

SPEAR: Selective Preference Elicitation with Adaptive Rewards for VLM-Guided Robotic Manipulation

Zewen Wu¹, Ruizhe Liu¹, Pei Zhou¹, Feng Chen¹, and Yanchao Yang^{1,2,*}

Abstract—Recent Vision-Language Models (VLMs) enable automated preference labeling for robotic manipulation, yet their deployment remains impractical since each query incurs substantial inference cost, while VLMs frequently abstain on visually challenging comparisons. We propose *SPEAR*, a framework that transforms VLMs into efficient preference oracles via two synergistic mechanisms. First, we design *distinguishability-aware query selection* to reduce abstention rates by targeting only informative and distinguishable comparisons. Second, we deploy an *adaptive reward transition* that bootstraps exploration with VLM-synthesized guidance before preference-based refinement. Experiments on the manipulation benchmarks demonstrate that *SPEAR* achieves 88.9% average success, outperforming baselines by 23.3 percentage points, while significantly reducing abstention and wall-clock training time on contact-rich tasks. *SPEAR* learns three Mobile ALOHA manipulation tasks fully online on real hardware, demonstrating end-to-end real-world applicability. By addressing query efficiency and exploration challenges, *SPEAR* establishes VLMs as practical preference oracles for scalable robot learning without manual reward engineering.

I. INTRODUCTION

Reinforcement learning (RL) offers a principled way to acquire robotic manipulation skills in demonstration-scarce settings, yet practical deployment remains constrained by reward specification. Preference-based Reinforcement Learning (PbRL) addresses this by learning from pairwise trajectory comparisons rather than numerical rewards [1], [2]. However, the scalability of PbRL is constrained by the need for frequent human-in-the-loop preference queries. Recent advances in Vision-Language Models (VLMs) present an appealing solution: modern VLMs can naturally serve as automated preference oracles [3] due to their encoding of extensive world knowledge and robust visual-language grounding. This enables automated robot learning from online interaction and language goals, without relying on demonstrations, manually designed rewards, or ongoing human labeling.

However, VLMs introduce a notable computational constraint – each preference query incurs a significant inference cost. This makes query efficiency critical, yet existing VLM-guided PbRL methods waste considerable budgets due to two fundamental inefficiencies. *First*, VLMs are frequently incapable of providing valid preferences for visually similar or indistinguishable trajectories, i.e., returning abstention. Naive query selection could misuse a significant portion of

expensive queries on comparisons that yield zero learning signal. Across backbones, higher visual distinguishability leads to both lower abstention and higher preference accuracy, motivating abstention-aware acquisition as a core design target. *Second*, meaningful preferences require behaviorally task-related trajectories, yet initial random policies produce only aimless motions. This creates a bootstrapping paradox: generating task-related behaviors requires rewards that need to be learned from comparing task-related behaviors. *Together*, these dilemmas make VLM-guided preference learning inefficient for automated learning.

We present *SPEAR* (Selective Preference Elicitation with Adaptive Rewards), a framework that addresses both challenges with two synergistic mechanisms designed specifically for VLM-based preference learning under computational constraints. *To overcome* the query efficiency challenge, we develop **Distinguishability-Aware Query Selection** that predicts VLM response likelihood before querying. Our approach combines global semantic embeddings with local spatial patterns, and further trains a calibrated distinguishability prediction model. By weighting information gain with predicted distinguishability, we ensure computational resources target only answerable and informative comparisons, while enhancing learning efficiency. *To resolve* the exploration bootstrapping challenge, we introduce **Adaptive Reward Transition** that leverages VLM-synthesized rewards to jumpstart exploration before transitioning to refined preference-based learning. While VLM-synthesized rewards lack precision for task completion, they successfully guide initial exploration toward semantically task-relevant regions. Combined with relay exploration that maintains distributional coverage, this mechanism ensures continuous generation of diverse yet task-relevant trajectories that are necessary for meaningful preference comparisons.

We evaluate our approach on six diverse manipulation tasks from Meta-World [4]. *SPEAR* achieves an 88.9% average success rate, surpassing the strongest VLM-based baseline by 23.3 percentage points, with substantially fewer wasted oracle queries and reduced wall-clock training time. Beyond simulation, we further validate *SPEAR* on a real Mobile ALOHA robot by running the full online pipeline directly in the real world, demonstrating VLM-guided preference learning can bootstrap manipulation skills from scratch under realistic sensing and actuation constraints. The ablation studies confirm that both mechanisms are essential for efficient online PbRL. Our contributions are threefold:

- *First*, we propose a distinguishability-aware query selection mechanism that predicts VLM answerability

¹HKU Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China. zwwu@connect.hku.hk, yanchao@hku.hk

²Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR, China.

*Corresponding author.

through multi-scale visual features and online-adapted modeling, effectively reducing wasted queries.

- *Second*, we introduce an adaptive reward transition strategy that bootstraps exploration with VLM-synthesized reward before transitioning to preference-based learning, ensuring the collection of informative trajectories.
- *Third*, we demonstrate substantial improvements across diverse manipulation tasks and on a real robot, validating that addressing query efficiency and exploration bootstrapping enables practical VLM-guided preference learning under realistic computational constraints.

II. RELATED WORK

A. VLM-guided Robot Policy Learning.

The difficulties in hand-crafting dense reward functions for robotic control have motivated a new paradigm leveraging recent large-scale models for reward specification or generation [3], [5], [6]. A typical current approach utilizes Vision-Language Models (VLMs) [7]–[11] as direct, zero-shot reward models (VLM-RMs) [5] by calculating the cosine similarity between the visual and the goal text embedding. This strategy has also been explored in works using CLIP-style models [12]–[14]. Other works, exemplified by Text2Reward [6], use Large Language Models (LLMs) to generate executable dense reward programs grounded in environment APIs [15], [16]. An alternative paradigm, which our work adopts, avoids direct reward specification. *Instead*, it leverages VLMs as automated preference oracles within a Preference-based Reinforcement Learning (PbRL) framework. For instance, RL-VLM-F [3] queries a VLM to provide preference labels for pairs of image observations, and then learns a reward function from this feedback. Our approach builds upon this VLM-driven preference-based paradigm, focusing on the query efficiency and effective preference elicitation under costly, abstaining VLM oracles.

B. Preference-based Reinforcement Learning.

PbRL [1], [2], [17]–[22] learns an implicit reward function from pairwise comparisons provided by an expert, often based on the Bradley-Terry model [23]. Traditionally, these comparisons are sourced from human experts [1], [24], [25]. *However*, this process is notoriously costly and slow, creating an annotation bottleneck that hinders scalability. *Meanwhile*, many existing PbRL systems and benchmarks assume access to an idealized scripted teacher. This oracle computes preferences by using privileged access to the ground-truth environment reward signals [19], an approach that constitutes an unrealistic training regime and contradicts the PbRL desideratum. The concurrent advancement of VLMs offers a promising direction, positioning VLMs as natural candidates to serve as automated preference oracles [3]. Moreover, AI feedback (RLAIF) has also been explored for aligning models [26], [27]. While this paradigm shift addresses the expertise bottleneck, it introduces new, open problems: how to query these powerful but fallible and non-negligibly costly VLM oracles efficiently and effectively remains challenging.

III. PRELIMINARIES

1) *Problem Formulation*: Formally, we model robotic manipulation with image observations as a Partially Observable Markov Decision Process (POMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma, \mathcal{O}, \Omega)$, where $\mathcal{S}$ denotes the (latent) state space, $\mathcal{A}$ the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ the transition dynamics, $\gamma \in [0, 1)$ the discount factor, $\mathcal{O}$ the observation space, and $\Omega(o | s)$ the observation model. In our setting, observations are images, denoted by $o_t = \mathcal{I}(s_t) \in \mathcal{O}$, and the policy acts on observations $\pi : \mathcal{O} \rightarrow \Delta(\mathcal{A})$. Unlike standard RL, we do not assume access to a reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Instead, we aim to learn policies through the preference feedback of the VLM while training a surrogate reward model.

2) *VLM Preference Acquisition*: Let $\tau = \{o_t\}_{t=1}^H$ denote an observation sub-trajectory of horizon H . Given a task description T and two observation sub-trajectories τ^a, τ^b , we query a VLM oracle to return a preference label $y \in \{0, 1, \emptyset\}$, where $y = 0$ indicates $\tau^a \succ \tau^b$ (sub-trajectory a is preferred over b), $y = 1$ indicates $\tau^b \succ \tau^a$, and $y = \emptyset$ denotes abstention (no clear progress difference or insufficient evidence).

3) *Reward Modeling*: To obtain a surrogate reward function, we adopt Bradley-Terry modeling [23]. We instantiate a neural reward model $r_\theta : o_t \rightarrow \mathbb{R}$ with parameters $\theta \in \Theta$ and model the likelihood of preference as:

$$P_\theta[\tau^b \succ \tau^a] = \sigma \left(\sum_t r_\theta(o_t^b) - \sum_t r_\theta(o_t^a) \right), \quad (1)$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function. Given a dataset $D_{\text{pref}} = \{(\tau^{a,n}, \tau^{b,n}, y^n)\}_{n=1}^N$ of labeled sub-trajectory pairs (with abstentions excluded), we optimize θ by maximizing the log-likelihood:

$$\mathcal{L}(\theta; D_{\text{pref}}) = \sum_{n=1}^N (y^n \log P_\theta[\tau^{b,n} \succ \tau^{a,n}] + (1 - y^n) \log P_\theta[\tau^{a,n} \succ \tau^{b,n}]). \quad (2)$$

4) *The General PbRL Learning Framework*: The whole training loop operates by iterating through three steps:

- *Exploration*: The current policy π trained with the most recent r_θ interacts with the environment while collecting candidate trajectories.
- *Query Selection and Preference Prediction*: A sampling mechanism selects a set of sub-trajectory pairs $\{(\tau^{a,n}, \tau^{b,n})\}_{n=1}^N$ for preference elicitation with a VLM oracle, forming $D_{\text{pref}} = \{(\tau^{a,n}, \tau^{b,n}, y^n)\}_{n=1}^N$. Since each VLM query incurs non-negligible inference cost and may abstain, the sampling mechanism critically determines query efficiency.
- *Reward Learning*: The preference dataset is augmented $D'_{\text{pref}} \leftarrow D'_{\text{pref}} \cup D_{\text{pref}}$, and the reward model r_θ is updated via Eq. 2.

The above steps are iterated until the policy performance converges or a pre-defined budget is reached.

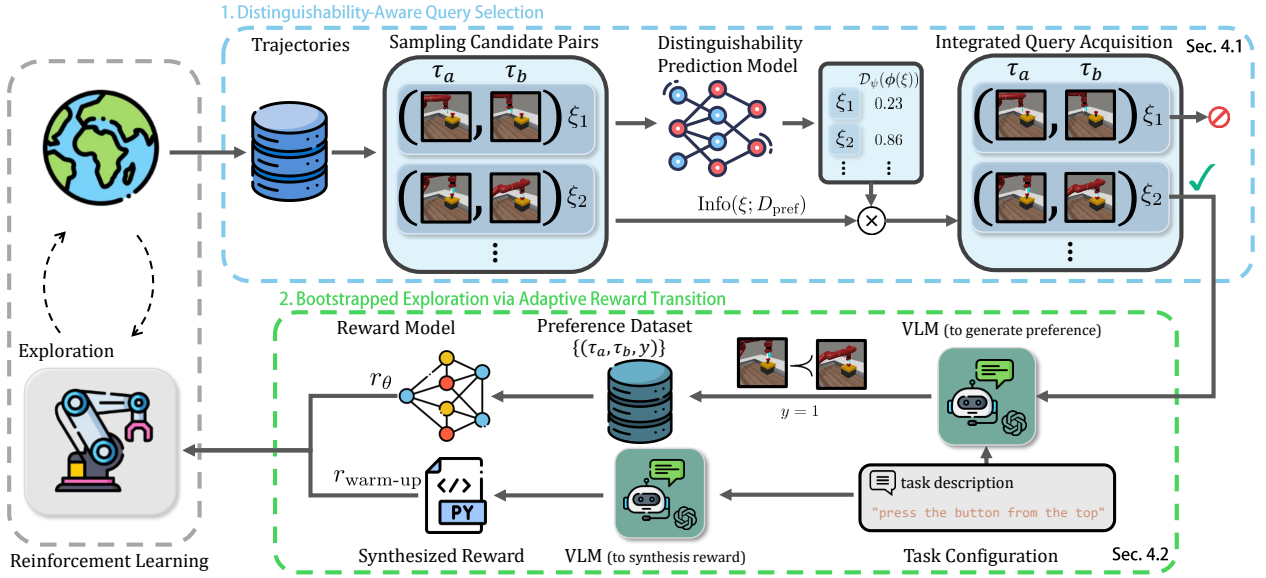


Fig. 1: **Overview of SPEAR.** (1) **Distinguishability-aware query selection** (top, blue) predicts VLM response likelihood via multi-scale features and weights information gain accordingly, selecting only answerable and informative trajectory pairs for querying; (2) **Adaptive reward transition** (bottom, green) bootstraps exploration using VLM-synthesized warm-up rewards before transitioning to refined preference-based learning. Together, these mechanisms enable efficient online learning by ensuring both high-quality query selection and sufficient behavioral diversity for meaningful preference elicitation.

IV. SPEAR: SELECTIVE PREFERENCE ELICITATION WITH ADAPTIVE REWARDS

Deploying VLMs as preference oracles in online PbRL is bottlenecked by two practical issues: (i) expensive queries frequently yield abstentions on visually indistinguishable comparisons, and (ii) early exploration lacks task progress, producing uninformative pairs. *SPEAR* addresses both via (a) **Distinguishability-Aware Query Selection** (Sec. IV-A), which spends queries only on answerable, informative pairs, and (b) **Adaptive Reward Transition** (Sec. IV-B), which bootstraps exploration with a coarse warm-up reward before switching to preference-based refinement.

A. Distinguishability-Aware Query Selection

Classical acquisition functions (e.g., BALD [28]) assume an ideal oracle that always provides effective labels. However, VLM oracles often abstain on trajectory pairs lacking clear visual differences, making expensive queries yield zero learning signal. Moreover, to enhance the accuracy of VLM preferences, we adopt a decomposed and conservative prompting protocol that mitigates positional bias and cross-image attribution errors. This makes abstention a systematic and unavoidable outcome for low-distinctiveness pairs. Empirically, increasing visual distinctiveness simultaneously reduces abstention and improves label accuracy (see Sec. V-B.3). Therefore, prioritizing distinguishable pairs effectively avoids both wasted queries and low-quality supervision.

We address the issue of avoiding queries that may lead to abstention and unreliable labels through *distinguishability-aware acquisition*: a mechanism that predicts VLM response likelihood via multi-scale visual features (Sec. IV-

A.1), maintains a distinguishability model (Sec. IV-A.2), and integrates these predictions with information gain to select only answerable, informative pairs (Sec. IV-A.3).

1) *Multi-Scale Visual Feature Extraction*: Predicting the distinguishability between two sub-trajectories conditioned on the VLM requires understanding both the task semantics and fine-grained manipulation progress. We therefore construct features that capture high-level semantic differences (which VLMs excel at) and subtle spatial variations (which VLMs may struggle with):

We extract global embeddings $\mathbf{e}^a, \mathbf{e}^b \in \mathbb{R}^c$ from the VLM’s vision encoder for the sub-trajectories τ^a, τ^b , then compute relational features:

$$\phi_{\text{rel}}(\xi) = [\|\mathbf{e}^a - \mathbf{e}^b\|; \mathbf{e}^a \odot \mathbf{e}^b; \cos(\mathbf{e}^a, \mathbf{e}^b)] \in \mathbb{R}^{2c+1}, \quad (3)$$

where $\xi = (\tau^a, \tau^b)$ denotes a sub-trajectory pair. These features capture complementary similarity measures: absolute differences indicate semantic distance, element-wise products capture feature co-activation, and cosine similarity measures directional alignment.

Global embeddings may miss subtle spatial differences crucial for manipulation tasks, e.g., whether a gripper is slightly open versus closed. We employ cross-attention between spatial token grids $\mathbf{T}^a, \mathbf{T}^b \in \mathbb{R}^{N \times D}$ to identify such discrepancies. After projecting spatial tokens into a lower-dimensional space $\mathbf{T}_p^a, \mathbf{T}_p^b \in \mathbb{R}^{N \times d}$, bidirectional cross-attention produces aligned representations $\tilde{\mathbf{T}}^a, \tilde{\mathbf{T}}^b$. We compute statistics capturing alignment quality: the mean and maximum of $|\mathbf{T}_p^a - \tilde{\mathbf{T}}^a|$ (misalignment strength) and $\mathbf{T}_p^a \odot \tilde{\mathbf{T}}^a$ (alignment confidence), yielding $\phi_{\text{loc}}(\xi) \in \mathbb{R}^{8d}$. The combined feature $\phi(\xi) = [\phi_{\text{rel}}(\xi); \phi_{\text{loc}}(\xi)]$ provides a

comprehensive signature of visual distinguishability across semantic and spatial scales.

2) *Distinguishability Prediction Model*: Given features $\phi(\xi)$, we train a calibrated predictor $\mathcal{D}_\psi : \mathbb{R}^{2c+8d+1} \rightarrow [0, 1]$ that estimates the probability that the VLM returns a non-abstention label (versus abstention). Crucially, this predictor must adapt online as the policy explores new regions where VLM behavior may differ.

We instantiate $\mathcal{D}_\psi$ as a lightweight MLP trained on accumulated VLM responses $\{(\phi(\xi_i), \mathbb{1}[y_i \neq \emptyset])\}$. Since raw neural predictors are often miscalibrated [29], we apply temperature scaling with parameter T . In practice, we periodically recalibrate T on a sliding window of recent queries to maintain reliable probabilities under distribution shift. The distinguishability model continuously learns from new VLM responses throughout training.

3) *Integrated Query Acquisition*: We combine distinguishability prediction with information-theoretic acquisition to maximize learning efficiency per VLM query. Standard BALD scores a pair by expected information gain to the reward model:

$$\text{Info}(\xi; D_{\text{pref}}) = \mathbb{H}[y \mid \xi] - \mathbb{E}_{p(\theta \mid D_{\text{pref}})}[\mathbb{H}[y \mid \xi, \theta]], \quad (4)$$

where $\xi = (\tau^a, \tau^b)$ and D_{pref} is the current preference dataset. We estimate $\text{Info}(\xi; D_{\text{pref}})$ following standard PbRL practice [19], and note that information gain becomes meaningless if the VLM abstains. We therefore weight by predicted distinguishability:

$$\text{Info}_{\text{SPEAR}}(\xi) = \mathcal{D}_\psi(\phi(\xi)) \cdot \text{Info}(\xi; D_{\text{pref}}). \quad (5)$$

This product balances two objectives: selecting pairs that resolve reward-model uncertainty (high Info) while avoiding pairs the VLM cannot distinguish (low $\mathcal{D}_\psi$).

Since many similar pairs may achieve high scores, directly selecting top-ranked pairs yields redundant queries. We apply facility-location optimization to select diverse batches that maximize coverage while maintaining high scores. Concretely, we perform diversification in the VLM embedding space $\phi(\xi)$ with cosine distance, so selected queries span diverse progress states instead of near-duplicates.

4) *VLM Query Protocol*: Given task T and two observation sub-trajectories τ^a, τ^b , our initial attempts start from the prompt template of RL-VLM-F [3], but we observe two systematic failure modes that corrupt preference learning: (i) *positional bias*, where the predicted preference correlates with presentation order rather than content, e.g., disproportionately favoring the latter candidate, and (ii) *cross-image attribution errors*, where jointly processing multiple images causes feature bleeding across the two candidates.

To mitigate these issues, we adopt a decomposed two-stage protocol that analyzes each sub-trajectory independently first, then makes a CoT-style comparison and encourages conservative decisions. Notably, this protocol does not *cause* abstention—uncertainty and abstentions are common for VLM oracles when visual evidence is insufficient—but makes it explicit under a conservative prompt, while

substantially improving label quality and enabling successful task learning where the original prompt fails.

B. Bootstrapped Exploration through Adaptive Reward Transition

Even with efficient query selection, preference learning faces a bootstrapping challenge: meaningful preferences require trajectories with varied task progress, yet generating such trajectories requires reward guidance that does not initially exist. Random initial policies often yield aimless motions, so comparing early trajectories provides little learning signal regardless of query efficiency.

We resolve this circular dependency through an *adaptive reward transition*: we use VLM-synthesized warm-up rewards to bias early exploration toward task-relevant behaviors, and then transition to the learned preference-based reward r_θ once exploration saturates. This two-phase procedure ensures the policy discovers comparable behaviors worth querying, while ultimately learning from genuine preferences rather than approximations.

1) *Warm-up Reward Synthesis*: We prompt the VLM with the task description $\mathcal{T}_{\text{task}}$ and the environment interface $\mathcal{E}$ to synthesize an executable warm-up reward function $r_{\text{warm-up}} = \text{VLM-Synthesize}(\mathcal{T}_{\text{task}}, \mathcal{E})$, following Text2Reward [6]. To ensure feasibility for real-world deployment, we allow synthetic rewards to use only realistically measurable geometric cues, such as distances between objects, while prohibiting simulator-privileged signals, e.g., goal coordinates provided by the simulator. Therefore, this warm-up reward is *not* intended to precisely represent task completion; rather, it aims to encourage coarse task-related progress (e.g., approach, alignment, and initial contact) that enables the collection of informative trajectory pairs for preference learning.

2) *Adaptive Reward Transition*: The policy initially trains under $r_{\text{warm-up}}$ while collecting preference queries in parallel. We monitor exploration saturation using a generic plateau criterion: when the warm-up return shows no improvement for K_{trans} consecutive episodes, we transition to the now-trained preference-based reward r_θ for further refinement.

3) *Relay Exploration for Coverage*: To prevent premature collapse to unduly narrow behaviors after transition, we implement relay exploration [30]: episodes alternate between the learned policy π and a random policy π_{rand} for sampled durations. We use an inverse-square-root decay schedule $p_{\text{relay}} \propto 1/\sqrt{n_{\text{epi}}}$ for the relay probability, where n_{epi} denotes the episode index. This provides substantial exploratory coverage after transition while progressively reducing relay intervention as learning proceeds.

V. EXPERIMENTS

A. Experimental Setup

1) *Tasks and Environment*: We evaluate *SPEAR* on six manipulation tasks from the Meta-World benchmark [4]: *drawer-open*, *drawer-close*, *button-press-topdown*, *soccer*, *door-close*, and *window-open*. All tasks use image observations for VLM queries and reward model training, with RL policies trained via Soft Actor-Critic (SAC) [31].

TABLE I: Success rates (mean $\pm$ std) across tasks. Middle = 45% - 55% steps.

Method	<i>Drawer-Open</i>	<i>Drawer-Close</i>	<i>Button-Press-Topdown</i>	<i>Soccer</i>	<i>Door-Close</i>	<i>Window-Open</i>	Overall
VLM-RM [5]	0.0 $\pm$ 0.0%	0.2 $\pm$ 0.5%	0.0 $\pm$ 0.0%	15.7 $\pm$ 15.3%	7.5 $\pm$ 10.5%	29.8 $\pm$ 25.9%	8.9 $\pm$ 8.7%
Text2Reward [6]	17.0 $\pm$ 30.4%	79.2 $\pm$ 6.4%	22.3 $\pm$ 17.7%	6.5 $\pm$ 5.9%	99.5 $\pm$ 0.9%	54.1 $\pm$ 24.0%	46.4 $\pm$ 14.2%
Pebble [19]	1.5 $\pm$ 2.0%	87.1 $\pm$ 5.5%	40.6 $\pm$ 9.5%	70.5 $\pm$ 26.1%	91.0 $\pm$ 6.8%	7.8 $\pm$ 5.5%	49.8 $\pm$ 9.2%
RL-VLM-F [3]	11.3 $\pm$ 11.2%	98.2 $\pm$ 3.6%	12.9 $\pm$ 13.4%	88.2 $\pm$ 7.9%	100.0 $\pm$ 0.0%	81.1 $\pm$ 2.7%	65.3 $\pm$ 6.5%
SPEAR (ours)	87.0$\pm$15.9%	100.0$\pm$0.0%	58.2$\pm$16.0%	89.9$\pm$11.7%	100.0$\pm$0.0%	94.3$\pm$4.3%	88.2$\pm$8.0%

TABLE II: Success rates (mean $\pm$ std) across tasks. Final = last 10% steps.

Method	<i>Drawer-Open</i>	<i>Drawer-Close</i>	<i>Button-Press-Topdown</i>	<i>Soccer</i>	<i>Door-Close</i>	<i>Window-Open</i>	Overall
VLM-RM [5]	0.0 $\pm$ 0.0%	0.0 $\pm$ 0.0%	0.0 $\pm$ 0.0%	14.1 $\pm$ 22.3%	3.7 $\pm$ 5.1%	71.6 $\pm$ 14.1%	14.9 $\pm$ 6.9%
Text2Reward [6]	34.8 $\pm$ 32.9%	95.8 $\pm$ 4.8%	53.7 $\pm$ 27.5%	8.9 $\pm$ 5.4%	99.8 $\pm$ 0.5%	87.7$\pm$12.7%	63.4 $\pm$ 14.0%
Pebble [19]	10.0 $\pm$ 11.6%	88.1 $\pm$ 3.2%	41.2 $\pm$ 7.7%	91.0 $\pm$ 13.3%	73.0 $\pm$ 11.5%	5.3 $\pm$ 4.1%	51.4 $\pm$ 8.6%
RL-VLM-F [3]	33.9 $\pm$ 21.4%	98.8 $\pm$ 3.1%	17.9 $\pm$ 16.6%	75.4 $\pm$ 11.6%	95.7 $\pm$ 7.6%	69.2 $\pm$ 5.3%	65.1 $\pm$ 10.9%
SPEAR (ours)	86.4$\pm$12.7%	100.0$\pm$0.0%	67.0$\pm$17.8%	95.1$\pm$3.9%	100.0$\pm$0.0%	84.9 $\pm$ 5.7%	88.9$\pm$6.7%

2) *Baselines*: We compare against four representative VLM-guided RL approaches: (1) **VLM-RM** [5] constructs rewards through cosine similarity between BLIP2 embeddings and task descriptions; (2) **Text2Reward** [6] leverages VLMs to generate executable reward functions; (3) **PEBBLE** [19] implements PbRL with entropy-based query selection and a scripted oracle (Meta-World’s ground-truth rewards); (4) **RL-VLM-F** [3] applies VLM preferences with random query selection. For fair comparison, we employ the same open-source preference oracle, InternVL3-78B [8], and set sub-trajectory length $H = 1$ for both RL-VLM-F and our method. In addition, we report VLM reliability diagnostics averaged over five VLM backbones in Sec. V-B.3 (including Qwen3-VL and InternVL3.5 variants [8], [9], [11]).

3) *Evaluation Schemes*: Each method is evaluated across 5 random seeds with mean success rates and standard deviations reported. Policies undergo evaluation every 10K environment steps over 50 episodes. Oracle queries are scheduled with a fixed step-coupled cadence: every 10K environment steps, we collect 40 preference comparisons and update the reward model (and our distinguishability predictor). Abstentions do not trigger additional queries; therefore, the total query budget is fixed by the training steps under the same schedule. All experiments utilize 8 NVIDIA A800 GPUs.

B. Main Results

1) *Task Performance*: Tables I and II present success rates at mid-training and final evaluation, respectively, while Figure 2 illustrates learning curves. **SPEAR** demonstrates consistent superiority across tasks of varying difficulty, achieving an overall success rate of 88.2% at mid-training and 88.9% at final evaluation – outperforming the RL-VLM-F baseline by a margin of 23.3%.

The performance gap is particularly significant in contact-rich tasks requiring precise reward shaping, such as *drawer-open* and *button-press-topdown*. This indicates that our distinguishability-aware query selection and reward transition enable the reward model to capture task progress more promptly than random sampling, and when equipped with efficient query selection, **SPEAR** can refine reward signals

to a level far exceeding that provided by the coarse VLM synthesis function.

On *window-open* task, Text2Reward achieves a slightly higher final success rate (87.7% versus 84.9%). Analysis reveals this task’s reward structure aligns particularly well with VLM-synthesized distance-based objectives, but considering the overall task performance, the synthesized reward is insufficient to support task completion stably and reliably. This suggests that our preference-based approach has a wider applicability and can be applied to various task semantics.

2) *Query Efficiency Analysis*: Beyond task success, we evaluate the practical efficiency of online training. Because oracle queries are scheduled at a fixed cadence (Sec. V-A.3), the total query count is proportional to environment steps; thus, improved success rate at the same number of environment steps directly implies improved query-efficiency. We choose *drawer-open* and compare **SPEAR** against RL-VLM-F at a matched 30% success-rate level in the setting without reward transition, see Table III. **SPEAR** uses 960 total queries with 29.6% abstention, compared to RL-VLM-F’s 1640 queries with 55.5% abstention, resulting in substantially fewer wasted comparisons while achieving faster learning. In terms of wall-clock, although **SPEAR** incurs additional overhead to train the distinguishability predictor at each query round, the total time to reach 30% success is reduced from 13.6 hours to 9.1 hours. This is because the savings from avoiding wasted queries more than offset the overhead of distinguishability-model training.

3) *VLM Reliability Diagnostics*: A key concern is whether VLM feedback can exhibit a “confident but incorrect” behavior when the VLM does not abstain. We therefore run a controlled diagnostic study on five VLM backbones (Qwen3-VL-8B/30B-A3B, InternVL3.5-8B/30B-A3B, and InternVL3-78B [8], [9], [11]) to quantify abstention versus correctness under different visual distinguishability.

For each task, we vary visual distinguishability by altering the frame gap in a single demo, and test both *radical* and *conservative* prompts (adding a style suffix to the original prompt). Importantly, these demos are used only for offline analysis and never for online training or query collection.

We find that our decomposed protocol markedly eliminates

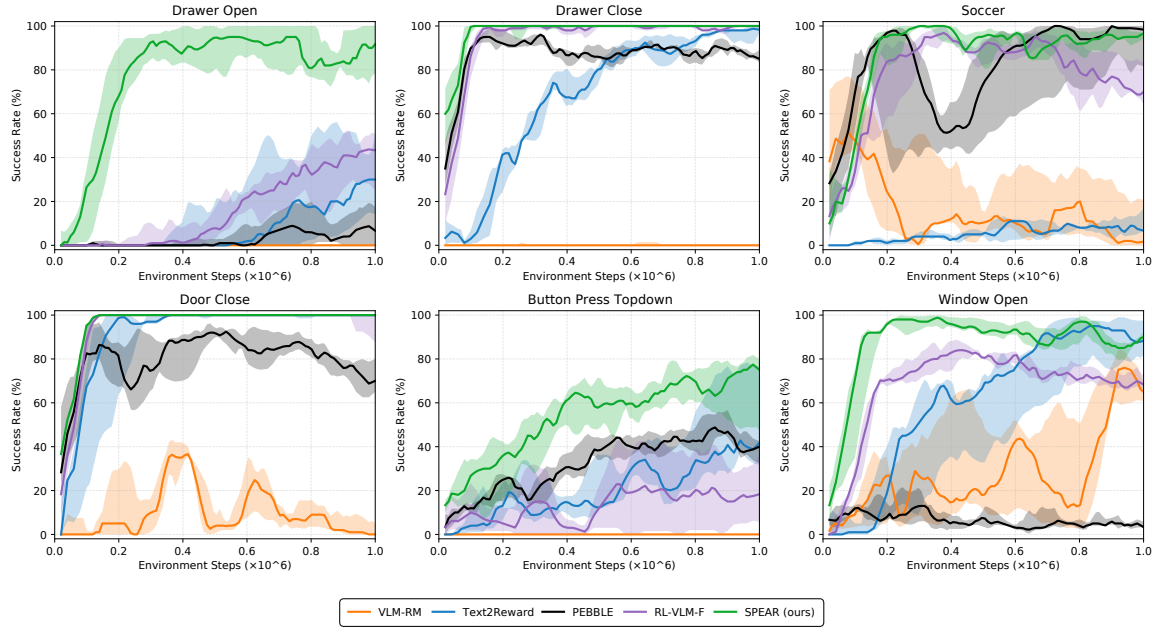


Fig. 2: **Learning curves across six Meta-World tasks.** Our method (green) consistently outperforms baselines in both mid-range and final performance. Shaded regions indicate standard deviation over 5 seeds.

TABLE III: **Query efficiency and wall-clock cost at 30% success.** *SPEAR* effectively reduces the abstention rate and lowers the training wall-clock time overhead.

Method	Query Efficiency			Time Consumption		
	Total Q	Abst. Rate	Effective Q	Aver. Q (s)	DistTrain (s)	Total (h)
RL-VLM-F	1640	0.555	729	29.87	–	13.6
<i>SPEAR (ours)</i>	960	0.296	676	31.42	114.5	9.1

TABLE IV: **Diagnostics of averaged 5 VLM backbones with conservative prompt.** Distinctiveness increases $\Rightarrow$ abstention decreases and accuracy increases.

Frame Gap	Metric (%)	<i>Drawer-Open</i>	<i>Door-Close</i>	<i>Soccer</i>	<i>Button-Press</i>	Avg
1–5	Abstain	61.2	88.8	77.2	76.6	75.9
	Acc	61.3	75.0	53.7	56.0	61.5
6–10	Abstain	61.6	87.5	64.6	68.1	70.5
	Acc	79.6	100.0	91.3	68.6	84.9
11–20	Abstain	49.2	82.0	72.7	65.6	67.4
	Acc	91.3	100.0	100.0	97.0	97.1
21–30	Abstain	29.5	60.7	–	54.7	48.3
	Acc	100.0	100.0	–	97.7	99.2
31–max	Abstain	11.6	–	–	45.0	28.3
	Acc	99.4	–	–	100.0	99.7

the systematic positional bias and feature bleeding caused by simple joint image pairwise cues. Table IV shows that the abstention rate consistently decreases with increasing visual distinguishability, while accuracy steadily improves, reaching 100% at sufficiently large frame gaps. Under the *conservative* prompting style, the average preference accuracy of the five backbones reached 92.3%, demonstrating that VLM preference feedback is reliably accurate in our setup with an appropriate query formula.

C. Ablation Studies

1) *Validity of Simulated Preference Oracle:* To run large-scale ablations without repeatedly invoking expensive VLM inference, we use a calibrated *simulated* preference oracle that emulates the aggregate behavior of five open-source

VLM backbones (as diagnosed in Sec. V-B.3). We model the oracle as a stochastic labeler: it abstains with probability $p_{\text{abst}}(\cdot)$; otherwise it outputs the correct preference label with probability $p_{\text{acc}}(\cdot)$ or flips it with probability $1 - p_{\text{acc}}(\cdot)$. We fit smooth functions $p_{\text{abst}}(g)$ and $p_{\text{acc}}(g)$ from aggregated diagnostic statistics via a binomial Generalized Additive Model (GAM) with B-spline smooths, as illustrated in Figure 3. During training, we use the return difference ΔR from simulation reward between two sub-trajectories as a practical proxy for distinguishability and linearly map ΔR to g using a per-task reward range estimated from demonstrations. While viewing ΔR as visual distinguishability is heuristic, this stochastic labeler accurately aligns with the monotonic trend of VLM behavior in Sec. V-B.3. This simulated oracle is used *only* in the ablation section.

2) *Query Acquisition Mechanisms:* Table V compares four query selection strategies on *drawer-open*. Random selection yields only 14.0% final success, confirming that uninformed sampling provides insufficient learning signal. Entropy-only (standard BALD) improves marginally to 21.0% but often selects pairs that are highly uncertain yet practically unanswerable (high abstention probability).

Distinguishability-only selection achieves 69.7% success by avoiding abstentions, but plateaus as it favors easily distinguishable pairs that may lack information gain. Our integrated approach synthesizes both criteria, reaching 80.7% final success with 58.4% at mid-training. This validates that jointly optimizing answerability and informativeness is essential for efficient preference-based learning.

3) *Distinguishability Model Validation:* We additionally isolate and evaluate the distinguishability predictor $\mathcal{D}_\psi$ using oracle-response data collected from *drawer-open* to validate its predictive accuracy. We report the model’s Mean Squared

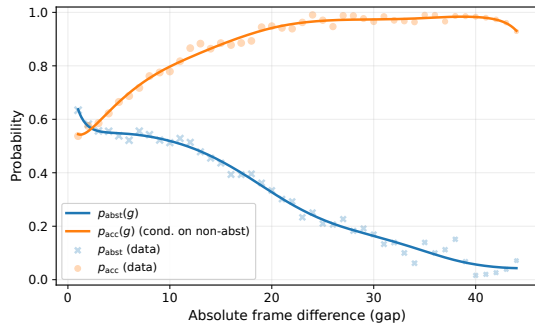


Fig. 3: **Average fit probability of 5 VLM backbone behaviors.** As visual distinguishability increases, the abstention rate decreases and the accuracy increases.

TABLE V: **Ablation on query acquisition mechanisms.** Integrated weighting (ours) achieves the best performance.

Method	Mid (450k–550k)	End (900k–1000k)
Random	$1.1 \pm 1.8 \%$	$14.0 \pm 15.9 \%$
Entropy-only	$3.3 \pm 4.2 \%$	$21.0 \pm 19.0 \%$
Distinguishability-only	$21.9 \pm 19.8 \%$	$69.7 \pm 24.3 \%$
Integrated (ours)	$58.4 \pm 23.8 \%$	$80.7 \pm 13.2 \%$

Error (MSE, i.e., Brier score) and Mean Absolute Error (MAE) to evaluate the model’s predictive performance.

Table VII shows that using multi-scale features achieves the lowest prediction error (Brier score 0.1508, MAE 0.2950), validating that the synergy between global semantic understanding and local spatial discrimination is critical, while applying temperature scaling improves calibration quality and reduces both MSE and MAE. The combined effect of both directly translates to more effective query selection and a lower abstention rate during online training.

TABLE VII: **Feature/calibration ablation.** Multi-scale features and temperature scaling improve prediction reliability.

Configuration	MSE↓	MAE↓
Global only (ϕ_{rel})	0.1542	0.2973
Local only (ϕ_{loc})	0.1662	0.3380
Combined (ours)	0.1508	0.2950
w/o calibration	0.1610	0.3232
w/ calibration (ours)	0.1515	0.2978

4) *Exploration Mechanisms:* Table VI exhibits our two-phase exploration strategy. The baseline configuration with random initialization and standard ϵ -greedy exploration achieves only 25.6% final success, illustrating the bootstrapping challenge. Adding reward transition alone improves mid-training performance to 39.8%, confirming that semantic bootstrapping jumpstarts exploration toward task-relevant regions. However, after the transition, the agent still faces the challenge of efficiently discovering better states beyond the guided regime.

Relay exploration alone maintains distributional coverage, but it cannot replace the directional guidance from warm-up rewards. The full method combining both mechanisms achieves 92.7% mid-training and 88.3% final success. This synergy emerges because $r_{\text{warm-up}}$ enables initial progress while relay exploration prevents the policy from collapsing

TABLE VI: **Ablation on exploration mechanisms.** Combining two mechanisms achieves the best performance.

Method	Mid (450k–550k)	End (900k–1000k)
Baseline	$1.7 \pm 2.5 \%$	$24.8 \pm 16.2 \%$
+ Relay Exploration	$26.1 \pm 21.6 \%$	$73.6 \pm 19.5 \%$
+ Reward Transition	$39.8 \pm 17.8 \%$	$67.0 \pm 21.9 \%$
Full Exploration (ours)	$92.7 \pm 8.9 \%$	$88.3 \pm 11.7 \%$

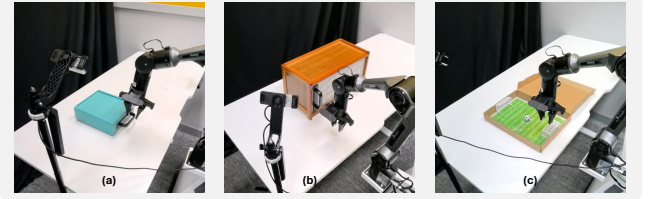


Fig. 4: **Real-robot experiment task settings.** (a) *drawer-open*. (b) *window-open*. (c) *soccer*.

to narrow behaviors, ensuring sufficient trajectory diversity for meaningful preference comparisons.

D. Real-Robot Experiments

We evaluate *SPEAR* on a real Mobile ALOHA robot, running the full online exploration and preference-based learning pipeline directly in the physical environment. Therefore, our experiment does not involve sim-to-real transfer.

1) *Platform and Tasks:* We use the left arm of Mobile ALOHA with a fixed side-view RGB camera. The action space matches our simulator setting: Cartesian end-effector increments in (x, y, z) and a gripper open/close command. We evaluate three manipulation tasks: *drawer-open*, *window-open*, and *soccer* (ball pushing), as shown in Figure 4.

2) *Online training protocol:* Each episode consists of 50 exploration steps. At each step, the end-effector motion is bounded by a maximum displacement of 2 cm, executing a step takes approximately 1 s, and we record the current RGB observation. At the end of each episode, the arm returns to an initial pose and waits for 5 s before the next episode begins. As a practical constraint of real-robot deployment, we manually reset the object to the initial configuration between episodes.

Every 10 episodes (500 steps), we enter the preference-query branch and request 20 preference comparisons to update the reward model (and the distinguishability model) following the same overall pipeline as Sec. V-A. We use Qwen3-VL-8B-Instruct [9] as the preference oracle, running on a local GPU workstation.

3) *Results:* *SPEAR* learns all three tasks directly on the physical robot with end-to-end online interaction. The policy converges in 18.2K steps for *drawer-open*, 16.6K steps for *window-open*, and 12.6K steps for *soccer*. Under our fixed real-robot query cadence, this corresponds to 720, 660, and 500 preference comparisons, respectively, demonstrating that the overall pipeline remains practical under real-time constraints.

4) *Safety and failure modes:* A primary failure mode in unconstrained exploration is unsafe contact, where the

robotic arm may apply excessive force to an object (e.g., pushing an object’s handle forcefully). To mitigate this, we (i) restrict the exploration workspace to a safe operating region while preserving sufficient coverage for learning, and (ii) add a small penalty when the commanded end-effector displacement cannot be realized due to collisions or environmental constraints, discouraging repeated unsafe actions. Overall, the experiments demonstrate that *SPEAR* can be deployed on real hardware and in real-world environments as a fully online learning system.

VI. CONCLUSION

We presented *SPEAR*, a practical framework for learning robotic manipulation via VLM-guided preference-based reinforcement learning. By addressing query efficiency and exploration bootstrapping as coupled challenges, our approach enables effective online learning despite the high cost of VLM queries. Experiments demonstrate that *SPEAR* significantly outperforms existing methods in both sample efficiency and final performance across diverse tasks, establishing a foundation for scalable robot learning without manual reward engineering or human annotation.

ACKNOWLEDGMENT

This work was supported by the Early Career Scheme of the Research Grants Council (RGC) under Grant # 27207224, the HKU-100 Award, a donation from the Musketeers Foundation, and in part by the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

REFERENCES

- [1] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.
- [2] C. Wirth, R. Akrou, G. Neumann, and J. Fürnkranz, “A survey of preference-based reinforcement learning methods,” *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.
- [3] Y. Wang, Z. Sun, J. Zhang, Z. Xian, E. Biyik, D. Held, and Z. Erickson, “RI-vlm-f: Reinforcement learning from vision language foundation model feedback,” *arXiv preprint arXiv:2402.03681*, 2024.
- [4] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.
- [5] J. Rocamonde, V. Montesinos, E. Nava, E. Perez, and D. Lindner, “Vision-language models are zero-shot reward models for reinforcement learning,” *arXiv preprint arXiv:2310.12921*, 2023.
- [6] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, “Text2reward: Reward shaping with language models for reinforcement learning,” *arXiv preprint arXiv:2309.11489*, 2023.
- [7] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
- [8] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, “InternV3: Exploring advanced training and test-time recipes for open-source multimodal models,” *arXiv preprint arXiv:2504.10479*, 2025.
- [9] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, “Qwen3 technical report,” *arXiv preprint arXiv:2505.09388*, 2025.
- [10] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, “A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges,” *arXiv preprint arXiv:2501.02189*, 2025.
- [11] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, “InternV3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” *arXiv preprint arXiv:2508.18265*, 2025.
- [12] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran, “Can foundation models perform zero-shot task specification for robot manipulation?” in *Learning for dynamics and control conference*. PMLR, 2022, pp. 893–905.
- [13] P. Mahmoudieh, D. Pathak, and T. Darrell, “Zero-shot reward specification via grounded natural language,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 14 743–14 752.
- [14] S. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Biyik, D. Sadigh, C. Finn, and L. Itti, “Roboclip: One demonstration is enough to learn robot policies,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 55 681–55 693, 2023.
- [15] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever, J. Humpik *et al.*, “Language to rewards for robotic skill synthesis,” *arXiv preprint arXiv:2306.08647*, 2023.
- [16] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, “Eureka: Human-level reward design via coding large language models,” *arXiv preprint arXiv:2310.12931*, 2023.
- [17] Y. Xu, R. Wang, L. Yang, A. Singh, and A. Dubrawski, “Preference-based reinforcement learning with finite-time guarantees,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 784–18 794, 2020.
- [18] K. Lee, L. Smith, A. Dragan, and P. Abbeel, “B-pref: Benchmarking preference-based reinforcement learning,” *arXiv preprint arXiv:2111.03026*, 2021.
- [19] K. Lee, L. Smith, and P. Abbeel, “Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training,” *arXiv preprint arXiv:2106.05091*, 2021.
- [20] J. Park, Y. Seo, J. Shin, H. Lee, P. Abbeel, and K. Lee, “Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning,” *arXiv preprint arXiv:2203.10050*, 2022.
- [21] R. Liu, F. Bai, Y. Du, and Y. Yang, “Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [22] Y. Liu, G. Datta, E. Novoseller, and D. S. Brown, “Efficient preference-based reinforcement learning using learned dynamics models,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2921–2928.
- [23] R. A. Bradley and M. E. Terry, “Rank analysis of incomplete block designs: I. the method of paired comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
- [24] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, “Reward learning from human preferences and demonstrations in atari,” *Advances in neural information processing systems*, vol. 31, 2018.
- [25] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, “Scalable agent alignment via reward modeling: a research direction,” *arXiv preprint arXiv:1811.07871*, 2018.
- [26] Y. Bai, S. Kadavath, S. Kundu, A. Asbell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [27] M. Klissarov, P. D’Oro, S. Sodhani, R. Raileanu, P.-L. Bacon, P. Vincent, A. Zhang, and M. Henaff, “Motif: Intrinsic motivation from artificial intelligence feedback,” *arXiv preprint arXiv:2310.00166*, 2023.
- [28] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, “Bayesian active learning for classification and preference learning,” *arXiv preprint arXiv:1112.5745*, 2011.
- [29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.
- [30] Y. Fu, H. Zhang, D. Wu, W. Xu, and B. Boulet, “Furl: Visual-language models as fuzzy rewards for reinforcement learning,” *arXiv preprint arXiv:2406.00645*, 2024.
- [31] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International conference on machine learning*. Pmlr, 2018, pp. 1861–1870.