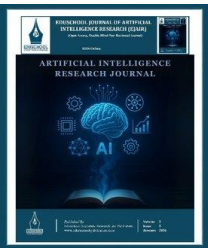
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Grounded Language Generation from Raw Sensory Input in Embodied AI Systems

Manasy Jayasurya

Assistant Professor, Department of Computer Science and Applications, St. Mary's College (Autonomous), Thrissur, India

Article information

Received: 5th August 2026
Received in revised form: 8th September 2026
Accepted: 10th October 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0011

Abstract

Language produced by a system that has never acted in the world describes that world from the outside. Embodied systems that perceive through their own sensors and change what they perceive by acting offer a different setting for language generation, one in which the meaning of a term can be tied to sensory evidence and to the actions the term licenses. This article examines what such grounding requires and how far current systems supply it. We distinguish three requirements, namely perceptual anchoring of reference, action conditioned meaning and situated deixis, and show that pipelines which caption an image and pass the caption to a planner satisfy only the first. End-to-end architectures that map sensory observations and instructions directly to actions and to language remove the intermediate text bottleneck, and we compare eight such systems by input, output and training signal. We then examine the data available, which is smaller than the corpora used for text pre-training by five orders of magnitude, and the evaluation problem, where reference based caption metrics reward fluency and cannot detect a description of an object that is not present. A taxonomy of grounding failures is proposed together with the diagnostic each one requires, and seven open problems are stated.

Keywords:- Symbol Grounding, Embodied Artificial Intelligence, Vision-Language-Action Models, Situated Language Generation, Robot Learning, Hallucination

.

I. INTRODUCTION

A system trained only on text learns the relations between words and other words. The question of whether such relations amount to meaning was posed precisely as the symbol grounding problem, which asks how the symbols manipulated by a formal system acquire any connection to what they are about other than through further symbols [1]. Work in cognitive science supplies one answer, namely that concepts are represented in the same systems that support perception and action, so that understanding a word about grasping involves the machinery used for grasping [2]. A position paper directed at the language processing community made the argument in operational form, holding that the scope of what a system can learn is bounded by the experience available to it, and that text alone occupies a narrow position on that scale [3]. Embodied systems change the terms of the problem. A robot that perceives through cameras and force sensors, and that alters its own observations by moving, has access to the correspondence between a word and an experience. The correspondence can be learned rather than stipulated, which was demonstrated in early work that acquired word forms and their visual referents from recordings of caregivers speaking to infants [39], and pursued further in systems that connected instructions to perceived objects and navigable routes [38], [40]. The scale available now is different, but the requirement is

unchanged. Two developments have made the topic active again. Large vision and language models trained on web data provide a starting point with broad recognition ability and fluent generation [17], [18], [19], and models that map images and instructions directly to robot actions have shown that this ability transfers to control [10], [11], [12], [14]. Yet the language these systems produce remains weakly tied to what their sensors report. Vision and language models describe objects that are not present at rates measurable on purpose built diagnostics [21], [22], and a system that misdescribes a scene it is acting in fails in a way that a captioning system does not, because the description informs the next action.This article examines grounded generation as a problem in its own right rather than as a component of control. Section II states three requirements for grounding and separates them from recognition. Section III compares modular and end-to-end architectures and reviews eight systems. Section IV examines the training data, whose scale is the binding constraint. Section V addresses evaluation, proposes a taxonomy of grounding failures and identifies the diagnostic each requires. Section VI states seven open problems and Section VII concludes.

II. WHAT GROUNDING REQUIRES

Grounding is used loosely in current writing, and separating its components clarifies what a given system achieves. The first requirement is perceptual anchoring, meaning that a referring expression corresponds to evidence in the current sensory input rather than to a prior over what is usually present. A system satisfies this requirement when its description of an object changes appropriately as the object changes, and fails it when the description reflects the statistics of the training corpus. Object hallucination measurements target exactly this failure, and reported rates on constructed probes remain substantial for systems that score well on caption benchmarks [21], [22].The second requirement is action conditioned meaning. Terms such as reachable, heavy, blocked and open describe relations between an object and the capabilities of the agent, and no observer without a body can verify them. A system that has attempted to lift objects has evidence about weight that no amount of image and caption data supplies. Affordance grounded language selection makes the requirement operational by scoring candidate instructions with a value function that estimates whether the described action can be executed from the current state, so that a plan is restricted to what the agent can actually do [9]. The third requirement is situated deixis, covering terms whose referent depends on the position and history of the speaker, including here, this one, behind you and the one I just put down. These cannot be resolved from a single frame, and they are the terms most common in natural instruction, which is why dialogue based benchmarks include them explicitly [27], [28]. The three requirements are ordered by the machinery they demand. Perceptual anchoring needs an encoder and a mechanism that ties generation to it. Action conditioned meaning needs a policy or a value function and the experience to fit it. Situated deixis needs persistent state covering the agent position and the recent history of the interaction. Almost all reported progress addresses the first requirement, which is also the one that transfers most readily from web data, and the third is addressed only in benchmarks designed for it.

III. ARCHITECTURES FOR GROUNDED GENERATION

Two designs are in use, and Fig. 1 contrasts them. The modular pipeline runs a perception module that produces text, passes the text to a language model that plans, and passes the resulting instruction to a controller. The arrangement is attractive because each component can be developed separately and the intermediate text is readable, and it has produced working systems for both planning [31] and composed reasoning across pre-trained models [32]. Its defect is the bottleneck at the text interface, since the caption discards everything the planner was not expected to need, including spatial detail, uncertainty and the appearance of objects that no caption template covers. When the planner asks a question the caption cannot answer, there is no path back to the sensor.

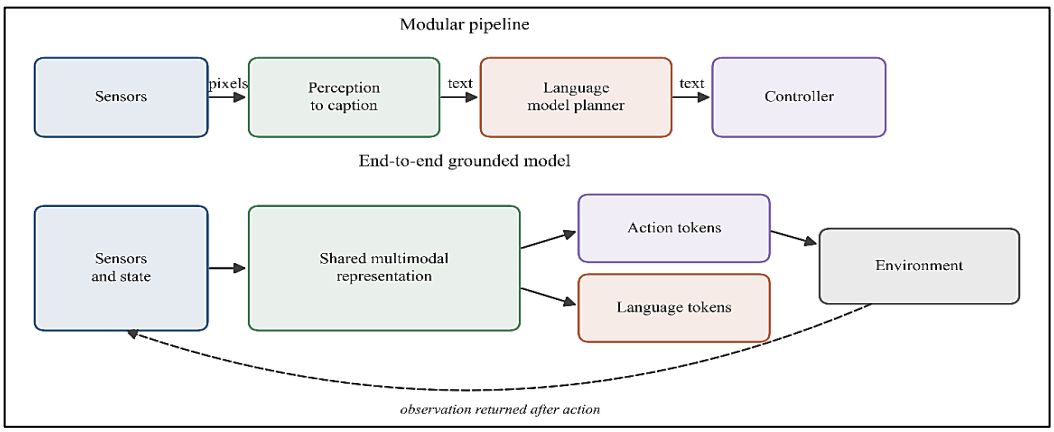


Fig. 1: Modular and end-to-end designs for grounded language generation. The modular pipeline passes information between components as text, which discards sensory detail. The end-to-end design conditions both action and language on a shared representation of observation, instruction and state.

The end-to-end design encodes observations, instruction and state into a shared representation from which both actions and language are produced. Robotics transformers established that actions can be emitted as discrete tokens by a sequence model trained on demonstration data [10], and the vision-language-action formulation extended this by initialising from a model pre-trained on web image and text data, so that the semantic knowledge acquired from the web is available for control [11]. An embodied multimodal language model injects continuous sensor embeddings directly into the token stream of a language model, which allows the same network to plan, answer questions and describe its situation without an intermediate caption [12]. Open implementations of the vision-language-action recipe have made the approach reproducible [14], and generalist sequence models trained across many tasks and embodiments show that a single set of parameters can serve varied outputs [13].

Between the two designs lies a family that keeps a pre-trained visual and linguistic representation and attaches a spatial mechanism to it. Manipulation policies that combine a semantic pathway from a contrastive vision-language model with a spatial pathway that predicts where to act follow this pattern [15], and methods that compose value maps in three dimensions from code produced by a language model achieve similar effects without task specific training [16]. Table 1 compares eight systems along the dimensions that matter for grounded generation, namely what the model consumes, what it emits, where its training signal comes from and whether it produces language at all during operation.

Table 1. Systems for embodied grounding compared by input, output and training signal

System	Year	Input	Output	Training signal	Emits language	Ref.
Gato	2022	Images, text, proprioception	Actions and text	Supervised across tasks	Yes	[13]
SayCan	2022	Image, instruction	Skill selection	Value function plus language model	Yes	[9]
CLIPort	2022	Image, instruction	Pick and place poses	Demonstrations	No	[15]
RT-1	2023	Image history, instruction	Discretised actions	Demonstrations	No	[10]
PaLM-E	2023	Sensor embeddings, text	Text and plans	Mixed robot and web data	Yes	[12]
RT-2	2023	Image, instruction	Action tokens	Web pre-training plus demonstrations	Partly	[11]
VoxPoser	2023	RGB-D, instruction	Value maps for planning	No task specific training	Yes	[16]
OpenVLA	2024	Image, instruction	Action tokens	Open cross-embodiment data	No	[14]

The final column of Table 1 identifies the gap this article addresses. Most systems that act do not speak, and most systems that speak do not act. The two that do both produce language conditioned on sensory input [12], [13], and neither has been evaluated for the accuracy of that language against what the sensors actually contained. A system that acts correctly while describing the scene incorrectly is a common and largely unmeasured outcome, and it matters wherever the description is presented to a person as an account of what the system is doing.

IV. THE TRAINING SIGNAL AND ITS SCALE

The constraint on embodied grounding is data. Fig. 2 shows the scale of the principal resources, and the comparison with text pre-training is instructive, since a corpus of a trillion tokens exceeds the largest robot demonstration collection by roughly five orders of magnitude when both are measured in units of experience. Simulated environments supply volume at the cost of realism, with photorealistic indoor scans supporting navigation and instruction following [6], [8] and interactive simulators supporting manipulation of household objects [7]. Instruction following benchmarks built on these environments provide paired language and trajectory data at the scale of tens of thousands of episodes [4], [5]. Real robot data are scarcer, and the response has been aggregation, with a collaborative collection joining datasets from many laboratories and embodiments into one resource [33].

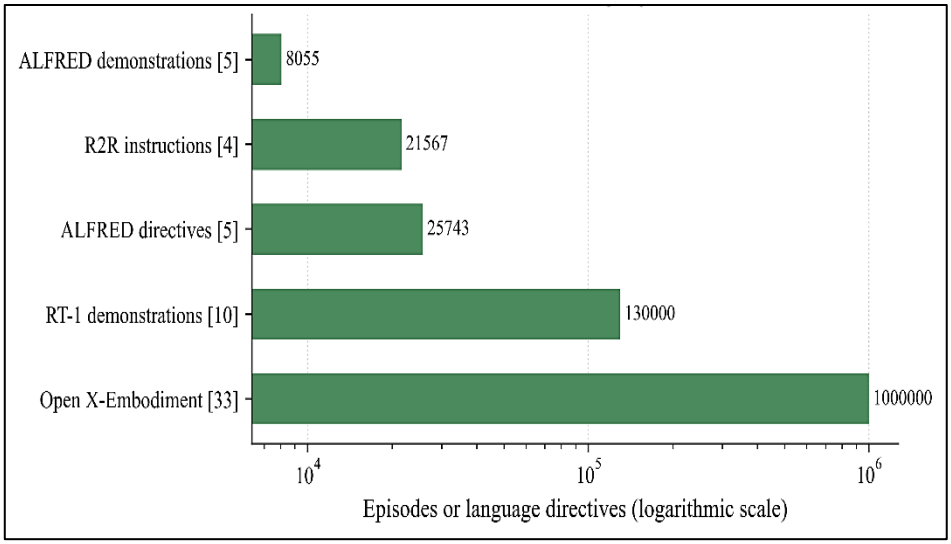


Fig. 2: Scale of paired language and experience resources for embodied learning, in episodes or instructions on a logarithmic scale, as reported by the authors of each resource: room-to-room navigation instructions [4], household task directives [5], real robot demonstrations [10] and the aggregated cross-embodiment collection [33].

Two responses to the shortage are visible in current work. The first is transfer from web data, on the argument that recognition and commonsense knowledge acquired from images and text carry to embodied settings and that only the mapping to action must be learned from robot data. Reported results support this for object identification and for instruction interpretation [11], [12], and they do not establish transfer for the action conditioned terms of Section II, which have no counterpart in web text. The second response is egocentric recording at scale, where thousands of hours of first person video capture human activity with narration, providing experience that is embodied even though the recorder was not an agent under the system's control [34].

Neither response addresses the deeper asymmetry. Language data are abundant and passive, while action data are scarce and active, and the terms that most require grounding are those tied to action. A system can learn from a billion captions what a cup looks like and cannot learn from any of them how heavy a full one is. World models offer a partial route, since a model that predicts future observations conditioned on actions can be trained from unlabelled interaction and can supply the counterfactual experience that language about affordances requires [36], [37]. Multimodal prompting, where a task is specified by a mixture of text and images rather than by text alone, reduces the amount of language that must carry the specification [35], which is a practical mitigation rather than a solution.

V. EVALUATION AND A TAXONOMY OF GROUNDING FAILURE

Caption evaluation was designed for a setting in which several human written references are available and the task is to produce something similar. Overlap based measures reward fluency and phrasing typical of the reference set [24], consensus weighted measures improve the correlation with human judgment [23], and scene graph based measures compare propositions rather than strings [25]. All of them share a defect for the present purpose, which is that a description mentioning an object absent from the scene may still score well when the rest of the sentence matches the references. Reference free measures that score the description against the image with a contrastive model avoid the dependence on references [26] and inherit the weaknesses of the model they use, including the compositional insensitivity documented for such models.

Fig. 3 sets out four failure modes and the diagnostic each requires. Fabricated reference is the description of an entity with no sensory support, which purpose built probes detect by asking about the presence of objects selected according to their frequency in the training data [21], [22]. Frozen reference is a description that remains correct for an earlier state and is not updated as the scene changes, which is detected by presenting a sequence in which one attribute changes and measuring the lag. Capability misattribution is the assertion that an action is possible when the agent cannot perform it, detected by comparing generated claims against the value function or against execution outcomes [9]. Deictic failure is the incorrect resolution of a term whose referent depends on position or history, detected in dialogue settings where the referent is established in an earlier turn [27], [28].

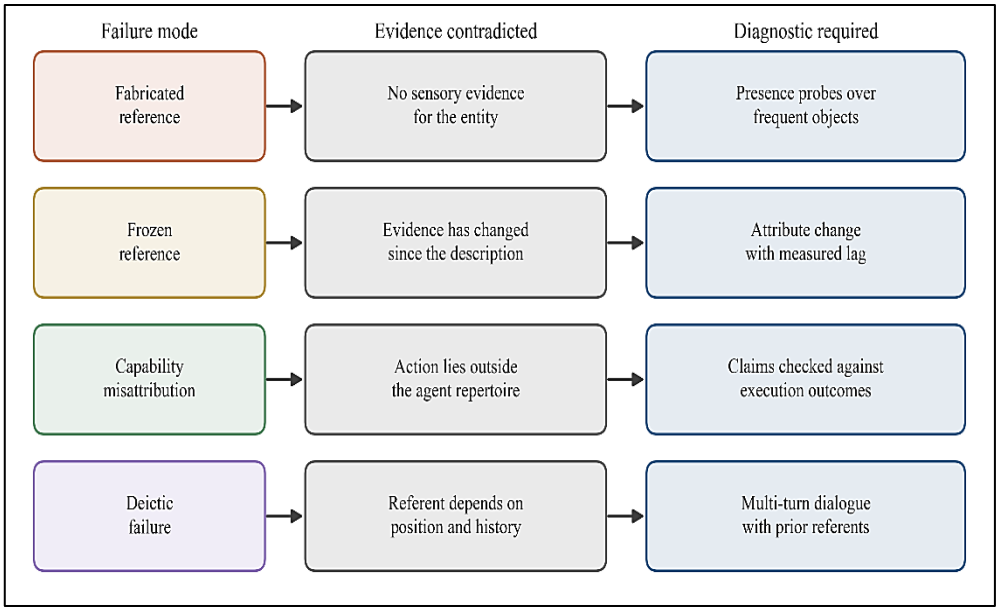


Fig. 3: Four modes of grounding failure in embodied language generation, with the sensory or behavioural evidence each contradicts and the diagnostic procedure required to detect it.

The taxonomy has a practical use beyond classification. Each mode calls for a different remedy. Fabricated reference responds to training objectives that penalise unsupported statements and to decoding constrained by detected entities. Frozen reference responds to architectural changes that force recomputation from current observations rather than reuse of a cached representation. Capability misattribution responds to the affordance scoring already demonstrated for action selection [9], applied to generation rather than to planning. Deictic failure responds to explicit state, meaning a record of agent position and of referents introduced in earlier turns. Treating all four as one phenomenon under the heading of hallucination, as much of the current literature does [20], obscures these distinctions and leads to remedies aimed at the wrong mode.

VI. OPEN PROBLEMS

Seven problems follow from the analysis. The first is an evaluation suite for embodied description that reports the four failure modes separately rather than a single accuracy figure. The second is a training objective in which the correctness of generated language is checked against the environment, which is available in simulation where a scene graph exists and which would supply the supervision that caption corpora cannot [6], [7], [29]. The third is action conditioned vocabulary, requiring corpora in which terms for weight, reachability and obstruction are paired with the outcomes of attempts rather than with images.

The fourth is persistent situated state, without which deictic terms cannot be resolved and which no current architecture maintains explicitly. The fifth is the latency budget, since a model large enough to generate fluent language is too slow for a control loop, and the two rate hierarchy suggested by current systems, with a fast policy and a slower language process, needs a principled treatment. The sixth is transfer measurement, distinguishing what web pre-training genuinely contributes from what it merely appears to contribute through benchmark overlap [11], [12], [30]. The seventh is safety in description, since a system whose account of its own actions is unreliable cannot be supervised by a person reading that account, and the requirement is more stringent than for a captioning system because the consequences are physical.

VII. CONCLUSION

Grounding is not one property. Perceptual anchoring, action conditioned meaning and situated deixis make different demands, and current systems supply the first well, the second partly and the third almost not at all. The architectures that act do not generally describe, the architectures that describe have not been evaluated on whether their descriptions match what their sensors reported, and the data available for the terms that most require embodiment remain smaller than text corpora by several orders of magnitude.

The evaluation position is the most immediately addressable. Caption metrics measure similarity to reference text and cannot detect a fluent description of a scene the system is not in. Separating grounding failure into fabricated reference, frozen reference, capability misattribution and deictic failure identifies four distinct diagnostics, each of which can be built from existing simulators and benchmarks. Beyond evaluation, the substantive requirements are objectives that check generated language against the environment, corpora that pair

action outcomes with the vocabulary describing them, and architectures that maintain the situated state on which the most ordinary instructions depend.

REFERENCES

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.
[2] L. W. Barsalou, "Grounded cognition," *Annual Review of Psychology*, vol. 59, pp. 617–645, 2008.
[3] Y. Bisk *et al.*, "Experience grounds language," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 8718–8735.
[4] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3674–3683.
[5] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "ALFRED: A benchmark for interpreting grounded instructions for everyday tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10740–10749.
[6] M. Savva *et al.*, "Habitat: A platform for embodied AI research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9339–9347.
[7] E. Kolve *et al.*, "AI2-THOR: An interactive 3D environment for visual AI," *arXiv preprint arXiv:1712.05474*, 2017.
[8] A. Chang *et al.*, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vision (3DV)*, 2017, pp. 667–676.
[9] M. Ahn *et al.*, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. Conf. Robot Learning (CoRL)*, 2022, pp. 287–318.
[10] A. Brohan *et al.*, "RT-1: Robotics transformer for real-world control at scale," in *Proc. Robotics: Science and Systems (RSS)*, 2023.
[11] A. Brohan *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Conf. Robot Learning (CoRL)*, 2023, pp. 2165–2183.
[12] D. Driess *et al.*, "PaLM-E: An embodied multimodal language model," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 8469–8488.
[13] S. Reed *et al.*, "A generalist agent," *Trans. Mach. Learn. Res.*, 2022.
[14] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," in *Proc. Conf. Robot Learning (CoRL)*, 2024.
[15] M. Shridhar, L. Manuelli, and D. Fox, "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learning (CoRL)*, 2022, pp. 894–906.
[16] W. Huang *et al.*, "VoxPoser: Composable 3D value maps for robotic manipulation with language models," in *Proc. Conf. Robot Learning (CoRL)*, 2023, pp. 540–562.
[17] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 34892–34916.
[19] J.-B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 23716–23736.
[20] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
[21] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2018, pp. 4035–4045.
[22] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 292–305.
[23] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.
[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.
[25] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 382–398.
[26] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 7514–7528.
[27] 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 *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2017–2025.
[28] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Proc. Conf. Robot Learning (CoRL)*, 2019, pp. 394–406.
[29] R. Krishna *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[30] T. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.
[31] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 9118–9147.
[32] A. Zeng *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[33] [A. Padalkar *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 6892–6903.
[34] [34] K. Grauman *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18995–19012.
[35] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "VIMA: General robot manipulation with multimodal prompts," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 14975–15022.
[36] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.
[37] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 2450–2462.
[38] [38] R. J. Mooney, "Learning to connect language and perception," in *Proc. AAAI Conf. Artif. Intell.*, 2008, pp. 1598–1601.
[39] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.
[40] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 1507–1514.