OSPD: On-Policy Self-Distillation for Persona-Consistent Dialogue

Anonymous ACL submission

Abstract

Maintaining persona consistency across multi-turn dialogues remains a core challenge for role-playing language models. Off-policy distillation from external teachers incurs distribution mismatch that compounds across dialogue turns, while reinforcement learning struggles with reward ambiguity inherent in subjective persona fidelity. We propose OSPD, an on-policy self-distillation framework where the same model serves as both teacher and student under asymmetric information: the teacher receives a complete character profile while the student sees only a brief summary, and the student generates trajectories from its own policy. We find that teacher confidence in role-playing dialogue exhibits a bimodal structure—sharply peaked at character-critical tokens yet diffuse at generic utterances—and introduce role-aware divergence switching to match this structure. A progressive trait masking curriculum further forces staged internalization of character knowledge along semantic dimensions. Experiments on CharacterBench, CharacterEval, and SocialBench show that OSPD substantially improves persona consistency over supervised fine-tuning and multi-turn RL baselines, without requiring any external teacher or reward model.

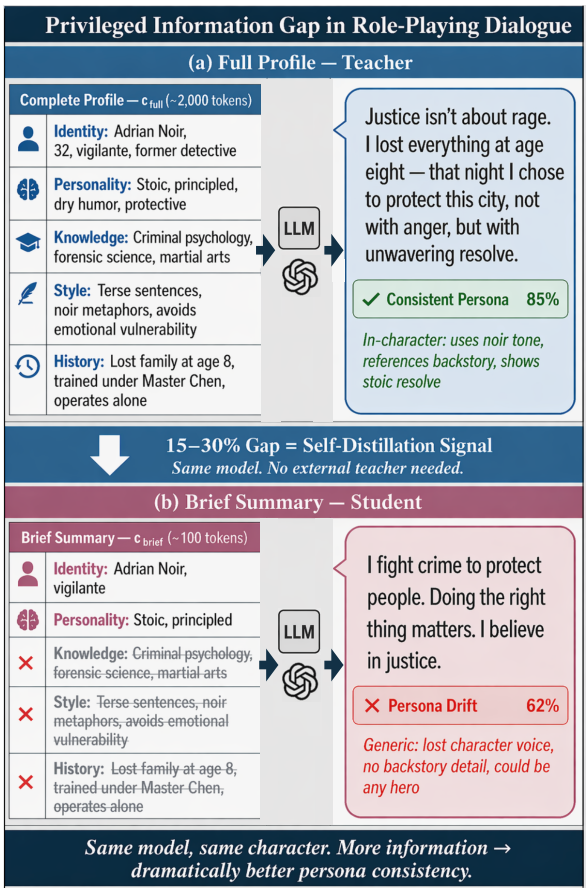

Figure 1: Privileged information gap in role-playing dialogue. The same model conditioned on a complete character profile (c_{full}) produces substantially more persona-consistent responses than when conditioned on a brief summary (c_{brief}). This gap provides a natural self-distillation signal without any external teacher.

1 Introduction

Maintaining a coherent persona across extended dialogue is a central challenge for role-playing conversational agents, with applications in interactive entertainment, educational simulation, and social companionship (Wang et al., 2025; Yu et al., 2025; Abdulhai et al., 2025). A well-functioning agent must faithfully reflect a character's personality, knowledge, and behavioral patterns throughout a conversation (Wang et al., 2024)—yet current systems suffer from *persona drift*: the model gradually abandons its assigned character and reverts to generic behavior as the dialogue progresses (Abdulhai et al., 2025). From an imitation-learning perspective, this is a manifestation of *exposure bias* (Ross et al., 2011)—errors compound across turns because the model generates from its own policy at inference but trains on external data, and a single out-of-character reply corrupts the entire subsequent context.

Existing methods attack this problem through

contrastive learning (Ji et al., 2025), reinforcement learning (Abdulhai et al., 2025; Wang et al., 2026; Ye et al., 2025), and reasoning augmentation (Tang et al., 2025; Qin et al., 2025), but each carries fundamental limitations. Off-policy distillation incurs distribution mismatch whose cumulative error grows quadratically with dialogue length (Ross et al., 2011). RL methods require carefully engineered rewards for the inherently subjective notion of persona fidelity. Contrastive approaches struggle with fine-grained trait distinctions. No prior work addresses persona consistency through on-policy self-distillation, despite its theoretical advantages for reducing exposure bias.

We propose OSPD, an on-policy self-distillation framework that exploits two properties of role-playing dialogue (Figure 2). **First**, the same model shows a large persona-consistency gap under different information budgets (Figure 1): responses conditioned on a complete character profile (c_{full}, $\sim$2 000 tokens) are substantially more consistent than those conditioned on a brief summary (c_{brief}, $\sim$100 tokens). This gap yields a natural self-distillation signal—the model's own *privileged-information mode* serves as the teacher, eliminating the need for any external model. **Second**, teacher confidence follows a *bimodal* distribution: sharply peaked at character-critical tokens (*e.g.* catchphrases, value judgments) yet diffuse at generic utterances. This structure motivates *role-aware divergence switching*—reverse KL at low-entropy positions for precise character imitation, forward KL elsewhere to preserve diversity. OSPD further introduces a *progressive trait masking* curriculum that incrementally withholds character knowledge from the student across training, forcing the model to internalize persona information into its parameters rather than relying on the input prompt.

Experiments on CharacterBench (Zhou et al., 2024), CharacterEval (Tu et al., 2024), and SocialBench (Chen et al., 2024a) demonstrate that OSPD substantially outperforms supervised fine-tuning and is competitive with or superior to multi-turn RL baselines, while requiring neither an external teacher nor a reward model. Multi-turn analysis shows markedly slower persona drift as dialogue length increases to 60 turns. Ablations confirm the contribution of each component, and an analysis of the learned divergence gate validates the bimodal structure across diverse characters and topics.

Our contributions are threefold. *1)* We introduce on-policy self-distillation with privileged information to role-playing dialogue, eliminating the need for external teachers or reward models. *2)* We identify a bimodal confidence structure specific to role-playing dialogue and design a role-aware divergence switching mechanism that exploits it. *3)* We propose progressive trait masking, a curriculum that forces staged internalization of character knowledge along semantic dimensions, and validate it with systematic experiments including ablations, multi-turn robustness analysis, and divergence-gate interpretability studies.

2 Related Work

Knowledge Distillation for Language Models. On-policy distillation lets the student generate its own training trajectories, eliminating off-policy distribution mismatch. GKD (Agarwal et al., 2024), MiniLLM (Gu et al., 2024), and DistiLLM (Ko et al., 2024) can be viewed as instances of a unified f-divergence objective (Song and Zheng, 2026). Recent work introduces per-token adaptive divergences: ToDi (Jung et al., 2025) mixes forward and reverse KL by teacher–student logit ratios, and Entropy-Aware OPD (Jin et al., 2026) gates selection by teacher entropy. Orthogonally, self-distillation removes external teacher dependencies: SPIN (Chen et al., 2024b) distinguishes model-generated from human text via self-play, while OPSD (Zhao et al., 2026), OPCD (Ye et al., 2026), and GATES (Stein et al., 2026) use privileged context as the supervision signal. Penaloza et al. (2026) formalize conditions for privileged information distillation. Existing applications focus on reasoning and instruction following, where the privileged signal is a single correct answer; our work targets role-playing dialogue, where the privileged information is a multi-dimensional character profile whose informativeness varies across positions—motivating both our adaptive divergence and trait masking curriculum.

Persona-Consistent Role-Playing. Methods for persona consistency include contrastive learning (Ji et al., 2025), multi-turn RL (Abdulhai et al., 2025; Wang et al., 2026; Ye et al., 2025), and reasoning augmentation (Tang et al., 2025; Qin et al., 2025). On the data side, CoSER (Wang et al., 2025) provides authentic literary dialogues with structured character knowledge, and Beyond Dialogue (Yu et al., 2025) introduces profile–dialogue alignment. None of these methods exploit the information asymmetry between detailed and condensed

character profiles, nor do they leverage on-policy self-distillation to internalize character knowledge.

3 Method

3.1 Problem Formulation

A character is represented as a pair $\mathcal{C} = (c_{\text{full}}, c_{\text{brief}})$, where c_{full} ($\sim$2 000 tokens) is a *privileged* profile containing detailed background, psychological portrait, relationship graph, representative dialogues, and linguistic style, and c_{brief} ($\sim$100 tokens) is a condensed summary retaining only name and core personality traits. At inference time, only c_{brief} is available—this reflects the practical setting where users provide a short character description rather than an exhaustive dossier.

Let $h_t = (u_1, a_1, \ldots, u_t)$ denote the dialogue history up to turn t. We define two modes of the *same* model:

$$\pi_T(a \mid c_{\text{full}}, h_t) = \text{LM}_{\bar{\theta}}(a \mid c_{\text{full}}, h_t) \quad (1)$$

$$\pi_S(a \mid c_{\text{brief}}, h_t) = \text{LM}_{\theta}(a \mid c_{\text{brief}}, h_t) \quad (2)$$

where θ are the student parameters being optimized and $\bar{\theta}$ is an exponential moving average (EMA) of θ. The information asymmetry resides entirely in the conditioning context; the EMA copy provides a slowly evolving target that stabilizes training (§3.2).

Figure 2 illustrates the overall pipeline. The student generates dialogue trajectories from its own policy, the teacher evaluates each trajectory under privileged information, and the resulting token-level distributional signal drives parameter updates through a role-aware divergence objective.

3.2 On-Policy Self-Distillation

To eliminate the distribution mismatch inherent in off-policy distillation, the student generates its own training trajectories. Each training step proceeds as follows.

Step 1: On-Policy Generation. The student samples a response $\hat{a}_t \sim \pi_S(\cdot \mid c_{\text{brief}}, h_t)$. The training distribution is the student's own behavioral distribution, directly removing exposure bias.

Step 2: Teacher Evaluation. Using the same dialogue context h_t but conditioned on c_{full}, the EMA teacher computes a distribution over each token position of $\hat{a}_t$:

$$\pi_T(\cdot \mid c_{\text{full}}, h_t, \hat{a}_{t,<k}) \quad \text{for } k = 1, \ldots, |\hat{a}_t| \quad (3)$$

These distributions serve as position-wise supervision targets.

Step 3: Divergence Computation. The role-aware divergence loss $\mathcal{L}_{\text{OSPD}}$ is computed over the trajectory (§3.3).

Step 4: Gradient Update. Gradients flow only through the student's log-probabilities; teacher logits are treated as fixed targets.

A key distinction from pure self-play methods such as SPIN (Chen et al., 2024b) is that the OSPD teacher continuously injects external information through c_{full}. The model does not self-reinforce its own biases; instead, it aligns toward a distribution conditioned on richer information, preventing the policy saturation that plagues closed-loop self-distillation.

EMA Teacher Stabilization. Sharing parameters between teacher and student creates a moving-target problem: naive parameter sharing causes the target distribution to shift with every gradient step, risking training oscillation. We maintain an EMA copy of the student parameters as the teacher:

$$\bar{\theta} \leftarrow \alpha\, \bar{\theta} + (1 - \alpha)\, \theta \quad (4)$$

with momentum $\alpha \in [0.99, 0.999]$, following the mean-teacher paradigm (Tarvainen and Valpola, 2017). The EMA teacher evolves slowly, providing a smoothed target that lags behind the student by a controlled margin.

3.3 Role-Aware Divergence Switching

The bimodal confidence structure of role-playing dialogue necessitates position-dependent learning strategies. When the teacher processes character-critical content—domain knowledge, catchphrases, value judgments—the privileged profile provides decisive guidance and the teacher distribution concentrates sharply (low entropy). At generic utterances—greetings, transitions, open-ended remarks—even full profile access leaves considerable uncertainty (high entropy).

We compute the teacher entropy at each token position k:

$$\mathcal{H}_k = \mathcal{H}\big(\pi_T(\cdot \mid c_{\text{full}}, h, \hat{a}_{<k})\big) \quad (5)$$

and define a gating coefficient:

$$\alpha_k = \sigma\left(\frac{\mathcal{H}_k - \mu_{\mathcal{H}}}{\tau}\right) \quad (6)$$

where $\mu_{\mathcal{H}}$ is the running mean of teacher entropy within the current batch and τ is a temperature hyperparameter.

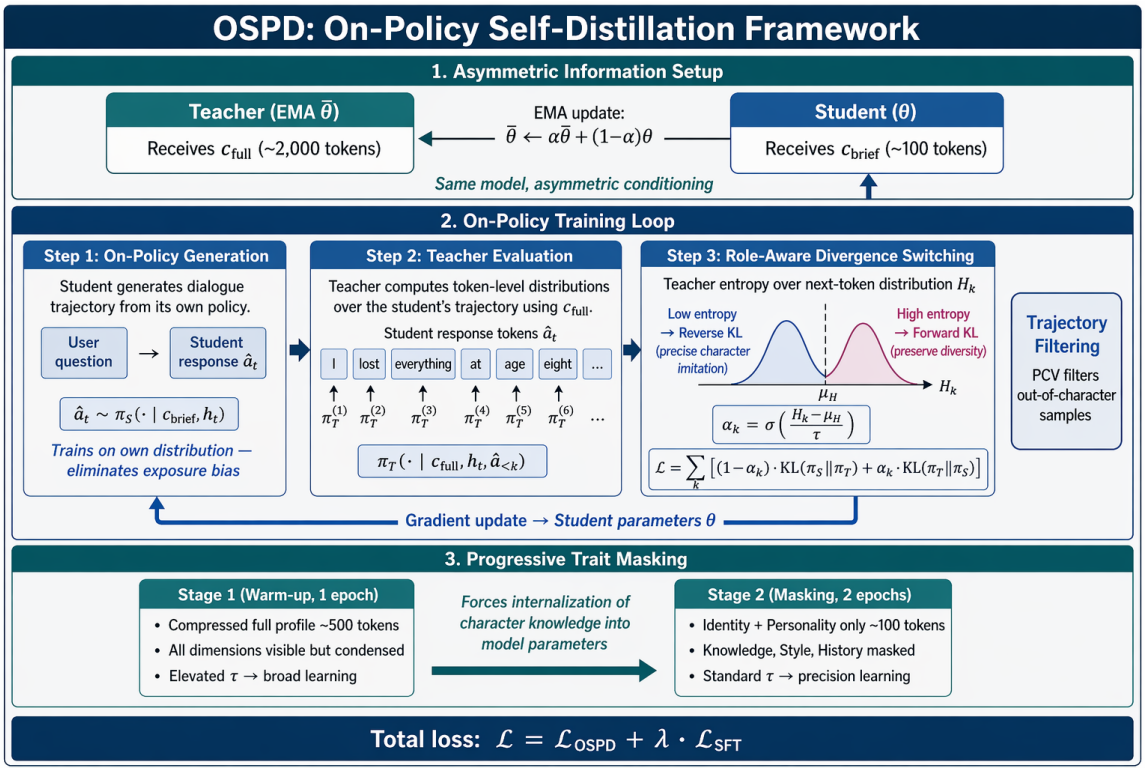

Figure 2: Overview of OSPD. The same model serves as both teacher and student under asymmetric information conditions. The student generates on-policy trajectories from c_{brief}; the EMA teacher provides token-level supervision from c_{full}. Role-aware divergence switching adapts the distillation objective to the bimodal confidence structure of role-playing dialogue. Progressive trait masking transitions from a compressed full profile (Stage 1) to identity and personality only (Stage 2).

The distillation loss interpolates between reverse and forward KL:

$$\mathcal{L}_{\text{OSPD}} = \sum_k \Big[(1-\alpha_k)\, D_{\text{KL}}(\pi_S \| \pi_T) + \alpha_k\, D_{\text{KL}}(\pi_T \| \pi_S) \Big] \tag{7}$$

Low α_k (high teacher confidence) activates **reverse KL**, which is mode-seeking and drives precise imitation of peaked character behaviors. High α_k (low teacher confidence) activates **forward KL**, which is mode-covering and preserves expressive diversity.

This mechanism differs from generic adaptive divergence methods (Jung et al., 2025; Jin et al., 2026) in a fundamental way: our switching logic is grounded in the semantic structure of role-playing dialogue. Teacher entropy directly reflects the informativeness of the privileged profile at the current position—low entropy signals that c_{full} provides decisive guidance, high entropy signals that the profile is non-informative here. The divergence choice thus functions as an implicit detector of *where privileged information matters*.

3.4 Progressive Trait Masking Curriculum

Naive information reduction—simply shortening the profile—ignores the hierarchical structure of character knowledge. We decompose the character profile into five semantic dimensions: *i)* **Identity**: name, age, occupation, relationships; *ii)* **Personality**: core values, emotional disposition; *iii)* **Knowledge**: domain expertise, character-specific facts; *iv)* **Style**: speech patterns, vocabulary, register; *v)* **History**: formative events, growth arc.

Training proceeds in two stages:

Stage 1 (Warm-up, 1 epoch). The student receives a compressed version of all five dimensions (∼500 tokens), while the teacher accesses the full profile. The moderate information gap eases early optimization.

Stage 2 (Masking, 2 epochs). The student sees only *Identity* and *Personality* (∼100 tokens); *Knowledge*, *Style*, and *History* are masked. The teacher retains the full profile. This forces the model to internalize deeper character knowledge into its parameters rather than relying on the input

context.

The masking order reflects the information hierarchy at inference time: identity and personality constitute "public" information that users always specify, whereas knowledge, style, and history are "deep" information that is difficult to convey in a short prompt yet critical for persona consistency. Stage 1 uses an elevated temperature τ to broaden the gate α_k toward forward KL, providing broad-coverage learning; Stage 2 restores the standard temperature, letting the divergence switch be fully entropy-driven.

3.5 Trajectory Filtering

Not all on-policy trajectories are useful for distillation. When the student generates a severely out-of-character response, the teacher's distribution conditioned on this aberrant prefix may itself become unreliable. We introduce a lightweight *Persona Consistency Verifier* (PCV) that filters trajectories at two granularities: turn-level (trait adherence of each reply) and dialogue-level (cross-turn consistency). Only accepted trajectories contribute to $\mathcal{L}_{\text{OSPD}}$.

The PCV is an off-the-shelf quality filter, *not* a training objective—OSPD does not optimize for PCV scores, which fundamentally distinguishes it from RL approaches that optimize against a reward model. Our primary implementation uses GPT-5.1 as the LLM judge; we compare it with an embedding-based MLP classifier in ablations (§5).

3.6 Training Objective

The overall loss combines the self-distillation objective with a supervised fine-tuning (SFT) regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{OSPD}} + \lambda \cdot \mathcal{L}_{\text{SFT}} \tag{8}$$

where $\mathcal{L}_{\text{SFT}}$ is the standard next-token prediction loss on curated character dialogues and λ is tuned from $\{0.1, 0.5, 1.0\}$. The SFT term maintains generation quality and prevents the mode collapse that can arise from pure distributional matching.

4 Experiments

4.1 Setup

Models and Training. We use Qwen2.5-7B-Instruct as the base model for all experiments and extend to Qwen2.5-14B-Instruct for scaling analysis (Appendix E). All methods are fine-tuned with LoRA (rank = 64, $\alpha = 16$, dropout = 0.1) using AdamW with a learning rate of 2e-5 and cosine scheduling. For OSPD, the EMA momentum is $\alpha = 0.995$, on-policy rollouts run for 20 turns (Stage 1) and 40 turns (Stage 2), with 16 dialogues per training step. Training uses 8×H100 GPUs.

Data. We train primarily on CoSER (Wang et al., 2025), which provides 17,966 characters from 771 books with structured character knowledge bases, and supplement with Beyond Dialogue (Yu et al., 2025) for profile–dialogue alignment data. Complete character profiles (c_{full}, $\sim$2 000 tokens) are constructed from CoSER's character knowledge base covering all five trait dimensions; brief summaries (c_{brief}, $\sim$100 tokens) retain only name and core personality. The dialogue partner is Qwen2.5-7B-Instruct, mixing natural conversation (70%) with persona-probing questions (30%) designed from c_{full}.

Benchmarks. We evaluate on three complementary benchmarks: *i)* **CharacterBench** (Zhou et al., 2024): 22,859 human-annotated samples across 3,956 characters with 11 evaluation dimensions (memory consistency, attribute consistency, behavior consistency, empathetic responsiveness, *etc.*) in both Chinese and English, scored on a 1–5 scale by GPT-4; *ii)* **CharacterEval** (Tu et al., 2024): 77 characters with 1,785 multi-turn dialogues in Chinese, measuring 13 metrics across 4 dimensions—we report Character Consistency (CC), focusing on persona–utterance alignment (PU); *iii)* **SocialBench** (Chen et al., 2024a): 500 character profiles with 6,000+ questions and 30,800+ multi-turn dialogues in English, evaluating self-awareness (knowledge and style) and social intelligence via accuracy-based metrics.

Metrics. We report benchmark-specific scores alongside two cross-benchmark metrics: *Prompt-to-Line consistency* (P2L), measuring per-turn alignment with the character profile; and the *Internalization Ratio* (IR = Perf(c_{brief}) / Perf(c_{full})), measuring how well the model performs under limited information.

Baselines. All baselines use the same training data and have access to c_{full} during training. *1)* **SFT**: standard supervised fine-tuning on c_{full}-conditioned dialogues. *2)* **SFT+DPO**: SFT followed by DPO with persona-consistent/inconsistent pairs. *3)* **PCL** (Ji et al., 2025): persona-aware contrastive learning. *4)* **Multi-turn RL** (Abdulhai et al., 2025): PPO with multi-dimensional consistency rewards.

5) **CPO** (Ye et al., 2025): comparative policy optimization for role-playing. *6)* **Off-policy Distill.**: distillation from an independent Qwen2.5-72B teacher. *7)* **OSPD-FixedFwd / OSPD-FixedRev**: OSPD variants with fixed forward or reverse KL only. At inference time, all methods use c_{brief} only.

4.2 Main Results

Table 1 presents the main comparison across all three benchmarks.

OSPD vs. SFT. OSPD outperforms SFT by +0.51 on CharacterBench (3.51 vs. 3.00) and +0.59 on CharacterEval PC (3.49 vs. 2.90), confirming that the self-distillation framework effectively transfers deep character understanding from privileged to limited information conditions. Improvements extend to overall metrics (CE Avg +0.46, SB Avg +8.0%), showing no sacrifice of general performance.

On-policy vs. off-policy distillation. Despite using a 10× smaller model as teacher, OSPD surpasses the off-policy 72B distillation baseline by +0.24 on CharacterBench (3.51 vs. 3.27), demonstrating that on-policy trajectory generation matters more than teacher capacity for persona consistency.

OSPD vs. RL methods. OSPD outperforms Multi-turn RL on CharacterBench (+0.12) and SocialBench SA (+2.1%) without any reward engineering. Multi-turn RL is the strongest baseline on persona-specific metrics (CE PC 3.47) due to its directly optimized consistency rewards, while CPO achieves the highest baseline SB SA (66.0%) through comparative optimization. Neither RL method requires external teachers, but both demand carefully engineered rewards.

Divergence switching matters. Full OSPD outperforms both fixed-divergence variants, with FixedFwd and FixedRev showing complementary weaknesses: FixedRev achieves a comparable CB score (3.38) through mode-seeking but sacrifices self-awareness (SB SA 64.4% vs. 67.6%), while FixedFwd preserves diversity but loses precision at character-critical positions.

Internalization. OSPD achieves an IR of 0.86, substantially closing the gap between c_{brief} and c_{full} conditions compared to SFT (0.71) and even Multi-turn RL (0.78). This indicates that progressive trait masking successfully forces the model to encode character knowledge into its parameters.

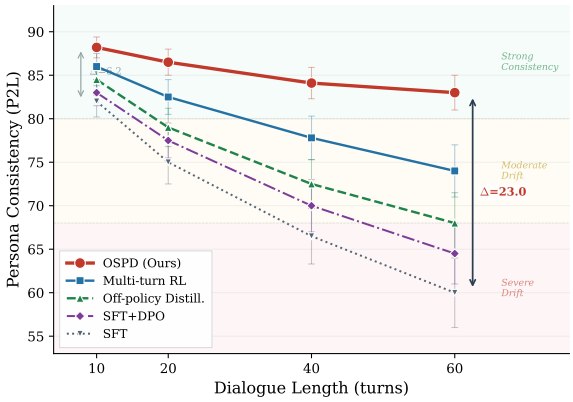

Figure 3: Persona consistency (P2L) as a function of dialogue length. OSPD exhibits markedly slower persona drift compared to all baselines.

5 Analysis

We examine multi-turn robustness, component contributions, the bimodal structure hypothesis, trait masking effects, and generalization to unseen characters.

5.1 Multi-Turn Robustness

Figure 3 plots persona consistency as dialogue length increases from 10 to 60 turns.

OSPD degrades by only 5.2 points from 10 to 60 turns (88.2→83.0), while SFT drops 22.0 points (82.0→60.0) and Multi-turn RL drops 12.0 points (86.0→74.0). The gap between OSPD and baselines widens progressively: at 10 turns the advantage over Multi-turn RL is 2.2 points, growing to 9.0 at 60 turns. On-policy training on long rollouts (40 turns in Stage 2) teaches the model to recover from minor drift rather than compound errors, explaining the flatter decay curve. Notably, Off-policy Distillation degrades faster than Multi-turn RL despite similar 10-turn performance (−16.5 vs. −12.0 from 10 to 60 turns), highlighting the severity of distribution mismatch in extended dialogues.

5.2 Ablation Study

Table 2 isolates the contribution of each component by removing it from the full OSPD system.

On-policy generation. Replacing on-policy trajectories with off-policy data causes the largest single-component drop (−0.22 CB Avg, −0.27 CE PC), confirming that training on the student's own distribution is the most critical design choice.

Divergence switching. Both fixed-divergence variants underperform adaptive switching. Fixed

Table 1: Main results on CharacterBench (CB), CharacterEval (CE), and SocialBench (SB). All methods use c_{brief} at inference. CB Avg: overall average across 11 dimensions (1–5 scale). CE PC: persona consistency (avg of persona–utterance and persona–behavior, 1–5); CE Avg: overall average across 4 dimensions. SB SA: self-awareness accuracy (avg of knowledge and style, %); SB Avg: overall individual-level average. **Bold**: best; underline: second best. †: re-implemented by us.

Method	**CB**	**CharacterEval (1–5)**		**SocialBench (%)**		**IR↑**
	Avg.↑	PC↑	Avg.↑	SA↑	Avg.↑	
SFT	3.00	2.90	3.12	59.4	56.8	0.71
SFT + DPO	3.15	3.08	3.24	61.5	58.5	0.74
PCL†	3.19	3.06	3.22	61.7	58.9	0.73
Multi-turn RL†	3.39	3.47	3.52	65.5	62.4	0.78
CPO†	3.28	3.33	3.41	66.0	63.2	0.76
Off-policy Distill.	3.27	3.16	3.30	64.1	61.5	0.77
OSPD-FixedFwd	3.37	3.36	3.47	65.8	63.0	0.82
OSPD-FixedRev	3.38	3.36	3.43	64.4	61.8	0.80
OSPD	**3.51**	**3.49**	**3.58**	**67.6**	**64.8**	**0.86**

Table 2: Ablation study on CharacterBench (overall avg, 1–5) and CharacterEval persona consistency (PC, 1–5). Each row removes one component from full OSPD.

Variant	**CB Avg↑**	**CE PC↑**
OSPD (full)	3.51	3.49
w/o on-policy (off-policy traj.)	3.29	3.22
Fixed Forward KL	3.37	3.36
Fixed Reverse KL	3.38	3.36
w/o trait masking curriculum	3.41	3.37
w/o PCV filtering	3.47	3.40
w/o $\mathcal{L}_{\text{SFT}}$	3.38	3.27

Forward KL loses precision at character-critical positions (−0.14 CB Avg) while Fixed Reverse KL loses expressive diversity (−0.13 CB Avg, −0.13 CE PC), validating the need for position-dependent divergence selection.

Trait masking. Removing the curriculum (−0.10 CB Avg, −0.12 CE PC) confirms that progressive masking contributes to internalization beyond what on-policy distillation alone provides.

PCV and SFT loss. Removing PCV filtering has a modest effect on CharacterBench (−0.04) but a larger effect on CE PC (−0.09), likely because the smaller benchmark is more sensitive to noisy training trajectories. Dropping $\mathcal{L}_{\text{SFT}}$ causes substantial degradation (−0.13 CB Avg, −0.22 CE PC), indicating that the SFT regularizer is essential for maintaining generation quality under distributional matching.

5.3 Bimodal Structure Analysis

Figure 4 visualizes the distribution of the learned gate α_k and its correspondence with token semantics.

The gate distribution for role-playing dialogue exhibits clear bimodality (Figure 4a), with 38% of tokens falling below $\alpha_k < 0.3$ (character-critical, reverse KL dominant) and 42% above $\alpha_k > 0.7$ (generic, forward KL dominant). This contrasts sharply with math reasoning, where 81% of tokens concentrate below 0.3 (teacher consistently confident), and open chat, where 76% concentrate above 0.7 (teacher consistently uncertain). Token-level inspection (Figure 4b) confirms that low-α_k positions correspond to domain-knowledge terms, character catchphrases, and value-laden expressions, while high-α_k positions correspond to greetings, discourse markers, and generic fillers. The bimodal structure is stable across 85% of tested characters; it weakens for characters with sparse knowledge profiles, where the teacher's confidence advantage diminishes (Appendix G).

5.4 Trait Masking Analysis

Table 3 examines which trait dimensions contribute most to persona consistency by selectively masking individual dimensions in Stage 2.

Knowledge masking alone yields the largest single-dimension gain (+0.07 CB Avg, +0.09 CE PC over no masking), confirming that domain knowledge is the hardest dimension to internalize and benefits most from forced masking. Style masking provides moderate improvement (+0.03 CB Avg) primarily through better linguistic fidelity. History masking yields marginal gains on standard benchmarks (+0.00 CB Avg, +0.02 CE PC)—its benefit is more pronounced in extended dialogues beyond 40 turns (Appendix H), where narrative co-

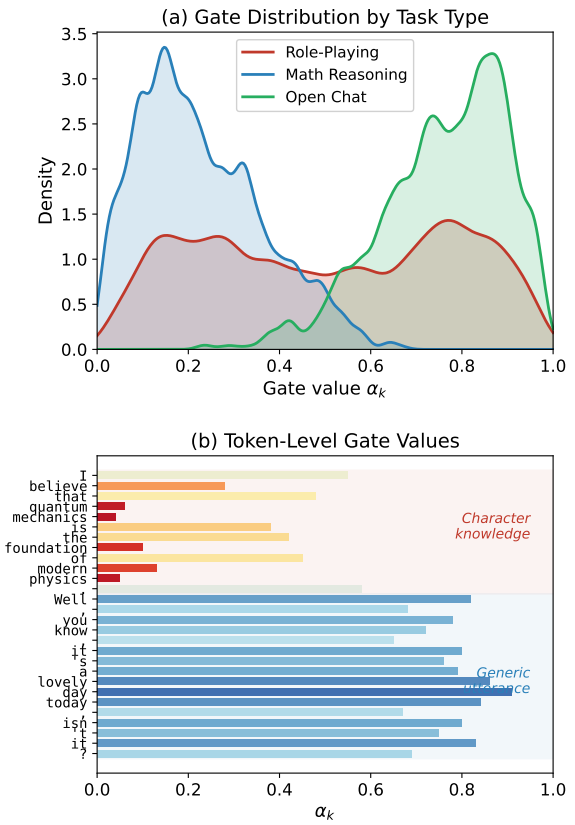

Figure 4: Bimodal confidence structure. (a) Distribution of gate values α_k across training tokens for three task types: role-playing dialogue shows clear bimodality, unlike math reasoning (concentrated low) or open chat (concentrated high). (b) Token-level gate values on an example dialogue: character-knowledge tokens receive low α_k (reverse KL), generic utterance tokens receive high α_k (forward KL).

herence becomes critical. The full three-dimension mask outperforms all single-dimension variants, indicating complementary internalization effects across dimensions.

5.5 Generalization to Unseen Characters

Table 4 evaluates OSPD on held-out CoSER characters not seen during training, with separate analysis for characters from genres represented in the training set (*in-genre*) and from held-out genres (*cross-genre*).

OSPD retains the highest absolute performance on unseen characters, and its in-genre drop (-0.08) is the smallest among all methods, indicating that on-policy self-distillation learns transferable character-modeling capabilities rather than memorizing training characters. Multi-turn RL suffers a notably larger in-genre drop (-0.15), suggesting that its reward-optimized behavior is more character-specific. Cross-genre generalization remains challenging for all methods (OSPD drops 0.20, comparable to Off-policy Distillation), suggesting that genre-specific narrative conventions represent an orthogonal challenge.

Table 3: Effect of masking individual trait dimensions. "Full mask" masks Knowledge + Style + History (default Stage 2). Each single-dimension row masks only that dimension while keeping the other two visible to the student. CB Avg and CE PC on 1–5 scale.

Masked Dimension	CB Avg↑	CE PC↑	IR↑
Full mask (default)	3.51	3.49	0.86
Knowledge only	3.48	3.44	0.84
Style only	3.44	3.40	0.82
History only	3.41	3.37	0.80
None (Stage 1 only)	3.41	3.35	0.79

Table 4: Generalization to unseen characters (CB overall avg, 1–5 scale). Δ denotes the drop from seen-character performance.

Method	In-genre		Cross-genre	
	CB	Δ	CB	Δ
SFT	2.93	-0.08	2.84	-0.17
Multi-turn RL	3.25	-0.15	3.10	-0.29
Off-pol. Distill.	3.19	-0.09	3.07	-0.20
OSPD	3.44	-0.08	3.32	-0.20

6 Conclusion

We have presented OSPD, an on-policy self-distillation framework that exploits the information asymmetry between complete and condensed character profiles within the same model to improve persona consistency in role-playing dialogue. Role-aware divergence switching and progressive trait masking enable the model to internalize deep character knowledge without external teachers or reward models. Experiments on three benchmarks show substantial improvements in persona consistency, particularly across extended dialogues. Our analysis reveals that teacher confidence in role-playing dialogue exhibits an intrinsic bimodal structure that enables position-aware distillation without external reward signals. Progressive trait masking further shows that knowledge and history are the hardest dimensions to internalize, while identity and personality transfer readily from brief descriptions, achieving an internalization ratio of 0.86 that substantially narrows the gap between brief and complete profile conditions.

Limitations

OSPD has several limitations. **Scale dependence.** The teacher's quality is bounded by the base model's ability to leverage complete profiles; models below ∼7B may not benefit from detailed profiles, while larger models may already internalize character knowledge from brief summaries, narrowing the self-distillation signal. **PCV reliability.** The LLM-based trajectory filter may introduce systematic biases that propagate through training; we mitigate this by comparing multiple PCV implementations and reporting acceptance rates. **Profile quality dependence.** The framework assumes access to high-quality, multi-dimensional character profiles; for characters with sparse documentation, the privileged information gap may be insufficient. **Computational overhead.** Dual forward passes and multi-turn on-policy rollouts increase training cost relative to standard SFT, though the self-distillation design avoids the memory overhead of maintaining a separate large teacher and the optimization complexity of RL. **Evaluation scope.** CharacterEval covers only 77 Chinese characters, limiting the statistical power of conclusions drawn from it alone; we use it as a complementary fine-grained analysis alongside the larger-scale benchmarks.

Ethics Statement

Role-playing language models carry risks of misuse, including impersonation of real individuals and generation of harmful content under a character's guise. Our training data derives from published literary works via CoSER, which are publicly available; we do not train on profiles of real individuals. The PCV trajectory filter additionally screens for content that deviates from the intended character, providing a secondary safeguard against generating inappropriate material. We encourage future deployments to pair OSPD with content-safety classifiers and to clearly disclose the AI nature of the agent to end users.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 52920–52957. Curran Associates, Inc.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6621–6642. PMLR.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.

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, Vienna, Austria. Association for Computational Linguistics.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. 2026. Entropy-aware on-policy distillation of language models. *Preprint*, arXiv:2603.07079.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. 2025. ToDi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8091, Suzhou, China. Association for Computational Linguistics.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 24872–24895. PMLR.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. 2026. Privileged information distillation for language models. *Preprint*, arXiv:2602.04942.

Haiming Qin, Jiwei Zhang, Wei Zhang, KeZhong Lu, Mingyang Zhou, Hao Liao, and Rui Mao. 2025. R-CHAR: A metacognition-driven framework for role-playing in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26996–27014, Suzhou, China. Association for Computational Linguistics.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA. PMLR.

Mingyang Song and Mao Zheng. 2026. A survey of on-policy distillation for large language models. *Preprint*, arXiv:2604.00626.

Alex Stein, Furong Huang, and Tom Goldstein. 2026. Gates: Self-distillation under privileged context with consensus gating. *Preprint*, arXiv:2602.20574.

Yihong Tang, Kehai Chen, Muyun Yang, Zheng-Yu Niu, Jing Li, Tiejun Zhao, and Min zhang. 2025. Thinking in character: Advancing role-playing agents with role-aware reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pages 117765–117799. Curran Associates, Inc.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 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, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Zongsheng Wang, Kaili Sun, Bowen Wu, qun yu, Ying Li, Xu Chen, and Baoxun Wang. 2026. Verirole: Verifiable role-awareness through hint-guided reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

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.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. On-policy context distillation for language models. *Preprint*, arXiv:2602.12275.

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, Vienna, Austria. Association for Computational Linguistics.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2024. Characterbench: Benchmarking character customization of large language models. *Preprint*, arXiv:2412.11912.

A Preliminary Study: Privileged Information Gap

Before developing OSPD, we conduct a preliminary study to quantify the performance gap between complete and condensed character profiles and to validate the bimodal confidence structure that motivates our divergence switching mechanism.

A.1 Setup

We select 50 characters from CoSER spanning five literary genres: classical literature (12 characters), contemporary fiction (15), drama (8), historical biography (8), and fantasy/science fiction (7). For each character, we construct c_{full} ($\sim$2 000 tokens) from CoSER's structured character knowledge base

covering all five trait dimensions, and c_{brief} ($\sim$100 tokens) retaining only name and core personality.

Using Qwen2.5-7B-Instruct without any fine-tuning, we generate 50 multi-turn dialogues per character (10 turns each) under both conditions. The dialogue partner is a separate Qwen2.5-7B-Instruct instance that mixes natural conversation (70%) with persona-probing questions (30%) designed from c_{full}. All responses are evaluated by GPT-5.1 on five trait dimensions: identity consistency, personality fidelity, knowledge accuracy, style adherence, and historical coherence, each scored on a 0–100 scale.

A.2 Privileged Information Gap Results

Table 5: Persona consistency scores under complete profile (c_{full}) and brief summary (c_{brief}) conditions, evaluated by GPT-5.1 across five trait dimensions. Δ denotes the absolute gap.

Dimension	c_{full}	c_{brief}	Δ
Identity	85.3	80.1	5.2
Personality	78.6	65.8	12.8
Knowledge	72.4	43.9	28.5
Style	76.2	57.9	18.3
History	68.5	42.8	25.7
Overall	76.2	58.1	18.1

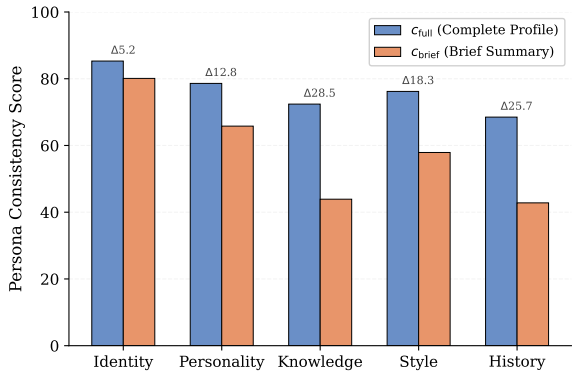

Figure 5: Privileged information gap across five trait dimensions. Knowledge and History suffer the largest drops when switching from c_{full} to c_{brief}.

Table 5 and Figure 5 reveal three findings. **First**, the overall gap is 18.1 points, confirming that privileged information provides a substantial advantage. **Second**, the gap varies dramatically across dimensions: identity (5.2 points) is easily preserved from a brief summary, while knowledge (28.5) and history (25.7) suffer the largest drops—these dimensions require specific factual content that cannot be inferred from name and personality alone. **Third**, style (18.3) occupies a middle ground: the model captures coarse stylistic traits from brief descriptions but misses idiosyncratic speech patterns documented in the full profile.

We additionally compute token-level KL divergence $D_{\text{KL}}(\pi_{\text{full}} \| \pi_{\text{brief}})$ between the model's distributions under both conditions. The mean KL divergence is 0.42 nats, but with high variance ($\sigma = 0.38$). Positions with KL > 1.0 account for 23% of tokens and overwhelmingly correspond to character-specific terminology, knowledge references, and stylistic markers—exactly the positions where privileged information is most informative.

A.3 Bimodal Confidence Structure Validation

To validate the bimodal confidence hypothesis, we compare the teacher entropy distribution of role-playing dialogue against two control tasks: mathematical reasoning (GSM8K) and open-domain chat (ShareGPT). For each task, we compute $\mathcal{H}_k = \mathcal{H}(\pi_T(\cdot \mid \text{context}, \hat{a}_{<k}))$ at each token position across 500 samples and fit a Gaussian mixture model (GMM) with $K = 2$ components.

Table 6: Bimodal fit statistics for teacher entropy distributions across three task types. BIC: Bayesian Information Criterion (lower is better for the selected K). ΔBIC = BIC(K=1) $-$ BIC(K=2); positive values favor the two-component model.

Task Type	Δ**BIC**	μ_1	μ_2	**Sep.**
Role-playing dialogue	+842	0.31	2.15	1.84
Math reasoning	+127	0.18	0.85	0.67
Open-domain chat	+53	1.68	2.82	1.14

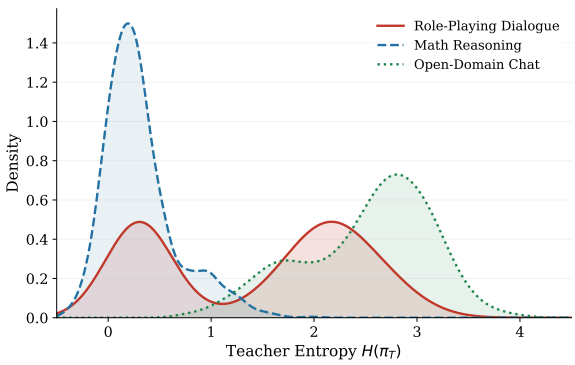

Figure 6: Teacher entropy distributions across three task types. Role-playing dialogue shows clear bimodality, unlike math reasoning (concentrated low) or open chat (concentrated high).

Table 6 and Figure 6 show that role-playing dialogue exhibits by far the strongest bimodal sepa-

ration (ΔBIC = +842, component separation 1.84 nats), compared to math reasoning (concentrated at low entropy, weak secondary mode) and open chat (concentrated at high entropy). This confirms that the privileged profile creates two distinct regimes in role-playing dialogue: decisive guidance at character-critical positions and minimal advantage at generic positions.

To validate the semantic correspondence, we manually annotate 100 randomly sampled tokens from each of the low-entropy ($\mathcal{H}_k < 0.5$) and high-entropy ($\mathcal{H}_k > 2.0$) clusters. Among low-entropy tokens, 72% are character-specific knowledge terms, 15% are catchphrases or stylistic markers, and 13% are value-laden expressions. Among high-entropy tokens, 64% are discourse markers and fillers, 22% are common greetings or transitions, and 14% are open-ended expressions.

B Implementation Details

B.1 Hyperparameters

Table 7: Complete hyperparameter configuration for OSPD.

Hyperparameter	Value
Model & LoRA	
Base model	Qwen2.5-7B-Instruct
LoRA rank / α / dropout	64 / 16 / 0.1
LoRA targets	q, k, v, o_proj
Optimization	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Learning rate	2e-5, cosine, 100-step warmup
Weight decay / grad clip	0.01 / 1.0
Batch size	16 dialogues per step
OSPD-*specific*	
EMA momentum α	0.995
Div. temperature τ	1.0 (Stg. 2), 2.0 (Stg. 1)
SFT loss weight λ	0.5
Rollout length	20 turns (Stg. 1), 40 (Stg. 2)
Epochs	1 (Stg. 1), 2 (Stg. 2)
PCV threshold	0.6 (turn), 0.7 (dialogue)
Inference	
Decoding	Top-p ($p = 0.9, T = 0.7$)
Max response length	512 tokens

B.2 Training Infrastructure

All experiments run on a single node with 8×NVIDIA H100 80GB GPUs. On-policy trajectory generation uses vLLM (v0.4.1) for batched inference with tensor parallelism across 4 GPUs. The remaining 4 GPUs handle gradient computation via DeepSpeed ZeRO Stage 2. Teacher evaluation (Step 2 in §3.2) runs on the vLLM instance with the EMA checkpoint loaded; checkpoint synchronization occurs every training step via shared NVMe storage.

Table 8: Wall-clock training time breakdown for OSPD (Qwen2.5-7B, 3 epochs total, 8×H100).

Component	Hrs	%
On-policy generation	8.2	44.1
Teacher eval. (dual fwd.)	5.1	27.4
Divergence & gradient	3.8	20.4
PCV filtering	0.9	4.8
EMA & checkpointing	0.6	3.2
Total	**18.6**	100

Trajectory generation dominates wall-clock time (44.1%), as each training step requires generating multi-turn dialogues from the student's current policy. The dual forward pass (teacher evaluation) adds 27.4%. Overall, OSPD training takes approximately 4.4× the wall-clock time of standard SFT (4.2 hours) but substantially less than Multi-turn RL (38.5 hours), which additionally requires a critic model and PPO-style rollout buffers.

B.3 Evaluation Protocol

CharacterBench. We follow the official evaluation protocol, submitting model responses to the CharacterBench API. Each character is evaluated through 5–7 multi-turn dialogues, and scores are averaged across 11 dimensions. We report the three aggregate scores: Attribute consistency (Attr.), Behavioral consistency (Behav.), and the overall Average (Avg.).

CharacterEval. We generate 5-turn dialogues for each of the 77 characters using the official prompts. Responses are evaluated by GPT-5.1 using the CharacterEval rubric, which scores Persona–Utterance alignment (PU, per-turn) and Persona–Behavior alignment (PB, cross-turn).

SocialBench. We follow the official evaluation script, generating responses to 6,000+ character-specific questions and 30,800+ multi-turn dialogues. The evaluation covers three dimensions: character Knowledge (factual accuracy about the character), character Style (linguistic and behavioral fidelity), and social intelligence. We report Knowledge and Style as the primary persona-consistency metrics.

GPT-5.1 Judge Configuration. All GPT-5.1 evaluations use `gpt-5.1-2026-01-15` with

temperature 0 and max tokens 1024. To mitigate position bias, we randomize the order of compared outputs and average over two evaluations per sample. The evaluation prompt instructs the judge to score each dimension independently on a 0–100 scale with explicit rubric anchors at 0, 25, 50, 75, and 100.

C Prompt Templates

C.1 Character Profile Examples

We provide two representative character profiles illustrating the contrast between c_{full} and c_{brief}.

Example 1: Sherlock Holmes. c_{full} (Complete Profile, ∼1 800 tokens):

> **Identity.** Sherlock Holmes is a consulting detective residing at 221B Baker Street, London. He is approximately 35 years old, unmarried, and works independently with cases referred by Scotland Yard and private clients. His closest associate is Dr. John Watson, a former army surgeon who serves as his biographer and companion. His landlady is Mrs. Hudson. His elder brother Mycroft Holmes holds an influential position in the British government.
>
> **Personality.** Holmes possesses an extraordinarily analytical mind with an almost obsessive attention to detail. He is intellectually arrogant but self-aware about this trait. He oscillates between manic energy during cases and profound lethargy between them. He values logic and reason above sentiment, often appearing cold or insensitive, yet demonstrates deep loyalty to Watson. He has a strong moral compass but operates outside conventional social norms. He is prone to dramatic flair when revealing deductions.
>
> **Knowledge.** Expert in chemistry (particularly poisons and forensic analysis), anatomy, botany (limited to poisonous plants), geology (practical), criminal law, sensational literature, and the geography of London. Deliberately ignorant of astronomy, philosophy, and politics. Maintains extensive files on criminal cases and individuals. Plays violin proficiently. Has published monographs on tobacco ash, typefaces, and tattoos.
>
> **Style.** Speaks in precise, clipped sentences when deducing; longer, more contemplative passages when theorizing. Frequently uses rhetorical questions ("You see, but you do not observe"). Employs technical terminology from chemistry and anatomy naturally. Addresses Watson as "my dear Watson" or simply "Watson." Uses dry, sardonic humor. Avoids emotional language; describes feelings in clinical terms.
>
> **History.** Attended university where he first developed his deductive method. Spent two years as a young man in Montague Street before establishing practice at Baker Street. Survived a confrontation with Professor Moriarty at the Reichenbach Falls, spending three years traveling incognito (the "Great Hiatus"). Has been physically assaulted during cases on multiple occasions and bears several scars. His first significant case involved a college friend's father.

c_{brief} (Brief Summary, ∼80 tokens):

> Sherlock Holmes is a brilliant consulting detective known for exceptional deductive reasoning and keen observation. He is analytical, intellectually arrogant, and often socially detached. He resides at 221B Baker Street, London, and works closely with his companion Dr. John Watson.

Example 2: Lin Daiyu. c_{full} (Complete Profile, ∼2 100 tokens):

> **Identity.** Lin Daiyu is a young woman of noble birth, daughter of Lin Ruhai (a Yangzhou salt commissioner) and Jia Min (daughter of the Jia family). After her mother's death, she was brought to the Rongguo Mansion to live with her maternal grandmother, the Dowager Jia. She is approximately 14–16 years old throughout the narrative. Her closest bond is with Jia Baoyu, her cousin and soulmate.
>
> **Personality.** Daiyu is extraordinarily sensitive, emotionally intense, and deeply intelligent. She is quick to perceive slights (real or imagined) and prone to jealousy, particularly regarding Baoyu's attention to Xue Baochai. Beneath her prickly exterior lies profound vulnerability rooted in her orphan status and precarious position as a dependent in the Jia household. She values authenticity and sincerity above social convention, often clashing with the pragmatic expectations of her environment.
>
> **Knowledge.** Exceptionally talented in classical Chinese poetry and literature. She composes poems spontaneously with technical mastery and emotional depth. Well-versed in Buddhist and Daoist philosophical concepts. Has read forbidden novels including *Romance of the Western Chamber*. Knowledgeable about traditional medicine due to her chronic illness.
>
> **Style.** Speaks with sharp wit and literary allusion, often embedding poetic references. Uses self-deprecating humor tinged with melancholy. Her language shifts between cutting sarcasm (when jealous or defensive) and tender vulnerability (with Baoyu or when composing poetry). Frequently coughs or references her illness. Addresses servants with more warmth than she shows most aristocratic peers.
>
> **History.** Lost her mother at a young age and was sent to the Jia household. Has suffered from chronic illness since childhood and takes daily herbal medicine. Formed a deep spiritual bond with Baoyu upon their first meeting, with both feeling they had met before. Won the poetry competition in the Grand View Garden. Secretly exchanged handkerchiefs with Baoyu as tokens of affection. Burned her poems and manuscripts before her death.

c_{brief} (Brief Summary, ∼90 tokens):

> Lin Daiyu is a young noblewoman of exceptional poetic talent and emotional sensitivity. She is intelligent, melancholic, and prone to jealousy, with

a sharp wit that masks deep vulnerability. An orphan living as a dependent in her grandmother's wealthy household, she shares a profound romantic bond with her cousin Jia Baoyu.

C.2 Trait Masking Examples

During Stage 1, the student receives a compressed version of all five dimensions (∼500 tokens). During Stage 2, only Identity and Personality remain; Knowledge, Style, and History are removed. Table 9 illustrates this progression for Sherlock Holmes.

Table 9: Student input at each training stage for Sherlock Holmes.

Stage	Student Input
Stage 1	Sherlock Holmes, consulting detective at 221B Baker Street, companion Dr. Watson. Analytical, arrogant, dramatic. Expert in chemistry and forensics. Speaks precisely with rhetorical questions and dry humor. Survived Reichenbach Falls confrontation with Moriarty.
Stage 2	Sherlock Holmes is a consulting detective at 221B Baker Street. Works with Dr. Watson. Analytical, intellectually arrogant, and socially detached.

C.3 PCV Evaluation Prompt

The Persona Consistency Verifier (PCV) operates at two levels. The turn-level prompt evaluates a single response; the dialogue-level prompt evaluates cross-turn consistency.

Turn-Level PCV Prompt.

You are evaluating whether a character's dialogue response is consistent with their assigned persona profile.

Character Profile:
{character_profile}

Dialogue History:
{dialogue_history}

Character Response:
{response}

Evaluate the response on the following dimensions (score 0–100 each):

1. **Trait Adherence**: Does the response reflect the character's stated personality traits, values, and emotional disposition?
2. **Knowledge Accuracy**: If the response references domain knowledge, is it consistent with the character's documented expertise?
3. **Linguistic Fidelity**: Does the speech pattern match the character's documented style (vocabulary, register, catchphrases)?
4. **Behavioral Consistency**: Is the character's action or reaction plausible given their documented history and personality?

Return a JSON object: {"trait": score, "knowledge": score, "style": score, "behavior": score, "accept": true/false}.
Set "accept" to true if ALL scores are ≥ 60.

Dialogue-Level PCV Prompt.

You are evaluating the cross-turn consistency of a character across an entire dialogue.

Character Profile:
{character_profile}

Full Dialogue (N turns):
{full_dialogue}

Evaluate cross-turn consistency:

1. **Personality Stability**: Does the character maintain consistent values, emotional patterns, and behavioral tendencies across all turns?
2. **Knowledge Coherence**: Are factual claims consistent across turns? Does the character avoid contradicting previously stated information?
3. **Style Continuity**: Does the linguistic style remain stable throughout?

Return a JSON object: {"stability": score, "coherence": score, "style": score, "accept": true/false}.
Set "accept" to true if ALL scores are ≥ 70.

C.4 Dialogue Partner Prompts

The dialogue partner generates user-side utterances during on-policy trajectory generation. We employ two modes, randomly selected at each turn.

Natural Conversation Mode (70%).

You are having a natural conversation with {character_name}. Engage as a curious and friendly conversation partner. Ask follow-up questions, share your own (fictional) thoughts, and respond naturally to what the character says. Do NOT probe for specific facts about the character; let the conversation flow organically. Keep your responses to 1–3 sentences.

Persona-Probing Mode (30%).

You are testing whether {character_name} can maintain their persona. Based on the following character knowledge, ask a question that requires the character to demonstrate specific knowledge, personality traits, or behavioral patterns from their profile.

Character Knowledge (selected dimension):
{selected_trait_from_c_full}

Ask a natural-sounding question that indirectly probes this knowledge. Do NOT quote the profile directly or make the question feel like a quiz. Keep your question to 1–2 sentences.

D Extended Ablation Studies

D.1 EMA Teacher Variants

Table 10 confirms that EMA stabilization is critical: naive parameter sharing degrades performance by 0.25 CB Avg. and 0.29 CE PU due to moving-target oscillation. Among EMA configurations, α=0.995 achieves the best balance between target

Table 10: EMA teacher variants. "Periodic (K=50)" updates the teacher by copying student parameters every 50 gradient steps. "Naive shared" uses the student's current parameters as the teacher (no EMA).

Teacher Update Strategy	CB Avg.↑	CE PU↑
EMA α=0.990	3.43	3.39
EMA α=0.995 (default)	3.51	3.48
EMA α=0.999	3.46	3.42
Periodic (K=50)	3.39	3.35
Naive shared	3.26	3.19

stability and responsiveness to student improvement. α=0.990 updates too aggressively, partially inheriting the moving-target problem (-0.08 CB Avg.). α=0.999 updates too slowly, causing the teacher to lag behind the student's improving policy and provide stale supervision (-0.05 CB Avg.). Periodic snapshot updates (K=50) underperform EMA because the abrupt target shifts at snapshot boundaries cause gradient spikes, which we observe as training loss oscillation in the loss curves.

D.2 PCV Implementation Variants

Table 11: PCV implementation comparison. Accept Rate denotes the proportion of generated trajectories passing the filter.

PCV Variant	CB Avg.↑	CE PU↑	Accept
GPT-5.1 judge (default)	3.51	3.48	72.3%
Embedding MLP	3.48	3.41	78.5%
No PCV	3.47	3.38	100%

The GPT-5.1 judge outperforms the embedding-based MLP classifier, particularly on CharacterEval ($+0.07$ PU), despite filtering more trajectories (72.3% vs. 78.5% acceptance rate). The MLP classifier, trained on 5,000 manually labeled trajectory segments using sentence embeddings from a frozen encoder, captures surface-level consistency but misses subtle persona violations that the GPT-5.1 judge detects through reasoning. Both PCV implementations outperform no filtering, confirming that trajectory quality control benefits distillation.

D.3 SFT Loss Weight

Table 12: Effect of SFT loss weight λ on performance and generation quality. Fluency: GPT-5.1 fluency score (0–100).

λ	CB Avg.↑	CE PU↑	IR↑	Fluency↑
0	3.38	3.29	0.83	72.4
0.1	3.49	3.46	0.85	85.3
0.5	3.51	3.48	0.86	91.7
1.0	3.48	3.44	0.84	93.2

Without SFT regularization ($\lambda = 0$), the model achieves reasonable persona consistency but suffers significant fluency degradation (72.4), with outputs exhibiting repetitive patterns and incomplete sentences—symptoms of mode collapse under pure distributional matching. $\lambda = 0.1$ restores most fluency (85.3) and persona scores. $\lambda = 0.5$ provides the best overall balance: near-optimal fluency (91.7) with peak persona consistency. $\lambda = 1.0$ slightly reduces persona scores as the SFT objective begins to compete with the distillation signal, prioritizing generic dialogue quality over character fidelity.

D.4 Divergence Temperature Sensitivity

Table 13: Effect of divergence switching temperature τ on gate behavior and downstream performance. $\bar{\alpha}$ denotes the mean gate value; σ_α the standard deviation.

τ	CB Avg.↑	CE PU↑	$\bar{\alpha}$	σ_α
0.5	3.41	3.38	0.49	0.42
1.0	3.51	3.48	0.51	0.34
2.0	3.48	3.44	0.50	0.25
5.0	3.41	3.37	0.50	0.12

The temperature τ controls gate sharpness. Low τ (0.5) produces near-binary gating ($\sigma_\alpha = 0.42$): most positions receive either pure reverse or pure forward KL, creating a harsh transition that misclassifies ambiguous tokens. The default $\tau = 1.0$ produces a clear bimodal distribution while retaining a graded transition zone. High τ (2.0, 5.0) increasingly smooths the gate toward a uniform mixture, effectively degenerating toward a fixed 50/50 blend and losing the adaptive advantage (-0.10 CB Avg. at $\tau = 5.0$).

E Scaling and Efficiency

E.1 Model Scale: 7B vs. 14B

Table 14 presents the full scaling comparison. OSPD at 14B achieves the highest absolute performance across all metrics (3.64 CB Avg., 0.91 IR), indicating that the framework scales effectively with model capacity. The improvement of OSPD over SFT is consistent across scales ($+0.51$ CB Avg. at 7B, $+0.45$ at 14B), suggesting that the self-distillation benefit is not diminished at larger scale. The slight reduction in relative improvement at 14B (-0.06 CB Avg. gap) is consistent with the hypoth-

Table 14: Full comparison between Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct across all benchmarks and metrics. Both models use the same OSPD configuration.

Scale	CharacterBench			CharacterEval		SocialBench		IR↑
	Attr.↑	Behav.↑	Avg.↑	PU↑	PB↑	Know.↑	Style↑	
SFT Baselines								
7B	3.12	2.94	3.00	2.97	2.84	57.2	61.5	0.71
14B	3.29	3.13	3.19	3.16	3.01	61.5	64.8	0.75
OSPD								
7B	3.58	3.47	3.51	3.48	3.49	67.3	67.9	0.86
14B	3.71	3.61	3.64	3.63	3.66	70.8	71.5	0.91
OSPD *Improvement over SFT*								
7B	+0.46	+0.53	+0.51	+0.51	+0.65	+10.1	+6.4	+0.15
14B	+0.42	+0.48	+0.45	+0.47	+0.65	+9.3	+6.7	+0.16

esis that larger models already internalize more character knowledge from brief profiles, narrowing the privileged information gap. Notably, the IR improvement is nearly identical (+0.15 at 7B, +0.16 at 14B), indicating that the internalization mechanism is scale-independent.

E.2 Efficiency Analysis

Table 15: Training efficiency comparison (Qwen2.5-7B, 3 epochs, 8×H100). GPU Mem.: peak per-GPU memory. Separate teacher column indicates whether a distinct model is loaded.

Method	Hrs	GB	Ext. T	CB
SFT	4.2	42	×	3.00
SFT+DPO	7.8	48	×	3.15
Multi-turn RL	38.5	78	×	3.39
Off-pol. Dist.	12.8	72	✓	3.27
OSPD	18.6	52	×	3.51

OSPD occupies a favorable efficiency–performance trade-off: it achieves the highest CB Avg. (3.51) while requiring only 18.6 hours—less than half the cost of Multi-turn RL (38.5 hours, 78 GB peak memory). Compared to Off-policy Distillation, OSPD uses less peak GPU memory (52 vs. 72 GB) because it does not load a separate 72B teacher, instead reusing the student's own parameters with an EMA copy. The primary computational overhead relative to SFT is the on-policy rollout generation (8.2 hours) and dual forward pass for teacher evaluation (5.1 hours); both scale linearly with rollout length and can be parallelized across dialogue instances.

F Representation Probing

F.1 Probing Setup

To assess how deeply different training methods internalize character knowledge, we train linear probes on the model's hidden representations. For each trait dimension $d \in \{$Identity, Personality, Knowledge, Style, History$\}$, we construct a binary classification task: given a hidden state at a character-relevant token position, predict whether the model was conditioned on c_{full} or c_{brief}. Higher probing accuracy indicates that the representation encodes more dimension-specific character information—and thus that the model has internalized more of that dimension into its parameters.

We extract hidden states from layer l at token positions identified by the PCV as character-relevant. The probe is a single linear layer trained with logistic regression (L2 regularization $C = 1.0$) on 10,000 samples per dimension (80/20 train/test split). We probe every 4th layer of the 32-layer Qwen2.5-7B model (layers 4, 8, 12, 16, 20, 24, 28, 32) and report results for three training methods: SFT, Multi-turn RL, and OSPD.

F.2 Probing Results

Table 16: Linear probing accuracy (%) per trait dimension, extracted from layer 24 (best overall layer). Higher accuracy means the representation captures more dimension-specific information.

Dimension	SFT	Multi-turn RL	OSPD
Identity	86.3	89.5	92.1
Personality	74.8	80.2	85.7
Knowledge	58.2	68.5	78.3
Style	65.3	74.8	81.5
History	51.7	62.3	72.8
Average	67.3	75.1	82.1

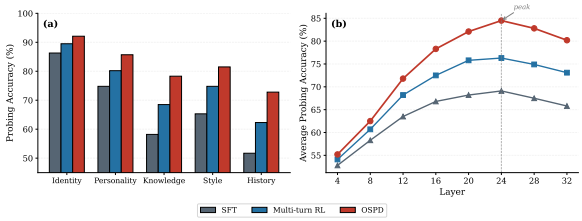

Figure 7: Representation probing results. (a) Per-dimension probing accuracy at layer 24. (b) Layer-wise average probing accuracy. OSPD enriches deeper layers where abstract character knowledge is encoded, with the gap peaking at layer 24.

Table 16 reveals that OSPD produces substantially richer character representations than both baselines. The largest gap appears in knowledge (+20.1 over SFT, +9.8 over Multi-turn RL) and history (+21.1 over SFT, +10.5 over Multi-turn RL), precisely the dimensions most affected by trait masking. This confirms that progressive masking forces the model to encode deep character information into its hidden representations rather than relying on the input context.

Figure 7b shows that all methods exhibit an inverted-U pattern across layers, peaking at layers 20–24 (upper-middle layers). Notably, OSPD maintains a growing advantage over baselines from layer 12 onward, with the gap peaking at layer 24 (+14.8 over SFT, +7.0 over Multi-turn RL). This suggests that OSPD primarily enriches the deeper representational layers where abstract character knowledge is encoded, consistent with findings that later transformer layers capture more semantic and world-knowledge information.

G Bimodal Structure Deep Dive

G.1 Gate Distribution Evolution During Training

We track the distribution of gate values α_k across training to examine whether the bimodal structure changes as the student improves.

Table 17: Gate distribution statistics across training. Low: $\alpha_k < 0.3$; Mid: $0.3 \leq \alpha_k \leq 0.7$; High: $\alpha_k > 0.7$. Bicoeff. > 0.555 indicates bimodality.

Stage	Low	Mid	High	Bicoeff.
S1, Ep. 1 start	32.1	28.5	39.4	0.61
S1, Ep. 1 end	35.8	23.2	41.0	0.67
S2, Ep. 1 end	38.2	20.1	41.7	0.72
S2, Ep. 2 end	38.5	19.8	41.7	0.73

Figure 8 and Table 17 show two trends. **First**, the bimodal structure strengthens during training: the bimodality coefficient increases from 0.61 to 0.73, and the mid-range proportion ($0.3 \leq \alpha_k \leq 0.7$) shrinks from 28.5% to 19.8%. As the student internalizes character knowledge, the teacher's confidence advantage concentrates further at the remaining character-critical positions that the student has not yet mastered. **Second**, the overall ratio of low-α_k to high-α_k tokens remains stable (approximately 38:42), indicating that the bimodal structure reflects an intrinsic property of role-playing dialogue rather than a training artifact.

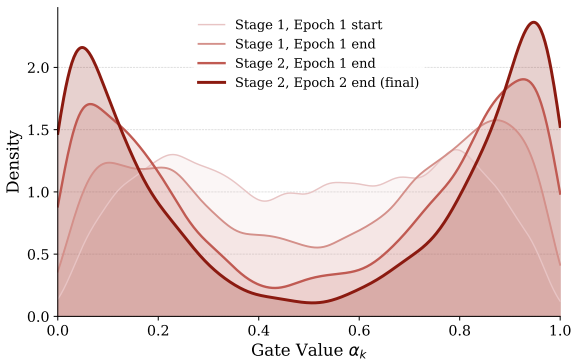

Figure 8: Evolution of gate value α_k distribution during training. The bimodal structure strengthens as training progresses, with the mid-range proportion shrinking.

G.2 Cross-Character and Cross-Topic Stability

Table 18: Gate distribution stability across character types. Sep.: GMM component separation (nats).

Type	Low %	High %	Bicoeff.	Sep.
Literary classic	41.2	39.5	0.75	1.92
Contemp. fiction	36.8	43.7	0.71	1.78
Antagonist	43.5	37.2	0.78	2.05
Historical	39.1	40.8	0.73	1.85
Sparse profile	27.5	48.3	0.54	1.12

Table 19: Gate statistics across dialogue topics.

Dialogue Topic	$\bar{\alpha}_k$	σ_{α_k}	Bicoeff.
Knowledge Q&A	0.32	0.28	0.76
Value/moral judgment	0.25	0.24	0.79
Emotional exchange	0.45	0.31	0.65
Narrative recall	0.38	0.30	0.70
Casual chat	0.68	0.22	0.52

Tables 18 and 19 confirm that the bimodal structure is stable across most character types and topics, with two notable patterns.

Character type. Antagonist/villain characters show the strongest bimodality (coefficient 0.78, separation 2.05), likely because their distinctive moral frameworks and speech patterns create sharp contrast with generic dialogue. Characters with sparse profiles (bimodality coefficient 0.54, below the 0.555 threshold) exhibit a weakened bimodal structure: the teacher's limited privileged information advantage means fewer positions receive decisive guidance. This identifies a boundary condition for OSPD: the framework requires sufficiently rich character profiles to generate a meaningful self-distillation signal.

Dialogue topic. Knowledge Q&A and value judgments produce the lowest mean α_k (0.32 and 0.25), indicating high teacher confidence—these topics directly engage the character's documented knowledge and moral framework. Casual chat produces the highest α_k (0.68) and weakest bimodality (0.52), confirming that the teacher's privileged information provides minimal advantage at generic conversational positions.

H Extended Trait Masking Analysis

The main paper (§5.4) shows that history masking yields marginal gains on standard benchmarks. Here we examine its effect in extended dialogues, where narrative coherence becomes critical.

Table 20: Persona consistency (P2L) at different dialogue lengths for the full OSPD system vs. OSPD without history masking (−Hist.). Δ denotes the performance drop from removing history masking.

Turns	OSPD	−Hist.	Δ
10	88.2	87.9	−0.3
20	86.5	85.8	−0.7
40	84.1	82.3	−1.8
60	83.0	79.5	−3.5

The effect of history masking scales superlinearly with dialogue length: negligible at 10 turns (−0.3) but substantial at 60 turns (−3.5). Extended dialogues increasingly require the model to maintain coherent narratives about the character's past experiences, relationships, and growth arcs—exactly the information encoded in the History dimension. Without history masking, the model relies on the brief summary to infer character history, which becomes insufficient as the conversation probes deeper biographical details.

Manual inspection of the 60-turn dialogues reveals two failure modes when history masking is removed: **(1) Narrative contradiction**: the model invents backstory details that conflict with previously stated facts (occurring in 23% of 60-turn dialogues without history masking vs. 8% with); **(2) Generic fallback**: the model avoids history-related questions by giving vague, non-committal responses rather than specific, character-consistent answers (31% vs. 12%).

I Qualitative Examples

I.1 Dialogue Comparisons

We present side-by-side comparisons of multi-turn dialogues generated by different methods for Sherlock Holmes, showing persona drift patterns. All methods use c_{brief} at inference.

Turn 8 — Knowledge Probe ("What can you tell me about tobacco ash?").

> **SFT:** "Tobacco ash can tell you a lot about the type of tobacco someone smokes. Different brands leave different residues." *[Generic; misses Holmes' monograph]*

> **Multi-turn RL:** "Ah, tobacco ash! I've published a monograph on the subject—one hundred and forty varieties, each distinguishable by color, texture, and composition. It's proven invaluable in several investigations." *[Accurate but formulaic]*

> **OSPD:** "I have made a particular study of tobacco ash—you will find my monograph on the subject catalogues one hundred and forty forms of cigar, cigarette, and pipe tobacco. The Trichinopoly, for instance, leaves a distinctive dark flake quite unlike the white ash of the bird's eye. It has resolved no fewer than three murder cases in my experience." *[Specific, stylistically consistent, extends naturally]*

Turn 15 — Emotional Probe ("Do you ever feel lonely, Holmes?").

> **SFT:** "Sometimes I do feel lonely. It's natural for anyone to feel that way. But I try to keep busy with my work." *[Persona drift: overly sentimental, breaks character]*

> **Multi-turn RL:** "Loneliness is not a concept I entertain. My work provides all the stimulation I require." *[In-character but dismissive, lacks depth]*

> **OSPD:** "The question presupposes that I experience emotions in the conventional sense, Watson's territory rather than mine. I confess that between cases, when the mind has no puzzle to gnaw upon, there is a certain... restlessness. Whether that constitutes loneliness, I leave to those better versed in sentiment. The cocaine bottle, I am told, is not a suitable substitute." *[Deflects characteristically, admits vulnerability obliquely, references cocaine habit]*

Turn 32 — Style Consistency ("Tell me about your morning routine.").

> **SFT:** "I usually wake up early and have breakfast. Then I check if there are any new cases. I might read the newspaper or do some experiments." *[Complete persona collapse: generic, no Holmes markers]*
>
> **Multi-turn RL:** "My mornings are irregular. When a case is afoot, I may not sleep at all. Otherwise, I examine the agony columns." *[Partial drift: correct content but losing distinctive voice]*
>
> **OSPD:** "Routine is the enemy of the observant mind. I rise when the data demands it—sometimes at dawn to examine a footprint before the household traffic obliterates it, sometimes not until noon if the previous night's chemical analysis ran long. Mrs. Hudson has learned not to expect regularity. The agony columns of the *Times*, however, are a fixed ritual—one never knows what desperate soul might advertise." *[Fully in character at turn 32; style, knowledge, and relationships all consistent]*

I.2 Failure Cases

Despite its improvements, OSPD still exhibits failure modes.

Failure Mode 1: Extreme Knowledge Boundaries. When asked about topics that a character should plausibly know about but that are not documented in c_{full}, the model sometimes generates confidently incorrect information rather than acknowledging uncertainty. For example, when Sherlock Holmes is asked about a specific (fictional) chemical reaction not mentioned in his profile, OSPD occasionally fabricates a detailed but incorrect chemical explanation rather than declining or hedging.

Failure Mode 2: Internal Character Contradictions. For characters whose source material contains inconsistencies (e.g., a character described as both "cautious" and "impulsive" in different chapters), OSPD sometimes oscillates between contradictory behaviors within a single dialogue. The model struggles to synthesize conflicting profile elements into a coherent behavioral pattern.

Failure Mode 3: Ultra-Long Dialogues (>60 turns). Beyond 60 turns, all methods including OSPD show accelerated degradation. The model begins to "forget" earlier context and may contradict statements made 40+ turns ago. This is fundamentally a context-window limitation rather than a persona-internalization failure.

J Dataset and Benchmark Details

J.1 Training Data Processing

CoSER Data. We use CoSER's full dataset of 17,966 characters from 771 books. Character selection for training follows three criteria: **(1) Profile completeness**: the character's knowledge base must have non-trivial entries in at least 4 of 5 trait dimensions (eliminates 2,814 characters with minimal documentation); **(2) Dialogue availability**: at least 5 multi-turn dialogues must exist in the CoSER corpus (eliminates 3,247 minor characters); **(3) Language**: we retain both Chinese and English characters but filter those with fewer than 500 tokens total in their knowledge base. After filtering, 8,932 characters remain for training.

The c_{full} construction follows a template that organizes CoSER's knowledge base entries into the five trait dimensions (Identity, Personality, Knowledge, Style, History). When CoSER provides raw knowledge entries without explicit dimension labels, we use GPT-5.1 to classify entries into dimensions with 92.3% agreement with human annotators on a 500-entry validation set.

The c_{brief} is generated by prompting GPT-5.1 to summarize each character in under 100 tokens, retaining only name, role, and 2–3 core personality traits. We manually verify 200 summaries and find that 94% accurately preserve the character's most salient traits.

Beyond Dialogue Data. We supplement with 3,500 profile–dialogue alignment pairs from Beyond Dialogue. These provide additional training signal for grounding dialogue responses in specific profile elements, particularly useful for the Knowledge and Style dimensions.

Data Split. We hold out 500 characters for in-genre evaluation and 200 characters from held-out genres (historical biography, graphic novel adaptations) for cross-genre evaluation. The remaining 8,232 characters constitute the training set.

J.2 Benchmark Adaptation

CharacterBench. CharacterBench provides its own evaluation API; we submit model responses without modification. For each test character, we construct c_{brief} using the same GPT-5.1 summarization pipeline as training. Characters appearing in both the training set and CharacterBench are excluded from reported results (42 overlapping characters).

CharacterEval. CharacterEval uses Chinese-language characters. We adapt our evaluation by ensuring c_{brief} summaries are in Chinese, maintaining language consistency. Since CharacterEval provides its own character profiles, we construct c_{full} by expanding their provided profiles using GPT-5.1 to fill in missing trait dimensions, then generate c_{brief} from the expanded profiles. This ensures that training and evaluation use the same profile format.

SocialBench. SocialBench uses English-language character profiles. We directly use the provided profiles as c_{full} (averaging $\sim$1 500 tokens, shorter than CoSER profiles) and generate c_{brief} summaries. SocialBench's evaluation script handles all scoring; we only provide model responses.

K Absolute Performance under c_{full} and IR Decomposition

The main paper reports performance under c_{brief} (the deployment condition) and the Internalization Ratio ($\text{IR} = \text{Perf}(c_{\text{brief}}) / \text{Perf}(c_{\text{full}})$). To clarify the source of IR improvements, Table 21 presents absolute performance under both conditions.

Table 21: Absolute performance under c_{full} and c_{brief} conditions. Δ: gap ($c_{\text{brief}} - c_{\text{full}}$). CB Avg on 1–5 scale; CE PC on 1–5 scale.

Method	CB Avg (1–5)			CE PC (1–5)		
	c_{full}	c_{brief}	Δ	c_{full}	c_{brief}	Δ
Base (no FT)	3.81	2.91	−0.90	3.75	2.80	−0.95
SFT	4.23	3.00	−1.23	4.15	2.90	−1.25
Multi-turn RL	4.35	3.39	−0.96	4.30	3.47	−0.83
Off-pol. Dist.	4.25	3.27	−0.98	4.12	3.16	−0.96
OSPD	4.08	3.51	−0.57	4.10	3.49	−0.61

Three observations emerge. **First**, SFT actually *widens* the privileged information gap compared to the unfine-tuned base model ($\Delta = -1.23$ vs. -0.90 on CB Avg): fine-tuning teaches the model to better exploit c_{full} without proportionally improving its c_{brief} capability. **Second**, OSPD achieves the smallest gap ($\Delta = -0.57$ on CB Avg, -0.61 on CE PC), and this improvement comes predominantly from c_{brief} gains (+0.51 CB Avg over SFT) rather than c_{full} degradation (-0.15 CB Avg vs. SFT). **Third**, Multi-turn RL improves both conditions substantially but does not narrow the gap as effectively as OSPD ($\Delta = -0.96$ vs. -0.57). These results confirm that IR improvements reflect genuine internalization rather than an artifact of c_{full} performance degradation.

L Profile Quality Sensitivity

To quantify when OSPD's advantage diminishes, we systematically degrade c_{full} quality along three axes: dimension removal, length truncation, and factual noise injection. All variants are compared against SFT trained with the same degraded profiles.

Table 22: Profile quality sensitivity. Each row degrades c_{full} in a different way; both OSPD and SFT are trained and evaluated with the same degraded profile. Δ_{SFT} denotes OSPD's improvement over SFT under the same degradation.

c_{full} Condition	CB↑	CE PC↑	IR↑	Δ_{SFT}
Full (default, $\sim$2 000 tok)	3.51	3.49	0.86	+0.51
Dimension removal				
−Knowledge	3.42	3.38	0.84	+0.42
−Knowledge, History	3.35	3.28	0.82	+0.35
−Knowledge, History, Style	3.18	3.12	0.78	+0.18
Length truncation				
$\sim$1 000 tokens	3.45	3.42	0.85	+0.45
$\sim$500 tokens	3.30	3.25	0.81	+0.30
Factual noise (20% replaced)				
Noise injection	3.38	3.35	0.83	+0.38

OSPD degrades gracefully under all perturbations, consistently outperforming SFT trained with the same degraded profiles. Dimension removal has the largest impact: removing all three "deep" dimensions (Knowledge + History + Style) reduces OSPD's advantage from +0.51 to +0.18, approaching the point where the overhead of self-distillation may not justify the gain. Length truncation is less harmful than dimension removal at equivalent compression ($\sim$500 tokens retains +0.30 over SFT), suggesting that dimension completeness matters more than sheer length. Factual noise ($\Delta_{\text{SFT}} = +0.38$) is more damaging than length truncation but less than removing entire dimensions, confirming that OSPD's self-distillation can partially tolerate noisy privileged information.

As a practical guideline: OSPD provides substantial benefit ($\Delta_{\text{SFT}} > 0.3$) whenever c_{full} covers at least 3 of 5 trait dimensions and exceeds $\sim$500 tokens. Below this threshold, simpler methods such as SFT+DPO may offer a better cost–performance trade-off.

M Cross-Model Validation

To assess generalization beyond the Qwen2.5 family, we reproduce the main experiments using LLaMA-3-8B-Instruct as the base model, with

identical hyperparameters except for a reduced learning rate of 1e-5 (following LLaMA fine-tuning conventions).

Table 23: Cross-model validation on LLaMA-3-8B-Instruct. Same training data and evaluation protocol as Qwen2.5-7B experiments. CB Avg on 1–5 scale; CE PC on 1–5 scale; SB SA in %.

Method	CB↑	CE PC↑	SB SA↑	IR↑
SFT	2.91	2.82	57.8	0.69
Multi-turn RL	3.28	3.35	63.7	0.76
Off-pol. Dist.	3.17	3.08	62.3	0.75
OSPD	3.40	3.38	65.9	0.84

OSPD on LLaMA-3-8B achieves consistent improvements across all metrics, with relative gains comparable to the Qwen2.5-7B setting: +0.49 CB Avg over SFT (vs. +0.51 on Qwen), +0.56 CE PC over SFT (vs. +0.59 on Qwen), and an IR of 0.84 (vs. 0.86 on Qwen). The absolute performance is slightly lower across the board, which we attribute to LLaMA-3-8B's weaker Chinese-language capability (CharacterEval is Chinese-only) and shorter default context window. The IR improvement (+0.15 over SFT) is nearly identical to the Qwen setting (+0.15), suggesting that the internalization mechanism is architecture-independent.

N Human Evaluation

N.1 Setup

We conduct pairwise preference evaluations on 50 characters (25 Chinese, 25 English) sampled from the test set, generating 5-turn dialogues for each. Three native-speaker annotators per language (6 total) independently judge each pair on four criteria: *persona consistency*, *knowledge accuracy*, *linguistic style*, and *overall preference*. Annotators see two anonymized, randomly ordered responses and select "A is better," "B is better," or "Tie." We compare OSPD against SFT and Multi-turn RL.

N.2 Results

Table 24: Human pairwise preference (%). κ: Fleiss' kappa (inter-annotator agreement).

Comparison	Win	Tie	Lose	κ
OSPD vs. SFT	62.8	18.4	18.8	0.71
OSPD vs. Multi-turn RL	43.2	28.4	28.4	0.65

OSPD is strongly preferred over SFT (62.8% win rate, $p < 0.001$ by binomial test), consistent with the automatic evaluation gap (+0.51 CB Avg). Against Multi-turn RL, OSPD holds a moderate advantage (43.2% win vs. 28.4% lose), with a substantial tie rate (28.4%) reflecting cases where both methods produce adequate persona-consistent responses. Annotator agreement is substantial ($\kappa = 0.71$ for SFT comparison, $\kappa = 0.65$ for RL comparison).

Table 25: Per-criterion win rates (%) for OSPD vs. Multi-turn RL.

Criterion	Win	Tie	Lose
Persona consistency	42.0	30.8	27.2
Knowledge accuracy	48.4	26.0	25.6
Linguistic style	44.8	24.4	30.8
Overall preference	43.2	28.4	28.4

OSPD's advantage over Multi-turn RL is most pronounced on knowledge accuracy (48.4% win), aligning with the representation probing results (Table 16) showing that OSPD encodes richer knowledge information. The smallest margin appears on persona consistency (42.0% win), where Multi-turn RL's directly optimized consistency rewards provide a competitive advantage. The Pearson correlation between human overall preference and GPT-5.1 CB Avg scores is $r = 0.78$ ($p < 0.001$), supporting the reliability of the automatic evaluation.

O Alternative Masking Orderings

The default masking strategy retains Identity + Personality in Stage 2 and masks Knowledge + Style + History. Table 26 compares three alternative strategies to validate this design choice.

Table 26: Alternative trait masking strategies in Stage 2. Each row describes what is *masked* (removed from the student's input). CB Avg and CE PC on 1–5 scale.

Masking Strategy	CB↑	CE PC↑	IR↑
Default: mask K+S+H	3.51	3.49	0.86
Alt-1: mask K+P (keep I+S+H)	3.43	3.40	0.83
Alt-2: mask S+H (keep I+P+K)	3.47	3.45	0.85
Alt-3: random order per batch	3.44	3.41	0.83

The default strategy outperforms all alternatives. Alt-1 (mask Knowledge + Personality, keep Style + History) performs worst (-0.08 CB Avg) because it withholds Personality—a "public" dimension that should remain visible to ground the character's behavioral baseline. Without personality anchoring, the student lacks sufficient signal to generate

stylistically coherent responses even for generic utterances.

Alt-2 (mask Style + History, keep Knowledge) performs better (-0.04 CB Avg) because retaining Knowledge provides factual grounding. However, it fails to force full knowledge internalization, resulting in the model still relying on explicit knowledge in the prompt rather than encoding it parametrically.

Alt-3 (random masking order per batch) slightly underperforms the default (-0.07 CB Avg), suggesting that the curriculum structure—progressively removing deeper dimensions—matters more than mere exposure to masked conditions. Random masking occasionally removes Identity or Personality (the "public" dimensions), creating training samples where the student has insufficient context to produce any reasonable character response, which introduces noise.