Demonstrating EMMA: Embodied MultiModal Agent for Language-guided Action Execution in 3D Simulated Environments

Alessandro Suglia, Bhathiya Hemanthage, Malvina Nikandrou, Georgios Pantazopoulos, Amit Parekh, Arash Eshghi, Claudio Greco, Ioannis Konstas, Oliver Lemon, Verena Rieser

{a.suglia, hsb2000, mn2002, gmp2000, amit.parekh, a.eshghi, c.greco, i.konstas, o.lemon, v.t.rieser}@hw.ac.uk

Abstract

We demonstrate EMMA, an embodied multimodal agent which has been developed for the Alexa Prize SimBot Challenge[1]. The agent acts within a 3D simulated environment for household tasks. EMMA is a unified and multimodal generative model aimed at solving embodied tasks. In contrast to previous work, our approach treats multiple multimodal tasks as a single multimodal conditional text generation problem. Furthermore, we showcase that a single generative agent can solve tasks with visual inputs of varying length, such as answering questions about static images, or executing actions given a sequence of previous frames and dialogue utterances. The demo system will allow users to interact conversationally with EMMA in embodied dialogues in different 3D environments from the TEACh dataset.

1 Introduction

Robots that perform tasks in human spaces can benefit from natural language interactions that provide both high and low-level instructions, as well as the ability to resolve ambiguities. The Alexa Prize SimBot Challenge aims to propel research efforts to develop embodied agents that learn to execute household tasks from instructions, such as *"Please clean all the tableware"*.

Transformers (Vaswani et al., 2017) coupled with joint vision-and-language pretraining have become the standard approach for tasks with single image inputs, where available object-detectors are used produce image features. We demonstrate how this approach can also benefit embodied agents for object manipulation tasks. While representing the scene in terms of object representations (object-centric) can also benefit embodied agents performing tasks involving object manipulation, this approach is not as widely adopted due to the increased computational overhead.

To complete a task, an embodied agent may be required to perform multiple successive actions. Each predicted action is conditioned on all previous observations that yields a new observation. From an object-centric point-of-view, each observation corresponds to a set of detected objects which must remain accessible by the agent to predict the next action. Therefore, even for smaller action trajectories, the resulting input length can become prohibitively large as the number of frames increases.

In this work, we present Embodied MultiModal Agent (EMMA), a language-enabled embodied agent capable of executing actions conditioned on historical dialogue interactions. To address the long-horizon input, we adopt advances from tasks involving processing long-documents (Beltagy et al., 2020). Existing embodied agents in similar environments treat action prediction as a classification task (Suglia et al., 2021; Pashevich et al., 2021). On the other hand, EMMA is a unified, visually-conditioned, autoregressive text generation model that accepts visual (observations) and textual (dialogue) tokens as input, and produces natural language text and executable actions.

2 Background

TEACh The Task-driven Embodied Agents that Chat (TEACh) dataset (Padmakumar et al., 2021) consists of gameplay sessions where two participants must complete household tasks in the AI2-THOR simulator (Kolve et al., 2017). Each session consists of a *Commander* with oracle information, and a *Follower* that interacts with the environment and communicates with the *Commander* to complete the task. This work focuses on Execution from Dialogue History (EDH), which is the reference task for the Alexa Prize SimBot Challenge. EDH instances are created by segmenting game sessions. Each instance is defined by an initial state S^E, action history A_H, set of interaction actions during the session A_I^R, and the goal environment

[1] https://amazon.science/alexa-prize/simbot-challenge

Proceedings of the SIGdial 2022 Conference, pages 649–653
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

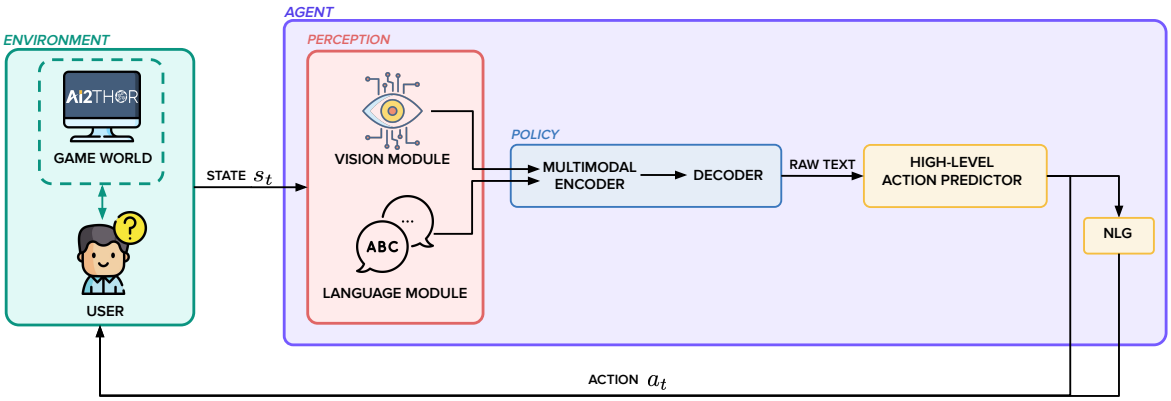


Figure 1: High-level architecture of EMMA. The *Perception* component processes new visual and language input at each timestep. Both streams are then processed by the *Policy* component to output raw text, which is mapped to actions that are executable in the environment. The resulting action a_t can be either a physical action or text (as utterances generated from the dedicated NLG component).

state F^E. The agent models the *Follower* who has to generate the actions leading to the goal state.

Training and Evaluation During training, the A_I^R are used for supervision. At inference time, the model is expected to generate a sequence of interaction actions which would result in F^E. The model is evaluated by comparing the simulator state resulted from inferred actions against F^E.

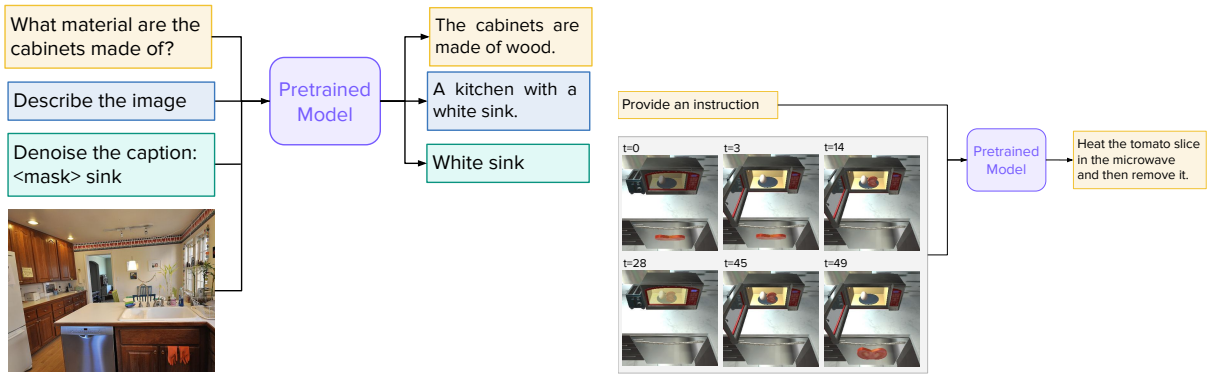
3 System Architecture

As shown in Figure 1, EMMA consists of three components: *Perception*, *Policy*, and *Action Predictor*. At each timestep, the agent generates the next action after receiving information regarding the current and previous states of the environment—including any executed actions and interactions. The agent receives a new observation and has to predict a follow-up action. The process is repeated until the agent outputs a stop action.

Perception This module is responsible for processing the state of the environment—encoding past actions, frames, and dialogue to create the model input. The current state for the EDH task consists of observations obtained after executing an action, or a dialogue utterance from the *Follower* or *Commander*. We extract local and global information from the visual scenes using the VinVL object detector (Zhang et al., 2021), after fine-tuning on the ALFRED images (Shridhar et al., 2020). From each scene, we obtain up to 36 regional features. We obtain the global representation as the mean pooled features from the backbone of the detector. In the second case, the dialogue utterance is concatenated with the dialogue history. We include special tokens to distinguish between *Follower* and *Commander* utterances.

Policy The core component of EMMA is a unified autoregressive text generation model. Given the current state, the previous observations and interactions, the model generates raw textual output. Assuming the input sequence consists of V frames—with each encoded into N_V scene and object tokens—and L language tokens, the total sequence length $V \times N_V + L$ will be dominated by the number of visual tokens. To reduce the impact of having a large V, we adapt the sparse attention pattern following Beltagy et al. (2020). Each token attends to its neighbouring tokens within a local window, and a subset of tokens are regarded as global to aggregate information from longer contexts. Global tokens act as a bottleneck of relevant information over the entire sequence. These tokens can attend to, and are attended by, all other tokens in the input sequence under causal masking.

To infuse our agent with knowledge about objects and their properties, we pretrain the model several image-text and video-text tasks. We use COCO (Lin et al., 2014), VisualGenome (Krishna et al., 2016), and GQA (Hudson and Manning, 2019) to learn an alignment between language and vision. Furthermore, we incorporate ALFRED (Shridhar et al., 2020), and EPIC-KITCHENS (Damen et al., 2018), two video-based datasets involving action execution and recognition to enable temporal reasoning.



(a) Output for three different tasks when given the same image.

(b) Output for the Instruction Prediction task, given an input video from the agent's trajectory.

Figure 2: Example of generated output for various pretraining tasks, showing how EMMA can be prompted for the task using Natural Language prefixes.

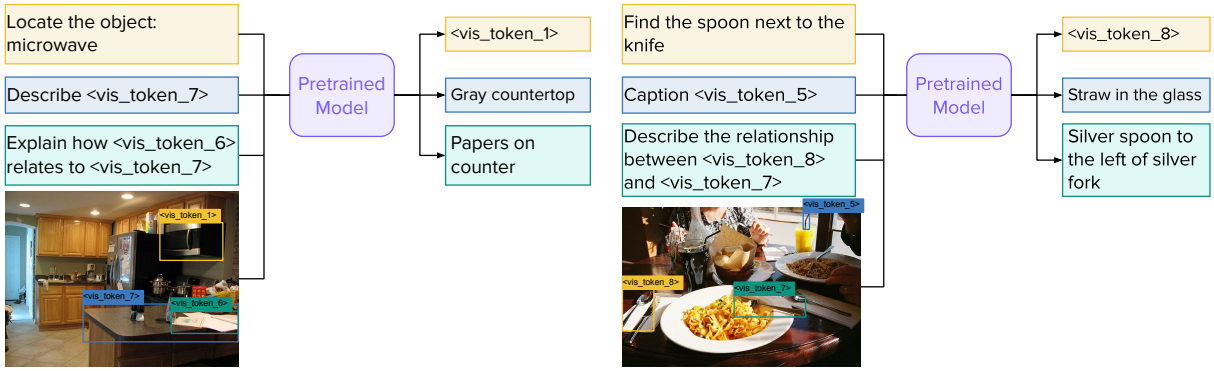


Figure 3: Example of generated output for pretraining tasks showing the use of visual tokens in order to reference specific objects. Visual tokens follow the format <vis_token_i> to refer to the i-th predicted bounding box.

Action Predictor The final component of EMMA is responsible for converting generated raw text into actions which are executable in the environment. We parse the raw text and map it to either a navigation (e.g., Forward) or interaction action (e.g., Pickup Mug). For interaction actions, we also select the associated object using its coordinates available from the *Perception* module.

4 System Demonstration

We demonstrate the ability of our model to solve several downstream tasks ranging from captioning to embodied action execution after casting all tasks into the same sequence-to-sequence framework. After training EMMA, we can use natural language *task prompts* to trigger specific behaviours, following literature on prompting for text-only models (Raffel et al., 2020; Brown et al., 2020).

4.1 Pretraining Tasks

Figures 2-3 show examples of outputs generated for various pretraining tasks. Figure 2a illustrates outputs of a model with the same weights for three image-based tasks: Visual Question-Answering (VQA), Image Captioning, and Masked Language Modelling (MLM). Figure 3 demonstrates the pretraining tasks that require referencing specific objects in the image: Visual Grounding, Dense Captioning and Relationship Detection. Without any special task-specific tokens, EMMA can infer the target task to generate summary descriptions for images, and can also respond to queries regarding attributes of specified objects. Figure 2b shows an example of a video pretraining task using a trajectory from the ALFRED (Shridhar et al., 2020) dataset. Given the task prefix "Provide an instruction" and a sequence of frames, EMMA learns to generate an high-level description of the action trajectory.

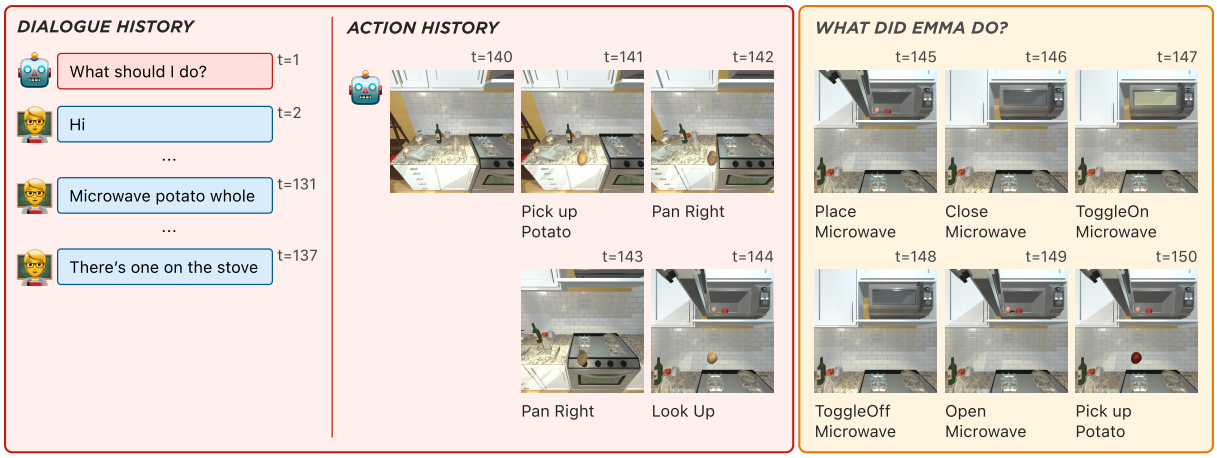


Figure 4: Example of action execution in the AI2Thor 3D environment. EMMA conditions the action generation on both the dialogue and the visual history.

4.2 Action Execution

Figure 4 provides an example of action execution from dialogue history using an episode from the TEACh dataset. The goal of the episode is to microwave a potato. The initial input to the model consists of the dialogue between the *Commander* and the *Follower* as well as the frames corresponding to the previously executed actions. Up to that point, the *Commander* has expressed the end goal and helped the agent locate a potato. Based on this input, EMMA executes a sequence of actions that successfully complete the task. At each step the initial input is augmented with the agent's egocentric observation after executing the most recent action. The process is repeated until the timestep 49, where EMMA predicts a stop action. For this particular example, the human follower completed the task in 10 steps including redundant actions such as looking up and down. EMMA's action trajectory is more efficient than the human demonstration by performing only the necessary actions.

5 Conclusion

In this work we presented EMMA, an embodied agent that learns to execute actions from dialogue, developed for the Alexa Prize SimBot Challenge. EMMA is based on a unified text generation model that is pretrained on multiple image and video-based tasks using natural language prompts. We will provide a conversational web-based demonstration of interaction with EMMA in 3D environments.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Kumar Gupta, and Ali Farhadi. 2017. AI2-THOR: An interactive 3d environment for visual ai. *ArXiv*, abs/1712.05474.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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. *arXiv preprint arXiv:2110.00534*.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15942–15952.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. 2021. Embodied BERT: A transformer model for embodied, language-guided visual task completion. *arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.