Harmonizing Dense and Sparse Signals in Multi-turn RL: Dual-Horizon Credit Assignment for Industrial Sales Agents

Haojin Yang
Peking University
yhj@stu.pku.edu.cn

Ai Jian
Meituan

Xinyue Huang
Meituan

Yiwei Wang
University of California at Merced

Weipeng Zhang
Meituan

Ke Zeng
Meituan

Xunliang Cai
Meituan

Jingqing Ruan*
Meituan
ruanjingqing@meituan.com

Abstract

Optimizing large language models for industrial sales requires balancing long-term commercial objectives (e.g., conversion rate) with immediate linguistic constraints such as fluency and compliance. Conventional reinforcement learning often merges these heterogeneous goals into a single reward, causing high-magnitude session-level rewards to overwhelm subtler turn-level signals, which leads to unstable training or reward hacking. To address this issue, we propose **Dual-Horizon Credit Assignment (DuCA)**, a framework that disentangles optimization across time scales. Its core, **Horizon-Independent Advantage Normalization (HIAN)**, separately normalizes advantages from turn-level and session-level rewards before fusion, ensuring balanced gradient contributions from both immediate and long-term objectives to the policy update. Extensive experiments with a high-fidelity user simulator show DuCA outperforms the state-of-the-art GRPO baseline, achieving a 6.82% relative improvement in conversion rate, reducing inter-sentence repetition by 82.28%, and lowering identity detection rate by 27.35%, indicating a substantial improvement for an industrial sales scenario that effectively balances the dual demands of strategic performance and naturalistic language generation.

1 Introduction

Large language models have demonstrated remarkable capabilities in open-domain conversation (Cheng et al., 2026; Feng et al., 2025b) and instruction following (Guo et al., 2025) tasks. However, deploying LLMs for high-stakes industrial applications, such as professional sales, remains a significant challenge. Unlike generic assistants, sales dialogues are inherently long-horizon and strictly goal-driven. An effective sales agent must not only maintain conversational fluency but also strategically guide interactions toward conversion, all while adhering to stringent compliance constraints (Dong et al., 2025). While supervised fine-tuning (Ouyang et al., 2022) can impart stylistic nuances and tone, it often fails to capture the long-term planning required to maximize conversion rates across multi-turn exchanges.

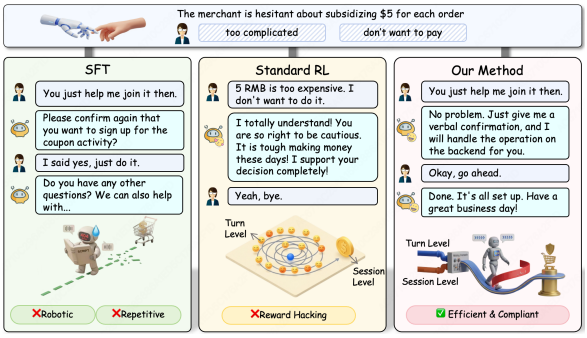


Figure 1: Comparison of dialogue strategies between SFT, Standard RL, and our proposed DuCA method.

Reinforcement learning (RL) offers a promising avenue for optimizing such goal-oriented policies (Shao et al., 2024). Yet, applying RL in sales dialogues exposes a fundamental temporal credit assignment dilemma. The primary business objectives (conversion rate) are sparse, session-level signals that are only revealed at the end of a conversation, while linguistic quality and engagement are dense, turn-level signals. Simultaneous optimization of these conflicting objectives is notoriously unstable. Naive reward aggregation frequently suffers from gradient dominance: high-magnitude, sparse rewards can overshadow nuanced conversational skills, whereas dense, local signals may encourage reward hacking, causing the agent to prioritize short-term gains over ultimate conversion.

To address these challenges, we propose a robust multi-turn RL framework tailored for industrial sales agents. We construct a high-fidelity user simulator to facilitate extensive multi-turn interactions,

*Corresponding author.

mitigating the risks and costs of online exploration. At its core is Dual-Horizon Credit Assignment (DuCA), which employs Horizon-Independent Advantage Normalization (HIAN) to disentangle optimization across time scales. Unlike traditional scalarization methods, DuCA treats turn-level guidance (e.g., linguistic heuristics) and session-level objectives (e.g., conversion and compliance) as distinct supervision signals. We normalize advantages independently for each granularity before fusion, ensuring balanced policy updates that integrate both immediate interaction patterns and long-term strategic incentives. As exemplified in Fig1, unlike SFT and standard RL which often lead to robotic or sycophantic responses, DuCA achieves a more strategic and compliant dialogue flow by effectively balancing these dual-horizon objectives. Our contributions are summarized as follows:

- **Industrial multi-turn training framework**: We propose a comprehensive training system for multi-turn dialogues, designing a high-fidelity user simulator to provide high-quality, interactive conversational data, which facilitates extensive multi-turn policy exploration and bridges the gap between static supervised learning and dynamic real-world deployment.
- **Dual-horizon credit assignment mechanism**: We introduce the horizon-independent advantage normalization, which independently normalizes advantages from turn-level and session-level rewards, effectively addressing optimization instability caused by multi-scale reward signals.
- **Empirical effectiveness**: Extensive experiments demonstrate that our approach substantially outperforms standard SOTA RL baselines, achieving a 6.82% relative improvement in conversion rate, reducing inter-sentence repetition by 82.28%, and lowering identity detection rate by 27.35%, validating its practical value for large-scale industrial deployment.

2 Related Work

Multi-turn RL for dialogue. Recent advancements in aligning LLMs for multi-turn interactions have moved beyond simple SFT. We categorize existing reinforcement learning (RL) approaches into two main streams: **Process-based Credit Assignment:** Traditional methods typically rely on Process Reward Models (PRMs) or critic models to assign credit to intermediate states (Schulman et al., 2017; Jian et al., 2026). However, these entail significant training overhead and heavy reliance on PRM quality. To eliminate critic dependency, recent studies like TARL (Tan et al., 2025) and GiGPO (Feng et al., 2025a) adapt preference optimization to multi-turn contexts using fine-grained rules or state clustering. A critical limitation of these methods is the assumption that step-level rewards align consistently with the final outcome. In industrial sales, immediate linguistic constraints (e.g., compliance) often conflict with aggressive long-term objectives (e.g., conversion), rendering simple dense reward integration insufficient.

Credit assignment in Multi-turn RL. Effective credit assignment is pivotal for learning from mixed signals in multi-turn training. Existing research primarily explores reward granularity and fusion architectures: MT-GRPO (Wei et al., 2025) propagates outcome rewards backward using GAE (Schulman et al., 2018), while MGR (Anonymous, 2026) employs turn-level rewards as a gating mechanism. PURE (Cheng et al., 2025) adopts a conservative approach by minimizing future rewards to penalize worst-case outcomes. Despite these strategies, most prior works merge different reward signals into one value, usually by summing or using gating mechanisms, and then normalize the result (Yang et al., 2025). This coupling leads to gradient dominance, where high-variance trajectory signals overwhelm subtle turn-level signals. Unlike the aforementioned works, our DuCA framework decouples these horizons during the *normalization* phase. This prevent strategic goals from being diluted by dense interaction constraints, achieving a robust balance that avoids the passive behaviors typical of conservative frameworks like PURE.

3 Method

In this section, we introduce **Du**al-Horizon **C**redit **A**ssignment mechanism (**DuCA**), a robust multi-turn RL framework for sales agents in Figure 2. We elaborated the problem formulation, high-fidelity user simulator, and our core contribution: **Horizon-Independent Advantage Normalization (HIAN)** for balancing dense and sparse signals.

3.1 Problem Formulation

We formulate the multi-turn sales dialogue as a Markov Decision Process, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \pi, \mathcal{R}, \gamma)$. At each turn t, the state $s_t \in \mathcal{S}$ encompasses the full dialogue history $s_t = \{u_1, a_1, \ldots, u_t\}$,

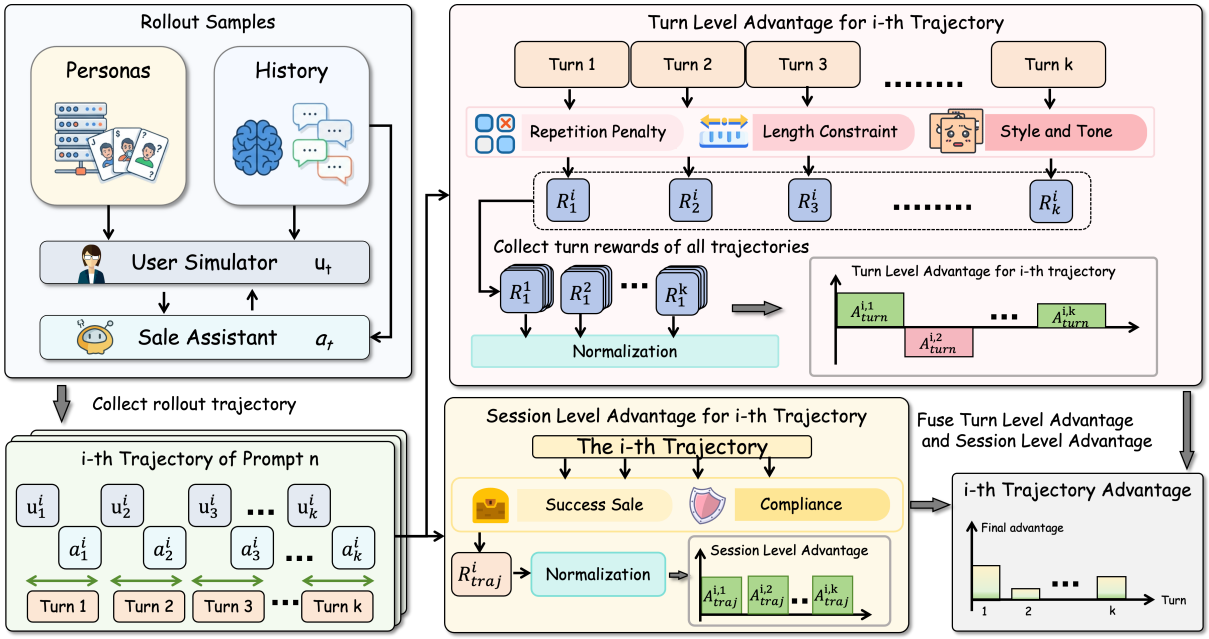


Figure 2: Overview of the DuCA framework. The system generates interaction trajectories via a user simulator conditioned on personas and history. It independently calculates and normalizes: (1) Turn-level advantages from dense heuristic constraints, and (2) Session-level advantages from sparse business outcomes. These decoupled signals are fused to provide a balanced final advantage for robust policy updates.

where u_t is the user query generated by a simulator $\pi_{sim}(u_t|s_{t-1}, \psi)$ with the persona ψ and $a_t \sim \pi_\theta(a_t|s_t)$ is the agent's response. The reward $R(s_t, a_t)$ is decomposed into dense turn-level rewards $r_{turn}^{(t)}$ for linguistic quality and a sparse session-level reward $R_{session}$ for final conversion and compliance. Our objective is to maximize the expected multi-granularity reward:

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau\sim(\pi_\theta,\pi_{sim})}\left[\sum_{t=0}^{T-1}\gamma^t r_{turn}^{(t)} + \gamma^T R_{session}(\tau)\right], \tag{1}$$

where τ represents a trajectory influenced by the strategic interaction between the agent and the personalized user simulator.

3.2 Environment: High-fidelity User Simulator with Personas

Direct training via interaction with real customers entails high operational risks and suffers from sparse feedback. To address this, we construct a high-fidelity user simulator $\mathcal{E}$ based on a powerful LLM, serving as a proxy for real-world customers.

We formulate the user simulator as a conditional generative policy $\pi_{sim}(u_t \mid s_t, \psi)$, which generates a user utterance u_t at turn t. This process is conditioned on three key components:

- **Dialogue context** (s_t): The interaction history $s_t = \{u_1, a_1, \ldots, u_{t-1}, a_{t-1}\}$, representing shared grounding and information state between the agent and the simulator.
- **User Persona** (ψ): A static attribute set $\psi \in \Psi$ extracted for each episode (e.g., *Price-sensitive*, *skeptical*), which conditions the simulator's linguistic style and decision logic to mimic diverse real-world customer behaviors.

This framework enables $\mathcal{E}$ to mimic complex, non-stationary customer decision-making behaviors, effectively bridging the simulation-to-reality gap for policy exploration.

3.3 Multi-Granularity Reward Design

To address the dual objectives of sales dialogues, we decompose the reward signal R into two distinct granularities: dense **turn-level rewards** and sparse **session-level rewards**.

3.3.1 Turn-level Rewards (r_{turn})

Turn-level rewards provide immediate and dense feedback, aimed at guiding fundamental conversational skills and maintaining dialogue consistency. Given the dense and immediate nature of turn-level interactions, we employ predefined heuristic rules to model these intermediate rewards. This approach provides stable and explicit value signals, thereby stabilizing the training trajectory. We define $r_{\text{turn}}(s_t, a_t)$ based on the following heuris-

tic rules and utility functions: repetition penalty, length constraint, and style and tone, elaborated in the Appendix A.1.1.

For each turn t, these sub-rewards are aggregated via a gating mechanism (e.g., penalties override style rewards) to form the final scalar $r_{\text{turn}}^{(t)}$.

3.3.2 Session-level Rewards (R_{session})

Session-level rewards reflect ultimate business objectives and strictly enforced safety constraints. These signals are sparse and typically determined only at the end of the dialogue (step T). We define the session reward for a completed episode τ as:

$$R_{\text{session}}(\tau) = \alpha \cdot \mathbb{I}(\text{Conversion}) + \beta \cdot S_{\text{compliance}}(\tau),$$

where $\mathbb{I}(\cdot)$ is an indicator function for successful conversion, and $S_{\text{compliance}}$ is a scoring function that penalizes regulatory violations, detailed in the Appendix A.1.2.

3.4 Dual-Horizon Credit Assignment (DuCA)

A naive scalarization of rewards (i.e., simply summing $r_t^{\text{total}} = r_{\text{turn}}^{(t)} + r_{\text{session}}^{(t)}$) often leads to optimization instability. This is primarily due to the **gradient dominance** problem: high-magnitude, high-variance sparse rewards (e.g., a large bonus for a successful sale) can overwhelm subtle dense signals, or conversely, dense rewards can induce reward hacking where the agent ignores long-term goals. To resolve this, we propose the **DuCA** mechanism, which disentangles the credit assignment process into three strategic steps.

Step 1: Independent Advantage Estimation. We maintain two separate value heads, $V_{\phi_{\text{turn}}}$ and $V_{\phi_{\text{session}}}$, to estimate expected returns for turn-level and session-level objectives respectively. We compute the Generalized Advantage Estimation (GAE) separately:

$$A_{\text{turn}}^{(t)} = \text{GAE}(r_{\text{turn}}, V_{\phi_{\text{turn}}}, \lambda_{turn}, \gamma_{turn}), \tag{2}$$

$$A_{\text{session}}^{(t)} = \text{GAE}(r_{\text{session}}.V_{\phi_{\text{session}}}, \lambda_{session}, \gamma_{session}), \tag{3}$$

This decoupling enables tailored temporal dynamics. We employ standard GAE parameters ($\gamma_{\text{turn}} = 0.99, \lambda_{\text{turn}} = 0.95$) for local fluency to balance bias and variance. Conversely, we set $\gamma_{\text{session}} = \lambda_{\text{session}} = 1.0$, ensuring the sparse terminal reward is propagated uniformly to all actions without decay, effectively bridging the long-horizon gap.

Step 2: Horizon-Independent Advantage Normalization (HIAN). To ensure balanced gradient contributions, we normalize advantages independently within each mini-batch $\mathcal{B}$:

$$\hat{A}_{\text{turn}} = \frac{A_{\text{turn}} - \mu_{\mathcal{B}}(A_{\text{turn}})}{\sigma_{\mathcal{B}}(A_{\text{turn}}) + \epsilon}, \tag{4}$$

$$\hat{A}_{\text{session}} = \frac{A_{\text{session}} - \mu_{\mathcal{B}}(A_{\text{session}})}{\sigma_{\mathcal{B}}(A_{\text{session}}) + \epsilon}. \tag{5}$$

Theoretical Justification. We analyze why HIAN outperforms naive scalarization. In naive methods, the effective gradient is scaled by the total variance: $\Delta\theta \propto (A_{\text{turn}} + A_{\text{session}})/\sqrt{\sigma_{\text{turn}}^2 + \sigma_{\text{session}}^2}$. Since sparse business rewards typically have high variance ($\sigma_{\text{session}} \gg \sigma_{\text{turn}}$), the contribution of the turn-level signal becomes $\approx A_{\text{turn}}/\sigma_{\text{session}} \to 0$. This **gradient suppression** halts the learning of linguistic skills. HIAN decouples this dependency, scaling A_{turn} by its own deviation σ_{turn}, ensuring robust learning of both objectives simultaneously.

Step 3: Strategic Fusion and Optimization. The final advantage is a weighted combination: $A_{\text{total}}^{(t)} = w_{\text{turn}}\hat{A}_{\text{turn}}^{(t)} + w_{\text{session}}\hat{A}_{\text{session}}^{(t)}$. The policy π_θ is updated via the PPO objective:

$$\begin{aligned}\mathcal{L}(\theta) = &\mathbb{E}_t\Big[\min\big(\rho_t(\theta)A_{\text{total}}^{(t)}, \\ &\text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)A_{\text{total}}^{(t)}\big)\Big],\end{aligned} \tag{6}$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio. This ensures strategic planning without sacrificing conversational quality.

4 Experiments

4.1 Experimental Setup

Datasets and simulator. We constructed our training dataset based on 31,000 anonymized real-world business sales dialogues. Additionally, we captured 10,000 high-quality online samples to fine-tune an internal LLM-based user simulator, serving as a high-fidelity interactive environment. The simulator uses heterogeneous user personas (e.g., price-sensitive, skeptical) to mimic complex real-world decision-making.

Implementation Details and Baselines. Our sales agent policy π_θ is initialized from a proprietary model after supervised fine-tuning. We

Table 1: Main Results on Sales Dialogue Evaluation. CVR and Compliance represent primary business outcomes, while the remaining metrics evaluate fine-grained interaction quality. Best results across all models are **bolded**.

Method	CVR(↑)	Compliance(↑)	Avg. Turn	Intra-R(↓)	Inter-R(↓)	IDR(↓)	Prefix-R(↓)	Filler(↓)	PATR(↑)
Foundation Models									
DeepSeek-R1	22.85%	65.13	9.85	35.19%	**1.01%**	**2.50%**	**7.69%**	59.82%	42.74%
Longcat-Flash-Chat	19.99%	70.17	9.27	3.02%	23.13%	5.92%	39.10%	**33.35%**	43.65%
Training Methods									
SFT (Base)	22.23%	59.11	11.41	9.37%	35.87%	10.59%	52.30%	51.70%	43.59%
REINFORCE++	21.51%	61.32	11.18	8.34%	30.17%	9.99%	50.34%	54.69%	43.62%
GRPO	22.88%	65.40	12.49	5.48%	15.29%	9.28%	47.76%	61.19%	43.30%
GDPO	22.51%	**68.53**	12.19	6.48%	17.22%	8.41%	50.03%	55.96%	44.11%
DuCA (Ours)	**24.44%**	66.72	**9.91**	**4.20%**	2.71%	6.11%	34.44%	**49.21%**	**44.61%**

evaluate DuCA against four RL baselines initialized from the same base model: **SFT** (behavioral cloning lower bound); **REINFORCE++** (linear reward scalarization); **GRPO** (group-wise advantage normalization); and **GDPO** (independent normalization for heterogeneous sources). Training is optimized via PPO with a KL-divergence penalty.

Metrics and Protocol. We focus on **Conversion Rate (CVR)** for business outcome and **Compliance Score** for regulatory adherence. Evaluations use an LLM-as-a-Judge paradigm on generated multi-turn trajectories from a held-out test set.

4.2 Main Results

Table 1 summarizes the performance across primary business outcomes and fine-grained interaction quality metrics. Overall, our DuCA method significantly outperforms all baseline algorithms, demonstrating a superior Pareto balance between conversion and compliance.

While foundation models like Longcat-Flash-Chat and DeepSeek-R1 exhibit high compliance (up to 77.23%) and stylistic diversity, they often lack the domain-specific strategic depth required for industrial sales. DeepSeek-R1's CVR (22.85%) is surpassed by DuCA, indicating that standard models, while fluent, are not fully optimized for goal-oriented industrial conversion.

DuCA achieves a 24.44% CVR, which is a 6.82% relative improvement over the strongest baseline GRPO, showing the superiority of our mwthod in exploring effective sales strategies. Additionally, the 7.61% improvement in compliance over SFT demonstrates that our framework effectively maintains safety boundaries while pursuing aggressive conversion goals.

Our method achieves a 2.71% Inter-turn Repetition rate, representing an 82.28% relative reduction compared to GRPO, which proves that the dual-horizon assignment effectively solves the "repetition loop" common in standard RL. Furthermore, the reduction of the Identity Detection Rate by 27.35% relative to GDPO indicates that DuCA produces more professional interactions.

DuCA reduces the Average Turn by 11.36% relative to REINFORCE++, while increasing the Positive Attitude Transfer Rate (PATR) to 44.61%, illustrating the superiority of our method in steering user intent toward conversion with higher operational efficiency.

These consistent gains across heterogeneous metrics suggest that the bottleneck in sales-oriented RL is the gradient interference between dense linguistic rewards and sparse business signals. By decoupling these horizons via HIAN, DuCA successfully mitigates gradient dominance, ensuring that turn-level constraints are not overwhelmed by high-variance session outcomes. This allows the agent to learn a balanced policy that is both persuasively strategic and linguistically compliant.

In summary, the experimental evidence confirms that DuCA establishes a new state-of-the-art for industrial sales agents, delivering a 6.82% CVR boost and a dramatic 82.28% drop in conversational redundancy. These results validate that hierarchical credit assignment is a robust and scalable paradigm for aligning autonomous agents with complex, multi-granular industrial constraints.

4.3 Ablation Study

To verify the contribution of each core component, we designed the following variants:

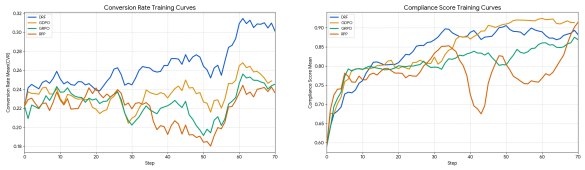
- **w/o HIAN**: Removes the decoupled normalization strategy and uses standard reward summation for advantage calculation.
- **w/o Multi-turn**: Trains on single-turn interaction scenarios only, where rewards consist of a weighted sum of single-turn compliance and conversion scores.

Table 2: Ablation Study Results.

Method	CVR (%)	Compliance(↑)	Avg. Turn
DuCA (Full)	**24.44**	**66.72**	**9.91**
w/o HIAN	24.13	64.09	10.31
w/o Multi-turn	21.64	60.09	12.02

The results demonstrate that the multi-turn interaction environment is fundamental to training quality. Without long-term dependency, the model loses business reward guidance, and optimization degenerates to turn-level dominance. Furthermore, without horizon-independent advantage normalization, the gradients from successful conversion cases are diluted by relatively stable turn rewards, leading to unstable convergence.

4.4 Training Dynamics



(a) Conversion Score (b) Compliance Score

Figure 3: Training dynamics of DuCA compared with baselines over 70 steps.

Figure 3 visualizes the optimization stability across all methods. As shown in Figure 3b, REINFORCE++ and GRPO exhibit rapid initial gains in compliance but suffer from significant instability or eventual collapse (e.g., the catastrophic drop in REINFORCE++ after step 35). This illustrates the **reward hacking trap**: in standard scalarized reward structures, the high-frequency dense turn-level rewards dominate the policy gradient, causing the agent to over-fit to local formatting rules while failing to sustain an upward trend in the sparse, long-term conversion rate.

In contrast, DuCA maintains a remarkably stable and superior learning trajectory in both metrics. By calculating advantages for different temporal horizons independently, our hierarchical decoupling mechanism prevents high-variance session signals from being diluted or overwhelmed by dense rewards. While GDPO achieves slightly higher compliance after step 50, it suffers from a "strategic collapse" in conversion rate due to unconstrained objective competition. DuCA achieves a superior Pareto balance, delivering state-of-the-art conversion performance while strictly adhering to industrial compliance standards.

4.5 Simulator Fidelity and Feasibility

To validate the simulator as a reliable proxy for real-world sales exploration, we conduct a multi-dimensional feasibility analysis. The core objective is to ensure the environment is consistent, realistic, and predictive of business outcomes.

We employ GPT-4.1 to evaluate the simulator (v4.7.1) against human expert baselines. As shown in Table 3, the simulator achieves an average score of **4.897**, slightly outperforming human experts (4.885). Specifically, the simulator excels in *Tone and Style* (4.969) and *Basic Judgment* (4.910). This indicates that the simulator provides a highly stable and professional conversational environment, which is essential for reducing gradient noise during RL policy updates.

Table 3: Consistency Evaluation: Simulator vs. Human Level. Scores (1-5) are generated via GPT-4.1 based on persona and context adherence. Best results are **bolded**.

Metric	Simulator	Human Level
Receiver Identity	4.906	**4.948**
Decision Subject	4.750	**4.824**
Merchant Profile	**4.899**	4.830
Busy State	4.957	**4.980**
Acceptance Level	**4.888**	4.830
Basic Judgment	**4.910**	4.843
Tone and Style	**4.969**	4.941
Average Score	**4.897**	4.885

Beyond linguistic consistency, the simulator demonstrates high behavioral realism. In a blind Turing Test, senior auditors mistakenly identified **52.0% of actual human interactions** (detailed in Appendix B)as being generated by the simulator, suggesting it has captured the professional "gold standard" of sales dialogues.

5 Conclusion

We present DuCA, a robust framework addressing the temporal credit assignment challenge in industrial sales agents. By explicitly decomposing rewards into dual horizons and employing HIAN, our method effectively resolves the gradient dominance and reward hacking issues inherent in multi-granularity optimization. Our work provides a scalable paradigm for aligning LLM agents with complex industrial constraints, ensuring a superior balance between strategic performance and linguistic compliance in real-world deployments.

Limitations

Dependency on High-Quality Real-World Data and Human Effort. The effectiveness of DuCA heavily relies on the high-fidelity user simulator. Constructing such a simulator requires a substantial amount of anonymized real-world business sales dialogues (e.g., 31,000 sessions in our study) and additional high-quality online interaction samples for fine-tuning. This dependency introduces significant human effort in data collection, cleaning, and professional auditing to ensure the simulator captures the "gold standard" of sales interactions. Consequently, deploying DuCA in new industrial domains where such large-scale, domain-specific data or expert human resources are scarce may pose practical challenges for initial model training and simulator calibration.

Ethical Statement

This paper proposes a reinforcement learning framework for industrial sales agents designed to balance strategic performance with compliance. While our approach offers significant benefits in terms of conversion efficiency, it also raises ethical considerations. The use of such agents could lead to unintended consequences, such as bias amplification, where the synthetic agents might inadvertently reinforce existing stereotypes or present skewed sales arguments due to biases in the historical training data.

Additionally, there is a risk of manipulation of user preferences, as the strategic credit assignment could be used to subtly influence customer behavior without explicit consent. We emphasize that our system incorporates a compliance scoring function to penalize prohibited terms and false promises to mitigate these risks. Furthermore, we suggest that synthetic user simulators should not be a substitute for real human feedback in the long-term design process. Rather, these agents should be leveraged to explore strategies in early stages or high-risk scenarios where direct online exploration with real customers is impractical. By adhering to these principles, we aim to ensure that the deployment of autonomous sales agents is ethical and socially responsible.

References

Anonymous. 2026. Both local validity and global effectiveness matter: Decoupled credit assignment for long-horizon agentic learning.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. 2025. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *Preprint*, arXiv:2504.15275.

Xuxin Cheng, Ke Zeng, Zhiquan Cao, Linyi Dai, Wenxuan Gao, Fei Han, Ai Jian, Feng Hong, Wenxing Hu, Zihe Huang, Dejian Kong, Jia Leng, Zhuoyuan Liao, Pei Liu, Jiaye Lin, Xing Ma, Jingqing Ruan, Jiaxing Song, Xiaoyu Tan, and 49 others. 2026. Higher satisfaction, lower cost: A technical report on how llms revolutionize meituan's intelligent interaction systems. *Preprint*, arXiv:2510.13291.

Wenjie Dong, Sirong Chen, and Yan Yang. 2025. ProTOD: Proactive task-oriented dialogue system based on large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9147–9164, Abu Dhabi, UAE. Association for Computational Linguistics.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025a. Group-in-group policy optimization for llm agent training. *Preprint*, arXiv:2505.10978.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2025b. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *Preprint*, arXiv:2505.19630.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Ai Jian, Jingqing Ruan, Xing Ma, Dailin Li, Weipeng Zhang, Ke Zeng, and Xunliang Cai. 2026. Patarm: Bridging pairwise and pointwise signals via preference-aware task-adaptive reward modeling. *Preprint*, arXiv:2510.24235.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-dimensional continuous control using generalized advantage estimation. *Preprint*, arXiv:1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Weiting Tan, Xinghua Qu, Ming Tu, Meng Ge, Andy T. Liu, Philipp Koehn, and Lu Lu. 2025. Process-supervised reinforcement learning for interactive multimodal tool-use agents. *Preprint*, arXiv:2509.14480.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *Preprint*, arXiv:2505.11821.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. 2025. Treerpo: Tree relative policy optimization. *Preprint*, arXiv:2506.05183.

A Implementation Details

A.1 Reward Function Formulations

To ensure the sales agent balances conversational quality with long-term business objectives, we define the reward functions for two distinct temporal granularities.

A.1.1 Turn-Level Reward ($r_{turn}^{(t)}$)

The turn-level reward provides dense supervision to maintain linguistic constraints and basic dialogue fluency. We employ a **Gating Fusion** mechanism to prevent the agent from "hacking" simple rewards (e.g., length rewards) by producing repetitive or rigid scripted content. The final reward for turn t is formulated as:

$$r_{turn}^{(t)} = \mathbb{1}_{\text{valid}}(a_t) \cdot r_{\text{len}}(a_t) + (1 - \mathbb{1}_{\text{valid}}(a_t)) \cdot R_{\text{penalty}} \tag{7}$$

where the indicator function $\mathbb{1}_{\text{valid}}(a_t)$ is defined by the following gating criteria:

$$\mathbb{1}_{\text{valid}}(a_t) = \begin{cases} 1 & \text{if } \text{Rep}(\tau) \leq \delta_1 \text{ or } \text{Sim}(a_t, \mathcal{S}) \leq \delta_2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The specific components and thresholds are:

- **Rep**(τ): Measures the token-level repetition within the current response and the lexical overlap against the previous three dialogue turns. The threshold δ_1 is set to 0.4.
- **Sim**$(a_t, \mathcal{S})$: Calculates the maximum cosine similarity between the response a_t and the standard professional sales script library $\mathcal{S}$. The threshold δ_2 is set to 0.85 to penalize verbatim template usage without substantive content.
- $r_{\textbf{len}}(a_t)$: A Gaussian-shaped length incentive defined as $r_{\text{len}} = \exp(-\frac{|len(a_t) - L_{\text{target}}|^2}{2\sigma^2})$, where $L_{\text{target}} = 30$ tokens.
- $R_{\textbf{penalty}}$: A constant penalty value of -2.0 applied when gating criteria are violated.

A.1.2 Session-Level Reward ($R_{\textbf{session}}$)

The session-level reward (session-level reward) reflects the ultimate industrial goals and is assigned only at the terminal step T:

$$R_{\text{session}}(\tau) = \alpha \cdot I(\text{Conversion}) - \beta \cdot S_{\text{violation}}(\tau) \tag{9}$$

Where:

- $I(\textbf{Conversion})$: A binary indicator (1 for success, 0 otherwise) determined by an LLM-as-a-Judge evaluating purchase intent.
- $S_{\textbf{violation}}(\tau)$: The cumulative penalty score derived from regulatory audits for prohibited terms or false promises. The *Compliance Score* reported in the main results is the transformed metric $100 - S_{\text{violation}}$.
- α, β: Scaling coefficients used to balance the magnitude of conversion and compliance signals.

A.2 Hyper-parameter and Training Settings

For the Strategic Credit Assignment via DRF, the optimization parameters are configured as follows:

- **GAE Configuration**: We set $\lambda = 1.0$ and $\gamma = 1.0$ for both value functions ($V_{\phi_{\text{turn}}}$ and $V_{\phi_{\text{session}}}$). This ensures the session-level reward is propagated back to every turn t without decay, which is critical for long-horizon credit assignment in sales.
- **Fusion Weights**: In our main experiments, we use a balanced ratio of $w_1 = 1.0$ and $w_2 = 1.0$ to ensure both immediate constraints and long-term goals contribute equally to the policy gradient.

- **PPO Setup**: We utilize the PPO algorithm with a KL-divergence penalty term (coefficient 0.05) to constrain policy drift from the SFT initialization.
- **Training Horizon**: All models are trained for a maximum of 70 steps.

A.3 Computing Infrastructure

Training was conducted on a cluster of 16 $\times$ NVIDIA A100 GPUs (80GB). The sales agent policy π_θ was initialized from a proprietary Large Model pre-trained on 31,000 anonymized real-world business sales dialogues.

B Evaluation Reliability and Validity

To ensure that our reinforcement learning policy is optimized against a representative environment, we evaluate our LLM-based User Simulator across two dimensions: linguistic realism and behavioral consistency.

B.1 Turing Test Evaluation

We conducted a blind Turing Test to assess the indistinguishability of our simulator compared to real-world human customers. Three senior sales experts were presented with 250 dialogue snippets (half human-human, half agent-simulator) and asked to identify whether the "customer" was a human or the simulator.

Table 3 summarizes the experts' judgment on actual human trajectories. Notably, only 36.0% of human trajectories were correctly identified, while 52.0% of human interactions were mistakenly classified as being generated by the simulator. This "inverse" confusion suggests that the simulator has captured the core characteristics of professional sales interactions so effectively that it defines the "standard" behavior in the eyes of the auditors.

Table 4: Human-vs-Simulator Turing Test Results. Results indicate the classification of actual human trajectories by experts.

Metric	Percentage (%)
Correctly identified as Human	36.0%
Indistinguishable / Hard to distinguish	12.0%
Mistakenly identified as Simulator	52.0%

B.2 Simulation-to-Reality (Sim-to-Real) Consistency

A high-fidelity simulator must not only sound realistic but also provide a reliable signal for business outcomes. We compared the Conversion Rates (CVR) obtained in our simulated environment against the actual performance observed in online outbound call scenarios.

We calculated the **Pearson Correlation Coefficient** across 50 different product categories and persona initializations. The correlation between the simulator's predicted conversion intent and the real-world conversion rate reached:

$$\rho = 0.9733 \tag{10}$$

This exceptionally high correlation ($\rho > 0.97$) validates that our simulator serves as a robust proxy for real-world customer decision-making processes. Consequently, policy improvements observed in the simulator are highly likely to translate into tangible business gains in production.

B.3 Persona Adherence

The simulator's ability to maintain a consistent persona (e.g., *Price-Sensitive, Skeptical*) is monitored via an internal LLM-based auditor. The auditor checks the consistency between the assigned persona prompts and the simulator's responses throughout the multi-turn interaction. Our version 4.6.1 maintains a high degree of adherence, ensuring that the policy is exposed to diverse and stable customer behaviors during training.

C Ablation Study Details

In this section, we provide a more granular analysis of the ablation experiments to further elucidate the impact of the Decoupled Advantage Normalization and the Multi-turn training environment on both business objectives and conversational quality.

C.1 Detailed Conversational Metrics

Table 5 presents the fine-grained metrics for the ablation variants. This table complements the high-level results in the main text by showing how each component influences the microscopic behavior of the agent.

C.2 Training Dynamics and Stability

To visualize the optimization process, Figure 4 illustrates the evolution of Conversion Rate (CVR) and Compliance scores over the 80-step training trajectory for all variants.

Table 5: Detailed Ablation Study on Conversational Quality. "w/o Decoupled Norm" refers to the variant using standard reward summation (rpp_baseline). "w/o Multi-turn" denotes training on single-turn scenarios. Intra-R and Inter-R: Intra-turn and Inter-turn Repetition; IDR: Identity Detection Rate; PATR: Positive Attitude Transfer Rate. Best results are **bolded**.

Variant	Intra-R↓	Inter-R↓	IDR↓	Prefix-R↓	Filler↓	PATR↑
DRF (Full)	4.20%	2.71%	**6.11%**	**34.44%**	49.21%	**44.61%**
w/o Decoupled Norm	**1.87%**	**1.68%**	7.35%	36.64%	52.49%	43.83%
w/o Multi-turn	8.94%	41.75%	10.34%	57.41%	**45.71%**	43.84%

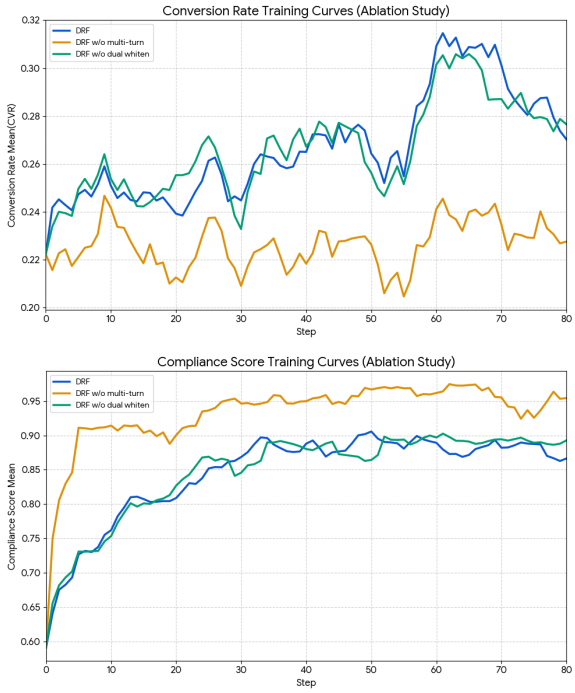


Figure 4: Training dynamics of DRF compared with ablation variants. (a) Conversion Rate (CVR) evolution. (b) Compliance Score evolution.

C.3 Impact Analysis

C.3.1 Effectiveness of Decoupled Normalization (Dual Whiten)

The comparison between **DRF (Full)** and **w/o Decoupled Norm** reveals a critical trade-off in multi-objective sales optimization. As shown in Figure 4(a), while the variant without decoupled normalization initially tracks with the full model, it exhibits higher variance and fails to sustain its peak CVR after step 60.

Furthermore, Table 5 shows that while this variant achieves the lowest repetition rates, its **Identity Detection Rate (IDR)** is significantly higher (7.35%) and its **Positive Attitude Transfer Rate (PATR)** is lower (43.83%) than the full model. This suggests that without independent advantage normalization, high-variance session-level signals are diluted by stable turn-level rewards, leading the model to converge toward a "safe" but rigid policy that produces robotic responses, ultimately failing to effectively persuade users.

C.3.2 Necessity of Multi-turn Interaction

The **w/o Multi-turn** variant serves as a critical baseline, demonstrating the strategic collapse when long-term dependencies are removed.

- **Pseudo-Compliance**: In Figure 4(b), this variant reaches high compliance scores faster than others. However, Table 5 clarifies that this is a result of "safe" but repetitive behavior, with *Inter-turn Repetition* (Inter-R) surging to **41.75%**.
- **Strategic Stagnation**: The CVR curve for this variant remains significantly lower and stagnant throughout training. Without a multi-turn interaction loop, the agent cannot capture the causal links between its actions and user intent shifts, validating that our framework requires a multi-turn environment to align dense constraints with sparse business objectives.