

LSH-RL: Long-Short Horizon Reinforcement Learning for Persona-Level Role-Playing Agents

Wenlong Shi¹, Jianxun Lian², Haiming Qin¹, Jiwei Zhang¹,
Wei Zhang¹, Mingyang Zhou¹, Hao Liao^{1,3}*,

¹College of Computer Science and Software Engineering, Shenzhen University, China

²Microsoft Research Asia

³Provincial Key Laboratory of Intelligent Communication and
Digital Society Governance, Shenzhen University

haoliao@szu.edu.cn

Abstract

Role-playing agents (RPAs) must sustain persona fidelity, natural interaction, and scene progression across multi-turn trajectories. However, supervised fine-tuning, preference optimization, and coarse episode-level RL can over-reinforce safe stylized responses that appear in-character but are repetitive, low-information, or weakly grounded in the evolving scene. We propose **LSH-RL**, a **Long-Short Horizon Reinforcement Learning** framework for persona-level RPAs. LSH-RL trains a shared policy through dynamic multi-role self-play, combining short-horizon utterance/action rewards with long-horizon character-trajectory rewards. A gated trajectory-to-turn reward fusion mechanism prevents locally degenerate turns from receiving undeserved long-horizon credit, while a dialogue-action curriculum supports staged optimization. Experiments on PersonaArena and CharacterBox show that LSH-RL improves over SFT and DPO in both dialogue-only and full interaction settings. Further analyses show reduced repetition and higher human preferences, suggesting a path toward more coherent and interaction-rich persona-level RPAs. The source code is available at <https://github.com/Shiwenlong-source/LSH-RL>.

1 Introduction

Role-playing agents (RPAs) aim to enable large language models (LLMs) to act as specific personas in immersive multi-turn interactions, supporting applications such as virtual companions, interactive games, narrative dialogue, and social simulation. Recent work has advanced RPAs through role-playing benchmarks and elicitation pipelines (Wang et al., 2024a), large-scale persona corpora (Wang et al., 2025c,b), psychology-grounded character modeling (Ran et al., 2024), and trajectory-level evaluation in social or virtual



Figure 1: Challenge of over-reinforced stylized responses. SFT, DPO, or episode-level RL may amplify safe in-character responses, leading to repeated, low-information interactions with weak scene progression.

environments (Chen et al., 2024; Tu et al., 2024; Wang et al., 2025a; Shi et al., 2026). Despite these advances, RPAs still suffer from trajectory-scale failures: emotional expression may deviate from the target character (Feng et al., 2025), scenes may stall without narrative progress (Zhang et al., 2025), long-context decisions may become persona-inconsistent (Xu et al., 2025), and generated behaviors may become generic, repetitive, or weakly grounded in the evolving scene.

SFT and preference optimization provide useful warm starts for RPAs, but they mainly learn role expression from static demonstrations or paired preferences. When safe, fluent, and stylistically aligned responses receive favorable signals, the model may learn a shortcut: producing utterances that look in character without substantially advancing the scene. RL is a natural alternative for interactive role-playing, but coarse reward design can further amplify this shortcut. Local rewards may miss repetition and weak scene progression, while trajectory-level rewards can reinforce poor local

*Corresponding author.

turns inside a globally acceptable rollout. As illustrated in Figure 1, optimization may over-reinforce stylized responses that remain scene-relevant but repeatedly rely on similar rhetorical forms.

These limitations motivate a reward design that separates and recombines local and trajectory-level supervision. Short-horizon rewards are needed to detect local degeneration, including repeated question forms, low-information responses, implausible actions, and action–dialogue inconsistency. Long-horizon rewards are needed to evaluate whether a character remains coherent, socially responsive, and persona-consistent across the whole trajectory. Crucially, trajectory-level feedback should not be assigned uniformly to every turn; otherwise, a globally acceptable scene may still reinforce shallow local shortcuts.

To address this challenge, we propose **LSH-RL**, a **Long-Short Horizon Reinforcement Learning** framework for persona-level role-playing agents. Built on a dynamic multi-character simulation setting inspired by PersonaArena (Shi et al., 2026), LSH-RL trains a shared policy through multi-role self-play, where the model generates both dialogue and action behaviors under evolving scene states. LSH-RL combines short-horizon utterance/action rewards with long-horizon character-trajectory rewards, and uses a gated trajectory-to-turn-to-token reward assignment mechanism to prevent locally poor turns from receiving undeserved long-horizon credit. Experiments under both dialogue-only and full interaction settings show that LSH-RL improves persona fidelity, naturalness, behavioral coherence, interaction richness, and repetition control over strong SFT and DPO baselines.

Our contributions are summarized as follows:

- We propose **LSH-RL**, a long-short horizon RL framework that optimizes persona-level role-playing in dynamic multi-role interaction rather than static response imitation.
- We introduce a gated long-short reward mechanism that combines local utterance/action quality with trajectory-level persona coherence while blocking degenerate turns from receiving credit.
- We evaluate LSH-RL on PersonaArena, CharacterBox, and RoleBench, showing consistent gains over SFT and DPO and narrowing the gap to closed-source models.

2 Related Work

2.1 LLM Role-Playing

Role-playing agents have been studied from both modeling and evaluation perspectives. On the modeling side, RoleLLM (Wang et al., 2024a) systematizes benchmark construction, elicitation, and fine-tuning; CoSER (Wang et al., 2025c) builds literary persona data with scenes, memory, and inner thought; and OpenCharacter (Wang et al., 2025b) provides a synthetic-persona SFT pipeline for customizable characters. Other studies improve persona representation through dynamic LoRA (Yu et al., 2024), psychology-grounded supervision (Ran et al., 2024), persona-aware contrastive learning (Ji et al., 2025), and trait–utterance alignment (Yu et al., 2025). These works show that role-playing requires more than surface style imitation, since character behavior must remain consistent with persona traits and scene context.

Evaluation has also moved from single-turn matching toward multi-dimensional and trajectory-level assessment, including personality fidelity (Wang et al., 2024c), social interaction (Chen et al., 2024), emotional consistency (Feng et al., 2025), plot progression (Zhang et al., 2025), persona-driven decisions (Xu et al., 2025), and trajectory-level behavior evaluation (Wang et al., 2025a; Wu et al., 2024; Shi et al., 2026). These developments motivate trajectory-aware training, but most existing pipelines still rely on static response imitation or utterance-level preferences rather than explicitly assigning rewards across local turns and whole trajectories.

2.2 Reinforcement Learning for Role-Playing Agents

Reinforcement learning is a natural direction for RPAs because role-playing is sequential, interactive, and feedback-driven. Recent methods improve reward design from different perspectives: CPO (Ye et al., 2025) addresses reward ambiguity with comparative trajectory ranking, MOA (Liao et al., 2026) optimizes multiple fine-grained role-playing objectives, and RAIDEN-R1 (Wang et al., 2025d) and Character-R1 (Tang et al., 2026) use verifiable role-awareness rewards. These methods strengthen reward construction, but they often concentrate supervision either at the response level or at the trajectory/preference level, leaving the connection between local degeneration and long-horizon persona coherence underexplored.

Broader multi-turn and multi-agent RL systems provide useful substrates for interactive LLM training. LMRL-Gym (Abdulhai et al., 2025) studies multi-turn RL with language models, SOTOPIA- π (Wang et al., 2024b) trains socially goal-oriented agents, OMAR (Jiang et al., 2026) uses one model to play multiple social roles, and Agent Lightning (Luo et al., 2025) decomposes complex agent trajectories into trainable transitions. In contrast to outcome-driven or dialogue-centric settings, LSH-RL focuses on persona-level role-playing quality itself: it jointly considers both dialogue and action behaviors, explicitly decomposes rewards into short-horizon local quality and long-horizon trajectory coherence, and gates trajectory-level credit through locally qualified turns.

3 Methodology

Role-playing requires consistency across both behavioral modalities and temporal scales. A character must not only speak in persona, but also take plausible actions that advance the scene. Meanwhile, an individual interaction may appear locally acceptable while the full trajectory drifts from the target persona, and a globally coherent trajectory should not justify weak local turns. LSH-RL addresses these challenges by training a shared policy in multi-character dialogue-action simulation with gated long-short horizon reward assignment.

3.1 Role-playing Environment

We formulate persona-level role-playing as sequential interaction in a structured social scene, where multiple characters interact through both dialogue and actions under an evolving narrative context.

Training scenes. Each training sample is a scene $E = (D, \mathcal{R}, A_0, Q)$, where D denotes the scene description, $\mathcal{R} = \{R_i\}_{i=1}^N$ denotes the set of role specifications, A_0 denotes the initial actions or utterances that instantiate the scene, and Q denotes the plot description that defines the narrative background and interaction premise. This formulation treats role-playing as scene-grounded interaction rather than isolated response generation.

To stabilize optimization, we organize scenes by difficulty and draw training batches from a stratified scene pool. This preserves coverage over easy, medium, and hard cases without over-concentrating on any single subset. Detailed scene construction and sampling settings are provided in Appendix A.

Multi-role self-play. Our environment is inspired by PersonaArena (Shi et al., 2026), which casts persona-level role-playing as dynamic multi-character interaction rather than static prompt-response evaluation. Unlike the original evaluation setting, where the evaluated model typically plays one focal character while other agents are controlled separately, LSH-RL assigns all characters to a single shared policy π_θ . Different role specifications serve as conditioning contexts, so the policy parameters are shared across roles while each generation remains grounded in the corresponding character profile and local observation. This yields a multi-role self-play process in which the model improves from dialogue-action experience generated by its own role-conditioned behaviors.

At round t , character c_i receives an observation $o_{i,t}$, including the current scene state, visible interaction history, narrator-provided scene understanding, and the local state of c_i . The role specification R_i provides the identity and behavioral prior of c_i , including persona traits, background, relationships, and role-specific constraints. Conditioned on $(o_{i,t}, R_i)$, the shared policy generates the next interaction:

$$u_{i,t} \sim \pi_\theta(\cdot \mid o_{i,t}, R_i) \quad (1)$$

where $u_{i,t}$ can be either an utterance or an action. In practice, $u_{i,t}$ is typed as either a dialogue transition or an action transition, depending on whether the character produces spoken content or performs a behavior in the scene. This typed-transition view allows LSH-RL to assign modality-specific feedback while still optimizing a single shared policy.

3.2 Long-Short Horizon Reward

The central difficulty in role-playing reward design is that local quality and global role fidelity are related but not interchangeable. Turn-level judgments can detect awkward, repetitive, or off-character utterances and actions, but they cannot fully determine whether a character remains coherent throughout the whole scene. Trajectory-level judgments capture holistic persona fidelity, but they are too coarse to assign reliable credit to individual turns. LSH-RL therefore combines short-horizon and long-horizon rewards. In our implementation, Qwen3-32B serves as the judge for both reward horizons and as the turn verifier.

Short-horizon reward. For each interaction $u_{i,t}$, we compute a turn-level reward using an LLM

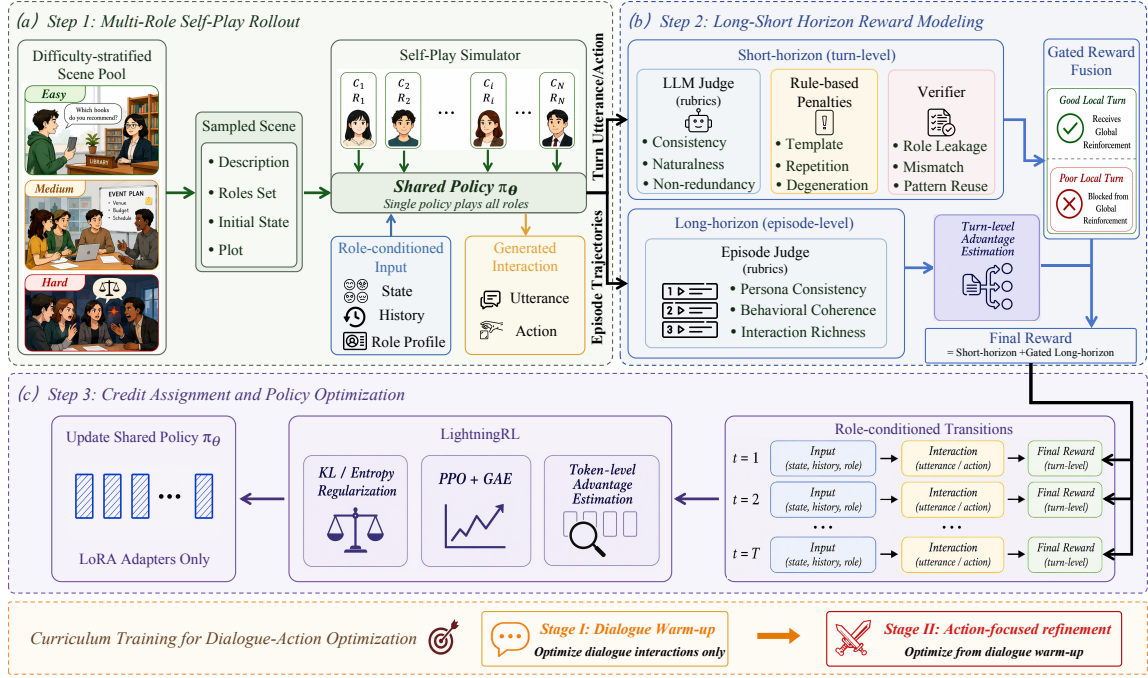


Figure 2: Overview of LSH-RL. A shared policy generates role-conditioned dialogue/action transitions through multi-role self-play, and long-short horizon rewards are fused through gating for policy optimization.

judge under a compact rubric covering character consistency, naturalness, and non-redundancy. Rule-based penalties are further applied to templated continuations, excessive repetition, and other local degeneration patterns. A verifier down-weights reward-hacking-friendly turns, such as role leakage, strong persona mismatch, or obvious pattern reuse. The short-horizon reward is

$$r_{i,t}^s = v_{i,t} \left(\sum_{k \in \mathcal{K}} w_k s_{i,t}^{(k)} - p_{i,t}^{\text{turn}} \right) \quad (2)$$

where $s_{i,t}^{(k)}$ is the judge score on dimension k , w_k is its weight, $p_{i,t}^{\text{turn}}$ is the local penalty, and $v_{i,t} \in [0, 1]$ is the verifier factor.

Long-horizon reward. After a scene ends, an LLM judge evaluates the complete trajectory of each character. For character c_i , its trajectory is $\tau_i = \{u_{i,1}, u_{i,2}, \dots, u_{i,T_i}\}$, together with the scene states and interaction history. The judge scores whether c_i maintains persona consistency, behavioral coherence, and interaction richness across the scene. The long-horizon reward is

$$R_i^l = \sum_{m \in \mathcal{M}} \alpha_m q_i^{(m)} \quad (3)$$

where $q_i^{(m)}$ denotes the trajectory-level score on dimension m , α_m is its weight, and $\mathcal{M}$ contains long-horizon dimensions such as persona consistency,

behavioral coherence, and interaction richness. Unlike a scene-level scalar reward, R_i^l is character-specific, allowing different roles in the same scene to receive different trajectory-level feedback.

Gated reward fusion. The key design in LSH-RL is that trajectory-level credit should reinforce good local behavior rather than wash out weak turns. We introduce a quality gate based on the verifier-adjusted short-horizon reward:

$$g_{i,t} = \begin{cases} 1, & r_{i,t}^s \geq \eta \\ 0, & r_{i,t}^s < \eta \end{cases} \quad (4)$$

where η is the minimum local-quality threshold. The final reward for interaction $u_{i,t}$ is

$$r_{i,t} = \lambda r_{i,t}^s + (1 - \lambda) g_{i,t} R_i^l \quad (5)$$

where $\lambda \in [0, 1]$ balances short-horizon and long-horizon feedback. Before fusion, short- and long-horizon rewards are normalized to comparable scales. If $g_{i,t} = 0$, the interaction receives only its short-horizon reward; if $g_{i,t} = 1$, it is additionally reinforced by the episode-level trajectory reward. This gated fusion prevents locally poor utterances or actions from benefiting from a high trajectory score, while still allowing globally coherent role-playing to strengthen qualified local behaviors. Detailed rubrics and penalties are provided in Appendix B.

3.3 Credit Assignment and Policy Optimization

LSH-RL adopts a gated trajectory-to-turn-to-token credit assignment scheme. After each scene, Eq. 5 produces an explicit reward $r_{i,t}$ for every role-conditioned interaction $u_{i,t}$. This reward is attached to the corresponding generated segment before optimization, so each utterance or action receives turn-specific feedback rather than inheriting credit from a single terminal score.

Each interaction is represented as a token sequence and optimized with PPO in Agent Lightning (Luo et al., 2025). The fused reward is assigned only to the generated utterance or action segment, while input-context tokens are excluded from reward assignment; token-level advantages are estimated using GAE and used in the PPO updates. Thus, LSH-RL decomposes credit at the turn level but does not manually assign identical reward to every token. The shared policy π_θ is updated with the clipped PPO objective, KL regularization toward the warm-start policy, and entropy regularization for exploration. For efficiency, we update only LoRA adapters.

3.4 Curriculum Training for Dialogue-Action Optimization

LSH-RL optimizes both dialogue and action behaviors, but directly training both from the beginning can be unstable. Dialogue quality is easier to judge from local text, whereas action quality depends on the evolving scene state, preceding dialogue, persona constraints, and trajectory development.

We adopt a dialogue-to-action curriculum. The first stage performs dialogue warm-up using the long-short horizon reward and gated credit assignment, stabilizing persona expression and reducing local degeneration in utterances. The second stage performs action-focused refinement on top of the dialogue-warmed policy. It uses the same reward-fusion formulation as the dialogue stage, with an action-specific short-horizon rubric for local plausibility and action–dialogue consistency while retaining the character-level long-horizon reward.

4 Experiments

4.1 Experimental Setup

Data and training. We use 800 of the 1,000 persona profiles released by PersonaArena (Shi et al., 2026) for training and reserve the remaining 200 profiles for evaluation. For each training persona,

GPT-5.1* generates a structured multi-character scene with a narrative premise and 3–4 interacting roles, resulting in 800 training scenes. The scenes are further stratified by difficulty for curriculum learning. Details of scene construction and sampling are provided in Appendix A.

For fair comparison, all trainable variants are based on the same Qwen3-8B backbone. The first RL stage is initialized from the DPO-trained model, which serves as the warm-start policy for reinforcement learning, and optimizes dialogue interactions to produce **LSH-RL-D**, where “D” denotes the dialogue-stage variant. The second stage performs action-focused refinement on top of LSH-RL-D and produces the final **LSH-RL** model. This training follows the curriculum training for dialogue-action optimization described in Section 3.4. PPO is implemented with Agent Lightning (Luo et al., 2025), and only LoRA adapters are updated. Additional training details are provided in Appendix C.

Benchmarks. We evaluate LSH-RL on three role-playing benchmarks: PersonaArena, CharacterBox (Wang et al., 2025a), and RoleBench, introduced as part of the RoleLLM framework (Wang et al., 2024a). PersonaArena and CharacterBox serve as our primary benchmarks because they evaluate multi-turn, scene-grounded role-playing trajectories. We evaluate both benchmarks under dialogue-only and dialogue-plus-action settings. Unless otherwise specified, Qwen3-32B[†] serves as the evaluator and scores each complete trajectory according to the corresponding benchmark rubrics. In contrast, RoleBench serves as a complementary single-turn benchmark. We evaluate models on both the general and role-specific subsets of its English instruction-generalization split using reference-based automatic metrics.

Baselines. To evaluate the effectiveness of LSH-RL, we compare it with both open-source training baselines and strong closed-source baselines. For the open-source comparison, all models are built on the same Qwen3-8B[‡] backbone, including **Base**, **SFT**, and **DPO**. Base denotes the original backbone. SFT is trained on 1,200 PersonaArena-style trajectories, and DPO is further trained with 660 preference pairs.

Closed-source baselines include DeepSeek-

*<https://developers.openai.com/api/docs/models/gpt-5.1>

[†]<https://huggingface.co/Qwen/Qwen3-32B>

[‡]<https://huggingface.co/Qwen/Qwen3-8B>

Model	PersonaArena-Dial.									CharacterBox-Dial.							
	KA	BA	EE	PT	IM	BC	AD	IR	Avg	KA	BA	EE	PT	IM	BC	AD	Avg
<i>Open-source training baselines</i>																	
Base	2.74	2.63	2.31	2.82	2.50	2.50	1.86	1.88	2.40	3.68	3.00	2.48	3.56	2.48	2.80	2.32	2.90
SFT	2.79	2.71	2.22	2.86	2.56	2.62	2.00	1.99	2.47	3.60	3.04	2.40	3.56	2.66	2.74	2.38	2.91
DPO	3.12	2.94	2.42	3.21	2.87	2.99	2.26	2.32	2.77	3.36	3.24	2.50	3.70	2.66	2.94	2.84	3.03
LSH-RL-D	3.26	3.29	2.86	3.63	3.10	3.10	2.59	2.60	3.05	3.76	3.74	2.86	3.82	3.68	3.48	2.86	3.46
<i>Closed-source baselines</i>																	
DeepSeek-v3.2	3.89	3.86	3.32	4.25	3.71	3.73	3.06	3.16	3.62	3.78	3.80	2.78	4.04	3.06	3.36	3.12	3.42
GPT-4.1	4.07	3.85	3.28	4.29	3.68	3.56	3.08	3.14	3.62	4.16	4.06	3.08	4.20	3.18	3.58	3.24	3.64
Doubao-Seed-Character	3.56	3.56	3.15	3.93	3.22	3.44	2.89	3.00	3.34	4.10	3.80	2.70	3.90	2.80	3.40	2.90	3.37
Gemini-3-Flash-Preview	3.47	3.30	3.17	3.53	3.27	3.13	2.80	2.93	3.20	3.70	3.80	3.20	4.10	3.20	3.40	3.50	3.56
GPT-4o-mini	3.79	3.64	2.91	4.09	3.33	3.06	2.52	2.52	3.23	4.60	3.70	2.90	4.20	3.00	3.20	3.00	3.51
Grok-4.3	3.30	3.37	2.73	3.83	3.23	3.13	2.57	2.50	3.08	4.00	3.80	2.80	3.80	2.90	3.40	3.10	3.40

Table 1: Dialogue-only evaluation on PersonaArena and CharacterBox. The light-blue block denotes open-source training baselines built on Qwen3-8B, and the light-beige block denotes closed-source baselines. Bold indicates the best result within each block.

v3.2 (DeepSeek-AI et al., 2025), GPT-4.1[§], GPT-4o-mini[¶], Gemini-3-Flash-Preview^{||}, Grok-4.3^{**}, and Doubao-Seed-Character^{††}.

Metrics. For PersonaArena and CharacterBox, we follow the official evaluation dimensions and report benchmark-aligned averages, covering Knowledge Accuracy (KA), Behavioral Accuracy (BA), Emotional Expression (EE), Personality Traits (PT), Immersion (IM), Behavioral Coherence (BC), Adaptability (AD), and Interaction Richness (IR) when available. For RoleBench, we report Rouge-L scores for English instruction generalization using three types of references: raw general responses (RAW), role-customized general responses (CUS), and role-specific responses (SPE). We further compute exact repetition and near repetition rate to diagnose the local degeneration problem discussed in Figure 1.

4.2 Main Results

We report results under two evaluation settings. The dialogue-only setting evaluates **LSH-RL-D** and the full interaction setting evaluates the final **LSH-RL** model.

Dialogue-only results. Table 1 reports the dialogue-only results. In the controlled compar-

ison, SFT yields limited gains over Base, while DPO provides a stronger warm start. LSH-RL-D further improves Avg from 2.77 to 3.05 on PersonaArena and from 3.03 to 3.46 on CharacterBox, showing that long-short horizon reward learning is effective even before action-focused refinement.

The gains are consistent across dimensions. On PersonaArena, LSH-RL-D improves all eight metrics over DPO, with particularly clear gains in PT, EE, BC, AD, and IR, indicating stronger persona expression, multi-turn coherence, and interaction quality. On CharacterBox, LSH-RL-D also improves most dimensions, especially BA, IM, and BC. Compared with closed-source LLMs, LSH-RL-D still trails the strongest models on PersonaArena but narrows the overall gap to the closed-source baselines on CharacterBox.

Full interaction results. Table 2 reports the full interaction results, where each trajectory contains both utterances and scene-grounded actions. We first compare **LSH-RL-D** with the final **LSH-RL** model to examine the effect of our dialogue-to-action curriculum. LSH-RL-D reaches 3.78 Avg on PersonaArena and 3.22 Avg on CharacterBox, while LSH-RL further improves these scores to 3.85 and 3.31, respectively. This confirms that the second stage is useful for complete interaction quality beyond dialogue-stage optimization. The reward curves in Figure 3 show broadly improving within-run trends in both stages.

Compared with DPO, LSH-RL achieves the best average score among open-source models on both benchmarks. It improves Avg from 3.79 to 3.85 on PersonaArena and Avg from 2.88 to 3.31 on

[§]<https://developers.openai.com/api/docs/models/gpt-4.1>

[¶]<https://developers.openai.com/api/docs/models/gpt-4o-mini>

^{||}<https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview>

^{**}<https://docs.x.ai/developers/models/grok-4.3>

^{††}<https://console.volcengine.com/ark/region:cn-beijing/model/detail?Id=doubao-seed-character>

Model	PersonaArena-Full									CharacterBox-Full								
	KA	BA	EE	PT	IM	BC	AD	IR	Avg	KA	BA	EE	PT	IM	BC	AD	Avg	
<i>Open-source baselines and LSH-RL variants</i>																		
Base	3.78	3.50	3.24	3.75	3.38	3.39	3.21	2.69	3.37	2.98	2.65	2.98	3.19	2.70	2.61	2.40	2.79	
SFT	3.87	3.56	3.27	3.82	3.42	3.46	3.41	2.91	3.47	2.75	2.58	2.61	2.92	2.78	2.60	2.34	2.65	
DPO	4.22	3.83	3.52	4.11	3.77	3.76	3.80	3.30	3.79	2.91	2.81	2.76	3.27	2.93	2.88	2.62	2.88	
LSH-RL-D	4.10	3.80	3.61	4.33	3.78	3.82	3.69	3.08	3.78	2.99	3.10	3.24	3.58	3.40	3.22	3.00	3.22	
LSH-RL	4.31	3.88	3.65	4.17	3.76	3.88	3.79	3.33	3.85	3.09	3.20	3.29	3.66	3.50	3.32	3.13	3.31	
<i>Closed-source baselines</i>																		
DeepSeek-v3.2	4.48	4.08	3.82	4.37	4.04	4.01	4.04	3.61	4.06	3.28	3.48	3.71	4.02	4.01	3.62	3.46	3.66	
GPT-4.1	4.65	4.03	3.67	4.22	3.95	4.00	3.99	3.44	3.99	3.03	3.20	3.37	3.83	3.70	3.54	3.06	3.39	
Doubao-Seed-Character	4.37	3.78	3.67	3.74	3.70	3.70	3.78	3.37	3.76	3.05	2.84	3.15	3.38	3.02	2.83	2.62	2.98	
Gemini-3-Flash-Preview	4.37	3.93	3.87	4.23	3.80	3.67	3.83	3.43	3.89	3.00	3.22	3.48	3.83	3.61	3.54	3.05	3.39	
GPT-4o-mini	4.39	3.97	3.54	4.12	3.85	3.88	3.88	3.39	3.88	2.86	2.81	2.99	3.38	3.35	2.92	2.56	2.98	
Grok-4.3	4.60	3.87	3.43	4.13	3.80	3.83	3.57	3.20	3.80	2.83	2.88	2.98	3.35	3.35	3.08	2.75	3.03	

Table 2: Full interaction evaluation on PersonaArena and CharacterBox. The light-blue block denotes open-source baselines and LSH-RL variants built on Qwen3-8B, and the light-beige block denotes closed-source baselines. Bold indicates the best result within each block.



Figure 3: Smoothed reward curves for the two-stage curriculum.

CharacterBox. The gains are cross-dimensional: LSH-RL improves six of eight PersonaArena dimensions over DPO and improves all seven CharacterBox dimensions over both DPO and LSH-RL-D. These results show that LSH-RL improves broader persona-level role-playing quality, including naturalness, interaction richness, behavioral coherence, and long-horizon persona consistency.

SFT and DPO improve more clearly on PersonaArena than on CharacterBox. For example, DPO increases PersonaArena Avg from 3.37 to 3.79, but only increases CharacterBox Avg from 2.79 to 2.88. Since they are trained on PersonaArena-style data, this suggests that imitation and preference optimization may partly learn dataset-specific role expressions. In contrast, LSH-RL improves both benchmarks, especially CharacterBox, supporting that long-short horizon RL improves transferable interaction quality rather than only surface style.

Comparison with episode-level RL. We further examine whether an episode-level reward alone is sufficient for role-playing optimization. We construct an episode-level RL baseline that starts from

Model	KA	BA	EE	PT	IM	BC	AD	IR	Avg
Base	2.73	2.58	2.24	2.76	2.39	2.45	1.73	1.79	2.33
SFT	2.79	2.67	2.21	2.76	2.52	2.45	2.12	2.00	2.44
DPO	2.97	2.88	2.52	3.15	2.82	2.94	2.21	2.27	2.72
Episode-level RL	1.64	1.82	1.73	2.46	1.82	2.27	1.64	1.55	1.87
LSH-RL-D	3.52	3.46	3.03	3.67	3.36	3.30	2.79	2.79	3.24

Table 3: Comparison with episode-level RL on a PersonaArena dialogue-only subset.

the same DPO-initialized policy as LSH-RL-D and replaces our turn-level reward assignment with a single terminal reward for the entire dialogue. Table 3 reports the results on a shared 10-scene subset of the PersonaArena dialogue-only evaluation.

The episode-level baseline reaches an average score of 1.87, substantially below both DPO (2.72) and LSH-RL-D (3.24), with consistent degradation across all evaluation dimensions. This result suggests a potential limitation of assigning one coarse reward to an entire dialogue: locally weak turns may receive the same positive credit as high-quality turns, causing optimization to deviate from the DPO warm start. LSH-RL-D instead assigns trajectory-level feedback only to locally qualified turns, yielding more reliable credit assignment and substantially stronger dialogue quality.

4.3 Repetition Control

Figure 4 examines whether LSH-RL effectively mitigates repetitive degeneration on PersonaArena. ExactRep and NearRep measure exact and near-duplicate expressions, respectively. Lower values indicate less repetition. In the dialogue-only setting, LSH-RL-D eliminates exact repetition and reduces

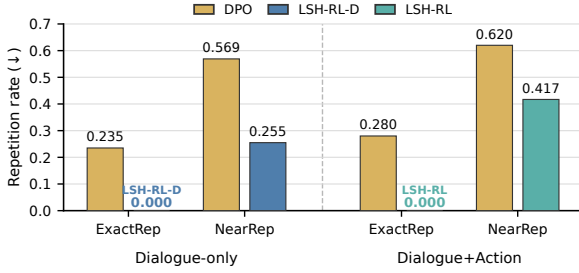


Figure 4: Repetition control on PersonaArena under dialogue-only and dialogue-plus-action settings. Lower is better.

NearRep from 0.569 to 0.255. In the dialogue-plus-action setting, LSH-RL likewise substantially reduces ExactRep from 0.280 to 0.000, and NearRep from 0.620 to 0.417.

These indicators are reported separately from benchmark averages because they diagnose degeneration patterns rather than role-playing quality dimensions directly. The consistent reductions across both evaluation settings suggest that LSH-RL improves interaction quality without reinforcing repetitive response patterns.

4.4 RoleBench Generalization

RoleBench (Wang et al., 2024a) serves as a complementary single-turn instruction-generalization benchmark rather than our primary evaluation of dynamic role-playing. As shown in Table 4, SFT and DPO reduce RAW and CUS relative to Base, whereas LSH-RL-D recovers both scores over DPO, improving RAW from 7.1 to 11.3 and CUS from 14.6 to 20.0. The final LSH-RL achieves the highest SPE score among the controlled variants, increasing it from 13.2 to 16.3 over DPO, although its RAW and CUS scores remain below Base.

This pattern suggests that SFT and DPO may amplify stylistic role expression, sometimes producing longer or more explanatory responses that are less aligned with Rouge-L references. LSH-RL appears to regularize this tendency by improving RAW/CUS over DPO while preserving role-specific performance on SPE. The results therefore suggest a better balance between role-playing style and task-focused correctness, rather than a uniform improvement on every aspect of role-playing. Full RoleBench results including closed-source baselines are provided in Appendix D.

Model	RAW	CUS	SPE
Base	11.9	22.9	13.9
SFT	9.1	18.6	15.5
DPO	7.1	14.6	13.2
LSH-RL-D	11.3	20.0	15.0
LSH-RL	9.8	19.9	16.3

Table 4: RoleBench English instruction generalization using Rouge-L. Bold marks the best controlled result.

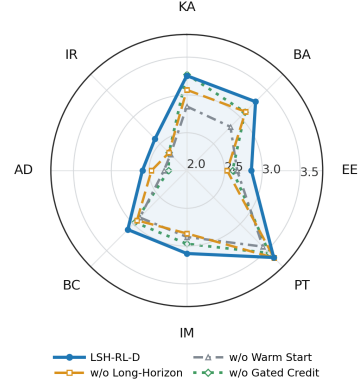


Figure 5: Ablation trends on PersonaArena dialogue-only evaluation. Exact scores are shown in Appendix E.

4.5 Robustness to Evaluator Choice

We further examine whether the full-interaction results are sensitive to the choice of LLM-based evaluator by re-scoring the same model outputs with GPT-4o-mini. Table 5 compares the resulting scores with those obtained using the original Qwen3-32B evaluator.

LSH-RL achieves the highest average score among the controlled variants on both benchmarks under both evaluators. Under GPT-4o-mini, it scores 3.63 on PersonaArena-Full and 3.18 on CharacterBox-Full, outperforming Base by 0.43 and 0.41, respectively. It also maintains clear gains over DPO, particularly on CharacterBox-Full (3.18 versus 2.92). The consistent improvements across evaluator models demonstrate the robustness of the full-interaction results to evaluator choice.

4.6 Ablation Study

Figure 5 summarizes the PersonaArena dialogue-only ablations, with scores reported in Appendix E. Removing the long-horizon reward lowers Avg from 3.05 to 2.86, showing that turn-level feedback alone is insufficient for sustained persona-level role-playing. Removing the DPO warm start causes the largest drop, reducing Avg to 2.77, which indicates that a preference-aligned initialization is important before reinforcement learning.

Model	PersonaArena-Full		CharacterBox-Full	
	Qwen3-32B	GPT-4o-mini	Qwen3-32B	GPT-4o-mini
Base	3.37	3.20	2.79	2.77
SFT	3.47	3.44	2.65	2.86
DPO	3.79	3.55	2.88	2.92
LSH-RL-D	3.78	3.50	3.22	3.11
LSH-RL	3.85	3.63	3.31	3.18

Table 5: Full-interaction results under two LLM-based evaluators.

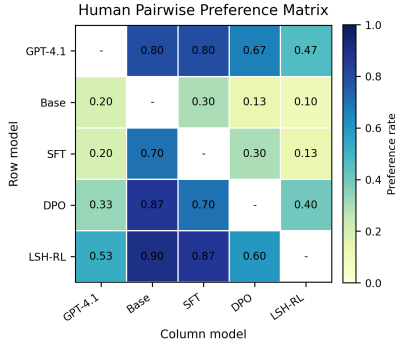


Figure 6: Pairwise human preference matrix.

Removing gated credit assignment reduces Avg to 2.88, with particularly clear drops in adaptability and interaction richness. This supports the central design of LSH-RL: trajectory-level feedback should not be assigned uniformly to all turns, because locally weak turns may otherwise receive undeserved global credit. Overall, the ablations demonstrate the complementary importance of long-short horizon rewards and gated credit assignment in LSH-RL.

4.7 Human Evaluation

We further conduct a human evaluation with three graduate-student annotators with relevant backgrounds in LLMs or role-playing agents. As shown in Figure 6, LSH-RL is preferred over Base, SFT, and DPO at rates of 0.90, 0.87, and 0.60, respectively. Against GPT-4.1, LSH-RL obtains a preference rate of 0.53, providing preliminary evidence of competitive human-perceived trajectory quality in this limited evaluation. Compared with the automatic evaluation results, the human preference margin is more conservative, especially against DPO, partly due to the subjective nature of role-playing evaluation and moderate inter-annotator agreement (Fleiss’ $\kappa = 0.393$). Detailed annotation settings, agreement analysis, and pairwise win/tie/loss results are provided in Appendix F.

5 Conclusion

In this work, we presented **LSH-RL**, a long-short horizon reinforcement learning framework for persona-level role-playing agents. LSH-RL combines short-horizon utterance/action rewards with long-horizon character-trajectory rewards, and uses gated reward fusion to prevent locally weak turns from receiving undeserved global credit. Curriculum training further supports staged dialogue-action optimization in dynamic multi-role self-play. Experiments on PersonaArena and CharacterBox show that LSH-RL improves over SFT and DPO in both dialogue-only and full interaction settings, with reduced repetition, effective action-focused refinement, and higher human preferences. These results suggest that LSH-RL offers a path toward more coherent, natural, and interaction-rich RPAs.

6 Limitations

LSH-RL relies on LLM-based judges for both short- and long-horizon rewards. Although verifier gating and rule-based penalties help reduce reward hacking and local degeneration, the resulting signals may still inherit judge biases, prompt sensitivity, and inconsistency across evaluation settings. Human evaluation is also inherently subjective, as reflected by the moderate inter-annotator agreement. Future work could explore stronger judge calibration, multi-judge aggregation, and larger human studies with finer-grained criteria.

The training scenes are generated from PersonaArena profiles using our simulator. Although results on CharacterBox, RoleBench, and additional transfer settings indicate some generalization, broader validation is needed across more diverse roles, languages, narrative genres, and interaction environments. The current results should therefore be interpreted as evidence of improved trajectory-level role-playing within the evaluated settings rather than universal generalization.

Our primary experiments optimize LoRA adapters on a Qwen3-8B backbone and separate dialogue warm-up from action-focused refinement. Additional experiments on Llama-3.1-8B provide initial evidence of transfer across model families, but broader validation on larger backbones, additional architectures, full-parameter training, and more flexible joint curricula remains necessary.

Acknowledgments

This work was motivated by the Society Zero Universe platform’s need for agents with robust cognitive consistency and high fidelity in role-playing.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62673340, 62276171, 62476173, 62532007), the Shenzhen Science and Technology Program (Grant Nos. JCYJ20240813141503005, JCYJ20240813142610014, ZDCY20250901-110940006), and the Major Special Project for Philosophy and Social Sciences Research of the Ministry of Education (Grant No. 2025JZDZ010).

References

- Marwa Abdulhai, Isadora White, Charlie Victor Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2025. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. In *International Conference on Machine Learning*, pages 126–153. PMLR.
- Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024. Social-bench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Qiming Feng, Qiuji Xie, Xiao-Long Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. Emocharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6218–6240.
- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms’ role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238.
- Bowen Jiang, Taiwei Shi, Ryo Kamoi, Yuan Yuan, Camillo J Taylor, Longqi Yang, Pei Zhou, and Sihao Chen. 2026. One model, all roles: Multi-turn, multi-agent self-play reinforcement learning for conversational social intelligence. *arXiv preprint arXiv:2602.03109*.
- Chonghua Liao, Ke Wang, Yuchuan Wu, Ruoran Li, Fei Huang, and Yongbin Li. 2026. Moa: Multi-objective alignment for role-playing agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11567–11588.
- Xufang Luo, Yuge Zhang, Zhiyuan He, Zilong Wang, Siyun Zhao, Dongsheng Li, Luna K Qiu, and Yuqing Yang. 2025. Agent lightning: Train any ai agents with reinforcement learning. *arXiv preprint arXiv:2508.03680*.
- Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576.
- Wenlong Shi, Jianxun Lian, Mingqi Wu, Haiming Qin, Mingyang Zhou, Xing Xie, Naipeng Chao, and Hao Liao. 2026. Personaarena: Dynamic simulation for evaluating and enhancing persona-level role-playing in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9685–9719.
- Yihong Tang, Kehai Chen, Xuefeng Bai, Benyou Wang, Zeming Liu, Haifeng Wang, and Min Zhang. 2026. Character-r1: Enhancing role-aware reasoning in role-playing agents via rlvr. *arXiv preprint arXiv:2601.04611*.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. Characterbox: Evaluating the role-playing capabilities of llms in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391.
- Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.
- Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024b. Sotopia- π : Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940.
- Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025b. Open-character: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, et al. 2025c. Coser: Coordinating llm-based persona simulation of established roles. In *International Conference on Machine Learning*, pages 64822–64858. PMLR.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024c. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873.
- Zongsheng Wang, Kaili Sun, Bowen Wu, Qun Yu, Ying Li, and Baoxun Wang. 2025d. Raiden-r1: Improving role-awareness of llms via grpo with verifiable reward. *arXiv preprint arXiv:2505.10218*.
- Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An llm solution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290.
- Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025. [Character is destiny: Can persona-assigned language models make personal choices?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059, Suzhou, China. Association for Computational Linguistics.
- Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. [CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557.
- Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. 2025. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11992–12022.
- Pinyi Zhang, Siyu An, Lingfeng Qiao, Yifei Yu, Jingyang Chen, Jie Wang, Di Yin, Xing Sun, and Kai Zhang. 2025. Roleplot: A systematic framework for evaluating and enhancing the plot-progression capabilities of role-playing agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12337–12354.

A Scene Construction and Sampling

Motivation for difficulty stratification. During preliminary dialogue-stage training on a small subset of approximately 100 scenes, we observed that scene difficulty had a noticeable effect on reward stability. As shown in Figure 7, the reward curve exhibits clear fluctuations in the early training process, motivating us to investigate whether difficulty-aware scene organization could improve training stability. Simple scenes with clear social goals usually produced smoother reward feedback, whereas scenes involving implicit emotions, conflicting intentions, or difficult social choices often led to larger reward fluctuations. This observation motivated us to organize the training scenes by difficulty before reinforcement learning, so that each curriculum stage could expose the policy to scenes with appropriate interaction complexity.



Figure 7: Preliminary dialogue-stage reward curve. The observed reward fluctuation motivates difficulty-stratified scene organization.

Scene construction. We construct 1,000 scenes from the PersonaArena persona profiles used in Section 4.1. Each scene is stored as a structured JSON file and contains a title, event description, time and location, role specifications, initial action/dialogue script, plot summary, social purpose, and maximum interaction setting. Each generated scene is grounded in one source profile and instantiates a social situation with 3–4 interacting roles. The role specifications include persona traits, relationship information, local emotional states, and behavioral constraints, so that a rollout requires the model to coordinate individual character identity with the evolving multi-character context. Figure 8 shows a compact example of the structured scene format, where only representative fields are displayed for readability.

```
{
  "scenes": [
    {
      "id": 0,
      "event": "Lila and Nina take a quiet walk after a gathering where Lila felt excluded.",
      "time": "Evening",
      "location": "Denton Square",
      "description": "The two friends walk along the tree-lined path, the soft rustle of leaves and distant chatter from the gathering behind them creating a backdrop of subdued energy. Lila keeps her gaze on the ground, while Nina tries to engage her with lighthearted comments.",
      "characters": [
        {
          "id": 0,
          "name": "Lila Porter",
          "gender": "female",
          "description": "A 25-year-old college student who is introverted, sensitive, and highly empathetic. She is walking with her friend Nina, trying to process her feelings of exclusion from the gathering. Female, College Student, 25 years old; Occupation: College Student; Personality: Introverted, Sensitive, Rational, Highly Empathetic; Values: Emotional control, Friendship, Self-reflection; Interests: Walking in quiet places, Observing surroundings, Emotional introspection; Experiences: Feeling excluded at social gatherings, Struggling with unexpressed emotions, Seeking comfort in close friendships.",
          "position": "Left side of the path",
          "states": "Tense, emotionally reserved, quietly observant.",
        },
        {
          "id": 1,
          "name": "Nina Ramirez",
          "gender": "female",
          "description": "A 24-year-old college student with a more outgoing and warm personality. She is trying to comfort Lila and keep the mood light; Female; Age: 24; Occupation: Psychology major and part-time barista; Personality: Empathetic, optimistic, talkative; Values: Loyalty, kindness, emotional well-being; Interests: Coffee culture, music festivals, journaling; Experiences: Volunteered at a local mental health outreach program, organized a campus peer support group, recently navigated a difficult friendship breakup.",
          "position": "Right side of the path",
          "states": "Concerned, cheerful, attentive",
        },
        ... ]
      },
      "actions": [
        {
          "action_id": 0,
          "character": "Nina Ramirez",
          "character_id": 1,
          "action": "Smiles and tries to lighten the mood.",
          "dialogue": "Hey, Lila, did you see that guy from the chemistry class at the gathering? He was trying to flirt with everyone in the room!"
        },
        ... ]
      },
      "plot": "Lila and Nina take a walk in Denton Square after a gathering where Lila felt left out. Nina attempts to connect with Lila and understand her quiet mood, while Lila struggles to express her feelings but appreciates the support.",
      ...
    }
  ]
}
```

Figure 8: Example of a structured training scene. The figure shows representative fields from the original JSON scene file, including event context, character profiles, initial interactions, plot summary, social purpose, and difficulty cue.

Difficulty annotation. Based on the preliminary observations above, we use an LLM annotator to assign coarse difficulty labels to all 1,000 structured scenes. The labels reflect the number of interacting roles, the amount of persona and relationship information to track, the emotional ambiguity of the scene, and the degree of implicit social inference required. Among the 800 scenes used for training, low-, medium-, and high-difficulty scenes follow a ratio of 2:1:1. Low-difficulty scenes typically involve simple goals and clear role relations; medium-difficulty scenes introduce conflicting intentions or emotional context; and high-difficulty scenes involve dilemmas or socially constrained situations that require balancing competing goals while maintaining persona consistency.

Training schedule. Both the dialogue and action stages run for 200 optimization steps with a batch size of 4. Scenes are sampled without replacement

Reward scope	Dimension	Representative checklist items
<i>Short-horizon dialogue reward</i>		
Dialogue turn	Character Consistency	- The dialogue demonstrates the character’s unique voice and style. - The dialogue reflects the character’s current emotional state.
Dialogue turn	Content Non-redundancy	- The dialogue introduces incremental new information, perspective, or stance beyond recent turns. - The dialogue pushes the conversation forward rather than stalling in rhetorical loops.
Dialogue turn	Naturalness	- The dialogue sounds like authentic spoken language. - The dialogue avoids stiff, formal, written-style phrasing and trivia-dump behavior.
<i>Short-horizon action reward</i>		
Action turn	Character Consistency	- The action demonstrates the character’s unique personality traits. - The action aligns with the character’s goals and motivations.
Action turn	Content Non-redundancy	- The action introduces a fresh interaction beat rather than replaying recent motion. - The action creates non-trivial follow-up possibilities for other characters.
Action turn	Naturalness	- The action feels like believable in-the-moment behavior rather than staged prose. - The action avoids generic flourish that could be pasted into unrelated scenes.
<i>Long-horizon trajectory reward</i>		
Character trajectory	Persona Consistency	- The character demonstrates stable core traits throughout the scene. - The character remains distinguishable from other characters.
Character trajectory	Behavioral Coherence	- The character’s actions and dialogue reinforce each other. - The overall behavior supports the scene’s narrative arc.
Character trajectory	Interaction Richness	- The character contributes varied response strategies instead of repeating one interaction shell. - The character introduces meaningful new information over time.

Table 6: Representative checklist items used by the short-horizon and long-horizon reward heads. Each dimension is implemented with 10 checklist items in the actual judge prompt.

and shuffled within the difficulty-stratified pool, allowing each stage to cover all 800 scenes once. The dialogue stage focuses on stabilizing local utterance quality and persona expression, whereas the action stage refines action choice, state change, and action–dialogue consistency.

B Reward Design Details

LSH-RL uses a three-layer reward design: checklist-based LLM scoring, rule-based degeneration penalties, and verifier-based gating. The checklist layer evaluates short-horizon dialogue turns, short-horizon actions, and long-horizon character trajectories. For each reward dimension, the judge prompt contains 10 checklist items. Table 6 reports two representative checklist items for each dimension. Each item is judged as satisfied, un-

clear or not applicable, or violated, and the resulting dimension-level scores are normalized before reward fusion.

Rule-based penalties. In addition to checklist scoring, we apply lightweight rule-based penalties for degeneration patterns observed during preliminary rollouts. For dialogue, these penalties target repeated sentence openings, excessive question templates, low-information turns, near-duplicate semantic content, and over-dramatic phrasing. For actions, they target repeated action templates, generic filler movements, low incremental state change, and implausibly dramatic behavior descriptions. At the trajectory level, we additionally penalize repeated interaction patterns, question-heavy scene progression, and generic action habits that make characters appear interchangeable. These penal-

Sample level	Rubric reward	Penalty severity	Verifier gate
Turn	Spearman = 0.683	95.0%	85.0%
Trajectory	Spearman = 0.788	–	–

Table 7: Cross-judge agreement for reward components.

ties do not replace LLM-based scoring; instead, they constrain easily exploitable patterns that can receive superficially positive judge scores.

Turn verifier. The verifier is used as a second LLM-based filter. Its purpose is not to assign fine-grained quality scores, but to identify turns that should not receive strong trajectory-level credit even if the broader scene is judged favorably. It checks for format errors, role leakage, third-person narration, obvious template reuse, severe persona mismatch, scene-register mismatch, repetitive expression, and over-dramatic style. The severity level is mapped to a multiplicative gate factor: {none $\rightarrow$ 1.0, minor $\rightarrow$ 0.70, major $\rightarrow$ 0.40, fatal $\rightarrow$ 0.00.} Thus, locally problematic turns are down-weighted before long-horizon fusion.

Reward fusion. The short-horizon reward is computed for each generated dialogue or action turn, while the long-horizon reward is assigned after the scene ends and evaluates the complete character trajectory. In the main experiments, we use equal weighting for short- and long-horizon feedback: $r_{i,t} = 0.5 r_{i,t}^s + 0.5 g_{i,t} R_i^l$, where $r_{i,t}^s$ is the verifier-adjusted short-horizon reward, R_i^l is the character-level trajectory reward, and $g_{i,t}$ is the long-horizon eligibility gate. We set the gate threshold to $\eta = 0.45$, so only turns whose verifier-adjusted short reward is at least 0.45 are eligible to receive the trajectory-level bonus. This equal-weighted design treats local turn quality and trajectory-level persona coherence as equally important, while preventing locally degenerate turns from benefiting from high long-horizon scores.

Reward-side cross-judge analysis. We assess the stability of the reward signals by sampling 100 training-generated character trajectories and their corresponding turn-level responses. GPT-4o-mini re-scores the samples using the same rubrics and prompt structure as the primary Qwen3-32B judge. Table 7 reports agreement for rubric rewards, penalty severity labels, and verifier gate decisions. Correlation is higher at the trajectory level, while the turn-level severity and gate decisions also show

Judge	Lower-reward rate
Qwen3-32B	92.5%
GPT-4o-mini	87.5%

Table 8: Controlled degraded-response reward check.

substantial agreement.

We further construct paired repetitive or template-like variants of sampled responses. Table 8 reports how often each judge assigns the degraded variant a lower reward than the original response. Both judges detect the intended degeneration in most pairs.

C Training Configuration

Shared settings. Both training stages use LoRA-based PPO updates on a Qwen3-8B backbone. LSH-RL-D starts from the DPO-initialized policy, and the final LSH-RL model starts from LSH-RL-D. Training is run on a single node with 4 NVIDIA H100 GPUs. We use Agent Lightning with the VERL PPO backend, vLLM rollout workers, generalized advantage estimation, KL regularization, and entropy regularization. For memory efficiency, we enable gradient checkpointing, remove padding during model execution, and offload actor parameters and optimizer states under FSDP. The reward fusion settings follow Appendix B: $\lambda = 0.5$ and $\eta = 0.45$.

Adapter settings. Only the actor LoRA adapter is updated during RL; the backbone parameters remain frozen. The adapter uses rank $r = 8$, LoRA scaling $\alpha = 32$, dropout 0.05, no bias terms, and PEFT causal-language-model adapters. Dialogue training loads the DPO adapter as its warm start, while action training loads the dialogue-stage adapter checkpoint. Checkpoints are saved as lightweight inference adapters, so actor LoRA weights can be resumed without restoring full actor, critic, or optimizer shards.

D Additional Generalization Results

Table 9 reports the complete Rouge-L results on the English instruction-generalization split of RoleBench for both controlled variants and closed-source baselines. These results complement the main trajectory-level evaluations by measuring overlap with reference responses under role-conditioned instruction prompts rather than performance in dynamic multi-character interactions.

Model	RAW	CUS	SPE
Qwen3-8B Base	11.9	22.9	13.9
Qwen3-8B SFT	9.1	18.6	15.5
Qwen3-8B DPO	7.1	14.6	13.2
LSH-RL-D	11.3	20.0	15.0
LSH-RL	9.8	19.9	16.3
DeepSeek-v3.2	9.4	21.4	13.0
GPT-4.1	14.6	26.3	16.6
GPT-4o-mini	11.4	25.3	16.1
Gemini-3-Flash-Preview	4.7	11.8	8.4
Grok-4.3	16.6	24.8	15.4
Doubao-Seed-Character	16.0	29.1	16.0

Table 9: Full Rouge-L results on RoleBench English instruction generalization.

Model	Avg. Words	Avg. Characters
Qwen3-8B Base	93.7	533.1
Qwen3-8B SFT	152.3	874.2
Qwen3-8B DPO	226.8	1289.8
LSH-RL-D	108.7	606.4
LSH-RL	106.8	588.2
GPT-4.1	73.5	429.7

Table 10: Average response lengths on the RoleBench English instruction generalization.

RoleBench response length. Table 10 reports the average response lengths of the models evaluated on RoleBench. SFT and especially DPO produce substantially longer responses than the base model, whereas LSH-RL-D and LSH-RL produce more concise outputs. These statistics suggest that the recovery in RAW and CUS over DPO is not driven by increased response length and that the lower scores of SFT and DPO may partly reflect lexical mismatch with the RoleBench references.

Transfer across language and character sources. We additionally evaluate 50 scenes drawn from five Chinese cultural sources, with 10 scenes each from *Journey to the West*, *Romance of the Three Kingdoms*, *Dream of the Red Chamber*, *My Fair Princess*, and *The Smiling, Proud Wanderer*. For each source, we evaluate both Chinese-script scenes and English-version counterparts using the CharacterBox-style protocol. Table 11 shows a clear difference between the two settings. LSH-RL yields only a modest improvement in Chinese, indicating limited direct cross-lingual transfer, but improves substantially on the English versions, which change the character sources and narrative genres while remaining closer to the training language.

Transfer to a different backbone family. We further evaluate the dialogue-stage method on

Model	Chinese script	English version
Base	3.214	2.786
SFT	3.235	2.857
DPO	3.048	2.907
LSH-RL-D	3.071	3.519
LSH-RL	3.286	3.712

Table 11: Average transfer results on Chinese-source scenes.

Llama-3.1-8B using the same LoRA-based training and evaluation protocol. As shown in Table 12, LSH-RL-D improves the average score from 2.45 for DPO to 2.94 and outperforms the other controlled variants on most dimensions. These results provide preliminary evidence that the improvements of LSH-RL-D transfer across different 8B-scale backbone families.

E Ablation Results

Table 13 reports the full PersonaArena dialogue-only ablation results behind Figure 5.

F Human Evaluation Details

The human evaluation covers five models: Qwen3-8B Base, SFT, DPO, LSH-RL, and GPT-4.1. We recruit three graduate-student annotators with relevant research backgrounds in LLMs, role-playing agents, or NLP evaluation. For each of the 10 evaluation personas, every pair of models is compared independently by all three annotators, yielding 30 judgments per model pair (10 personas $\times$ 3 annotators). For each comparison, annotators select the trajectory that better demonstrates overall persona fidelity, naturalness, interaction coherence, and scene progression.

Agreement and aggregate trends. The five evaluated models form 10 unique model pairs. Evaluating each pair on 10 personas yields 100 pairwise items, each independently judged by three annotators, for a total of 300 judgments. Among the 100 items, 50 receive unanimous agreement, while the remaining 50 follow a two-versus-one majority pattern. Fleiss’ κ is 0.393, indicating modest inter-annotator agreement in open-ended role-playing evaluation. The aggregate win rates across all pairwise appearances are 0.725 for LSH-RL, 0.683 for GPT-4.1, 0.575 for DPO, 0.333 for SFT, and 0.183 for the base model, consistent with the preference trends summarized in Section 4.7.

Model	KA	BA	EE	PT	IM	BC	AD	IR	Avg
Base	2.25	2.17	2.03	2.62	2.17	2.16	2.08	2.02	2.19
SFT	2.23	2.33	2.21	2.89	2.33	2.22	1.67	1.89	2.22
DPO	2.28	2.25	2.30	3.37	2.42	2.56	2.24	2.18	2.45
LSH-RL-D	3.67	3.33	2.33	4.00	2.83	3.00	2.00	2.33	2.94

Table 12: Dialogue-only PersonaArena results with a Llama-3.1-8B backbone.

Variant	KA	BA	EE	PT	IM	BC	AD	IR	Avg
LSH-RL-D	3.26	3.29	2.86	3.63	3.10	3.10	2.59	2.60	3.05
w/o Long-Horizon Reward	3.07	3.10	2.53	3.63	2.83	2.93	2.47	2.33	2.86
w/o Warm Start	2.85	2.82	2.67	3.42	2.88	2.88	2.30	2.30	2.77
w/o Gated Credit Assignment	3.27	3.09	2.61	3.55	2.97	2.97	2.24	2.33	2.88

Table 13: Full ablation results on PersonaArena dialogue-only evaluation.

Opponent	LSH-RL wins	Opponent wins	LSH-RL preference
Qwen3-8B Base	27	3	0.90
Qwen3-8B SFT	26	4	0.87
Qwen3-8B DPO	18	12	0.60
GPT-4.1	16	14	0.53

Table 14: Pairwise human preference results. Each row contains 30 votes.

Aligned pairwise LLM evaluation. Automatic benchmark evaluation assigns independent dimensional scores, whereas the human study directly compares complete trajectories. To examine the effect of this protocol difference, Qwen3-32B and GPT-4o-mini evaluate the same trajectory pairs used in the human study under an aligned pairwise prompt. Table 15 reports the preferred model and preference rate for each comparison. Both LLM evaluators agree with the human direction for the Base, SFT, DPO, and LSH-RL comparisons. However, humans marginally prefer LSH-RL over GPT-4.1 (53.3%), whereas both LLM judges prefer GPT-4.1 (60.0%). Given this discrepancy and the limited scale of the human study, we treat the human results as complementary evidence rather than a definitive ranking.

G Case Study

Table 16 presents two qualitative case studies summarized from generated simulation trajectories. We report condensed behavior summaries rather than full transcripts to highlight trajectory-level differences while keeping the comparison readable. The examples illustrate the pattern behind the automatic and human-evaluation results: LSH-RL tends to maintain a concrete scene-level plan, introduce new evidence or actions across turns, and avoid common degeneration patterns such as hidden reason-

ing leakage, repeated procedural templates, and role-confused dialogue.

Observed baseline failure modes. The two cases reveal several recurring issues. The base model may violate the interaction protocol by exposing hidden deliberation or shifting speaker positions, which directly harms immersion. SFT usually preserves the topic and persona surface, but often turns the trajectory into a checklist-like plan with repeated sentence functions. DPO produces more plausible local turns, yet it can still over-reinforce safe patterns, such as repeatedly requesting another piece of evidence or adding another logistical artifact. These behaviors are not necessarily off-topic, but they provide limited new social or narrative information after the first few turns.

Why LSH-RL helps. In these examples, LSH-RL improves the trajectories by coupling local turn quality with trajectory-level progression. In the Scarlett case, the model builds a multi-step evidence chain involving public verification, private review, and the father’s gradual shift from suspicion to participation. In the Luke case, the model keeps both practical collaboration and institutional caution active, turning an abstract outreach idea into a negotiated implementation plan. These behaviors align with the improvements in behavioral coherence, interaction richness, and repetition control reported in the main experiments.

Relation to gated long-short reward assignment. The cases also illustrate why gated fusion is useful. A purely trajectory-level reward could still reinforce a mostly coherent scene containing leaked reasoning, repeated question shells, or generic actions, while a purely local reward may prefer fluent

Pair	Human preference	Qwen3-32B preference	GPT-4o-mini preference
Base vs. SFT	SFT (70.0%)	SFT (80.0%)	SFT (100.0%)
Base vs. DPO	DPO (86.7%)	DPO (100.0%)	DPO (100.0%)
Base vs. LSH-RL	LSH-RL (90.0%)	LSH-RL (100.0%)	LSH-RL (100.0%)
DPO vs. LSH-RL	LSH-RL (60.0%)	LSH-RL (90.0%)	LSH-RL (90.0%)
GPT-4.1 vs. LSH-RL	LSH-RL (53.3%)	GPT-4.1 (60.0%)	GPT-4.1 (60.0%)

Table 15: Pairwise preferences on the human-evaluation trajectories under aligned evaluation protocols.

persona-like turns without requiring scene progression. LSH-RL instead reinforces turns that are both locally qualified and useful for the long-horizon character arc, leading to fewer repeated interaction templates and more concrete and meaningful state changes in these examples.

Case	Model	Condensed trajectory behavior	Observation
Scarlett Chavez. Scarlett tries to clear an accusation in front of her father while Rose supports her.	LSH-RL	Scarlett begins with phone and location evidence, then progressively cross-checks Mr. Alvarez’s location history, library app activity, maintenance-light logs, camera-disable records, and staff access-code changes. Her actions alternate between public verification and private review, and the father gradually shifts from suspicion to participation.	The trajectory forms a coherent evidence chain. Each turn adds a new grounded detail while preserving Scarlett’s truth-seeking, controlled, and technology-oriented persona.
	GPT-4.1	Scarlett presents a recorded video, asks the librarian and security guard to verify the timeline, and then moves toward checking security footage with her father.	The trajectory is fluent and coherent, while its resolution relies more on external confirmation than on sustained multi-source state progression.
	Base	The trajectory exposes hidden reasoning markers and repeatedly describes the model’s own planning process before producing the character response. Several turns ask Scarlett-like questions from the wrong speaker position.	The interaction is disrupted by format leakage and role confusion, weakening immersion and persona consistency.
	SFT	Scarlett repeatedly aligns phones, printed metadata, and timestamps, and several turns state the “next step” directly instead of acting it out through the scene.	The trajectory remains on topic but becomes procedural and template-like, with limited social progression.
	DPO	Scarlett repeatedly pulls up new cross-referenced timestamps and asks similar questions about phone logs, app history, and maintenance logs.	The trajectory is locally plausible, but the repeated evidence-query pattern limits new social or narrative development.
Luke Cox. Luke proposes a faith-based youth outreach plan while balancing enthusiasm, church tradition, and community trust.	LSH-RL	Luke uses a whiteboard to turn a vague outreach idea into a concrete plan: elder consultation, youth feedback, mentor training, a youth-center liaison, and a community feedback loop. The plan develops through Jack’s practical suggestions and Mike’s cautious institutional concerns.	The trajectory balances idealism with constraints. It keeps the social tension active while converting it into specific, collaborative planning steps.
	GPT-4.1	Luke frames the project through church service, theological alignment, advisory review, and volunteer agreements.	The trajectory is coherent and well-structured, while placing greater emphasis on institutional approval than on reciprocal youth co-design and evolving social negotiation.
	Base	Several turns expose hidden reasoning markers and describe the model’s deliberation about how to answer before returning to the role-play.	The leakage breaks the simulated interaction and makes the character less believable.
	SFT	The plan repeats clipboard columns, reflection prompts, and cleanup-tracking structures across turns.	The interaction becomes mechanically structured and less socially adaptive.
	DPO	The trajectory repeatedly returns to QR-code signups, printed flyers, mission-statement cards, and sign-up-table adjustments.	The trajectory remains on task, but repeatedly converts the scene into logistical artifacts, weakening interpersonal tension and adaptive negotiation.

Table 16: Qualitative case studies summarized from generated trajectories. The summaries highlight trajectory-level behavior patterns rather than reproducing full transcripts.