Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Journal Track

Core Challenges in Embodied Vision-Language Planning (Extended Abstract)*

Jonathan Francis[1,2†], **Nariaki Kitamura**[3], **Felix Labelle**[2], **Xiaopeng Lu**[2], **Ingrid Navarro**[2], **Jean Oh**[2]
[1]Bosch Center for Artificial Intelligence
[2]School of Computer Science, Carnegie Mellon University
[3]Komatsu Limited
{jmf1, ingridn, jeanoh}@cs.cmu.edu, nariaki_kitamura@global.komatsu,
{flabelle, xiaopen2}@alumni.cmu.edu

Abstract

Recent advances in the areas of Multimodal Machine Learning and Artificial Intelligence (AI) have led to the development of challenging tasks at the intersection of Computer Vision, Natural Language Processing, and Robotics. Whereas many approaches and previous survey pursuits have characterised one or two of these dimensions, there has not been a holistic analysis at the center of all three. Moreover, even when combinations of these topics are considered, more focus is placed on describing, e.g., current architectural methods, as opposed to *also* illustrating high-level challenges and opportunities for the field. In this survey paper, we discuss Embodied Vision-Language Planning (EVLP) tasks, a family of prominent embodied navigation and manipulation problems that jointly leverage computer vision and natural language for interaction in physical environments. We propose a taxonomy to unify these tasks and provide an in-depth analysis and comparison of the current and new algorithmic approaches, metrics, simulators, and datasets used for EVLP tasks. Finally, we present the core challenges that we believe new EVLP works should seek to address, and we advocate for task construction that enables model generalisability and furthers real-world deployment.

1 Introduction

With recent progress in the fields of Artificial Intelligence (AI) and Robotics, intelligent agents are envisaged to interact with humans in shared environments. Such agents include any entities that can make decisions and take actions autonomously and are expected to understand semantic concepts in those environments, using, e.g., visual, haptic, auditory, or textual information perceived via sensors [Wooldridge and Jennings, 1995; Castelfranchi, 1998]. With the goal of developing intelligent agents equipped with these sensory and reasoning capabilities, Embodied AI (EAI), as a field, has become popular for studying the particular set of AI problems surrounding agents situated in a physical environment: recently, the number of papers and datasets for the tasks that require the agents to use both vision and language understanding has increased markedly [Das *et al.*, 2018; Gordon *et al.*, 2018; Anderson *et al.*, 2018; Krantz *et al.*, 2020; Thomason *et al.*, 2019; Nguyen and Daumé III, 2019; Majumdar *et al.*, 2020; Li *et al.*, 2020]. In this article, we conduct a survey of recent works on these types of problems, which we refer to as *Embodied Vision-Language Planning* (EVLP) tasks. In this article, we aim to provide a bird's-eye view of current research on EVLP problems, addressing their main challenges and future directions. Our main contributions are as follows: (i) We formally define the field of Embodied Vision-Language Planning and we propose a taxonomy that both unifies a set of related tasks in EVLP and serves as a basis for categorising new tasks; (ii) we survey recent EVLP tasks, compare their task properties, highlight modelling approaches used in those tasks, and analyse the datasets, simulators, and metrics used to evaluate the approaches on those tasks; finally, (iii) we identify open challenges that afflict existing works in the EVLP family, with an emphasis towards encouraging unseen generalisation and deploying algorithms to the real world. We refer readers to our full journal article for further details [Francis *et al.*, 2022b].

1.1 Problem Definition

We discuss a broad set of problems, related to an embodied agent's ability to make planning decisions in physical environments. Formally, let S and A denote sets of states and actions; V and L denote sets of vision and language inputs available to the agent. A planning problem is defined by the tuple $\Phi = \{S, A, s_{ini}, s_{goal}\}$, where $s_{ini}, s_{goal} \in S$ denote initial and goal states, respectively. A solution $\psi \in \Psi_\Phi$ to planning problem Φ is a sequence of actions to take in each state, starting from an initial state to reach a goal state, $\psi = [s_{ini}, a_0, ..., s_t, a_t, ..., a_T, s_{goal}]$, where $t \in T$ is a finite time-step in episode length T and Ψ_Φ is a set of possible solutions to Φ. Given a particular EVLP problem Φ, state $s_t \in S$ at time step t can be defined in terms of vision and language inputs up to the current time step, such that, $s_t = \{(v_0, l_0), (v_1, l_1), \ldots, (v_t, l_t), \ldots, (v_T, l_T)\}$, where $v_t \in V$ and $l_t \in L$. The agent's objective is to minimize the difference between an admissible solution $\psi \in \Psi_\Phi$ and its predicted one $\bar{\psi}$. This definition broadly captures the crux of

*The full version was published at JAIR [Francis *et al.*, 2022].
†Contact author.

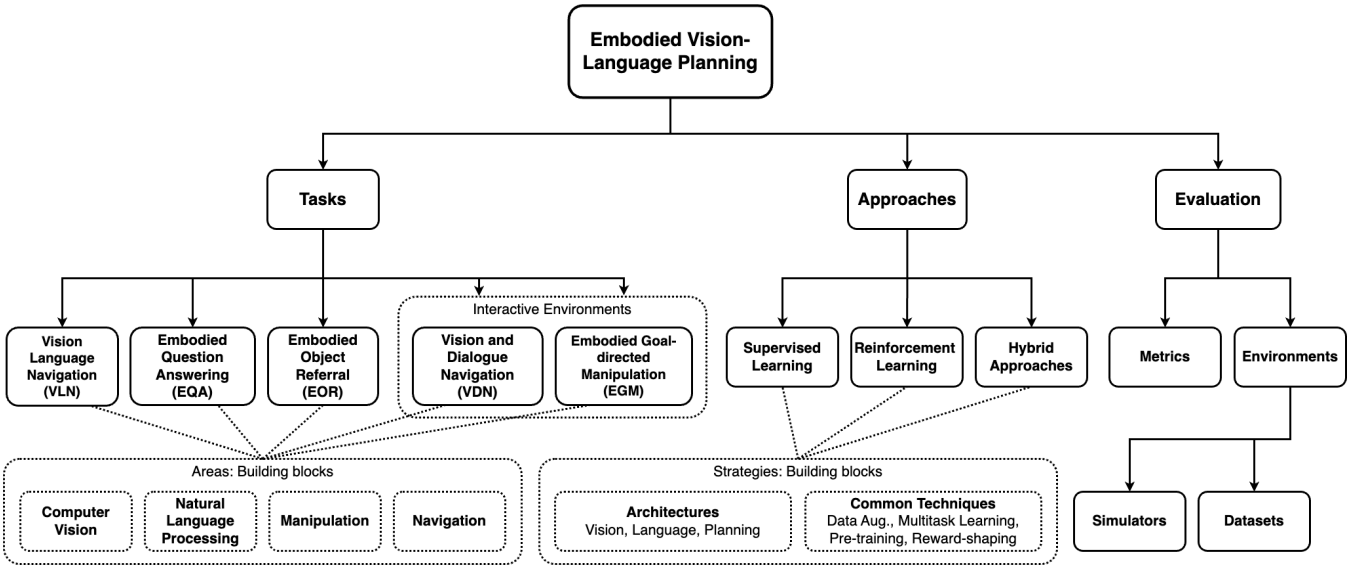

Figure 1: Taxonomy of *Embodied Vision-Language Planning*.

EVLP problems. Customized definitions are needed for specific tasks, where additional constraints or assumptions are added to focus on particular subareas of this general problem.

1.2 Taxonomy

We propose a taxonomy of EVLP research, illustrated in Figure 1, around which the rest of the paper is organized. The taxonomy subdivides the field into three branches; tasks, approaches, and evaluation methods. The Tasks branch proposes a framework to classify existing tasks and to serve as a basis for distinguishing new ones. The Approaches branch touches on the learning paradigms, common architectures used for the different tasks, as well as common tricks used to improve performance. The right-most branch of the taxonomy discusses task Evaluation Methodology, which is subdivided into two parts: metrics and environments. The metrics subsection references many of the common metrics, used throughout EVLP tasks, while the environments subsection presents the different simulators and datasets currently used.

2 Current Approaches

We provide a brief overview of the tasks, methodology, learning paradigms, datasets, simulators, and metrics used in the EVLP task family; we provide additional references, task-specific problem definitions, architectural descriptions and training objectives, dataset and simulator comparisons and statistics, and metric formulæ in [Francis *et al.*, 2022b].

2.1 Tasks, Methods, and Learning Paradigms

EVLP Tasks. Many EVLP tasks have been proposed, with each task focusing on different technical challenges and reasoning requirements for agents. Tasks vary on the basis of the action space (types and number of actions possible), the reasoning modes required (e.g., instruction-following, versus exploration and information-gathering), and whether or not the task requires interaction with another agent. Vision-Language Navigation and Vision and Dialogue Navigation require agents to use natural language instructions to navigate to goal locations in environments, where the latter provides agents with intermediate supervision and clarifications. In Embodied Question Answering tasks, an agent initially receives a language-based question, and must engage in guided exploration of the environment, in order to collect enough information about its surroundings to generate an answer. In Embodied Object Referral tasks, an agent navigates to an object mentioned in a given instruction, and has to identify (or select) it upon reaching its location. Embodied Goal-directed Manipulation tasks combine manipulation-based environment interactions with requirements from aforementioned tasks, such as navigation and path-planning, state-tracking, instruction-following, instruction decomposition, and object-selection.

Methods and Learning Paradigms. Technical approaches that pursue solutions to EVLP tasks must model various facets. Firstly, modelling vision typically involves building a lossless and predictive state representation of the agent's environment; because ego-centric visual observations change as the agent navigates or manipulates objects in the space, the agent must also include temporal modelling mechanisms, in order to represent observed state-changes in its environment over time and to monitor progress of task-execution. Next, modelling language in EVLP tasks typically requires using the instructions or question prompts provided to generate a rich description of the agent's goal; because language can be ambiguous in practice, challenges remain in obtaining unbiased representations. Next, the agent must be able to compare its progress in the task with its representation of the goal, typically requiring sophisticated strategies for multimodal representation, alignment, and fusion. Finally, in order to interact with their environments, agents must include mechanisms for action-generation and planning; inspired by classical approaches in robotics, many such approaches follow from early works in mapping and exploration strategies, search and topological planning,

and hierarchical task decomposition. To bias agents towards desired task-oriented behaviour, approaches leverage various learning paradigms (e.g., semi/self/fully-supervised learning, reinforcement learning, etc.) and strategies (e.g., pre-training, data augmentation, multitask learning, reward-shaping, cycle-consistency, etc.).

2.2 Datasets, Simulators, and Metrics

Datasets. EVLP datasets vary across three primary main dimensions: visual observations, natural language inputs, and expert demonstrations. Visual observations, in general, consist of RGB images often paired with depth data or semantic masks. These observations can represent both indoor and outdoor environments from both, photo-realistic or synthetic-based settings. In contrast, language varies in the type of prompt. Language prompts may come in the form of questions, step-by-step instructions, or ambiguous instructions that require some type of clarification through dialog or description. Language can also vary in terms of complexity of language sequences and scope of vocabulary. Finally, navigation traces differ in aspects like the granularity (or discretization) of the action-space and the implicit alignment that a provided action sequence has with the other two dimensions.

Simulators. Early simulation platforms for EAI research typically leveraged simple video game environments to create and train neural controllers. Human performance was quickly achieved on many of these platforms, as simplified environments generally lack the diversity and complexity of real-world settings. Recent works have addressed this lack of realism through the use of photo-realism and the use of interactive contexts where agents are able to modify the states of objects in the environment. Toward this end, there is also interest in developing frameworks focused on simulation-to-real transfer and evaluation, allowing the study of discrepancies between real settings and simulated ones. Finally, other platforms have also focused on encouraging reproducibility of work, flexibility of design, and benchmarking.

Metrics. Popular metrics in EVLP research can be grouped into categories—each measuring a different aspect of agent performance, such as distance (quantifies the manner in which an agent traversed a space), success (characterises extent to which the overall task is completed by an agent), path-path similarity (assesses the extent to which the agent's trajectory was similar to the ground-truth), instruction-based metrics (measures the alignment between natural language instructions and the agent's trajectory), and object-centric metrics (assess efficacy of object selection, identification, or manipulation). We illustrate the first three, in Figure 2.

3 Core Challenges

3.1 New Directions in EVLP Research

We highlight three promising directions, in the pursuit of more ubiquitous human-robot interaction and better agent generalisation. Firstly, we advocate for improved *social interaction*: we feel that a progression from static instructions to active dialogues would enable new collaborative and assistive capabilities to emerge. Next, to enable agents that accommodate more complexities of real-world deployment, we advocate for the introduction of *dynamic environments* in EVLP research, encouraging agents to incorporate reasoning strategies that are robust to environment uncertainty and non-stationarity. We discuss a vision for *cross-task robot learning*, wherein agents may acquire experience from related modality-centric tasks, before their deployment to shared multimodal settings with significant task overlap. Finally, we would highlight new directions in interactive object perception for transfer learning, where agents must physically interact with the environment in order to learn new concepts [Tatiya *et al.*, 2023b; Tatiya *et al.*, 2023a].

3.2 Use of Domain Knowledge

We further encourage the development of methods that utilise domain knowledge in a principled way, for guiding the learning and transfer of models; while this notion has seen a recent resurgence in other fields [Francis, 2022; Park *et al.*, 2020; Francis *et al.*, 2022a; Herman *et al.*, 2021; Andreas *et al.*, 2016; Francis *et al.*, 2019], we notice few such works in EVLP. Domain knowledge comes in many forms, e.g., graphical models, logical rules, constraints, pre-training, knowledge graphs, and others; and while domain knowledge holds the promise of improving agents' sample-efficiency, interpretability, safety, and generalisability, the challenge exists in how to effectively express and utilise this domain knowledge in an arbitrary learning problem. *Pre-training* and *commonsense knowledge*, in particular, serve as two manifestations that show promise for imbuing agents with the aforementioned attributes.

Pre-training tasks have been carefully designed and coupled with popular high-capacity models, for self-supervision in such domains as image classification [He *et al.*, 2016] and natural language processing [Devlin *et al.*, 2019; Yang *et al.*, 2019b; Ma *et al.*, 2021], in attempts to maximise the generalisability of transferred or fine-tuned approaches. While there is some progress in the context of specific multimodal problems [Majumdar *et al.*, 2020; Hao *et al.*, 2020; Lu *et al.*, 2019], challenges remain for developing generalisable pre-training strategies that encompass the scope of the broader EVLP task family.

Commonsense knowledge acquisition and injection in models remains an active research area in NLP [Talmor *et al.*, 2019; Ma *et al.*, 2019; Ma *et al.*, 2021; Li *et al.*, 2021], with some works proposing to ground observations with structured commonsense knowledge bases, directly, thereby improving downstream performance on relevant tasks. However, the use of commonsense knowledge in the context of EVLP tasks remains largely unexplored. As the ultimate goal of EVLP tasks is to develop intelligent agents that are capable of solving real-world problems in realistic environments, it is reasonable to consider providing models with structured external knowledge of the world [Yang *et al.*, 2019a; Tatiya *et al.*, 2022].

3.3 Agent Training Objectives

Selecting the appropriate training objective(s) for agents undertaking a given task has been a long-standing problem in machine learning and artificial intelligence; this selection depends on the nature of the available training signals (e.g.,

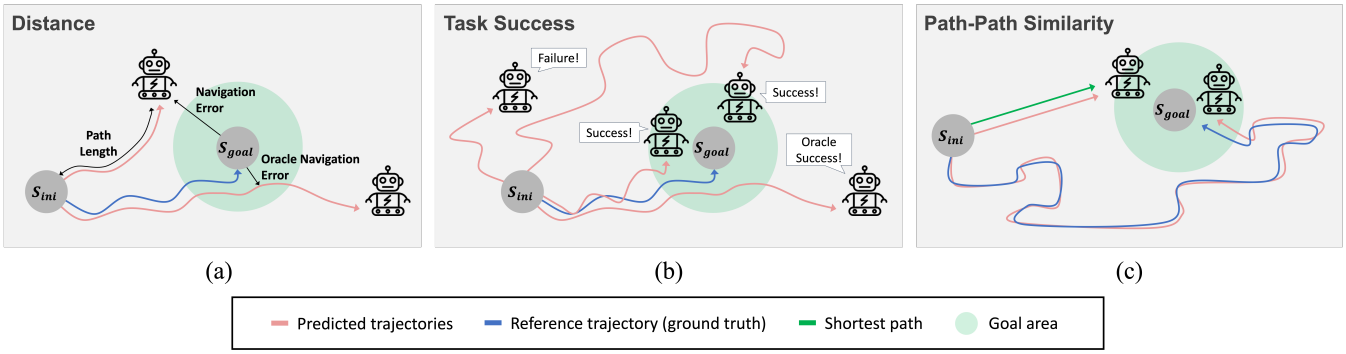

Figure 2: Phenomena measured by typical metrics used in *Embodied Vision-Language Planning* tasks.

reward or cost functions, level of supervision, environment observability) and on the degree to which external knowledge (e.g., auxiliary objectives, common sense, constraints) are deemed necessary for effectively biasing agent behaviour. For EVLP tasks, the selection of training objectives is made more challenging by the complex nature of the environments in which agents operate. This often necessitates frameworks that consist of more than one biasing strategy. Given the underlying motivation of optimising for generalisability and interpretability, explicit treatment should be given to finding the learning paradigm(s) that most effectively integrate information for various related sources and generalises agents' inductive biases to new environments; indeed, the training paradigms should include explicit mechanisms for encouraging the properties we hope to imbue.

3.4 Simulation-to-Real Gap

Datasets and simulation environments are the primary driving forces behind EVLP research, since the measure of model efficacy relies on the availability of strong testing scenarios and the appropriate evaluation criteria. Current EVLP tasks are implemented as a set of goals and metrics, atop pre-existing simulators or datasets. In this section, we urge the community to consider and prioritise the deployment of EVLP agents to real-world settings. Specifically, we assert that various EVLP tasks and metrics may be improved on the basis of three dimensions: *simulator realness*, *dataset realness*, and *tests for model generalisability*.

Simulation-based training and execution are especially attractive when modelling a sequential learning problem, since offline datasets do not allow for such recursive interaction with an environment. There are limitations, however, in how effectively scientists and practitioners can encourage the desired model behaviour to emerge for real-world use-cases. Because this dissonance reduces models' immediate viability for real-world deployment, we assert the importance of increased attention from the computer vision and robotics communities on the topics of simulation-to-real transfer, unseen generalisation, robustness to out-of-distribution settings. We encourage the definition of metrics that assess intermediate agent behaviours and task efficiency, as opposed to simply indicating in-domain task completion.

Dataset-based training can be highly effective, e.g., when providing models with strong priors on agent behaviour. Relying solely on datasets for training can present significant issues, however, not least of all causal confusion, limited observation of the environment's transition dynamics, and unrealistic priors due to class imbalance. Additionally, tasks metrics that have been defined on top of datasets may vary in their ability to truly assess whether agents are ready to be deployed in the real world. Despite these challenges, well-constructed datasets can prove instrumental in encouraging models to learn specialised skills; datasets that enable agents to be trained across multiple environments can lead to more generalisable behaviour.

We further discuss the issues in evaluating EVLP agents in real-world deployments, and we propose new assessment methods to address these challenges. We suggest that current evaluation paradigms should explicitly test the agents' generalisability across domains and tasks, as models need to be transferable to unseen environments and robust to model and environmental uncertainty. We also highlight the need for test beds with different (sub-)domain partitions, as current splits may not be representative of the variation in the intended real-world scenarios. Overall, our goal is to emphasise the importance of comprehensive evaluation paradigms for EVLP agents and to propose new methods to assess their generalisability and robustness in real-world deployment.

4 Conclusion

In this extended abstract, we proposed a taxonomy for the field of *Embodied Vision-Language Planning* (EVLP), which highlights: tasks, modelling approaches, learning paradigms, and evaluation settings; we provided a framework for discussing existing and future tasks, based on the skills required to solve them. We alluded to various learning paradigms, training strategies, and commonly-used optimisation objectives; and we considered different forms of evaluation—e.g., using datasets of expert demonstrations, simulators, and various task metrics. Finally, we focused on the challenges currently being tackled in the field, as well as those that still remain to be addressed. Specifically, we discuss issues that could prevent real-world deployment, such as a lack of generalisation, robustness, simulator realness, and lack of rich interaction; we highlight these as the most promising and fulfilling next directions to follow.

Acknowledgments

The authors thank Alessandro Oltramari, Yonatan Bisk, Eric Nyberg, and Louis-Philippe Morency for insightful discussions on the original journal article; the authors also thank Gyan Tatiya, Jimin Sun, Prasoon Varshney, Sahiti Yerramilli, and Jayant Tamarapalli for new/recent EVLP collaborations.

References

[Anderson *et al.*, 2018] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3674–3683. IEEE Computer Society, 2018.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48. IEEE Computer Society, 2016.

[Castelfranchi, 1998] Cristiano Castelfranchi. Modelling social action for ai agents. *Artificial intelligence*, 103(1-2):157–182, 1998.

[Das *et al.*, 2018] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1–10. IEEE Computer Society, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[Francis *et al.*, 2019] Jonathan Francis, Matias Quintana, Nadine von Frankenberg, Sirajum Munir, and Mario Berges. Occutherm: Occupant thermal comfort inference using body shape information. In *Proceedings of the 6th International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '19, New York, NY, USA, 2019. ACM.

[Francis *et al.*, 2022a] Jonathan Francis, Bingqing Chen, Siddha Ganju, Sidharth Kathpal, Jyotish Poonganam, Ayush Shivani, Vrushank Vyas, Sahika Genc, Ivan Zhukov, Max Kumskoy, Jean Oh, Eric Nyberg, and Sylvia L. Herbert. Learn-to-race challenge 2022: Benchmarking safe learning and cross-domain generalisation in autonomous racing. In *1st ICML Workshop on Safe Learning for Autonomous Driving*, 2022.

[Francis *et al.*, 2022b] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, 74:459–515, 2022.

[Francis, 2022] Jonathan Francis. *Knowledge-enhanced Representation Learning for Multiview Context Understanding*. PhD thesis, Carnegie Mellon University, 2022.

[Gordon *et al.*, 2018] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: visual question answering in interactive environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4089–4098. IEEE Computer Society, 2018.

[Hao *et al.*, 2020] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13134–13143. IEEE, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[Herman *et al.*, 2021] James Herman, Jonathan Francis, Siddha Ganju, Bingqing Chen, Anirudh Koul, Abhinav Gupta, Alexey Skabelkin, Ivan Zhukov, Max Kumskoy, and Eric Nyberg. Learn-to-race: A multimodal control environment for autonomous racing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9793–9802, 2021.

[Krantz *et al.*, 2020] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the navgraph: Vision-and-language navigation in continuous environments. *CoRR*, abs/2004.02857, 2020.

[Li *et al.*, 2020] Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1448–1458. IEEE, 2020.

[Li *et al.*, 2021] Yikang Li, Pulkit Goel, Varsha Kuppur Rajendra, Har Simrat Singh, Jonathan Francis, Kaixin Ma, Eric Nyberg, and Alessandro Oltramari. Lexically-constrained text generation through commonsense knowledge extraction and injection. In *Common Sense Knowledge Graphs at the 35th AAAI Conference on Artificial Intelligence (CSKGs@AAAI-21)*, 2021.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer,

Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019.

[Ma *et al.*, 2019] Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China, 2019. Association for Computational Linguistics.

[Ma *et al.*, 2021] Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[Majumdar *et al.*, 2020] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. *CoRR*, abs/2004.14973, 2020.

[Nguyen and Daumé III, 2019] Khanh Nguyen and Hal Daumé III. 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, 2019. Association for Computational Linguistics.

[Park *et al.*, 2020] Seong Hyeon Park, Gyubok Lee, Manoj Bhat, Jimin Seo, Minseok Kang, Jonathan Francis, Ashwin R Jadhav, Paul Pu Liang, and Louis-Philippe Morency. Diverse and admissible trajectory forecasting through multimodal context understanding. In *European Conference on Computer Vision (ECCV)*, 2020.

[Talmor *et al.*, 2019] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[Tatiya *et al.*, 2022] Gyan Tatiya, Jonathan Francis, Luca Bondi, Ingrid Navarro, Eric Nyberg, Jivko Sinapov, and Jean Oh. Knowledge-driven scene priors for semantic audio-visual embodied navigation. *arXiv preprint arXiv:2212.11345*, 2022.

[Tatiya *et al.*, 2023a] Gyan Tatiya, Jonathan Francis, and Jivko Sinapov. Cross-tool and cross-behavior perceptual knowledge transfer for grounded object recognition. *arXiv preprint arXiv:2303.04023*, 2023.

[Tatiya *et al.*, 2023b] Gyan Tatiya, Jonathan Francis, and Jivko Sinapov. Transferring implicit knowledge of non-visual object properties across heterogeneous robot morphologies. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2023.

[Thomason *et al.*, 2019] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. *CoRR*, abs/1907.04957, 2019.

[Wooldridge and Jennings, 1995] Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

[Yang *et al.*, 2019a] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[Yang *et al.*, 2019b] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764, 2019.