Ask-to-Act: Learning When Robots Should Clarify Ambiguous Grounded Instructions

Ekele A. Ogbadu[1], Stephanie Lukin[2], and Cynthia Matuszek[1]

***Abstract*— Robots that follow natural-language instructions must decide not only what action to execute, but also whether the current evidence is sufficient to act at all. We present Ask-to-Act, a framework for grounded robot instruction following that treats clarification as a cost-sensitive decision under uncertainty. The system either executes immediately or requests one minimal grounding cue before acting. We evaluate Ask-to-Act on SCOUT++, a grounded HRI benchmark with command-label-image examples from situated human-robot interaction. In a text-only setting, Ask-to-Act improves grounded action prediction from 0.8225 to 0.8308 accuracy while asking in 0.2350 of episodes. In a multimodal setting, command-time visual grounding improves the always-act baseline to 0.8503 accuracy, and Ask-to-Act further improves accuracy to 0.8545 while asking in 0.1277 of episodes. These results show that Ask-to-Act can adapt its interaction rate to the available evidence. It asks more often when language alone leaves action choices underdetermined and asks less often when visual grounding strengthens the base action predictor. To analyze this behavior, we define *ask-beneficial ambiguity* as cases where the gold action is in the model's top-k candidates but not ranked first, and evaluate uncertainty-based ambiguity detectors with AUROC and AUPRC. Overall, Ask-to-Act positions clarification as an adaptive component of grounded robot learning rather than a fixed requirement of language interaction.**

I. Introduction

Natural language is a compelling interface for robots because it allows users to specify goals without low-level programming. In practice, however, many commands are ambiguous or under-specified at the moment they are issued. A user may say "turn a bit left," "move it over there," or "go to the doorway" while leaving open key grounding variables such as direction, distance, target identity, or traversal mode. A robot that commits too early may execute the wrong action with high confidence. A robot that asks too often may turn a fluent interaction into an annoying one. The central problem is therefore not only *what* action to take, but also *whether the robot currently has enough grounded evidence to act without clarification*.

Most grounded instruction-following systems are evaluated as one-shot predictors: they map an utterance, and sometimes perceptual context, directly to an action label. This framing hides an important upstream decision. Even when a model can produce a top-ranked action, the robot still has to decide whether the current belief state is reliable enough for execution or whether one additional piece of information would materially improve the outcome. That decision matters in embodied settings because acting incorrectly has physical consequences, while unnecessary clarification carries social and temporal cost.

We study this problem as a form of **robot learning under uncertainty**. Our framework, **Ask-to-Act**, augments a grounded action predictor with a clarification policy. Given an instruction and available context, the system either acts immediately or requests one minimal grounding cue before acting, as shown in Fig. 1. This formulation treats clarification as part of the robot's control pipeline rather than as a separate dialogue accessory. The key question becomes: when is asking worth it?

We also test a second hypothesis that is especially relevant: the value of clarification depends on the quality of perceptual grounding. If command-time visual evidence already resolves the ambiguity that language leaves open, then asking should become less useful. To evaluate this effect, we use SCOUT++, a benchmark that pairs natural-language commands with synchronized images captured at the time each command is issued. This allows us to compare a text-only regime against a multimodal regime while holding the task and label space fixed.

Our contributions are fourfold. First, we formalize ask-vs-act as a cost-sensitive decision problem layered on grounded action selection. Second, we define *ask-beneficial ambiguity*, an operational notion of ambiguity aligned with the interactive correction setting: the gold action is among the model's plausible candidates but not currently ranked first. Third, we show that selective clarification improves grounded action prediction in both text-only and multimodal settings, while requiring fewer questions when command-time visual grounding is available. Fourth, we show that simple uncertainty signals, especially probability margin, can help identify clarification-worthy episodes. Together, these findings position clarification as an adaptive component of grounded robot learning whose value depends on the robot's current evidence and grounding capability.

II. Related Work

a) Grounded language understanding for robots: Grounding language in robot action has long required combining linguistic input with environmental context, perceptual evidence, and action affordances [1], [2], [3]. Early work in navigation and manipulation showed that many instructions are only interpretable relative to a scene, an embodiment, and a task structure. More recent embodied and multimodal systems extend this paradigm through neural representations and vision–language models, including

[1]University of Maryland Baltimore County, Baltimore, MD United States eogbadu1@umbc.edu, cmat@umbc.edu

[2]DEVCOM Army Research Laboratory, Adelphi, MD, USA stephanie.m.lukin.civ@army.mil

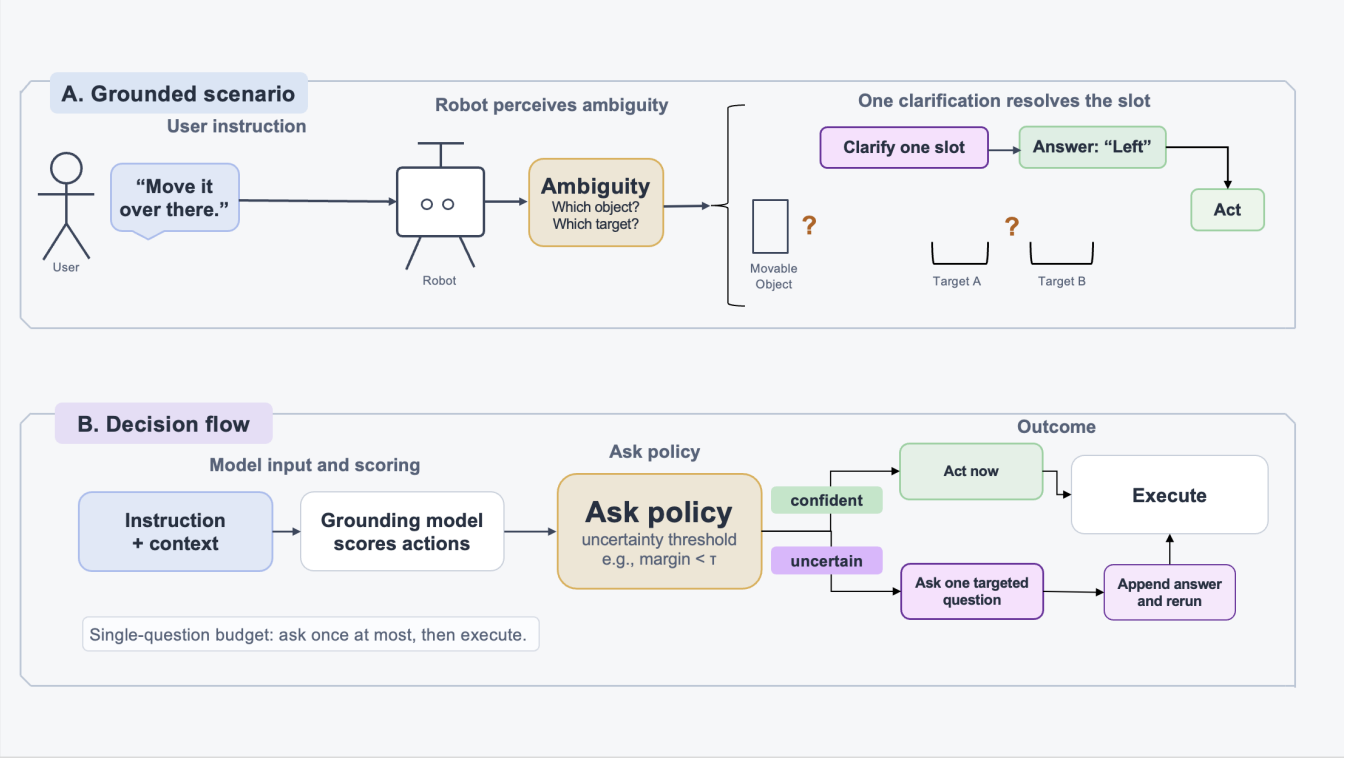

Fig. 1. The decision flow from instruction and context to (i) immediate action or (ii) a clarification query followed by reranking or rerun-based correction.

CLIP-style visual features [4]. Our work also builds on the SCOUT dataset, a Wizard-of-Oz multimodal human–robot dialogue corpus for situated interaction, and on our prior work introducing SCOUT++ as a benchmark derived from that source corpus for grounded instruction understanding [5], [6]. Unlike prior work focused primarily on improving grounded action prediction, we study a preceding decision: whether the currently available grounded evidence is sufficient for execution or whether the robot should first seek clarification.

b) Clarification, help-seeking, and mixed-initiative interaction: A second line of work studies when robots should request help, clarification, or disambiguation from users [7], [8]. These works show that interaction can improve collaboration and robustness when the robot is uncertain or lacks task-critical information. More directly, Deits et al. formulate clarification as an information-theoretic dialogue problem, showing how a robot can ask targeted questions to reduce uncertainty in ambiguous commands [9]. Related work on situated dialogue also emphasizes the need to detect and recover from miscommunication during task-oriented interaction [10]. Ask-to-Act is aligned with this perspective, but differs in two ways. First, we treat *non-interaction* as a valid and often preferred outcome, rather than assuming that asking is desirable whenever uncertainty is present. Second, we make the clarification decision measurable through task-oriented quantities such as grounded accuracy, question rate, and answerability under the model's candidate set.

c) Interactive embodied agents and selective questioning: Recent work has renewed interest in agents that ask questions only when needed. In embodied settings, dialogue-enabled benchmarks such as DialFRED and TEACh show that question asking can be integrated directly into instruction following and task completion [11], [12]. More recent language-agent work studies selective asking under unclear instructions, arguing that capable systems should gather additional information when needed instead of always acting or always asking [13], [14]. Our setting differs because clarification is tied directly to grounded robot action selection. A clarification is useful only if it changes the robot's action distribution in a way that improves downstream execution. This grounding constraint makes the problem more concrete and links the value of questioning directly to robot performance.

d) Uncertainty and reliability: Our work also connects to broader research on uncertainty-aware behavior and reliability in language-enabled systems. Prior work on uncertainty-resolving questions for robots shows that questioning can serve as an explicit mechanism for reducing action uncertainty before execution [15]. Ask-to-Act follows a similar intuition, but operationalizes it as a decision policy over acting versus asking. Rather than using confidence only as a diagnostic signal, we use it to decide whether the robot should request additional evidence before committing to an action. This framing turns uncertainty into an information-gathering decision that can be evaluated directly through downstream task performance.

III. SCOUT++ Benchmark

SCOUT++ is a grounded benchmark derived from the original SCOUT (Situated Corpus of Understanding Trans-

actions) dataset [5], a Wizard-of-Oz multimodal human–robot dialogue corpus collected in a collaborative exploration setting. From this source corpus, we construct SCOUT++ examples by aligning each instance with (i) a commander utterance, (ii) a canonical grounded robot response label, and (iii) a time-synchronized image captured at the moment the command is issued. This reformulation turns the broader SCOUT dialogue resource into an instance-level benchmark for grounded decision making. It is particularly suitable for Ask-to-Act because it supports both text-only and multimodal versions of the same grounded, ambiguity-sensitive prediction problem. SCOUT++ and its initial grounded instruction understanding benchmark were introduced in our prior work [6].

In the configuration used here, the benchmark contains 12,000 command-label-image rows. After filtering for complete commander instruction, grounded robot response label, and file-name fields, the final experimental set contains 11,989 examples spanning 382 unique grounded action labels. The label space is long-tailed, reflecting realistic variation in how humans instruct a robot and in the resulting grounded action outcomes. To reduce leakage from repeated or closely related file references, we split by grouping on `file_name`, assigning all examples from a group to the same partition using an 80/10/10 train/validation/test split. This yields 9,591 training examples, 1,198 validation examples, and 1,200 test examples.

The benchmark is useful for two reasons. First, it supports a text-only setting in which clarification should plausibly help because language bears most of the burden of disambiguation. Second, it supports a multimodal setting in which command-time visual context may resolve part of that ambiguity before interaction becomes necessary. This makes SCOUT++ a natural testbed for studying whether clarification remains valuable as grounding improves, and for evaluating when asking a question is likely to provide enough additional information to justify delaying action.

IV. Method

Let x denote the instruction and available context, and let $y \in \mathcal{Y}$ denote a grounded action label. A base model produces a distribution $p_\theta(y \mid x)$ and predicts $\hat{y} = \arg\max_y p_\theta(y \mid x)$. Ask-to-Act adds a binary decision: *act* now, or *ask* one clarification question, receive an answer, update the input, and then act.

A. Cost-aware ask policy

Our main uncertainty signal is the probability margin between the top two actions,

$$m(x) = p_\theta(\hat{y}_1 \mid x) - p_\theta(\hat{y}_2 \mid x).$$

A threshold policy asks when $m(x) < \tau$. We choose τ on the validation split by maximizing

$$J(\tau) = \mathrm{Acc}(\tau) - \lambda\, \mathrm{Q/E}(\tau),$$

where Q/E is questions per episode and λ sets the interaction penalty. Unless otherwise stated, we use $\lambda = 0.02$. We sweep τ from 0.0 to 0.9 in increments of 0.05, select the threshold that maximizes $J(\tau)$ on validation, and report the selected operating point on the held-out grouped test split. This formulation makes explicit that clarification is neither free nor always desirable: it is valuable only when the expected gain in grounded accuracy outweighs the interruption cost.

The threshold policies use raw softmax probabilities from the trained classifier. We do not apply post-hoc temperature scaling or conformal calibration in the reported experiments; calibration-aware ask policies remain an important direction for future work.

We also evaluate a lightweight learned ask policy using logistic regression over uncertainty features including margin, entropy, and instruction length. This provides a simple alternative to a hand-tuned threshold while preserving interpretability.

B. Ask-beneficial ambiguity

We define an episode as *ask-beneficial ambiguous* at top-k if the gold action is in the model's top-k candidates but not ranked first:

$$\mathbb{1}[\mathrm{ambig}_k(x)] = \mathbb{1}[y^* \in \mathrm{TopK}(p_\theta(\cdot \mid x), k)] \wedge \mathbb{1}[y^* \neq \arg\max_y p_\theta(y \mid x)].$$

This matches the intended use of clarification. Asking is useful when the correct action is already plausible, but not currently selected. In contrast, asking is likely to be wasted when the gold action is not present in the model's candidate set at all.

C. Clarification mechanisms

We evaluate two mechanisms. The first is *structured candidate pruning*. The system asks a targeted question corresponding to a dominant confusion type, such as direction, angle, distance, traversal mode, image request, or confirmation. Question types are inferred by a deterministic rule-based procedure over the leading candidate confusion pair. We first mine frequent validation-set confusion pairs from the TF-IDF/logistic-regression baseline and map each pair to a dominant confusion type using surface cues in the label strings. At test time, if a candidate pair is not present in the saved map, the same regex-based fallback rules are applied. The system then receives a deterministic oracle answer consistent with the gold label and selects the highest-ranked candidate compatible with that answer. This mechanism isolates the value of the ask decision from the model's ability to consume follow-up text.

The second is *append-and-rerun*. Here, the clarification answer is appended to the original instruction in a structured form, for example `Clarification: direction = left`, and the grounding model is rerun on the augmented input. This mechanism treats clarification as an additional learned signal rather than a post hoc candidate filter.

To make append-and-rerun effective, we introduce *clarification augmentation* during training. Clarification augmentation is applied only to the training split. For each training

Algorithm 1 Ask-to-Act (single-question budget)

Require: grounding model $p_\theta(y \mid x)$, ask policy π, top-k size k=5

$p \leftarrow p_\theta(\cdot \mid x)$; $y_1 \leftarrow \arg\max_y p(y)$
if $\pi(x, p) = \text{ASK}$ **then**
    choose question type q from leading candidate confusion
    $a \leftarrow \textsc{Oracle}(q, y^*)$
    $x' \leftarrow x \oplus$ ``Clarification: q = a''
    return $\arg\max_y p_\theta(y \mid x')$
else
    return y_1
end if

example, with probability 0.55, we append a structured clarification string derived from task-relevant attributes extracted from the gold action label. The generated strings use slot-value templates such as `Clarification: direction = left`, `Clarification: angle = 90 degrees`, or `Clarification: traversal = through`. The procedure does not append the full action label; it appends only a small attribute value extracted using deterministic regular-expression rules. Validation and test instructions are left unmodified before the ask policy decides whether to append a clarification at evaluation time.

D. Models

Our text-only neural model is a RoBERTa-base sequence classifier over the grounded action label space, with TF-IDF plus logistic regression as a lightweight baseline. We tokenize commander instructions with the RoBERTa tokenizer using a maximum sequence length of 256. The text model is trained with learning rate 1×10^{-5}, batch size 8, evaluation batch size 16, gradient accumulation of 2, weight decay 0.01, warmup ratio 0.06, and 8 training epochs. We use random seed 42 and select the best checkpoint by validation accuracy. Our multimodal model uses late fusion. We encode the command text with RoBERTa-base and concatenate the RoBERTa text representation with a frozen CLIP image embedding from `openai/clip-vit-base-patch32`. The fused representation is passed through dropout 0.1 and a linear classification head. The multimodal model is trained with AdamW using learning rate 2×10^{-5}, batch size 8, evaluation batch size 16, gradient accumulation of 2, and 6 training epochs. We keep the CLIP encoder frozen and optimize the text encoder and fusion head. This design lets us test the effect of command-time visual grounding without introducing additional complexity from end-to-end vision-language finetuning.

V. Experiments

We ask three questions: (1) does selective clarification improve text-only grounded action prediction, (2) how does command-time perception change the ask-vs-act tradeoff, and (3) can simple uncertainty features detect ask-beneficial ambiguity? We report top-1 accuracy, questions per episode, answerable@top5, and wasted@top5. For ambiguity detection we report AUROC and AUPRC.

For append-and-rerun, each training example is independently augmented with probability 0.55 using a structured clarification string derived from task-relevant attributes extracted from the gold action label. Validation and test examples are not pre-augmented. We also train a lightweight logistic regression ask policy using margin, entropy, and instruction length as features. All operating points are selected on the validation split and reported once on the grouped held-out test split. The full experimental setup is summarized in Table I.

TABLE I
Dataset and evaluation setup used for all experiments.

Property	Value
Command-label-image rows	12,000
Usable examples	11,989
Grounded action labels	382
Split strategy	Grouped by file name
Train / validation / test	9,591 / 1,198 / 1,200
Random seed	42
Text model	RoBERTa-base classifier
Multimodal model	RoBERTa-base + frozen CLIP ViT-B/32
Ask budget	At most one clarification query
Interaction penalty λ	0.02
Threshold sweep	$\tau \in [0.0, 0.9]$, step 0.05
Primary metrics	Accuracy, Q/E, answerable@5, wasted@5

a) Evaluation intuition: The key distinction in our evaluation is between episodes where clarification could plausibly recover the correct action and episodes where asking would only delay an inevitable error. Answerable@top5 measures the former, while wasted@top5 measures the latter. This distinction is important for interpreting question rates: a system that asks often is only useful if those questions are concentrated on episodes where the correct action is already within reach.

VI. Results

A. Text-only Ask-to-Act improves grounded decision quality

The text-only RoBERTa baseline reaches 0.8225 accuracy as an always-act model. Under a single-question budget, Ask-to-Act improves this tradeoff. The best append-and-rerun operating point reaches 0.8308 accuracy while asking in 0.2350 of episodes, with answerable@top5 = 0.6773 and wasted@top5 = 0.3227. The selected text-only operating point uses $\tau = 0.75$ under $\lambda = 0.02$ and is summarized in Table II. These results show that a subset of errors are recoverable: the correct action is often already among the model's plausible candidates and can be recovered with a small amount of additional grounding information.

Structured pruning is useful as an analysis tool because it isolates the value of the ask decision from the model's ability

TABLE II

ASK-TO-ACT PERFORMANCE UNDER THE SELECTED OPERATING POINTS.

Method	Acc.	Q/E	Ans.@5	Waste@5
Text always act	0.8225	0.0000	–	–
Text Ask-to-Act	0.8308	0.2350	0.6773	0.3227
Multimodal always act	0.8503	0.0000	–	–
Multimodal Ask-to-Act	0.8545	0.1277	0.7020	0.2980

to consume appended clarification text. In contrast, append-and-rerun becomes effective only when the text model is trained with clarification augmentation; otherwise, the model tends to ignore short follow-up strings and the optimal policy shifts toward asking less. This supports the broader argument that clarification should be modeled as part of the learning setup rather than treated as a purely symbolic post-processing step.

B. Multimodal grounding reduces the amount of clarification needed

As shown in Fig. 2, multimodal grounding shifts the operating region toward higher accuracy with fewer questions. In the multimodal setting, late fusion reaches 0.8503 accuracy as an always-act policy. Under the same cost-sensitive model selection procedure, the best multimodal Ask-to-Act operating point reaches 0.8545 accuracy while asking in 0.1277 of episodes, using $\tau = 0.50$ under $\lambda = 0.02$. This improves over the multimodal always-act baseline while asking less often than the text-only Ask-to-Act policy.

This result shows that Ask-to-Act remains useful across grounding regimes while adapting its interaction rate to the available evidence. In the text-only setting, the policy asks more often because language alone leaves more action choices underdetermined. With command-time visual grounding, the base model becomes stronger, and Ask-to-Act selects a lower question rate while still improving accuracy over always acting. The key result is therefore not that clarification is always needed, but that the ask policy can adjust when additional interaction is worth the cost.

C. Uncertainty detects clarification-worthy episodes

TABLE III

AMBIGUITY DETECTION QUALITY USING SIMPLE UNCERTAINTY SIGNALS.

Setting	Ambig.@5	AUROC	AUPRC
Text-only	0.0783	0.8374	0.2332
Multimodal	0.0888	0.8718	0.3103

Using ask-beneficial ambiguity as the target, we evaluate uncertainty-based detectors. Probability margin remains the strongest simple signal in both settings. The resulting ambiguity detection performance is shown in Table III. In the text-only setting, the model reaches Ambig.@5 = 0.0783, with margin-based AUROC = 0.8374 and AUPRC = 0.2332. In the multimodal setting, Ambig.@5 = 0.0888, with AUROC = 0.8718 and AUPRC = 0.3103. Representative values are summarized in Table IV.

TABLE IV

MAIN QUANTITATIVE SUMMARY.

Setting	Acc.	Q/E	Ambig.@5	AUROC
Text-only always act	0.8225	0.0000	0.0783	0.8374
Text-only Ask-to-Act	0.8308	0.2350	–	–
Multimodal always act	0.8503	0.0000	0.0888	0.8718
Multimodal Ask-to-Act	0.8545	0.1277	–	–

These results show that uncertainty signals remain informative even when visual grounding is available. The multimodal model improves overall action accuracy and supports a lower selected question rate, but ask-beneficial ambiguity is not eliminated. This reinforces the role of Ask-to-Act as a selective decision layer: the policy does not assume clarification is always useful, but uses uncertainty to identify cases where an additional grounding cue may still improve the final action.

VII. DISCUSSION

Ask-to-Act reframes clarification as a decision layer inside grounded robot learning. Rather than treating clarification as always helpful or always costly, the policy evaluates whether an additional grounding cue is worth the interaction cost for a given episode. In the text-only setting, selective clarification improves execution reliability because language alone often leaves multiple grounded actions plausible. In the multimodal setting, command-time visual context strengthens the base action predictor and reduces the selected question rate, but Ask-to-Act still provides an additional improvement over always acting.

This pattern supports the main claim of the paper: clarification should be adaptive to the robot's current evidence. A weaker grounding regime benefits from asking more often, while a stronger multimodal regime can act more often without losing the ability to ask when uncertainty remains. The value of Ask-to-Act is therefore not only in increasing accuracy, but in exposing the tradeoff between task performance and interaction burden.

From a robot learning perspective, this matters because the decision to gather more information is itself part of the control problem. A robot should not ask merely because an instruction is linguistically underspecified; it should ask when the current belief over grounded actions suggests that clarification is likely to change the final decision. The ask-beneficial ambiguity formulation provides one way to measure this condition offline by identifying cases where the gold action is plausible but not top-ranked.

These results also show that perceptual grounding and clarification are complementary. Visual context resolves many action-selection errors directly, while selective clarification remains useful for cases where the available evidence is still insufficient. This suggests that future grounded instruction-following systems should integrate clarification policies with

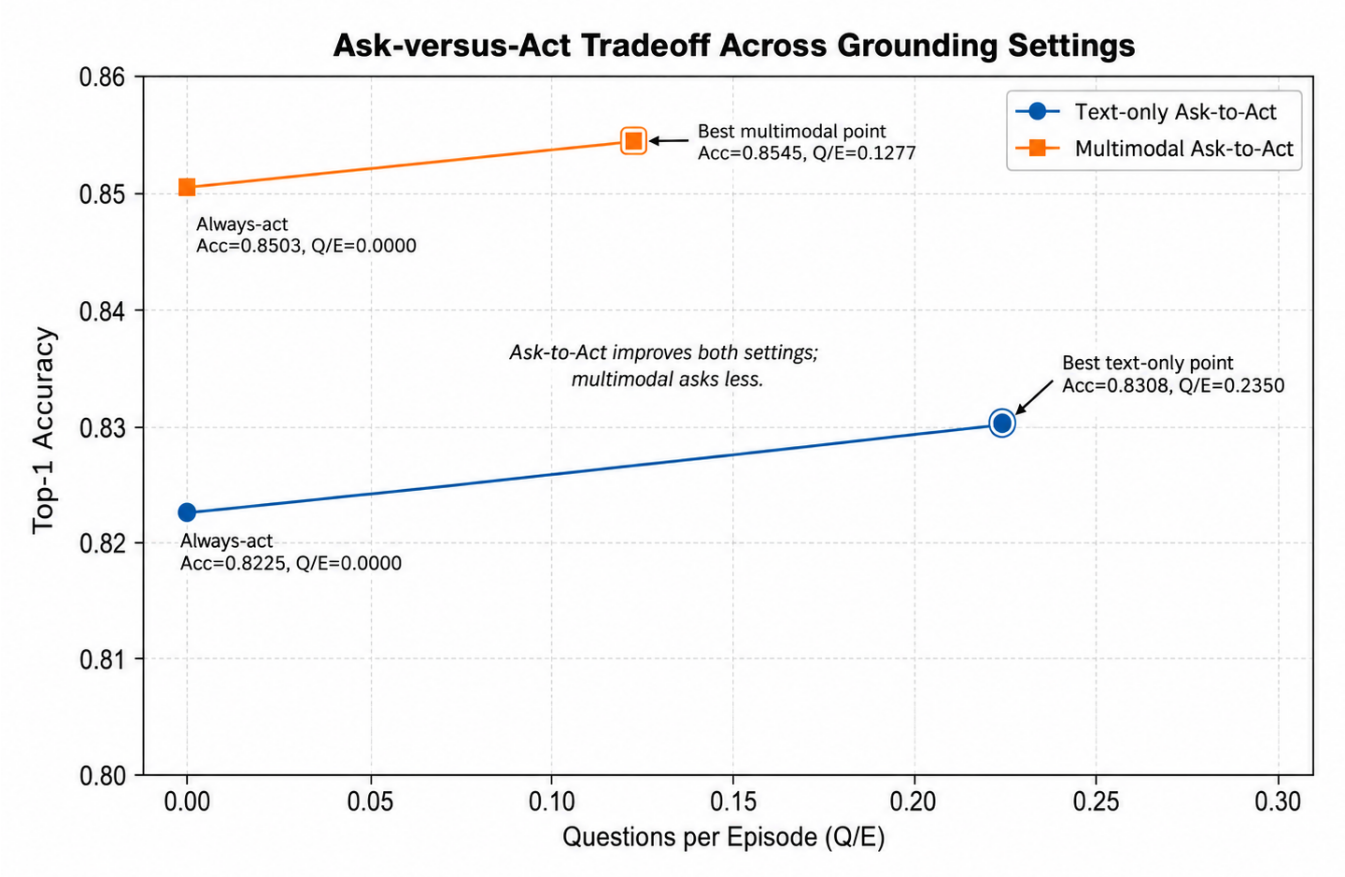

Fig. 2. Accuracy versus questions per episode for text-only and multimodal Ask-to-Act policies. Multimodal grounding improves the always-act baseline and shifts the selected Ask-to-Act operating point toward fewer questions, while selective clarification still provides an additional gain in both grounding settings.

the robot's sensing capabilities rather than treating dialogue as a separate module.

VIII. Limitations

Our evaluation uses a deterministic clarification oracle. This design makes the ask-vs-act tradeoff reproducible and isolates whether the policy asks in cases where an additional grounding cue can recover the correct action. However, it abstracts away noisy, incomplete, delayed, or differently phrased human responses. The reported clarification gains should therefore be interpreted as offline estimates under idealized user responses, not as deployed human-in-the-loop performance. Evaluating Ask-to-Act with human users or realistic dialogue models is an important next step.

Our experiments are also offline benchmark evaluations. We evaluate grounded action prediction and clarification decisions, but do not execute actions on a physical robot or measure task completion, recovery behavior, response latency, or user experience in an embodied interaction. The results support the value of selective clarification for grounded action selection, but physical robot execution remains necessary before making stronger HRI deployment claims.

The benchmark label space is long-tailed, with many infrequent grounded action labels. Aggregate top-1 accuracy may therefore obscure performance differences on rare but important actions. Future work should evaluate rare-label behavior more directly and study whether clarification provides larger benefits for low-frequency actions.

Finally, the ask policies use raw softmax probabilities rather than calibrated probabilities, and the reported neural results use a fixed random seed. Additional calibration methods, conformal or abstention-style baselines, noisy-oracle sensitivity analyses, and multiple-seed evaluations would further strengthen the robustness claims.

IX. Conclusions

We presented Ask-to-Act, a framework for learning when a robot should ask for clarification before acting on an ambiguous grounded instruction. Ask-to-Act treats clarification as a cost-sensitive decision under uncertainty rather than as a fixed dialogue behavior. On SCOUT++, selective clarification improves grounded action prediction in both text-only and multimodal settings. In the text-only setting, Ask-to-Act improves accuracy from 0.8225 to 0.8308 while asking in 0.2350 of episodes. In the multimodal setting, command-time visual grounding improves the always-act baseline to 0.8503, and Ask-to-Act further improves accuracy to 0.8545 while reducing the selected question rate to 0.1277.

These results show that clarification and perceptual grounding are complementary. Stronger visual grounding improves immediate action selection and reduces the need for interaction, while selective clarification remains useful when the available evidence is still insufficient. More broadly, Ask-to-Act positions clarification as an adaptive component of grounded robot learning: robots should ask not whenever language is ambiguous, but when asking is likely to improve the final grounded action decision enough to justify the interaction cost.

ACKNOWLEDGMENT

The authors used LLMs for limited editing & some literature search. All content was reviewed by the authors.

REFERENCES

[1] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 1507–1514, Aug. 2011. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7979

[2] S. R. Branavan, H. Chen, L. Zettlemoyer, and R. Barzilay, "Reinforcement learning for mapping instructions to actions," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 82–90.

[3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 740–10 749.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[5] S. M. Lukin, C. Bonial, M. Marge, T. A. Hudson, C. J. Hayes, K. Pollard, A. Baker, A. N. Foots, R. Artstein, F. Gervits, M. Abrams, C. Henry, L. Donatelli, A. Leuski, S. G. Hill, D. Traum, and C. Voss, "SCOUT: A situated and multi-modal human-robot dialogue corpus," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 14 445–14 458. [Online]. Available: https://aclanthology.org/2024.lrec-main.1259/

[6] E. Ogbadu, S. Lukin, and C. Matuszek, "Grounded instruction understanding with large language models: Toward trustworthy human-robot interaction," *Proceedings of the AAAI Symposium Series*, vol. 7, no. 1, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI-SS/article/view/36890

[7] C. D. Hayes and B. Scassellati, "Challenges in shared-environment human-robot collaboration," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013.

[8] R. A. Knepper, S. Tellex, A. Li, N. Roy, and D. Rus, "Single assembly robot in search of human partner: Versatile grounded language generation," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013, pp. 167–168.

[9] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy, "Clarifying commands with information-theoretic human-robot dialog." *J. Hum. Robot Interact.*, vol. 2, no. 2, pp. 58–79, 2013.

[10] M. Marge and A. I. Rudnicky, "Miscommunication detection and recovery in situated human–robot dialogue," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 1, Feb. 2019. [Online]. Available: https://doi.org/10.1145/3237189

[11] X. Gao, Q. Gao, R. Gong, K. Lin, G. Thattai, and G. S. Sukhatme, "Dialfred: Dialogue-enabled agents for embodied instruction following," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 049–10 056, 2022.

[12] 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," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2017–2025, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20097

[13] W. Wang, S. Juluan, Z. Ling, Y.-K. Chan, C. Wang, C. Lee, Y. Yuan, J.-t. Huang, W. Jiao, and M. R. Lyu, "Learning to ask: When LLM agents meet unclear instruction," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 21 773–21 784. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1104/

[14] M. J. Zhang and E. Choi, "Clarify when necessary: Resolving ambiguity through interaction with LMs," in *Findings of the Association for Computational Linguistics: NAACL 2025*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 5541–5558. [Online]. Available: https://aclanthology.org/2025.findings-naacl.306/

[15] M. Shin, M. Jang, M. Cho, and J.-K. Ryu, "Uncertainty-resolving questions for social robots," in *Companion Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI Companion '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 226–230. [Online]. Available: https://doi.org/10.1145/3568294.3580077