# Mitigating Context Divergence in Multi-Turn Embodied Reasoning via Hierarchical Decay

Gangao Liu
*Institute of Software, Chinese Academy of Sciences*
*University of Chinese Academy of Sciences*
Beijing, China
liugangao2023@iscas.ac.cn

Mengna Wang
*Institute of Software, Chinese Academy of Sciences*
*University of Chinese Academy of Sciences*
Beijing, China
wangmengna2024@iscas.ac.cn

Peng Li*
*Institute of Software, Chinese Academy of Sciences*
*University of Chinese Academy of Sciences*
Beijing, China
lipeng@iscas.ac.cn

***Abstract*—The integration of Large Language Models (LLMs) into embodied agents has enabled impressive capabilities in zero-shot task planning. While Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) have proven effective for aligning LLMs in dialogue tasks, applying them to multi-turn embodied reasoning introduces a critical challenge: Context Divergence. In sequential decision-making, a divergent action alters all subsequent observations, rendering the comparison between chosen and rejected trajectories mathematically ill-posed due to state drift. In this paper, we propose H-D$^2$PO , a novel alignment framework specifically designed for embodied agents. H-D$^2$PO introduces a hierarchical temporal decay mechanism that dynamically calibrates the optimization objective, prioritizing the causal branching decision while robustly filtering the noise from the divergent trajectory tail. Furthermore, we propose a Priority-based Preference Construction strategy that explicitly prioritizes long-horizon failures over short ones, encouraging the agent to learn reasoning stability even from unsuccessful attempts. Extensive experiments on ALFWorld and LOGICWorld demonstrate that H-D$^2$PO significantly outperforms SFT and vanilla DPO baselines. Notably, on ALFWorld, our Qwen3-4B based agent achieves a success rate of 89.78%. Our analysis reveals that this performance gain stems from the agent's enhanced capability to persist through and solve complex, long-horizon tasks that baselines fail to complete.**

***Index Terms*—Embodied Agent, Large Language Model, Direct Preference Optimization**

## I. Introduction

The integration of Large Language Models (LLMs) [1], [2] into embodied agents has revolutionized the field of robotics and embodied AI [3]–[5]. By leveraging the vast world knowledge and reasoning capabilities of LLMs, embodied agents can now interpret complex natural language instructions, decompose long-horizon tasks, and interact with physical environments in a zero-shot or few-shot manner [6]–[10]. The dominant paradigm for training such agents typically involves Supervised Fine-Tuning (SFT) on expert demonstrations [11]–[14]. While effective for initializing basic capabilities, SFT-based agents often struggle with distribution shift and lack the ability to correct errors or optimize for long-term success, particularly in tasks characterized by long horizons and sparse rewards.

To address these limitations, Reinforcement Learning from Human Feedback (RLHF) [15], [16] and its more stable variant, Direct Preference Optimization (DPO) [17], have been adapted to align embodied agents with task objectives and physical constraints [18]–[20]. However, directly applying vanilla DPO originally designed for static dialogue contexts to multi-turn embodied reasoning tasks presents a unique and critical challenge: **Context Divergence**.

Vanilla DPO assumes that the chosen response $y_w$ and the rejected response $y_l$ share the same prefix context $x$. In single-turn dialogue generation, this assumption holds naturally as the optimization is restricted to the immediate branching step where histories are identical. However, embodied interaction is a sequential decision-making process where the environment's feedback or the agent observation depends strictly on the agent's previous actions. As illustrated in Fig. 1, once the agent's action diverges at a specific timestep $t^*$, the subsequent observations for the chosen and rejected trajectories become fundamentally different ($o^w_{t+1} \neq o^l_{t+1}$). Consequently, the history contexts for future steps drift apart ($h^w_{t+1} \neq h^l_{t+1}$).

This phenomenon leads to two detrimental effects. First, comparing actions under different environmental states is mathematically ill-posed, as the optimal policy is state-dependent. Second, as shown in Fig. 1 Entropy Phenomenon, we observe a **Local Entropy Spike** at the onset of each interaction turn following the divergence. These spikes correspond to the injection of divergent environmental observations, corroborating that context divergence introduces noise during multi-turn optimization. Vanilla Multi-Turn DPO, which treats

* indicates Corresponding author. Peng Li is also with University of Chinese Academy of Sciences, Nanjing and Hangzhou Institute for Advanced Study, UCAS.

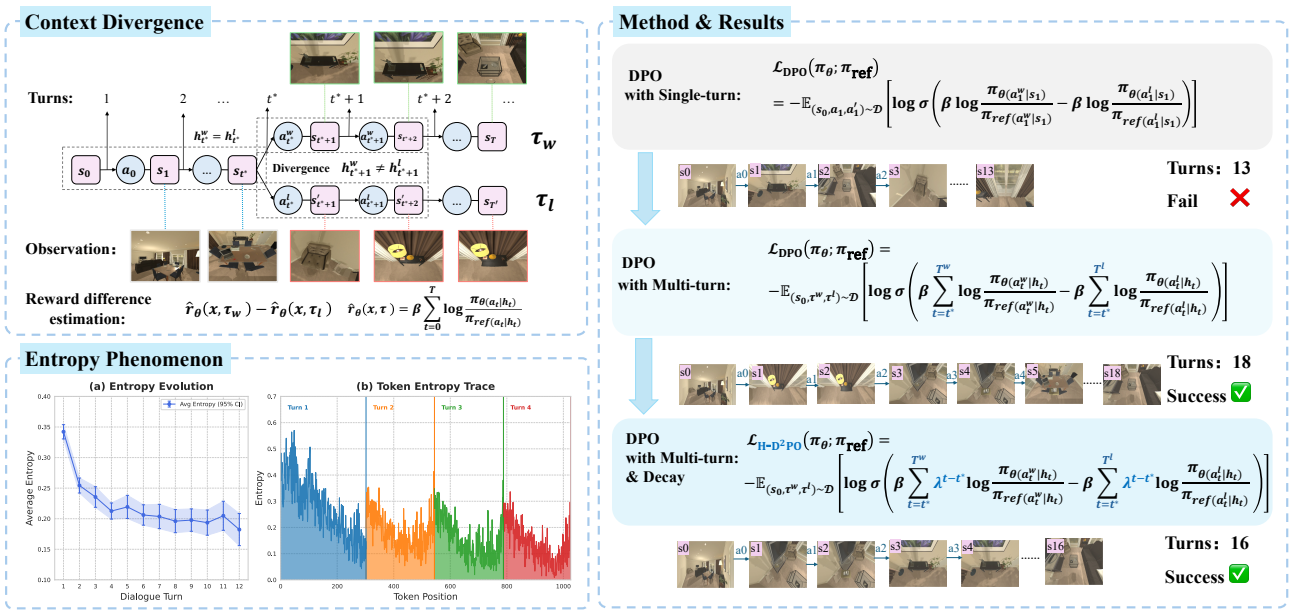


Fig. 1. **Overview of Context Divergence and the H-D$^2$PO Framework. Left:** The challenge of Context Divergence caused by inserted observations in multi-turn embodied reasoning. The Entropy Phenomenon reveal that while average sentence-level entropy decreases across turns, **local entropy spikes** occur at the onset of each interaction turn. These spikes correspond to the injection of environmental observations, corroborating that context divergence introduces noise during multi-turn optimization. **Right:** Comparison of loss functions and resulting trajectories. Single-Turn DPO fails to handle long horizons, while Multi-Turn DPO suffers from noise interference. By employing turn-wise weight decay, H-D$^2$PO effectively suppresses divergence noise, achieving the highest performance.

the tokens of each turn equally, forces the model to overfit this environmental noise rather than learning the robust policy logic.

In this paper, we propose **H-D$^2$PO (Hierarchical Temporal Decay DPO)**, a novel alignment framework designed specifically for multi-turn embodied reasoning. Our core insight is that the reliability of preference comparison decays as the context diverges. H-D$^2$PO introduces a Hierarchical Temporal Decay mechanism that dynamically re-weights the optimization objective at the turn level. Specifically, we identify the branching point $t^*$ where the trajectories first diverge and assign the highest weight to this critical decision. For subsequent steps, we apply a decay factor that exponentially reduces the gradient contribution based on the temporal distance from the branching point. This mechanism ensures that the optimization focuses on the root cause of the branching action while suppressing the noise from the divergent tail.

Furthermore, to facilitate effective learning under sparse rewards, we also propose a data construction strategy that prioritizes long failure trajectories over short ones, encouraging the agent to value deep exploration and reasoning stability even in failed attempts.

Our main contributions are summarized as follows:

- We identify the **Context Divergence** dilemma in embodied DPO: extending preference optimization to multi-turn trajectories allows for long-term credit assignment, but inevitably introduces significant gradient noise due to the drift in observation histories.
- We propose **H-D$^2$PO** , a turn-wise alignment framework that employs Hierarchical Temporal Decay to dynamically down-weight divergent steps, effectively isolating the branching decisions from environmental noise.
- We introduce a priority-based data construction strategy that leverages the exploration value of long failure trajectories, enhancing reasoning robustness in sparse-reward environments.
- Extensive experiments on ALFWorld [21] and LOGICWorld [22] demonstrate that H-D$^2$PO achieves state-of-the-art performance, surpassing vanilla multi-turn DPO and validating the importance of stabilizing divergence in long-horizon tasks.

## II. Method

### A. Preliminaries: Embodied DPO

We model the embodied task as a POMDP where the policy $\pi_\theta(a_t|h_t)$ maps the interaction history $h_t = \{I, o_0, a_0, \ldots, o_t\}$ to an action.The embodied alignment problem can be modeled as optimizing a policy $\pi_\theta$ to maximize the expected return. DPO [17] reparameterizes the optimal policy $\pi^*$ in terms of a reward function $r(x, y)$ and a reference policy $\pi_{ref}$:

$$r(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{ref}(y|x)} + Z(x) \tag{1}$$

The vanilla DPO objective minimizes the negative log-likelihood of the Bradley-Terry model:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l)\right)\right] \tag{2}$$

where the implicit reward estimator is $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$.

### B. Derivation of H-D$^2$PO

**The Context Divergence Problem.** In multi-turn embodied tasks, a trajectory $\tau$ consists of a sequence of state-action pairs $\{(s_t, a_t)\}_{t=0}^{T}$. The standard implicit reward $\hat{r}_\theta(x, \tau)$ accumulates the log-ratios over the entire sequence:

$$\hat{r}_\theta(x, \tau) = \beta \sum_{t=0}^{T} \log \frac{\pi_\theta(a_t|h_t)}{\pi_{ref}(a_t|h_t)} \tag{3}$$

However, for a preference pair $(\tau_w, \tau_l)$ diverging at $t^*$, the interaction histories $h_t$ for $t > t^*$ are sampled from different distributions induced by the environment dynamics $P(s_{t+1}|s_t, a_t)$. Specifically, the observation $o_{t+1}$ depends on $a_t$, leading to $h_t^w \neq h_t^l$. This *Context Divergence* implies that the reward contributions from the tail ($t > t^*$) are conditioned on disparate states, introducing high variance into the reward difference estimation $\hat{r}_\theta(x, \tau_w) - \hat{r}_\theta(x, \tau_l)$.

**Hierarchical Temporal Decay Objective.** To mitigate the impact of context divergence, we propose to model the preference probability as depending on a **Temporally Decayed Implicit Reward** $R_\gamma$. We postulate that the reliability of the reward signal decays as the trajectory drifts from the common branching point due to accumulated aleatoric uncertainty.

Formally, for a trajectory $\tau = \{(s_t, a_t)\}_{t=0}^{T}$ involved in a preference pair with branching point $t^*$, we define the decayed implicit reward $R_\gamma(\tau)$ as the reliability-weighted sum of step-wise implicit rewards:

$$R_\gamma(\tau) = \beta \sum_{t=0}^{T} \gamma(t) \log \frac{\pi_\theta(a_t|h_t)}{\pi_{ref}(a_t|h_t)} \tag{4}$$

Here, $\gamma(t) \in [0, 1]$ acts as a hierarchical confidence coefficient. Specifically, we define $\gamma(t)$ based on the temporal distance from the branching point $t^*$:

$$\gamma(t) = \mathbb{I}(t < t^*) \cdot 0 + \mathbb{I}(t \geq t^*) \cdot \lambda^{t-t^*} \tag{5}$$

where $\lambda \in (0, 1]$ is the decay hyperparameter. The term $\mathbb{I}(t < t^*) \cdot 0$ reflects that identical prefixes contribute zero information to the preference difference. For divergent steps $t \geq t^*$, the weight decays exponentially from $\lambda^0 = 1$ to suppress noise from the diverging context.

By substituting this decayed reward formulation into the Bradley-Terry preference model, i.e., $P(\tau_w \succ \tau_l) = \sigma(R_\gamma(\tau_w) - R_\gamma(\tau_l))$, we arrive at the **H-D$^2$PO objective**:

$$\begin{aligned} \mathcal{L}_{\text{H-D}^2\text{PO}} &= -\mathbb{E}_{(x,\tau_w,\tau_l)\sim\mathcal{D}}\Big[\log P(\tau_w \succ \tau_l)\Big] \\ &= -\mathbb{E}_{(x,\tau_w,\tau_l)\sim\mathcal{D}}\Big[\log \sigma\Big(R_\gamma(\tau_w) - R_\gamma(\tau_l)\Big)\Big] \end{aligned} \tag{6}$$

This formulation highlights that H-D$^2$PO effectively optimizes a discounted value function relative to the decision bifurcation, prioritizing the causal root of the outcome while robustly handling long-horizon drift.

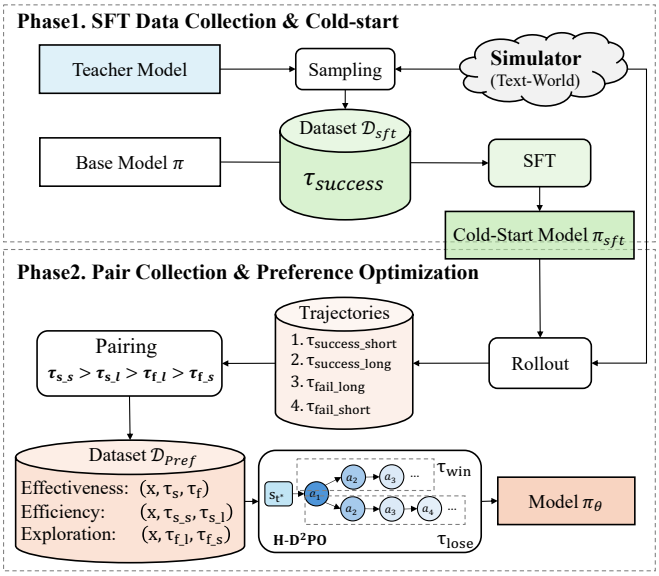


Fig. 2. **Overview of the H-D$^2$PO Training Pipeline.** The framework proceeds in two phases: (1) **Cold-Start SFT**, where the base model is initialized using expert demonstrations distilled from a teacher model; and (2) **Priority-based Preference Learning**. We generate diverse rollouts and construct preference pairs according to a strict hierarchy. Finally, the policy is aligned using **H-D$^2$PO** , which applies temporal decay to mitigate context divergence in the trajectory tail.

### C. Training Pipeline

As shown in Fig. 2, our alignment framework consists of two phases: *Cold-Start Supervised Fine-Tuning* (SFT) for capability initialization, and *Priority-based Preference Learning* for robust policy alignment. This two-stage training ensures the agent first learns valid primitives before optimizing for long-horizon stability.

*1) Phase 1: Cold-Start SFT via Distillation:* To equip the base LLM with fundamental instruction-following and environment interaction capabilities, we employ a knowledge distillation approach. We utilize a powerful teacher model (**DeepSeek-V3.2-Exp**) to generate expert trajectories. Given a task instruction $I$, the teacher interacts with the environment to produce a trajectory $\tau_{exp} = \{(s_t, a_t)\}_{t=0}^{T}$. We filter for successful trajectories and treat them as ground-truth demonstrations. The base model $\pi_\theta$ is then SFT to minimize the standard negative log-likelihood loss:

$$\mathcal{L}_{SFT} = -\mathbb{E}_{\tau\sim\mathcal{D}_{exp}}\left[\sum_{t=0}^{T} \log \pi_\theta(a_t|I, s_0, a_0, \ldots, s_t)\right] \tag{7}$$

The resulting policy, denoted as $\pi_{sft}$, serves as the reference policy $\pi_{ref}$ and the starting point for exploration in the subsequent phase.

*2) Phase 2: Priority-based Preference Construction:* Unlike standard dialogue tasks, embodied alignment requires the agent to understand not just "what is right," but "what constitutes a better attempt." We propose a priority-based strategy to construct the preference dataset $\mathcal{D}_{pref}$.

**Explorative Rollout.** We prompt $\pi_{sft}$ to interact with the environment using temperature sampling ($T = 1.5$) to generate 5 diverse trajectories $\{\tau_1, \ldots, \tau_5\}$ for each instruction. As shown in Fig. 3, we categorize these trajectories into four buckets based on their outcome (Success $S$ / Failure $F$) and horizon length (Short / Long): $T_{S\_S}, T_{S\_L}, T_{F\_S}, T_{F\_L}$. **Priority Pairing Logic.** We construct preference pairs $(\tau_w, \tau_l)$

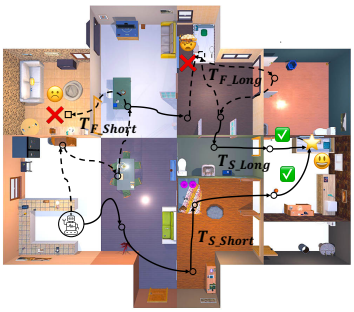


Fig. 3. **Visualization of Trajectory Types.**

according to a strict hierarchy of capabilities we wish to instill:

- **Effectiveness** ($S \succ F$)**:** The most fundamental requirement. Any successful trajectory is strictly preferred over a failed one to anchor goal completion.
- **Efficiency** ($T_{S\_S} \succ T_{S\_L}$)**:** To discourage redundant actions (e.g., navigation loops or indecisive planning), we prefer shorter successful paths over longer ones.
- **Exploration** ($T_{F\_L} \succ T_{F\_S}$)**:** This is a critical design choice for sparse-reward environments. A long failure ($T_{F\_L}$) typically implies the agent successfully grounded itself and reasoned through multiple steps before an error, whereas a short failure ($T_{F\_S}$) often indicates hallucination or immediate syntax errors. By optimizing $T_{F\_L} \succ T_{F\_S}$, we encourage *reasoning stability* and prevent the policy from collapsing into conservative "early-quit" behaviors.

## III. Experiment Setup

### A. Benchmarks and Metrics

We conduct comprehensive evaluations on two challenging embodied reasoning benchmarks that require multi-turn interaction and long-horizon planning.

**ALFWorld** [21] is a text-based environment aligned with the ALFRED benchmark. It encompasses six diverse task types and 274 total tasks(140 seen, 134 unseen).

**LOGICWorld** [22] is designed to evaluate complex logical reasoning in embodied scenarios. LOGICWorld introduces intricate constraints and dependencies (e.g., "Put the eggs in the refrigerator. If you see lettuce and tomatoes, wash them and put them in the refrigerator together."), demanding rigorous state tracking and causal reasoning over longer horizons. Our evaluation encompasses 1,082 tasks spanning four categories, each situated in unseen scenarios to assess the model's zero-shot generalization.

**Metrics.** The primary evaluation metric is **Success Rate (SR)**. To evaluate execution efficiency, we also report the **Average Turns** for successful episodes.

### B. Dataset Statistics

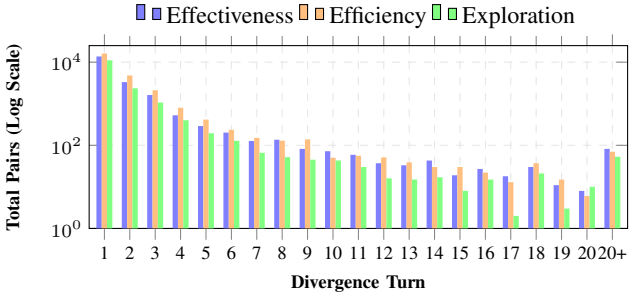


Fig. 4. **Distribution of Preference Pairs.**

**SFT Dataset.** From the raw rollouts generated by DeepSeek-V3.2-Exp, we filter out failed attempts and infinite loops. The final high-quality SFT dataset consists of approximately **23k** successful trajectories, covering all task types in ALFWorld and LOGICWorld. This dataset serves to initialize the instruction-following capability.

**Preference Pair Distribution.** Constructing a balanced preference dataset is crucial for H-D$^2$PO . Fig. 4 illustrates the distribution of the constructed preference pairs with respect to the Divergence Turn $t^*$ and pair types . We observe a natural long-tail distribution: over 70% of raw divergences occur at the very first turn ($t^* = 1$). After we downsampled the data for Turn 1 pairs, the final dataset comprises approximately **60k** preference pairs.

### C. Implementation Details

**Models.** We employ the **Qwen3** family as our backbone, specifically evaluating on **Qwen3-0.6B** and **Qwen3-4B** to verify the scalability of our approach across different model sizes. Due to the high computational cost of long-horizon embodied tasks, we utilize Low-Rank Adaptation (LoRA) [23] for parameter-efficient fine-tuning in both phases.

**Hyperparameters.** We utilize the **AdamW** optimizer with a cosine learning rate scheduler. Both SFT and DPO phases are trained for 3 epochs with a global batch size of 16, LoRA rank 8, and LoRA alpha 16. For *SFT*, we set the learning rate to 1e-4 and the maximum context length to **16k tokens** to accommodate extensive interaction histories. For *DPO*, we reduce the learning rate to 5e-6 and set the context length to **12k tokens**. The KL coefficient is $\beta = 0.1$. For H-D$^2$PO , the temporal decay factor is set to $\lambda = 0.9$ based on ablation studies. All experiments are conducted on 2×NVIDIA A100 (80GB) GPUs using DeepSpeed ZeRO-3.

### D. Baselines

We compare H-D$^2$PO against the following baselines to validate the effectiveness of our contributions:

TABLE I
PERFORMANCE COMPARISON ON ALFWORLD AND LOGICWORLD

| Models | ALFWorld SR(%) | | | | | | | LOGICWorld SR(%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Pick* | *Look* | *Clean* | *Heat* | *Cool* | *Pick2* | *All* | *Multi-goal* | *Condition* | *Priority* | *Logic* | *All* |
| Qwen3-0.6B | 6.45 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.73 | 0.00 | 0.00 | 6.00 | 6.39 | 4.81 |
| Qwen3-0.6B-SFT | 88.14 | 51.61 | 74.14 | 84.62 | 67.39 | 34.15 | 68.98 | 24.00 | 61.67 | 12.67 | 44.17 | 42.79 |
| Qwen3-0.6B-SFT+DPO(single-turn) | 91.53 | 48.39 | 89.66 | 82.05 | 67.39 | 70.73 | 77.74 | 32.00 | 65.33 | 24.67 | 60.15 | 54.07 |
| Qwen3-0.6B-SFT+DPO(multi-turn) | **98.31** | 54.84 | **93.10** | 76.92 | 86.96 | 58.54 | 81.39 | 34.00 | **67.67** | 24.67 | 60.53 | 55.08 |
| Qwen3-0.6B-SFT+H-D$^2$PO | 96.61 | **58.06** | **93.10** | **84.62** | **89.13** | **70.73** | **84.67** | **38.00** | **67.67** | **33.33** | **63.91** | **58.32** |
| Qwen3-4B | 83.05 | 16.13 | 58.62 | 26.09 | 23.08 | 39.02 | 45.62 | 17.00 | 32.00 | 6.67 | 15.23 | 18.85 |
| Qwen3-4B-SFT | 94.92 | 51.61 | 87.93 | 71.79 | 67.39 | 70.73 | 77.01 | 44.0 | 76.33 | 48.00 | 76.50 | 69.50 |
| Qwen3-4B-SFT+DPO(single-turn) | 96.61 | 74.19 | 91.38 | **82.05** | 80.43 | 82.93 | 86.13 | **52.00** | 82.33 | 58.00 | 76.32 | 73.20 |
| Qwen3-4B-SFT+DPO(multi-turn) | 98.31 | 77.42 | 89.66 | **82.05** | 78.26 | **90.24** | 87.23 | 47.00 | 81.33 | **61.33** | **78.57** | 74.03 |
| Qwen3-4B-SFT+H-D$^2$PO | **100** | **80.65** | **93.10** | 79.49 | **86.96** | **90.24** | **89.78** | 51.00 | **84.67** | **61.33** | 77.82 | **74.95** |

- **Zero-shot:** We evaluate the intrinsic capability of the base Qwen3 models without any task-specific fine-tuning, relying solely on prompt engineering.
- **SFT-Only:** The model fine-tuned on the expert demonstrations collected from the teacher model.
- **Single-Turn DPO ($\lambda = 0$):** A hard-cutoff DPO strategy where we strictly train on the branching step ($t = t^*$) and discard all subsequent steps in the trajectory. This baseline avoids context divergence noise but fails to utilize signals from the trajectory tail.
- **Multi-Turn DPO ($\lambda = 1$):** A naive application of DPO on the entire divergent trajectory ($t \geq t^*$) with uniform weights. This utilizes maximum data quantity but suffers from high gradient variance due to context divergence.

## IV. EXPERIMENT RESULTS

### A. Main Performance

Table I summarizes the success rates across ALFWorld and LOGICWorld benchmarks using Qwen3-0.6B and Qwen3-4B. On the ALFWorld benchmark, **Qwen3-4B-SFT-H-D$^2$PO** achieves a remarkable **89.78%** average success rate, surpassing the SFT baseline (77.01%) by nearly 13 absolute percentage points. More importantly, it consistently outperforms both Single-Turn DPO (86.13%) and Multi-Turn DPO (87.23%). This result validates our core hypothesis: while the trajectory "tail" contains valuable supervision signals, naively using it introduces significant variance. H-D$^2$PO effectively filters this noise via temporal decay, capturing the best of both worlds.

**Robustness in Complex Logic.** The advantages of H-D$^2$PO are even more pronounced in LOGICWorld, which demands rigorous state tracking. In the challenging "Condition" and "Logic" sub-tasks, Qwen3-4B-H-D$^2$PO achieves **84.67%** and **77.82%** respectively. Multi-Turn DPO struggles here (e.g., 78.57% in Logic is comparable, but significantly lower in Condition). We attribute this to the fact that context divergence in logical puzzles often leads to invalid reasoning chains in the tail. By down-weighting these divergent steps, H-D$^2$PO maintains logical consistency.

**Scalability across Model Sizes.** The performance gains are consistent across model scales. For the smaller Qwen3-0.6B, H-D$^2$PO improves SFT performance from 68.98% to 84.67% on ALFWorld. This suggests that the benefits of resolving context divergence are algorithmic and agnostic to model parameter scale.

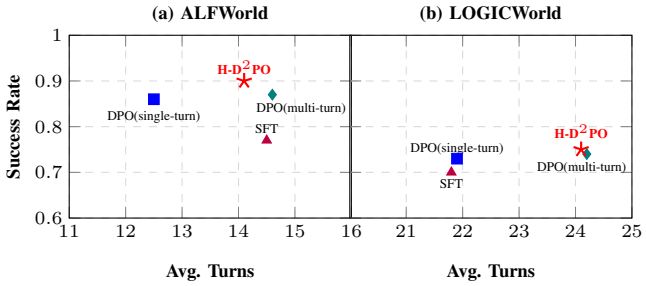


Fig. 5. **Success Rate vs. Efficiency Analysis (Qwen3-4B).** We visualize the trade-off between task performance (SR ↑) and execution cost (Avg. Turns ↓) on (a) ALFWorld and (b) LOGICWorld.

### B. Efficiency Analysis.

We analyze the behavioral characteristics of different alignment strategies by plotting Success Rate against Average Turns in Fig. 5.

- **Single-Turn DPO:** On ALFWorld, this baseline achieves reasonable success with the lowest turn count (12.5). This suggests it performs well on simpler tasks but lacks the robustness to handle complex, long-horizon scenarios found in LOGICWorld.
- **Multi-Turn DPO vs. H-D$^2$PO :** Notably, both Multi-Turn DPO and H-D$^2$PO exhibit higher average turn counts (∼14-15 and 24 turns), indicating that utilizing the trajectory tail encourages the agent to engage in longer reasoning chains. However, with similar path lengths, **H-D$^2$PO** consistently achieves higher success rates (e.g., +3% on ALFWorld). This implies that while Multi-Turn DPO attempts long tasks, it often fails due to the noise from context divergence. In contrast, H-D$^2$PO 's temporal decay mechanism effectively filters this noise, ensuring that the agent's persistence translates into actual task success rather than futile long-horizon wandering.

### C. Ablation Study: Data Strategy

We investigate the impact of our Priority-based Preference Data Construction strategy (Sec. II-C) in Table II. We conduct ablations on Qwen3-0.6B by removing specific types of preference pairs.

TABLE II
PERFORMANCE COMPARISON ON PREFERENCE PAIR COMPOSITION.

| Configuration | ALFWorld | | LOGICWorld | |
|---|---|---|---|---|
| | ***SR (%)*** | ***Turns*** | ***SR (%)*** | ***Turns*** |
| **H-D$^2$PO (All)** | **84.67** | 12.39 | **58.32** | 52.65 |
| - w/o Exploration ($F_L \succ F_S$) | 79.93 | **11.89** | 57.39 | **38.48** |
| - w/o Efficiency ($S_S \succ S_L$) | 80.29 | 14.92 | 55.40 | 63.10 |

**Preventing Premature Failure (Exploration Pairs).** Removing the exploration pairs ($F_L \succ F_S$) results in the most significant performance drop on ALFWorld (84.67% $\rightarrow$ 79.93%). Crucially, this drop is accompanied by a sharp decrease in Average Turns (15.57 $\rightarrow$ 13.83). This correlation strongly suggests that without learning that "long failures are better than short failures," the policy giving up early (e.g., outputting `END` action) when facing uncertainty. Including these exploration pairs is essential for incentivizing the agent to attempt deep reasoning chains. This learned deep exploration provides the necessary opportunity for the agent to eventually correct itself and solve complex, long-horizon tasks.

**Regularizing Path Redundancy (Efficiency Pairs).** Removing the efficiency pairs ($S_S \succ S_L$) leads to a notable increase in Average Turns (15.57 $\rightarrow$ 17.44), along with a drop in success rate. This confirms that explicitly preferring shorter successful paths effectively regularizes the agent. It discourages redundant actions (e.g., repeated navigation) which not only waste time but also increase the probability of execution errors over long horizons.

## V. CONCLUSION

In this paper, we mitigate the critical challenge of **Context Divergence** in embodied alignment. We propose **H-D$^2$PO** , which employs Hierarchical Temporal Decay to attenuate the optimization noise caused by state drift. This is coupled with a novel priority-based data strategy that leverages deep exploration failures to prevent policy collapse. Extensive experiments on ALFWorld and LOGICWorld demonstrate that H-D$^2$PO significantly outperforms existing baselines, establishing a new state-of-the-art in both success rate and reasoning robustness for long-horizon embodied tasks. Future work will extend this framework to online iterative learning and real-world robotic manipulation tasks.

## REFERENCES


[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[3] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, "Aligning cyber space with physical world: A comprehensive survey on embodied ai," *IEEE/ASME Transactions on Mechatronics*, 2025.

[4] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric *et al.*, "Embodied ai agents: Modeling the world," *arXiv preprint arXiv:2506.22355*, 2025.

[5] W. Liang, R. Zhou, Y. Ma, B. Zhang, S. Li, Y. Liao, and P. Kuang, "Large model empowered embodied ai: A survey on decision-making and embodied learning," *arXiv preprint arXiv:2508.10399*, 2025.

[6] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[7] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models," *arXiv preprint arXiv:2209.11302*, 2022.

[8] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on robot learning*. PMLR, 2023, pp. 287–318.

[9] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar, "Voyager: An open-ended embodied agent with large language models," *arXiv preprint arXiv:2305.16291*, 2023.

[10] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2998–3009.

[11] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[12] D. Driess *et al.*, "Palm-e: An embodied multimodal language model," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8469–8488.

[13] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, and P. Luo, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 081–25 094, 2023.

[14] Y. Chen, W. Cui, Y. Chen, M. Tan, X. Zhang, J. Liu, H. Li, D. Zhao, and H. Wang, "Robogpt: an llm-based long-term decision-making embodied agent for instruction following tasks," *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

[15] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[16] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[17] R. Rafailov *et al.*, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 53 728–53 741.

[18] Y. Yang, T. Zhou, K. Li, D. Tao, L. Li, L. Shen, X. He, J. Jiang, and Y. Shi, "Embodied multi-modal agent trained by an llm from a parallel textworld," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 275–26 285.

[19] S. Wang, Z. Fei, Q. Cheng, S. Zhang, P. Cai, J. Fu, and X. Qiu, "World modeling makes a better planner: Dual preference optimization for embodied task planning," *arXiv preprint arXiv:2503.10480*, 2025.

[20] K. Jiao, Z. Fang, J. Liu, B. Li, Q. Wang, X. Liu, J. Ruan, Z. Qiao, Y. Zhu, Y. Xu *et al.*, "Tcpo: Thought-centric preference optimization for effective embodied decision-making," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 9585–9599.

[21] M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht, "Alfworld: Aligning text and embodied environments for interactive learning," *arXiv preprint arXiv:2010.03768*, 2020.

[22] G. Liu, H. Xu, M. Wang, and P. Li. (2026) Logicworld: Automated deep logic generation for multi-turn interactive reasoning in embodied agents. [Online]. Available: https://github.com/iGangao/logicworld

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.