QuickLAP: Quick Language-Action Preference Learning for Semi-Autonomous Agents

Jordan Abi Nader, David Lee, Nathaniel Dennler, Andreea Bobu
{jordancn, dlee888, dennler, abobu}@mit.edu
MIT
United States of America

***Abstract*—Robots must learn from both what people do and what they say, but either modality alone is often incomplete: physical corrections are grounded but ambiguous in intent, while language expresses high-level goals but lacks physical grounding. We introduce QuickLAP: *Quick Language–Action Preference learning*, a Bayesian framework that fuses physical and language feedback to infer reward functions in real time. Our key insight is to treat language as a probabilistic observation over the user's latent preferences, clarifying which reward features matter and how physical corrections should be interpreted. QuickLAP uses Language Models (LMs) to extract reward feature attention masks and preference shifts from free-form utterances that are combined with physical feedback in a closed-form update rule. This enables fast, real-time, and robust reward learning that handles ambiguous feedback. In a robotic manipulation and a semi-autonomous driving simulator, QuickLAP reduces reward learning error by over 70% compared to physical-only and heuristic multimodal baselines. User studies further validate our approach: participants found QuickLAP significantly more understandable and collaborative, and preferred its learned behavior over baselines.**

I. Introduction

Humans reveal their preferences through both what they do and what they say. Consider the semi-autonomous vehicle approaching a construction zone in Fig. 1: the user could physically nudge the steering wheel away, or say "Stay away from the cones!". Each of these signals is only part of the story: physical corrections offer precise, reactive adjustments but are opaque in intent, while language clarifies *what* matters to the user but lacks the physical grounding needed to guide behavior on its own. To adapt effectively, robots must learn to interpret both together, in context and in real time.

Inverse Reinforcement Learning (IRL) methods that learn from physical corrections can adapt quickly and infer a reward function over task-relevant features (e.g. distance to obstacles, etc.) that reflects the user's underlying preferences [4]. But while these corrections are grounded and precise, they are inherently ambiguous: the same steering adjustment in Fig. 1 could reflect a desire to avoid construction zones, to change lanes, or to stay close to puddles. Natural language can help resolve this ambiguity. An utterance like "Stay far from the cones!" makes the user's intent explicit, revealing preferences that are hard to infer from motion alone. But language has its own limitations: it can be vague ("Stay away!"), underspecified, or context-dependent. On its own, it may fail to convey exactly how the robot should act. Crucially, we observe that these two feedback channels are complementary: language clarifies the intent behind a correction, and the correction grounds and disambiguates the language. Together, they offer a more complete signal than either can provide alone.

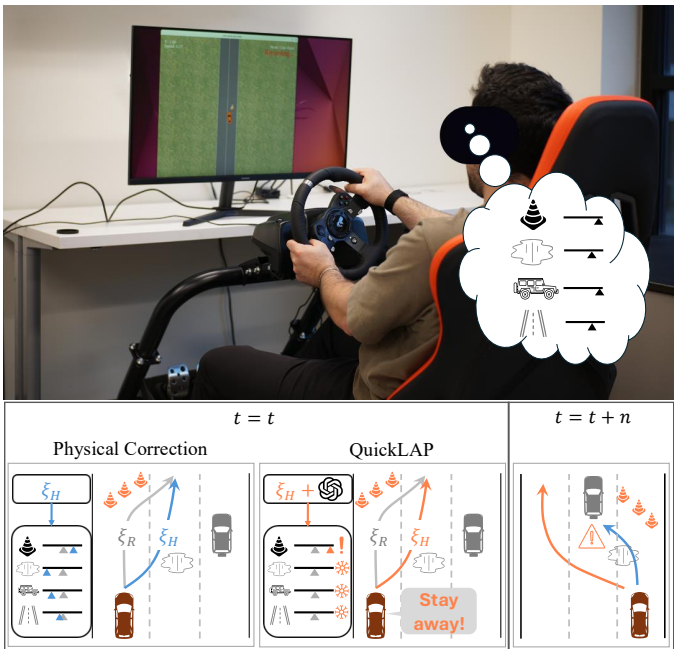


Fig. 1. (Top) User Study Setup. Participants controlled the virtual car using a gaming steering wheel. (Bottom) QuickLAP fuses physical corrections and natural language for improved reward learning. (Bottom-Left) A human's physical correction (blue) leads to an update that incorrectly changes the reward on multiple correlated features. (Bottom-Middle) QuickLAP combines this physical signal with concurrent language input, using an LLM to clarify intent and produce a more accurate reward update prioritizing cone avoidance. (Bottom-Right) In future scenarios, QuickLAP generates safer trajectories (orange) than the baseline (blue).

Yet few existing methods take full advantage of this complementarity. Some multimodal robot learning methods rely on large datasets of paired trajectories and language [37, 38, 55, 57, 11], limiting online adaptation. Others assume language inputs are precise, unambiguous, and self-contained [15, 31], rather than interpreting them in conjunction with behavior. As a result, these methods fall short in exactly the settings where joint feedback should be most useful: ambiguous, fast-changing, and underspecified environments.

Our key insight is that language can serve as a probabilistic

observation over the user's latent reward, not just *describing* what matters, but *modulating* how physical corrections should be interpreted. Language plays two roles: it identifies which features are relevant to the user's intent, and it suggests how those preferences should shift. This perspective enables a principled fusion of modalities that resolves ambiguity and supports rapid online learning.

We introduce **QuickLAP**: *Quick Language–Action Preference learning*, a closed-form Bayesian framework for online reward inference from joint physical and language feedback. For each intervention, a Language Model (LM) processes the user's utterance in context to produce a feature attention mask (what matters), a proposed reward shift (how behavior should change), and confidence weights (how certain the LM is). These signals are fused with a Boltzmann-rational model of the correction and an attention-weighted prior to produce a Gaussian posterior over reward parameters. This yields an efficient MAP update rule that extends IRL to multimodal feedback, robustly handles ambiguity, and supports real-time adaptation. Importantly, QuickLAP is robust to ambiguity in either channel: vague utterances are interpreted in the context of physical corrections, and unclear corrections are disambiguated by language. Rather than assuming clean or isolated inputs, QuickLAP exploits the natural interplay between language and action to infer what the user wants efficiently and in real time.

To summarize, our main contributions are: 1) an efficient Bayesian framework for jointly interpreting physical corrections and natural language feedback in real time; 2) a LM-based semantic parser procedure that maps free-form language to structured reward signals without task-specific training; 3) extensive simulation results showing large reductions in reward-inference error; 4) two user studies demonstrating improved understanding, collaboration, and user-preferred behaviors for QuickLAP over baselines. Beyond driving and robot manipulation, QuickLAP offers a general framework for preference learning in any domain where users can show and tell, from assistive robots to collaborative drones, enabling faster, more intuitive, and more personalized human-robot interaction.

II. Related Work

Learning from Physical Human Feedback. Early work in IRL focused on inferring reward functions from demonstrations, framing the problem as a linear program [44], or incorporating max-margin [2], maximum-entropy [67, 8], and Bayesian [50] principles to address ambiguity over optimal behavior. Although these methods traditionally assumed a fixed batch of demonstrations, later work recognized the value of *interactive* feedback provided during task execution [4, 28, 68, 19]. Bajcsy et al. [4] proposed learning from *physical corrections*, kinesthetic nudges that convey gradient-like information about user preferences in real time. Follow-up work extended these ideas to handle both offline and online feedback modalities through a unified framework [41], while Jain et al. [28] explored coactive learning from incremental physical corrections in manipulation tasks. While physical feedback offers grounded and precise input, it often affects multiple features simultaneously, making it difficult to infer the user's true intent [5, 9, 25]. Our method tackles this shortcoming by exploring richer feedback modalities that can clarify user intent, such as natural language.

Language in Robot Learning. Natural language is an expressive interface for communicating goals and preferences to robots. While early systems mapped language to symbolic actions [62], recent methods use LLMs to interpret and guide behavior [27, 26, 57]. However, most approaches assume access to large paired datasets of trajectories and utterances [37, 38, 55, 11, 48, 47, 56, 64] or rely on offline reward supervision [45]. Others adapt to language at test time [15, 31], but treat instructions as clean, self-contained inputs, ignoring physical context. These approaches typically require extensive offline training on paired datasets, and rely on pre-training with unambiguous instructions [15, 64]. Fusion strategies are often heuristic, with no probabilistic treatment of uncertainty or signal agreement. In contrast, we treat language as a probabilistic observation over the user's latent reward. Our method fuses it with physical feedback in real time, yielding a closed-form Bayesian update that supports ambiguity resolution, generalization, and efficient online learning.

Online Reward Learning from Human Input. QuickLAP builds on prior work for online reward adaptation during task execution. Interactive methods like TAMER [32] and COACH [39] showed how scalar feedback can shape behavior incrementally. Jain et al. [28] extended this to richer corrections in manipulation tasks, where humans suggest incremental trajectory improvements that are immediately incorporated. Related efforts in active preference learning [7, 54, 18] focus on generating informative queries to improve reward inference, though typically in offline or batch settings. More recent approaches incorporate language into the learning loop. Some require extensive pre-collected feedback [33, 15], while others use language for reward shaping [13, 59] but treat it as a standalone modality, decoupled from physical input. While recent work explores LLM-driven mode switching for assistive tasks [15, 61], it often assumes a discrete set of behaviors rather than operating in continuous control domains. QuickLAP enables online adaptation in continuous domains by interpreting physical and linguistic feedback jointly, within a unified Bayesian framework. This allows reward inference from sparse, ambiguous inputs without large datasets or perfect instructions.

III. Problem Formulation

We consider a robot that adapts to a human operator's preferences and plans under deterministic dynamics $x^{t+1} = f(x^t, u^t)$, where x denotes the state (e.g., position, velocity) and u the control input (e.g., acceleration, steering). To align its behavior with the human preference, the robot optimizes a linear reward function $R_\theta = \theta^\top \Phi(\xi)$, where $\Phi(\xi) = \sum_{j=0}^{T-1} \phi(x^j, u^j)$ aggregates features $\phi(x^j, u^j) \in \mathbb{R}^d$ over a trajectory $\xi = (x^0, u^0, \ldots, x^T)$. These features encode relevant aspects such as proximity to obstacles, collision risk, and speed. The parameter $\theta \in \mathbb{R}^d$ indicates the relative importance of these features. We assume the human's true preference weights $\theta^\star$ are *unknown* to the robot.

The robot plans using Model Predictive Control (MPC) [23]. At each timestep t the robot generates a trajectory ξ_R^t by solving $\xi_R^t = \arg\max_\xi \hat{\theta}^{t\top}\Phi(\xi)$ using its current preference estimate $\hat{\theta}^t$. A human may intervene with a physical correction u_H^t (e.g., a nudge) and/or a natural-language utterance l^t, producing a corrected trajectory ξ_H^t (e.g., via trajectory deformation [4, 21]. The robot's task is to infer $\theta^\star$ online from ξ_H^t and l^t.

Learning from Physical Corrections. When the human provides a physical correction, we follow prior work [4] and model the likelihood of observing their induced trajectory ξ_H given the robot's current trajectory ξ_R and hidden preference θ using a Boltzmann noisily-rational model [6, 67]:

$$P(\xi_H \mid \xi_R, \theta) \approx \exp\Big(\theta^\top\big(\Phi(\xi_H) - \Phi(\xi_R)\big) - \lambda\|\xi_H - \xi_R\|^2\Big). \tag{1}$$

Here, the term $\theta^\top(\Phi(\xi_H) - \Phi(\xi_R))$ represents the relative improvement induced by the human's correction over the robot's current trajectory, while the quadratic term $\lambda\|\xi_H - \xi_R\|^2$ models their reluctance to make large corrections (effort minimization), with $\lambda > 0$ controlling this trade-off.

In the physical-only setting, this likelihood is combined with a Gaussian prior to yield a MAP update that adjusts the reward parameters in the direction of the feature difference $\Delta\Phi = \Phi(\xi_H) - \Phi(\xi_R)$, following [4].

$$\hat{\theta}^{t+1} = \hat{\theta}^t + \Sigma_{prior}\Delta\Phi \ . \tag{2}$$

This update adjusts the preference weights $\hat{\theta}^t$ in the direction of the observed feature change $\Delta\Phi$, scaled by the prior covariance Σ_{prior}, reflecting learning from the physical intervention.

Unfortunately, relying solely on physical feedback has three critical limitations. First, physical input is often noisy [36]. Second, it can exhibit feature coupling, where an intended change to one feature inadvertently alters others – for example, a correction to reduce speed might also unintentionally alter the vehicle's lateral position. Third, physical actions lack semantic expressivity: they cannot easily disambiguate the underlying intent (e.g., was a slowdown near a construction zone due to general safety, concern for workers, or adherence to a perceived speed limit?) nor convey more abstract preferences related to comfort or style. These shortcomings create an ambiguity gap between what humans intend and what robots interpret. QuickLAP addresses this gap by fusing physical corrections with natural language explanations, creating a multimodal framework that captures both the precise *what* of physical corrections and the contextual *why* of verbal explanations.

IV. Quick Language-Action Preference Learning

QuickLAP is a Bayesian framework that jointly interprets physical and language input to infer user preferences θ in real time. Our approach extends established Bayesian treatments of physical interaction [4] to the multimodal domain through a LM-mediated language likelihood, achieving both computational tractability and enhanced expressivity. The core principle of QuickLAP leverages the complementary nature of these feedback types: physical actions demonstrate *what* the user wants through trajectory adjustments, while language clarifies *why* by revealing the intent behind those corrections. This interpretation allows QuickLAP to achieve preference learning that is both responsive to immediate corrections and grounded in understandable user rationales.

Formally, our goal is to compute the posterior distribution over preference parameters $P(\theta \mid \xi_H, \xi_R, l)$ given the robot's proposed trajectory ξ_R, the human's corrective trajectory ξ_H, and natural language l. Using Bayes' rule, this posterior becomes $P(\theta \mid \xi_H, \xi_R, l) \propto P(\xi_H, l \mid \xi_R, \theta)P(\theta)$. We factorize the joint likelihood into:

$$P(\theta \mid \xi_H, \xi_R, l) \propto \underbrace{P(\xi_H \mid \xi_R, \theta)}_{\text{Physical Likelihood}} \cdot \underbrace{P(l \mid \xi_H, \xi_R, \theta)}_{\text{Language Likelihood}} \cdot \underbrace{P(\theta)}_{\text{Prior}} \ . \tag{3}$$

The physical likelihood is given by the Boltzmann model in Eq. (1). We next define the language likelihood and prior models, then derive our closed-form MAP update rule for θ.

A. Learning from Language Input

To incorporate language input l, we interpret it as a structured signal about preference changes and define a likelihood function based on this interpretation. In our implementation, this interpretation is produced by a dual-language-model (LM) framework that is motivated by our Bayesian framework (described in Appendix A). However, the learning formulation itself is agnostic to the specific language model.

1) Language Processing with LLMs: We decompose language interpretation into two components: (1) identifying which features the utterance attends to, and (2) determining how those features should change.

Attention Language Model ($\mathrm{LM}_{\mathrm{att}}$). We use a language model to identify which features are relevant to the human's utterance, producing an attention mask $r \in \{0,1\}^d$. Our insight is that when humans give corrections, they typically focus on specific behavioral aspects rather than addressing all simultaneously [5]. For instance, "slow down near schools" addresses speed-related features while implicitly accepting current lane position and route choices. Given language correction l and interaction context $c = \big(\Delta\Phi, \theta^t, \textit{environment\_description}\big)$ capturing the induced feature difference, the robot's current reward estimate, and a description of the environment, we generate $r = \mathrm{LM}_{\mathrm{att}}(l, c)$. Each component r_i indicates whether feature ϕ_i is relevant to the current feedback. This attention mechanism will be crucial for modulating our prior (Sec. IV-B).

Preference Language Model ($\mathrm{LM}_{\mathrm{pref}}$). The second LM determines how attended features should change. Given the utterance l, and context c, we generate $(m, \mu) = \mathrm{LM}_{\mathrm{pref}}(l, c, r)$, where $m \in [0,1]^d$ is a confidence vector indicating how precisely the language specifies an update for each feature, and $\mu \in \mathbb{R}^d$ is a shift vector representing the update's magnitude and direction. Importantly, m (and μ) are distinct from r. The attention mask r identifies *which features to attend to* for a potential update, influencing the prior. The confidence m and shift μ determine *how much and in what direction* the language suggests θ should change for those attended features, influencing the likelihood.

For example, when an ambiguous utterance such as "Stay away!" is applied to a correction involving both a puddle and a cone, the attention mask may mark both features as relevant ($r_{\text{car}} = r_{\text{cone}} = 1$), but low specificity leads to a moderately confident m and a conservative shift μ. In contrast, a more specific utterance like "Stay away from cones" yields focused attention ($r_{\text{cone}} = 1$) with high confidence and a strong shift, enabling a decisive update aligned with the user's intent.

In this work, we instantiate both components using an off-the-shelf large language model (LLM). However, our framework does not depend on this choice: future work could replace or fine-tune these models while preserving the same probabilistic interface for integrating language with physical feedback.

2) Language Likelihood: Given the structured outputs of our dual-LLM framework, we now define a likelihood model for language in Eq. (3). Modeling $P(l \mid \xi_H, \xi_R, \theta)$ directly would require specifying how humans generate free-form utterances conditioned on trajectories and hidden preferences, which is not feasible. Instead, we introduce a latent *proxy variable* $\mu^t = \theta - \theta^t$, representing the reward shift that the utterance is intended to convey.

Intuitively, the human observes the robot's current trajectory ξ_R, which reflects the reward weights θ^t that the robot optimized. They compare this behavior to their own preferences θ, and select an utterance l that, if understood, would shift the robot's parameters from θ^t towards θ. In this sense, the utterance implies an average desired change $\mu = \theta - \theta^t$. Our LM_{pref} acts as an estimator, producing μ^t and confidence m^t describing this latent shift. We treat μ^t as a noisy observation of μ, which gives a Gaussian likelihood:

$$P(\mu^t \mid \theta, \theta^t) = \mathcal{N}(\mu^t; \theta - \theta^t, \text{diag}(\sigma_L^t(m^t))) \ , \qquad (4)$$

where σ_L modulates the standard deviation as a function of the confidence scores m_i^t. Intuitively, if confidence is high, we trust the LM-generated shift μ_i^t is close to the true corrective shift $\theta_i - \theta_i^t$, so the variance should be small, anchoring the distribution tightly to the mean. Conversely, if confidence is low, the variance should be large, reducing the influence of language in favor of relying on the physical correction. We model this by choosing $\sigma_L : [0,1] \to \mathbb{R}^+$ such that $\sigma_L(m) \to 0$ for $m \approx 1$, and $\sigma_L(m) \to 1/\epsilon_{var}$ for $m \approx 0$. One choice for σ_L^2 is, for instance, $\sigma_L(m) = k \cdot \frac{(1-m)}{\epsilon_{var}+m}$, where ϵ_{var} is a small constant for numerical stability. The diagonal structure reflects the assumption that language often addresses features independently (e.g., 'slow down and stay centered' addresses speed and lane position separately), which allows us to factorize the likelihood across dimensions: $P(\mu^t \mid \theta, \theta^t) = \prod_{i=1}^d P(\mu_i^t \mid \theta_i, \theta_i^t)$.

The likelihood in Eq. (4) provides the language component of the posterior in Eq. (3). Rather than modeling the raw utterance distribution $P(l \mid \xi_H, \xi_R, \theta)$ directly, we capture its effect through the proxy shift with tractable likelihood $P(\mu^t \mid \theta, \theta^t)$, which anchors free-form utterances in reward space and allows them to be combined with physical feedback in a unified Bayesian update.

B. Conditional Prior

One way to do multimodal fusion could be to gate the likelihood or update rule based on language. We propose a more principled approach: incorporating the attention mask r directly into the prior over θ, based on the intuition that humans selectively focus on relevant features *before* intervening.

Specifically, we interpret r as indicating which components of θ are likely to change in the current interaction. Before receiving feedback, we hold the marginal prior $\int P(\theta \mid r)p(r)\,dr$; once we parse the utterance and infer $\hat{r}$, we collapse this mixture to a conditional prior $P(\theta \mid \hat{r})$. Our formulation parallels Meta-IRL [63], where a latent variable selects a task-specific Gaussian prior over reward weights before a MAP update. Similarly, our mask $\hat{r}$ modulates only the prior, avoiding double-counting in the likelihood while enabling fast, structured adaptation. Formally, once LM_{att} provides $\hat{r}^t$ from utterance l^t, we use a conditional Gaussian prior:

$$P(\theta \mid \hat{r}^t) = \prod_{i=1}^d P(\theta_i \mid \hat{r}_i^t) = \prod_{i=1}^d \mathcal{N}(\theta_i; \theta_i^t, \sigma^2_{prior}(\hat{r}_i^t)), \quad (5)$$

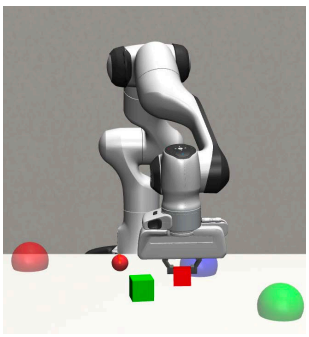
where θ_i^t is the current estimate for that feature's weight. The prior variance $\sigma^2_{prior}(\hat{r}_i^t)$ (or its inverse, precision $\Lambda_{prior,i}(\hat{r}_i^t)$) depends on the attention mask $\hat{r}_i^t$. We define the precision as $\Lambda_{prior,i}(\hat{r}_i^t) = \frac{1}{\alpha(\hat{r}_i^t + \epsilon_{prior})}$, where α is a base variance scale and $\epsilon_{prior} \ll 1$ ensures minimum precision. For low attention ($\hat{r}_i^t \to 0$), precision is high (small variance $\alpha\epsilon_{prior}$), anchoring θ_i to θ_i^t. For high attention ($\hat{r}_i^t \to 1$), precision is lower (variance $\alpha(1+\epsilon_{prior})$), allowing θ_i to adapt more freely. The prior model in Eq. (5) is the third term in our posterior from Eq. (3).

C. Full Posterior and MAP Update

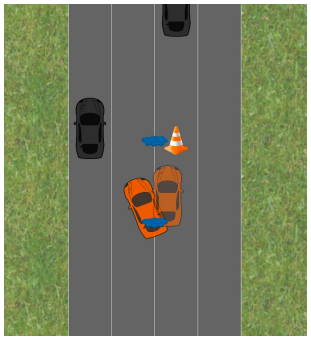
Before deriving the full posterior and MAP update, we summarize the key assumptions underpinning our framework for tractable inference: (1) a locally Gaussian posterior over preferences, induced by a log-quadratic physical likelihood and Gaussian language likelihood and prior; (2) stationary user preferences within each interaction episode; and (3) a reasonably calibrated dual-LLM, such that its predicted shift μ^t is an unbiased estimate of the true preference change and its confidence scores meaningfully reflect uncertainty.

Posterior Formulation and MAP Update. Substituting the physical likelihood (Eq. (1)), language likelihood (Eq. (4)), and conditional prior (Eq. (5)) into Eq. (3), we obtain the full posterior over θ. To compute the MAP estimate, we work in the log domain. Following Bajcsy et al. [4], we approximate the physical likelihood's contribution to the log-posterior near $\hat{\theta}^t$ using the feature difference $\Delta\Phi = \Phi(\xi_H) - \Phi(\xi_R)$, yielding a linear term $\theta^\top \Delta\Phi$. The resulting log-posterior is:

$$\log P(\theta \mid \cdot) = \theta^\top \Delta\Phi - \frac{\Lambda_{\text{prior}}(\hat{r}^t)}{2}(\theta - \theta^t)^2 - \frac{(\mu^t - (\theta - \theta^t))^2}{2(\sigma_L^t(m^t))^2} + \text{const.} \qquad (6)$$



(a) Two-Block Environment



(b) 4-lane CPC (CPC-4)

Fig. 2. Example scenarios created from one of our four experimental driving scenarios and one robotic arm scenario.

with Λ_{prior} the prior precision from Sec. IV-B and σ_L^2 the language likelihood variance from Sec. IV-A2. Setting the gradient with respect to θ to zero and solving for θ, we obtain the closed-form update:

$$\hat{\theta}_i^{t+1} = \hat{\theta}_i^t + \frac{\sigma_{L,i}^2}{\Lambda_{\text{prior,i}}\sigma_{L,i}^2 + 1}\Delta\Phi_i + \frac{1}{\Lambda_{\text{prior,i}}\sigma_{L,i}^2 + 1}\mu_i^t \ . \quad (7)$$

This is the QuickLAP MAP update rule, applied element-wise for each feature i.

Interpretation. Eq. (7) exhibits a Kalman-style fusion in which an *attention–adaptive precision* term decides the credit assignment between modalities. We define the scalar gain

$$\kappa_i(m_i^t, \hat{r}_i^t) = \left[\Lambda_{prior,i}(\hat{r}_i^t)\sigma_{L,i}^2(m_i^t) + 1\right]^{-1} \ .$$

The MAP update can be rewritten element-wise as:

$$\hat{\theta}_i^{t+1} = \theta_i^t + \kappa_i(m_i^t, \hat{r}_i^t)\left[\sigma_{L,i}^2(m_i^t)\,\Delta\Phi_i + \mu_i^t\right].$$

The behavior of this update depends on the attention gate $\hat{r}_i^t$ (via $\Lambda_{prior,i}$) and language confidence m_i^t (via $\sigma_{L,i}^2$). For features with high attention score (low prior precision $\Lambda_{prior,i}$) and high language confidence (low language variance $\sigma_{L,i}^2 \approx \epsilon_{var}$), $\kappa_i \approx 1$. The update is dominated by the language-suggested shift μ_i^t, while the physical correction $\Delta\Phi_i$ is suppressed by the small $\sigma_{L,i}^2$ multiplier. Conversely, if language confidence is low (large $\sigma_{L,i}^2$), then $\kappa_i\sigma_{L,i}^2 \approx 1/\Lambda_{prior,i}$. The update gives more weight to the physical correction $\Delta\Phi_i$, scaled by the prior variance. If attention score $\hat{r}_i^t$ is low (high prior precision $\Lambda_{prior,i}$), κ_i becomes small, and θ_i changes little from θ_i^t, primarily driven by $\Delta\Phi_i$ if μ_i^t is also small or language confidence is low. This adaptive weighting allows QuickLAP to react decisively to clear interventions while remaining robust to ambiguous signals.

V. Simulated Experiments

We first evaluate **QuickLAP** using simulated human feedback in two domains: a robotic manipulation scenario and an autonomous vehicle scenario.

A. Robot Manipulation Scenario

1) Experimental Setup: **Simulator.** We conduct our manipulation experiments in Robosuite [66], a physics-based simulation framework for robot learning. We model the robot with a 6D damped second-order end-effector dynamics model. The state is $x = [p_x, p_y, p_z, v_x, v_y, v_z]^\top$, where (p_x, p_y, p_z) denotes the end-effector position and (v_x, v_y, v_z) its velocity. Actions $u = [u_x, u_y, u_z]^\top$ correspond to desired Cartesian velocity commands. The robot plans motions using MPC.

Environment. The task is shown in Fig. 2a. The task requires the robot arm to transport an object from a start location (shown in red) to a goal region (shown in green) while avoiding obstacles in the *Two-Block Environment*. The green block represents an obstacle located along the nominal shortest path to the goal, while the blue region represents an area of indifference that is safe but increases path length. This environment induces feature coupling: avoiding the obstacle often requires moving away from the goal, creating ambiguity in how physical corrections should be interpreted.

We model a simulated human with a ground-truth reward function θ^* that prioritizes obstacle avoidance over path efficiency. This simulated human provides physical corrections and language feedback at predetermined timesteps whenever the robot's autonomous behavior conflicts with these true preferences. Physical feedback is applied as a corrective input that perturbs the robot's motion, often pulling the arm closer to the blue region and farther from the goal in order to avoid the obstacle. This introduces an inherent ambiguity: a physical correction that improves safety may simultaneously degrade task efficiency. Natural language feedback is used to clarify intent (e.g., obstacle avoidance versus goal progress), and is interpreted using the same dual-LLM pipeline and Bayesian update described earlier.

Independent Variables. We compare four approaches: 1) **Physical-Only (pHRI):** baseline that uses the physical correction to update $\hat{\theta}$ via Eq. (2), without any language input. We hypothesize that it incorrectly updates the weights for features that the human did not intend to correct. 2) **QuickLAP (Full Model):** Our proposed method, which uses both the attention mask r in the conditional prior *and* the language-derived shift μ and confidence m in the language likelihood (7). We additionally compare against two ablations of QuickLAP. 3) **Masked Physical Update (Masked pHRI):** ablation of our method that incorporates the attention mask r derived from language (via LM_{att}) into the conditional prior (5), effectively gating which features are updated based on the physical correction $\Delta\Phi$. However, it does *not* use the language-derived shift μ or confidence m from LM_{pref}; the update relies solely on $\Delta\Phi$ for the attended features. This ablation helps identify the benefit of the attention mechanism alone. 4) **QuickLAP Language Only:** ablation of our method that isolates the effect of language input on the update function by removing the physical component from the proposed MAP update (but the LLM framework still gets access to the feature difference induced by the physical correction). This ablation

is meant to test if just having the context about the person's physical input is enough for an LLM to reasonably estimate the full weight update.

During system development, we examined a "No-Context" baseline, where the LLM received the language utterance without any physical information. This baseline showed a drop in performance, detailed in Appendix N, because, without physical grounding, the model's semantic interpretations are not anchored to the actual feature changes occurring in the environment. Consequently, simulated utterances remained ambiguous or misaligned with the robot's state. To ensure consistent grounding, the language input in all our primary experiments is always accompanied by concurrent physical feedback. While QuickLAP is designed for this joint interpretation, we leave the exploration of entirely decoupled, language-only interventions for future work.

Evaluation Metrics. We evaluate the accuracy of the learned preference weights $\hat{\theta}$ against the ground-truth user preferences $\theta^\star$ using **Normalized Mean Squared Error (NMSE)**, calculated as $\text{NMSE}(\hat{\theta}, \theta^\star) = \frac{1}{d}\|\text{norm}(\hat{\theta}) - \text{norm}(\theta^\star)\|^2$. Lower NMSE indicates that the learned normalized preference vector is closer to the true normalized preference vector, signifying a more accurate capture of the *relative importance* and *direction* of the user's preferences, independent of their absolute scale.

By comparing these methods using NMSE, we aim to test two key points: (i) language provides a stronger signal than purely physical feedback for achieving reward alignment, and (ii) the combination of language and physical corrections yields the best results. In particular, the richer and more reliable interpretation of language through the shift μ and confidence m contextualized by physical interventions in QuickLAP provides significant advantages beyond a simpler attention-based gating of physical updates (as in the Masked pHRI baseline), enabling more accurate recovery of rewards.

2) Results: Fig. 3a reports the normalized reward-weight MSE, averaged over six utterances per method. Physical-only learning (pHRI) and Masked pHRI shows the highest error (0.1018 ± 0.0001 and 0.1000 ± 0.0002, respectively), indicating that physical corrections alone are insufficient to disambiguate coupled objectives such as obstacle avoidance and goal progress in this manipulation task. In contrast, methods that incorporate language achieve substantially lower error. Both **QuickLAP Language Only** (0.0427 ± 0.0043) and **QuickLAP** (0.0444 ± 0.0052) reduce reward-weight error by more than $2\times$ relative to physical-only baselines.

The strong performance of the language-only variant reflects the fact that, in this controlled setting, the language model is provided with rich physical-intervention context in its prompt, enabling it to infer the intended reward shift even without using the physical correction in the update. The full **QuickLAP** method achieves comparable accuracy while additionally grounding the update itself in physical interaction. This grounding does not improve accuracy in this simple domain, but is critical for robustness in settings with noisier feedback or more ambiguous language, as shown in later experiments. Overall, these results demonstrate that language is essential for resolving ambiguity in physical corrections, while physical grounding provides a principled mechanism for maintaining reliability beyond controlled scenarios.

B. Driving Scenario

1) Experimental Setup: **Simulator.** In our experiments, we build on the InterACT Lab Driving Simulator [53] that uses a 4D bicycle model for vehicle dynamics. The state $x = [x, y, \theta, v]^T$ includes vehicle coordinates (x, y), heading (θ), and speed (v). Actions $u = [\omega, a]^T$ are steering (ω) and linear acceleration (a).

Environments. We test **QuickLAP** in four scenarios of increasing complexity (Fig. 2). We model a simulated human with a ground truth reward function parameterized by $\theta^\star$ that highly prioritizes safety from traffic cones and other cars, is moderately concerned about avoiding puddles, and also values lane alignment (detailed in Appendix O). The simulated human would provide physical corrections and language feedback at predetermined points when the robot approached a cone too closely.

Our environments are as follows: 1) *2-lane Cone (C)*: A simple two-lane road with a single cone. Physical corrections away from the cone may be confounded with lane deviation, testing whether the system can isolate the true preference (cone avoidance) from correlated features (lane alignment). 2) *2-lane Cone + Puddle (CP)*: Adds a puddle to the scene. Avoiding the cone now requires entering a less-penalized puddle, making it harder to infer whether the human prefers cone avoidance over surface quality. 3) *3-lane Cone + Puddle + Car (CPC-3)*: A 3-lane road with a cone, a puddle, and another vehicle. Avoiding the cone may require moving closer to the puddle or the car, testing the system's ability to disambiguate trade-offs among multiple penalized features. 4) *4-lane Cone + Puddle + Cars (CPC-4)*: This world has 4 lanes and multiple cones, puddles, and several other vehicles, combining all complexities of the previous environments.

In each scenario, we assess robustness by testing QuickLAP across varied natural language phrasings with the same underlying intent to avoid traffic cones but with varying ambiguity, such as "Be careful.", "Avoid the obstacle.", "Stay away from construction zones.", or "Steer clear of the cone" (see Appendix N). Additionally, the number of human interventions were varied to observe convergence effects. We performed a total of 600 simulations to obtain the results in Fig. 4. We ran our experiments on a single CPU and used up to 6-dimensional driving features (distance to cone, distance to puddle, lane offset, off-road risk, distance to cars, speed).

We follow the same experimental protocol, learning framework, baselines, and evaluation metrics introduced in the robotic manipulation simulation, and adapt them to a driving environment. We also evaluated a "No-Context" baseline (Appendix N), which confirmed that without physical grounding, language-only updates fail to anchor semantic intent to specific feature changes, especially in complex environments.

2) Results: Our simulated driving experiments demonstrate that the trends observed in the robotic manipulation domain persist in a more complex setting, with QuickLAP consistently outperforming baseline methods in learning accurate user reward weights. The results, averaged across various language phrasings, random seeds, and distinct environmental scenarios, highlight QuickLAP's superior accuracy and convergence speed.

Accuracy Across Environments. Figure 4a reports normalized weight MSE across four environments. QuickLAP consistently achieves lower NMSE than Masked pHRI and pHRI, with reductions of more than 50% in moderate cases such as `CP` (0.0761 ± 0.1090 vs. 0.3034 ± 0.0236 for Masked pHRI and 0.5941 ± 0.0013 for pHRI).

The QuickLAP Language Only variant closely tracks the full model (e.g., 0.2020 ± 0.0375 vs. 0.2058 ± 0.0525 in `CPC-4`), reflecting the strength of the LLM-provided signal in the update. In contrast, the full QuickLAP model combining both language and physical corrections achieves comparable accuracy while maintaining consistent performance across environments. The physical grounding provided by QuickLAP contributes to stability and prevents spurious preference shifts, which may become critical when language inputs are less reliable. We further examine this hypothesis through targeted simulation ablations that remove physical grounding from language interpretation.

Convergence and Stability. Fig. 4b illustrates the convergence behavior across methods. QuickLAP converges more rapidly than all baselines. After the first intervention, QuickLAP achieves lower NMSE than Masked pHRI or pHRI after four interventions. By the third intervention, QuickLAP stabilizes below 0.25, while the baselines remain above 0.40. This behavior illustrates that jointly estimating both a language-conditioned shift (μ) and confidence (m) allows QuickLAP to incorporate human feedback efficiently while remaining grounded in physical context. In contrast, Masked pHRI provides only incremental gains over pHRI.

VI. User Evaluations

Our simulated results showed that QuickLAP performs well for idealized user inputs. We aim to evaluate QuickLAP's ability to generalize to realistic user inputs across different input devices. We first perform a pilot study with $N = 12$ expert users in the robotic manipulation scenario (Fig. 2a) using a Connexion 3D SpaceMouse [1] to control the robot. Next, we perform a more extensive evaluation with $N = 15$ non-expert users in the driving scenario (Fig. 2b) using a physical wheel to control the robot.

A. Pilot Study

In our pilot study, each expert user interacted with a Franka robot in the block manipulation scenario. In this study, participants were first shown a desired robot trajectory that avoided the green block while moving to the green goal zone. After the participant was familiar with the optimal trajectory, the robot began executing a path, and the expert user was instructed to provide simultaneous physical and language input to correct the robot. The physical input was collected using a SpaceMouse and the user's language input was collected through a microphone and transcribed using OpenAI's Whisper model [49]. The time for each intervention in our studies was comprised of (1) approximately one second for audio capture, (2) 1–2 seconds for speech-to-text transcription, and (3) 0.4–1.5 seconds each for 2 sequential LLM calls. While our code was not optimized for speed this latency can be reduced to sub-second levels via parallelized 'mini/nano' models and optimized, on-device speech to text transcription.

Each expert user performed three rounds of corrections, and the robot updated its estimate of the user's weights using three algorithms: pHRI, QuickLAP Language Only, and QuickLAP, described in Sec. V. These three methods were presented to the user in a randomized order to control for ordering effects.

We report the NMSE of the robot's learned weights during the pilot study compared to the ground-truth weights in Fig. 5. We found that both variants of QuickLAP outperformed the pHRI baseline. This result indicates that the QuickLAP approach is robust to realistic physical and language input from real users.

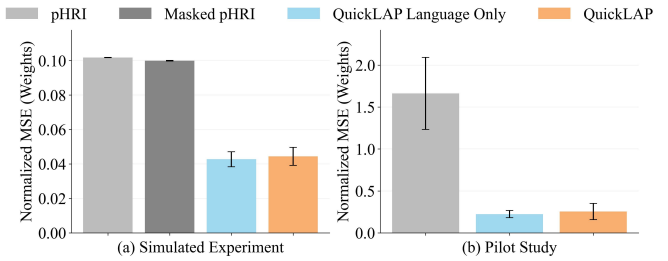


Fig. 3. a) Normalized reward-weight MSE for QuickLAP and baselines in the simulated robotic manipulation scenario, averaged over 6 language inputs across 5 seeds. (b) Normalized reward-weight MSE for QuickLAP and baselines from the pilot study, averaged over 12 participant interactions. In both figures, error bars indicate the mean $\pm$ standard error of the mean (SEM) calculated across the language inputs.

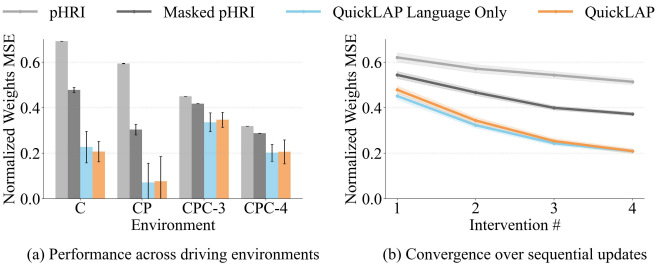


Fig. 4. Comparison of adaptation methods across different environments for 4 interventions per episode. (a) Bars represent the Normalized Mean Squared Error (NMSE) for each environment, averaged over 6 language inputs. Error bars indicate the mean $\pm$ standard error of the mean (SEM) calculated across the language inputs. (b) Solid lines represent the average NMSE. Shaded regions indicate the mean $\pm$ SEM, with the SEM calculated across the 4 environments. Lower NMSE indicates better performance.

B. Non-Expert User Study

Following our expert user pilot study, we conducted a more extensive analysis with non-expert users to determine if QuickLAP generalizes across different user populations. We elected to use the car driving scenario for this evaluation to avoid confounding effects from the steep learning curves associated with high-dimensional robot manipulation control [20]. We performed a within-subjects user study where participants controlled the car using a steering wheel input as depicted in Fig. 1. The experimental procedure was approved by the institutional review board.

Procedure. Each participant entered a room containing the user study setup shown in Fig. 1. Before beginning the main study, participants performed a familiarization task using the steering wheel and pedals in our setup until they were comfortable controlling the car in environment C (described in Sec. V-B). Participants were assigned to a sequence of three experimental blocks. Each block sampled one environment (CP, CPC-3, CPC-4) and one algorithm (pHRI, QuickLAP Language Only, or QuickLAP). These experimental blocks were presented in a randomized and counter-balanced order.

Each block consisted of three stages. In the first stage, the participant was shown the desired driving behavior to teach the robot. In the next stage, the participant performed interventions to correct the robot using language and physical input. Finally, the participant observed the car performing the learned behavior. After completing the three experimental blocks, participants filled out a questionnaire and ranked their overall preferences among the algorithms.

Evaluation Metrics. Our non-expert user study evaluated our algorithm in three ways. First, we adapted scales validated in previous work [4] to assess four self-reported constructs that reflect how well a system learns from user feedback on a 7-point scale: (1) *understandability*, (2) *ease of use*, (3) *predictability*, and (4) *collaborativity*. Second, participants additionally ranked each algorithm relative to the other algorithms according to their overall preference. Finally, we measured performance of the vehicle behavior learned from the participant using the NMSE as in our simulated experiments.

Hypotheses. Based on the prior literature that highlights the importance of learning from multimodal signals [29, 42, 22, 24, 60, 30] and our simulated results, we developed the following hypotheses about our measures:

> **(H1)** QuickLAP is (a) *more understandable*, (b) *easier to use*, (c) *more predictable*, and (d) *more collaborative* than pHRI.
> **(H2)** Participants will *prefer* QuickLAP to pHRI.
> **(H3)** QuickLAP will achieve *lower error* than pHRI.

C. Non-Expert User Study Results

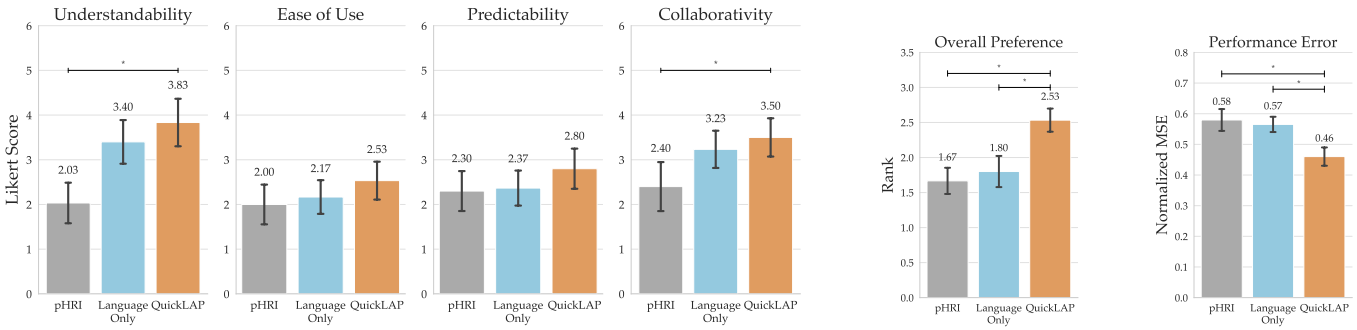
H1: User-reported Constructs. We performed a repeated measures ANOVA for each construct to determine significant effects. Fig. 5a illustrates an overview of our results. We found that QuickLAP was significantly more *understandable* than pHRI, $p = .023$, providing support for **H1a**. We found no significant differences for *Ease of Use* (H1b) or *Predictability* (H1c), but found an empirical trend that users rated QuickLAP as the easiest to use and the most predictable on average. QuickLAP was significantly more collaborative than pHRI, $p = .029$, providing support for **H1d**. Overall, these results partially support **H1**, illustrating that learning from both language and physical input results in more intuitive updates on robot behavior than learning from physical input alone, though providing two forms of input simultaneously was not necessarily easier to do.

H2: Overall Preference. Users ranked the three algorithms relative to each other at the end of the experiment. Because these ranking data are ordinal rather than continuous, we used a non-parametric Friedman test with the algorithm as a within-subjects factor and rank as the dependent variable. The average ranks of each algorithm are shown in Fig. 5b. Participants ranked QuickLAP significantly higher than both Language Only, $p = .048$, and pHRI, $p = .022$. This result supports **H2**, highlighting QuickLAP's ability to incorporate language input to provide more accurate preference updates.

H3: Performance Error. We assessed performance error by comparing the behavior the vehicle learned to the optimal behavior participants attempted to emulate, shown in Fig. 5c. We found that QuickLAP achieved significantly lower NMSE than both Language Only, $p = .015$, and pHRI, $p = .040$. This result supports **H3**, demonstrating that participants were able to teach the robot behaviors that were functionally better by using QuickLAP compared to the other algorithms.

VII. Discussion

Our results highlight key differences between *simulated* and *real* human feedback. In simulation, QuickLAP Language Only performs nearly as well as the full model, because our simulated utterances correspond directly to feature-level feedback (e.g., "avoid the obstacle" for the cone feature). We also saw that expert users showed similar trends, where NMSE was similar for both QuickLAP Language Only and QuickLAP. In the non-expert user study, however, participants often used higher-level or indirect phrasing, such as "move to the right lane", which does not map directly onto the robot's features. When this happens, our probabilistic model places more emphasis on learning from physical corrections than language input, resulting in better robustness to noisy behaviors. The user study also elicited realistic edge cases that occur when learning from multimodal feedback, such as contradictory inputs and imprecise language. For example, one user verbally commanded the robot to "stay away" from an obstacle while erroneously moving the robot toward the obstacle. QuickLAP's confidence-weighted formulation (Eq. 7) balanced correction and language updates to prevent the incorrect reward shifts that typically occur with physical-only updates. Another participant mispronounced 'cone' as 'coin'. Because the LLM likelihood is conditioned on physical context, where a cone is present but a coin is not, QuickLAP successfully resolved the user's true intent. Overall, the language-only



(a) Self-reported constructs (b) Algorithm Ranking (c) Objective Measure

Fig. 5. User study results. All error bars represent standard error. (a) Average ratings for *Understandability*, *Ease of Use*, *Predictability*, and *Collaborativity*. Higher values are better. (b) Average ranking of each algorithm. Three corresponds to the most preferred algorithm, and one corresponds to the least preferred algorithm. (c) Normalized MSE between the learned vehicle behavior and the optimal vehicle behavior. Lower values are better.

variant's strong simulated performance suggests that, when language is accurate, a few physical examples provided as in-context prompts could approximate the role of real corrections, potentially enabling purely voice-based preference updates. However, our user study indicates that real human feedback is often inconsistent or underspecified, underscoring the need for physical grounding in practice.

Limitations and Future Work. Despite QuickLAP's demonstrated benefits, several limitations suggest avenues for future exploration. First, we assumed that the outputs of our dual-LLM framework were reasonably calibrated, i.e., that the estimated shift μ^t and confidence m^t were unbiased reflections of user intent. While empirical results suggest this assumption is appropriate in practice, explicit calibration of LLMs remains an open challenge. Formally evaluating perturbed confidence values would determine the extent to which this assumption holds in general domains. Future work could explore in-context learning with labeled examples [52] or conformal prediction [35, 65] to systematically align model uncertainty with user intent.

Second, QuickLAP currently focuses on fusing physical corrections and language. Robots, however, can learn from many other human cues: gaze [3, 34], facial expression [16, 17], gesture [10], and more [58, 12]. Extending QuickLAP to incorporate additional user input modalities would enable richer, more intuitive interaction. In addition, QuickLAP focused on pre-defined features that are updated via multi-modal feedback. In many HRI settings, these features may be hard to specify, however future work may incorporate methods like language-guided abstraction [46] or pretrained segmentation models [51] to dynamically generate features.

Third, in our non-expert user study, participants did not rate QuickLAP as significantly easier to use or more predictable than baselines. We believe that using the same interaction to provide feedback for all methods led to this perception. While this was important for controlling the duration of feedback for each algorithm, future work may investigate the impact of the interface design for initiating corrections with respect to their ease of use and predictability. Additionally, formal quantitative analysis on the robustness to contradictory inputs remains a limitation of this study. Furthermore, while current LLM API latencies were sufficient for the interaction frequencies observed in our studies, future deployments on high-speed platforms could utilize local, distilled models to further minimize the delay between a user's utterance and the reward update.

Finally, our user evaluations assumed a single optimal behavior for users to emulate. While this is realistic for driving, where people often follow the same road rules, other HRI settings like social navigation [40], assistive teleoperation [14, 61], or manufacturing [43], involve diverse, subjective goals. Our experiments showed QuickLAP's ability to generalize across simple manipulation tasks and driving environments, but future work may further evaluate its applicability in different domains.

Conclusion. In this paper, we presented QuickLAP, a Bayesian framework that fuses physical corrections and natural language to infer human reward functions in real time. Across simulation and a user study, QuickLAP consistently improved reward learning compared to physical-only or heuristic multimodal baselines. Participants found our approach more understandable and collaborative, and its learned behaviors were closer to their intended goals. Together, these findings highlight the promise of learning from multiple feedback types, grounding abstract utterances directly in reward space.

Acknowledgments

This research was supported in part by the MIT Energy Initiative (MITEI) Seed Award. The authors would like to thank the participants of our user studies for their time and valuable feedback.

References

[1] 3Dconnexion. *SpaceMouse Enterprise Technical Specifications*, 2024. Available at https://3dconnexion.com/.

[2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi: 10.1145/

1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

[3] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[4] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning robot objectives from physical human interaction. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 217–226. PMLR, 2017. URL http://proceedings.mlr.press/v78/bajcsy17a.html.

[5] Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pages 141–149, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-4953-6. doi: 10.1145/3171221.3171267. URL http://doi.acm.org/10.1145/3171221.3171267.

[6] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[7] Erdem Bıyık, Malayandi Palan, Nicholas C Landolfi, Dylan P Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. *arXiv preprint arXiv:1910.04365*, 2019.

[8] A. Bobu, A. Bajcsy, J. F. Fisac, and A. D. Dragan. Learning under misspecified objective spaces. In *Conference on Robot Learning (CoRL)*, 2018.

[9] A. Bobu, A. Bajcsy, J. F. Fisac, S. Deglurkar, and A. D. Dragan. Quantifying hypothesis space misspecification in learning from human–robot demonstrations and physical corrections. *Transactions on Robotics (T-RO)*, 2020.

[10] Landon Brown, Jared Hamilton, Zhao Han, Albert Phan, Thao Phung, Eric Hansen, Nhan Tran, and Tom Williams. Best of both worlds? combining different forms of mixed reality deictic gestures. *J. Hum.-Robot Interact.*, 12(1), February 2023. doi: 10.1145/3563387. URL https://doi.org/10.1145/3563387.

[11] Arthur Bucker, Luis F. C. Figueredo, Sami Haddadin, Ashish Kapoor, Shuang Ma, Sai Vemprala, and Rogerio Bonatti. LATTE: language trajectory transformer. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 7287–7294. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10161068. URL https://doi.org/10.1109/ICRA48891.2023.10161068.

[12] Kate Candon, Nicholas C Georgiou, Helen Zhou, Sidney Richardson, Qiping Zhang, Brian Scassellati, and Marynel Vázquez. React: Two datasets for analyzing both human reactions and evaluative feedback to robots over time. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 885–889, 2024.

[13] Vanya Cohen, Geraud Nangue Tasse, Nakul Gopalan, Steven James, Matthew C. Gombolay, and Benjamin Rosman. Learning to follow language instructions with compositional policies. *CoRR*, abs/2110.04647, 2021. URL https://arxiv.org/abs/2110.04647.

[14] Maggie A Collier, Rithika Narayan, and Henny Admoni. The sense of agency in assistive robotics using shared autonomy. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 880–888. IEEE, 2025.

[15] Y. Cui, S. Karamcheti, R. Palleti, N. Shivakumar, P. Liang, and D. Sadigh. No, to the right: Online language corrections for robotic manipulation via shared autonomy. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 93–101, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399647. doi: 10.1145/3568162.3578623. URL https://doi.org/10.1145/3568162.3578623.

[16] Yuchen Cui, Qiping Zhang, Brad Knox, Alessandro Allievi, Peter Stone, and Scott Niekum. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, pages 604–626. PMLR, 2021.

[17] Nathaniel Dennler, Catherine Yunis, Jonathan Realmuto, Terence Sanger, Stefanos Nikolaidis, and Maja Matarić. Personalizing user engagement dynamics in a non-verbal communication game for cerebral palsy. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 873–879. IEEE, 2021.

[18] Nathaniel Dennler, Zhonghao Shi, Stefanos Nikolaidis, and Maja Matarić. Improving user experience in preference-based optimization of reward functions for assistive robots. *arXiv preprint arXiv:2411.11182*, 2024.

[19] Nathaniel Dennler, Stefanos Nikolaidis, and Maja Matarić. Contrastive learning from exploratory actions: Leveraging natural interactions for preference elicitation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 778–788, 2025. doi: 10.1109/HRI61500.2025.10974136.

[20] Varad Dhat, Nick Walker, and Maya Cakmak. Using 3d mice to control robot manipulators. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 896–900, 2024.

[21] A. D. Dragan, K. Muelling, J. Andrew Bagnell, and S. S. Srinivasa. Movement primitives via optimization. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2339–2346, May 2015. doi: 10.1109/ICRA.2015.7139510.

[22] Tesca Fitzgerald, Pallavi Koppol, Patrick Callaghan, Russell Quinlan Jun Hei Wong, Reid Simmons, Oliver Kroemer, and Henny Admoni. Inquire: Interactive querying for user-aware informative reasoning. In *6th Annual Conference on Robot Learning*, 2022.

[23] Carlos E. García, David M. Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335 – 348, 1989. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(89)90002-2. URL http://www.sciencedirect.com/science/article/pii/0005109889900022.

[24] Michael Hagenow and Julie A. Shah. Realm: Real-time estimates of assistance for learned models in human-robot interaction. *IEEE Robotics and Automation Letters*, 10 (6):5473–5480, 2025. doi: 10.1109/LRA.2025.3560862.

[25] Erin Hedlund-Botti, Julianna Schalkwyk, Nina Moorman, Sanne van Waveren, Lakshmi Seelam, Chuxuan Yang, Russell Perkins, Paul Robinette, and Matthew Gombolay. Learning interpretable features from interventions. In *Robotics: Science and Systems (RSS)*, 2025.

[26] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR, 2022. URL https://proceedings.mlr.press/v162/huang22a.html.

[27] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In Karen Liu, Dana Kulic, and Jeffrey Ichnowski, editors, *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 2022. URL https://proceedings.mlr.press/v205/ichter23a.html.

[28] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. Learning preferences for manipulation tasks from online coactive feedback. *Int. J. Robotics Res.*, 34(10):1296–1313, 2015. doi: 10.1177/0278364915581193. URL https://doi.org/10.1177/0278364915581193.

[29] Rajat Kumar Jenamani, Tom Silver, Ben Dodson, Shiqin Tong, Anthony Song, Yuting Yang, Ziang Liu, Benjamin Howe, Aimee Whitneck, and Tapomayukh Bhattacharjee. Feast: A flexible mealtime-assistance system towards in-the-wild personalization. In *Robotics: Science and Systems (RSS)*, 2025.

[30] Emily Jensen, Sriram Sankaranarayanan, and Bradley Hayes. Automated assessment and adaptive multimodal formative feedback improves psychomotor skills training outcomes in quadrotor teleoperation. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pages 185–194, 2024.

[31] Siddharth Karamcheti, Megha Srivastava, Percy Liang, and Dorsa Sadigh. LILA: language-informed latent actions. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 1379–1390. PMLR, 2021. URL https://proceedings.mlr.press/v164/karamcheti22a.html.

[32] W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: the TAMER framework. In Yolanda Gil and Natasha Fridman Noy, editors, *Proceedings of the 5th International Conference on Knowledge Capture (K-CAP 2009), September 1-4, 2009, Redondo Beach, California, USA*, pages 9–16. ACM, 2009. doi: 10.1145/1597735.1597738. URL https://doi.org/10.1145/1597735.1597738.

[33] Kimin Lee, Laura M. Smith, and Pieter Abbeel. PEBBLE: feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6152–6163. PMLR, 2021. URL http://proceedings.mlr.press/v139/lee21i.html.

[34] Anthony Liang, Jesse Thomason, and Erdem Bıyık. Visarl: Visual reinforcement learning guided by human saliency. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2907–2912. IEEE, 2024.

[35] Lars Lindemann, Matthew Cleaveland, Gihyun Shim, and George J Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 2023.

[36] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, 41(1): 20–44, 2022.

[37] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. In Dylan A. Shell, Marc Toussaint, and M. Ani Hsieh, editors, *Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021*, 2021. doi: 10.15607/RSS.2021.XVII.047. URL https://doi.org/10.15607/RSS.2021.XVII.047.

[38] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *CoRR*, abs/2210.06407, 2022. doi: 10.48550/ARXIV.2210.06407. URL https://doi.org/10.48550/arXiv.

2210.06407.

[39] James MacGlashan, Mark K. Ho, Robert Tyler Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2285–2294. PMLR, 2017. URL http://proceedings.mlr.press/v70/macglashan17a.html.

[40] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023.

[41] Shaunak A. Mehta and Dylan P. Losey. Unified learning from demonstrations, corrections, and preferences during physical human-robot interaction. *ACM Trans. Hum. Robot Interact.*, 13(3):39:1–39:25, 2024. doi: 10.1145/3623384. URL https://doi.org/10.1145/3623384.

[42] Amal Nanavati, Ethan K Gordon, Taylor A Kessler Faulkner, Yuxin Ray Song, Jonathan Ko, Tyler Schrenk, Vy Nguyen, Bernie Hao Zhu, Haya Bolotski, Atharva Kashyap, et al. Lessons learned from designing and evaluating a robot-assisted feeding system for out-of-lab use. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 696–707. IEEE, 2025.

[43] Heramb Nemlekar, Neel Dhanaraj, Angelos Guan, Satyandra K Gupta, and Stefanos Nikolaidis. Transfer learning of human preferences for proactive robot assistance in assembly tasks. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 575–583, 2023.

[44] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 663–670. Morgan Kaufmann, 2000.

[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[46] A. Peng, A. Bobu, B. Z. Li, T. Sumers, I. Sucholutsky, N. Kumar, T. Griffiths, and J. Shah. Preference-conditioned language-guided abstraction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, 2024. ISBN 979840079322. doi: 10.1145/3610977.3634930.

[47] Andi Peng, Andreea Bobu, Belinda Z Li, Theodore R Sumers, Ilia Sucholutsky, Nishanth Kumar, Thomas L Griffiths, and Julie A Shah. Preference-conditioned language-guided abstraction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 572–581, 2024.

[48] Andi Peng, Belinda Z. Li, Ilia Sucholutsky, Nishanth Kumar, Julie Shah, Jacob Andreas, and Andreea Bobu. Adaptive language-guided abstraction from contrastive explanations. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, pages 3425–3438. PMLR, 2024. URL https://proceedings.mlr.press/v270/peng25c.html.

[49] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[50] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2586–2591, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=1625275.1625692.

[51] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128, 2025.

[52] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 661–682. PMLR, 2023. URL https://proceedings.mlr.press/v229/ren23a.html.

[53] D. Sadigh, S. Sastry, S. Seshia, and Anca D. Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, 2016.

[54] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. In *Robotics: Science and systems*, 2017.

[55] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting robot plans

with natural language feedback. In Kris Hauser, Dylan A. Shell, and Shoudong Huang, editors, *Robotics: Science and Systems XVIII, New York City, NY, USA, June 27 - July 1, 2022*, 2022. doi: 10.15607/RSS.2022.XVIII.065. URL https://doi.org/10.15607/RSS.2022.XVIII.065.

[56] A. Sripathy, A. Bobu, Z. Li, K. Sreenath, D. S. Brown, and A. D. Dragan. Teaching robots to span the space of functional expressive motion. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[57] Simon Stepputtis, Joseph Campbell, Mariano J. Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9909794d52985cbc5d95c26e31125d1a-Abstract.html.

[58] Maia Stiber, Russell Taylor, and Chien-Ming Huang. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 676–683. IEEE, 2022.

[59] Theodore R. Sumers, Mark K. Ho, Robert X. D. Hawkins, Karthik Narasimhan, and Thomas L. Griffiths. Learning rewards from linguistic feedback. *CoRR*, abs/2009.14715, 2020. URL https://arxiv.org/abs/2009.14715.

[60] Tauhid Tanjim, Jonathan St George, Kevin Ching, and Angelique Taylor. Help or hindrance: Understanding the impact of robot communication in action teams. *arXiv preprint arXiv:2506.08892*, 2025.

[61] Yiran Tao, Jehan Yang, Dan Ding, and Zackory Erickson. Lams: Llm-driven automatic mode switching for assistive teleoperation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 242–251. IEEE, 2025.

[62] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In Wolfram Burgard and Dan Roth, editors, *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, pages 1507–1514. AAAI Press, 2011. doi: 10.1609/AAAI.V25I1.7979. URL https://doi.org/10.1609/aaai.v25i1.7979.

[63] Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, and Chelsea Finn. Learning a prior over intent via meta-inverse reinforcement learning. *CoRR*, abs/1805.12573, 2019. URL http://arxiv.org/abs/1805.12573.

[64] Zhaojing Yang, Miru Jun, Jeremy Tien, Stuart Russell, Anca Dragan, and Erdem Biyik. Trajectory improvement and reward learning from comparative language feedback. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 5389–5404. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/yang25e.html.

[65] Michelle Zhao, Reid Simmons, Henny Admoni, and Andrea Bajcsy. Conformalized teleoperation: Confidently mapping human inputs to high-dimensional robot actions. *arXiv preprint arXiv:2406.07767*, 2024.

[66] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, Yifeng Zhu, and Kevin Lin. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

[67] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1433–1438. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620270.1620297.

[68] Matthew Zurek, Andreea Bobu, Daniel S. Brown, and Anca D. Dragan. Situational confidence assistance for lifelong shared autonomy. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 2783–2789. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9561839. URL https://doi.org/10.1109/ICRA48506.2021.9561839.