited replication of No Op actions.

Method	SR	GC	SR w. Query	GC w. Query
Validation Seen				
ET	8.97	14.13	0.00	0.00
Symbiote	0.00	11.76	0.00	0.00
FILM	9.59	20.26	0.00	0.00
Validation Unseen				
ET	2.39	8.69	0.00	**0.49**
Symbiote	0.00	0.00	0.00	0.00
FILM	1.29	9.79	0.00	0.00

Table 5: We consider a task as involving "query utterances", if in its demonstration, a relevant object inside an originally closed receptacle was picked up. SR/GC measure the vanilla task success on tasks with "query utterances"; SR/GC w. Query measure if the success was achieved using information in the "query utterances."

While hard to quantify, we also note that the higher intention of seemingly unnecessary human demonstrations (e.g. to explore, to understand, etc) are not replicated by ET. This is backed by our observation that ET tends to be stuck in many (10 or more) repetitions of the same No Op/ unnecessary actions, until the end of the task or before resuming to perform other actions.

Note that even Symbiote is exhibiting some no-op behaviors, but as the model supervision/structure becomes more abstract (ET vs FILM) this disappears, leaving only object choice errors.

Grounding Queries Key to dialogue is language based information seeking. A target object may be located in a closed receptacle (cabinet, etc); in this case, the agent has to query the commander for its location, as a human would. We examine whether models ground *query utterances* into meaning and accurate actions, since this is one essential aspect of dialog grounding. While there are utterances with other essential intents, such as confirmation, we focus on query utterances since these are relatively easy to extract mechanically.

In Table 5 we consider a subset of tasks that involve "query utterances" that can be detected automatically. Specifically, we present the performance of models in terms of success rate and goal condition success on tasks that require opening a receptacle based on the answer to a question – and then measure if the models leverage the query. Not all query utterances will be of this type, but these tasks necessarily involve grounding query utterances for task success.

Queries are present in 23.05% and 25.31% of valid seen and unseen splits, respectively. This is a key challenge as it demonstrates a clear use case for dialogue and limitation of current models.

Given a statement like "the fork is *in the cabinet* left to the refrigerator", the evaluation mismatch occurs if an agent grabs a different fork on a table. This allows them to succeed, as measured by SR/GC, but not in SR/GC with Query. Notably, all models fail at query grounding, indicating they are simply ignoring the language instructions. This shows that enabling complex dialogue grounding is an important open problem for DIF. Especially, for the ultimate goal of two-agent task completion (TaTC), it is necessary that models can ground query and other essential utterances in a dialogue.

6 Conclusion and Next Steps

This paper is not an indictment of TEACh, nor an endorsement of a particular model, rather it seeks to lay out important questions and challenges that NLP will need to tackle as it moves into embodied dialogue. Unlike existing work in dialogue that looks to model human satisfaction (Ghandeharioun et al., 2019) or state-tracking, DIF has the advantage of explicit and verifiable semantic goals. We pose a challenge to the community: How can we build agents where success is not tied to specific actions yet language understanding and production are accurate and fluent? As a first step, we posit that imitation learning should be avoided.

7 Limitations

We focus on a new embodied benchmark – there is substantial work in dialogue (including goal directed) for non-embodied environments which we do not consider as aligned with the goals of embodied DIF, but may have important insights. Additionally, future work may overcome the issues raised and it is unclear how to transfer our findings back to the goal directed dialogue in the non-embodied setting. Additional insights may derive from research in social intelligence.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Valts Blukis, Dipendra Misra, Ross A Knepper, and Yoav Artzi. 2018. Mapping navigation instructions to continuous control actions with position-visitation prediction. In *Conference on Robot Learning*, pages 505–518. PMLR.

Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. 2021. A persistent spatial semantic representation for high-level natural language instruction execution. In *Proceedings of the Conference on Robot Learning (CoRL)*.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

David Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the National Conference on Artificial Intelligence*.

Ta-Chung Chi, Mihail Eric, Seokhwan Kim, Minmin Shen, and Dilek Hakkani-tur. 2019. Just ask:an interactive learning framework for vision and language navigation.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv:2007.14435*.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2022. Procthor: Large-scale embodied ai using procedural generation. *arXiv:2206.06994*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2012. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th International Symposium on Experimental Robotics (ISER)*.

So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. 2021. Film: Following instructions in language with modular methods. *arXiv preprint arXiv:2110.07342*.

Khanh Nguyen and Hal Daumé III. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 684–695, Hong Kong, China. Association for Computational Linguistics.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramithu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Conference on Artificial Intelligence (AAAI)*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. 2021. Episodic transformer for vision-and-language navigation. *arXiv preprint arXiv:2105.06453*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. RMM: A Recursive Mental Model for Dialog Navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Chai, and Ning Xi. 2014. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 89–97, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, and Jonghyun Choi. 2020. Moca: A modular object-centric approach for interactive instruction following. *arXiv preprint arXiv:2012.03208*.

Alane Suhr, Claudia Yan, Jacob Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. *arXiv preprint arXiv:1910.03655*.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):null.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence*.

Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. 2013. Toward information theoretic human-robot dialog. In *Robotics: Science and Systems*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. 2019. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, pages 6818–6827. PMLR.

Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. 2021. Confidence-aware imitation learning from demonstrations with varying optimality. *Advances in Neural Information Processing Systems*, 34.

Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022.

A More Discussion of Symbiote

Symbiote has a modular structure, which consists of language understanding, mapping, and low-level planning components. It is not trained with imitation learning of low-level demonstrations (e.g. move right, move left, etc.). Demonstrations are used only in the sense that they provide subgoals that suervise the training of the language understanding component.

More specifically, a pretrained T5 model (Raffel et al., 2020) fine-tuned with the ground truth subgoals (`edh_instance['future_subgoals']`), serves as the language understanding component. The model takes the driver and commander's dialogue and previous actions as input; it is trained to output a sequence of subgoals of the form "{action} {obj}", where {action} is either "navigate" or any of the primitive interactions commands "pickup", "cut", "toggle", etc, and {obj} is any of the object classes in ai2thor.

For the mapping component, a DETR detector (Carion et al., 2020) was finetuned on the train set scenes of TEACh and the depth prediction model from FILM was used off-the-shelf. Frontier based exploration is used for environment exploration. Similarly as in FILM, the agent navigates to object goals in the map using the fast marching method.

B How the Statistics of Section 5 were Obtained

We explain how the statistics that appear in each table of Section 4 were obtained. All analyses, except for TfD results in **Penalizing Agents for Accuracy**, were done on EDH tasks.

Irrelevant Actions The first table shows some representative unnecessary state changes that EDH tasks require for "task success' in evaluation. For example, in our common sense, it is not necessary that we leave the coffee machine on to successfully make coffee (indeed, it is better to turn it off after use). However, since EDH evaluation requires that the agent exactly follows state changes done in the demonstration, the agent will have to leave coffee machine turned on for a particular validation task, if this was done in its corresponding demonstration.

Each row shows unnecessary state changes that are exemplary and the average frequency of these noises across relevant tasks. More specifically,

- **Coffee Machine on/ off**: *'Coffee'* tasks
- **Picked up and not placed**: all tasks
- **Faucet on/ off**: all tasks that may involve using the faucet (*'Coffee', 'Clean All X', 'Boil X', 'Water Plant', 'Sandwich', 'Breakfast', 'Plate Of Toast', 'Salad'*)
- **Stove/ Microwave on/off**: all tasks that may involve using a heating appliance (*'Boil X', 'N Cooked Slices Of X In Y'*)

"Total" accounts for the percentage of EDH tasks that fall into any of the above criteria. Please refer to (Padmakumar et al., 2021) for the possible types (e.g. *'Coffee'*) of tasks.

While the first table shows statistics of irrelevant state changes of "relevant objects", the second table shows those of more random actions, at a lower level. Navigation No Op, the first kind, was simply obtained by detecting the existence of consecutive Turn Lef/Right x 4, Forward + Backward, Pan Right + Pan Left, Turn Right + Turn Left. The second kind, interaction No Op, was similarly detected. Whether an consecutive and opposite interactions were done on the same "object" was detected by replaying the `pred_actions` in the model outputs. Interaction w. unrelated objects denotes whether the demonstration an object that is completely unrelated from task type (e.g. picking up *saltshaker* for a task whose type is *'Coffee'*). Demonstrations unaligned with dialogue were counted manually since there is no automatic way to filter these.

Penalizing Agents for Accuracy The statistics in this subsection were straightforwardly obtained by averaging over the evaluation outputs (whose formats follow that of the original ET code from TEACh) of each task.

Behavior Cloning with Suboptimal Demonstrations The same procedures for the second table in **Irrelevant Actions** were used.

C TEACh Prefiltering

Only necessary state changes are checked in EDH evaluation, but all are present in training. https://github.com/alexa/teach#downloading-the-dataset mentions that the authors filtered the EDH tasks so that "the state changes checked for to evaluate success are only those that contribute towards task success in the main task of the gameplay session the EDH instance is created from." Our analysis is on data that has already been filtered and cleaned and yet still exhibits these problems.