WHEN TOKENS DECAY AND TURNS AMPLIFY: A DUAL-GRANULARITY FRAMEWORK FOR MULTI-TURN PREFERENCE OPTIMIZATION

Yangyi Fang[1,*], Jiaye Lin[1,*], Xiaoliang Fu[2,*], Cong Qin[3], Chaowen Hu[4], Haolin Shi[1,†]

[1]Tsinghua University [2]Fudan University [3]Peking University [4]Zhejiang University

ABSTRACT

Multi-turn dialogue alignment faces critical challenges where tokens and turns contribute heterogeneously to preference signals. Existing methods apply uniform token weighting or binary turn selection, overlooking fine-grained structures. We present **T^{3}PO**, a dual-granularity framework incorporating: (i) token-level temporal discounting prioritizing early high-signal tokens with provable partition function cancellation; (ii) turn-level self-evaluated weighting via multi-perspective scoring, eliminating external dependencies. Experiments across multiple benchmarks and model scales demonstrate consistent improvements over baselines, with ablations confirming independent contributions from both mechanisms.

1 INTRODUCTION

Aligning LLMs with human preferences is essential for reliable conversational systems (Ouyang et al., 2022). Direct Preference Optimization (DPO) (Rafailov et al., 2023) streamlines this by directly optimizing policies on preference data via Bradley-Terry objectives, eliminating explicit reward modeling from traditional RLHF (Christiano et al., 2017).

Extending DPO to multi-turn dialogues faces critical dual-granularity gaps (Figure 1). Session-level methods (ETO (Song et al., 2024), DMPO (Shi et al., 2024)) apply uniform token weighting despite position-dependent preference signals. SDPO (Kong et al., 2025) employs binary turn selection, discarding contextual information. At the token level, KL divergence decreases along positions within turns, yet existing methods weight uniformly. At the turn level, uniform weighting or binary selection fails to capture heterogeneous signal distribution across turns.

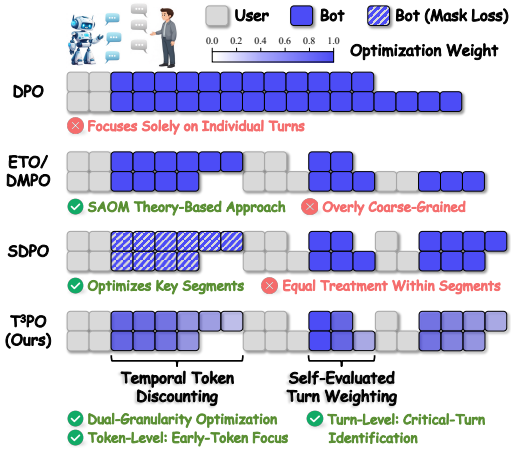


Figure 1: Multi-turn preference optimization methods. Color intensity indicates optimization weight.

We present **T^{3}PO**, operating at **T**oken and **T**urn levels in multi-**T**urn dialogues. Building on SAOM (Sutton et al., 1998), we establish segment-level formulation with dual-granularity mechanisms: (i) token-level temporal discounting prioritizes early high-signal tokens with provable partition cancellation; (ii) turn-level self-evaluated weighting uses multi-perspective scoring via the reference model, eliminating external dependencies.

Contributions: (1) Token-level temporal discounting with theoretical guarantees (partition cancellation, approximation-optimization trade-off). (2) Self-evaluated turn weighting eliminating external models. (3) Validation across benchmarks/scales with ablations confirming independent contributions.

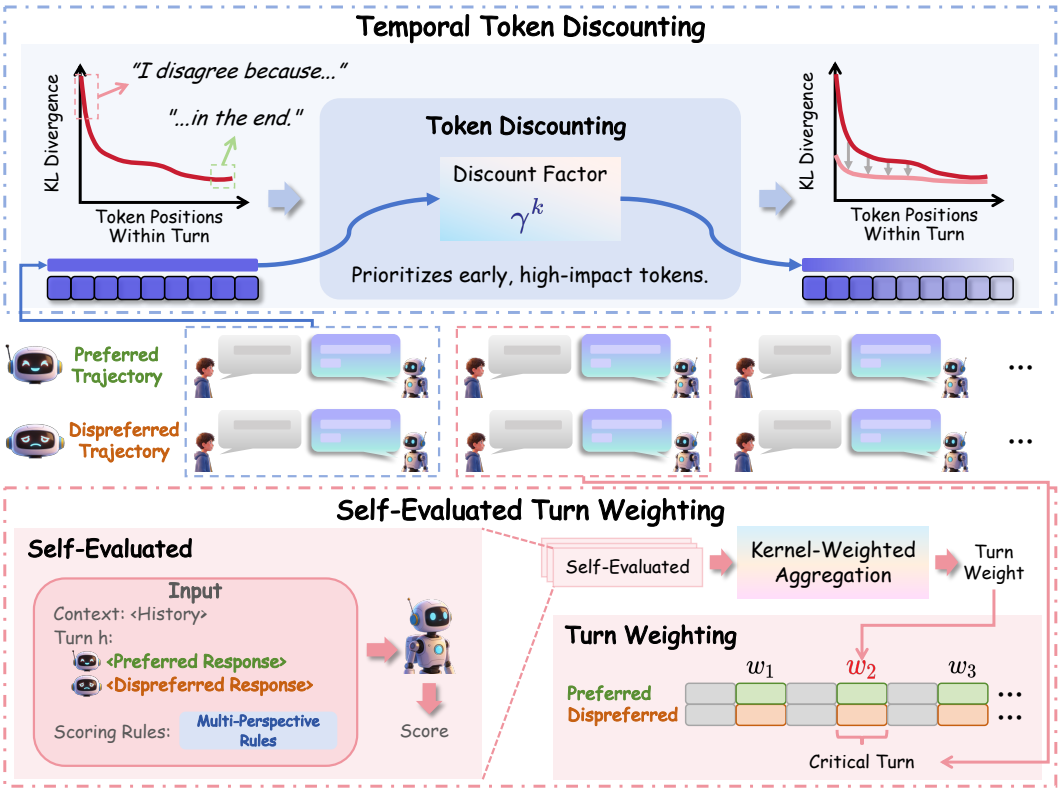


Figure 2: T^{3}PO framework: token-level temporal discounting and turn-level self-evaluated weighting.

2 METHODOLOGY

We extend DPO to multi-turn dialogues via SAOM framework (Appendix B). Following Kong et al. (2025), segment-level optimization for equal-length segments ($L^w = L^l = L$):

$$\mathcal{L}(\pi_\theta) = -\mathbb{E}_{(\tau^w,\tau^l)}\left[\log\sigma\left(\frac{\beta}{L}\sum_{h\in\mathcal{I}^w}\log\frac{\pi_\theta(a_h^w|s_h^w)}{\pi_{\text{ref}}(a_h^w|s_h^w)} - \frac{\beta}{L}\sum_{h\in\mathcal{I}^l}\log\frac{\pi_\theta(a_h^l|s_h^l)}{\pi_{\text{ref}}(a_h^l|s_h^l)}\right)\right]. \tag{1}$$

This assumes uniform weighting. We introduce dual-granularity mechanisms: token-level temporal discounting and turn-level self-evaluated weighting (Figure 2).

2.1 TEMPORAL TOKEN DISCOUNTING

We introduce discount factor $\gamma \in (0, 1)$ to prioritize early tokens within each turn. The discounted reward formulation applies exponential decay to token positions:

$$r_\gamma(\tau) = \sum_{h=0}^{H}\sum_{k=0}^{T_h-1}\gamma^k \cdot r(s_h, y_{h,k}). \tag{2}$$

Theorem 1 (Partition Cancellation). *For equal-turn pairs ($H^w = H^l$) from the same prompt, partition function cancels despite discounting.*

Theorem 2 (Approximation-Optimization Trade-Off). *Performance gap $\mathcal{E}(\pi_\gamma)$ satisfies:*

$$\mathcal{E}(\pi_\gamma) \le \underbrace{2R_{\max}\left[L_{\max} - \frac{1-\gamma^{L_{\max}}}{1-\gamma}\right]}_{\textit{Approximation Bias}} + \underbrace{2R_{\max}\bar{d}_{TV}(\pi_\gamma,\pi^*)\frac{(1-\gamma^{L_{\max}})^2}{(1-\gamma)^2}}_{\textit{Optimization Error}}. \tag{3}$$

[*] Equal contribution. [†] Corresponding author.

Table 1: Performance comparison across different model scales and methods on MT-Bench-101. All results are reported as scores. The best results are highlighted in bold, and the second-best results are underlined, respectively. ▲: Uniform token weighting, ☆: Uniform turn weighting, ●: Binary turn selection, ✪: Fine-grained token & turn modeling. GR: General Reasoning, AR: Anaphora Resolution, MR: Mathematical Reasoning, CC: Content Confusion, TS: Topic Shift, CM: Context Memory, SC: Self-Correction.

Method	GR	AR	MR	CC	TS	CM	SC	Avg.
Qwen2.5-3B-Instruct								
Base	3.21	6.09	3.37	6.85	4.25	4.18	7.09	5.01
DPO Rafailov et al. (2023) ▲	4.42	7.25	4.01	7.79	6.23	6.09	8.16	6.28
ETO Song et al. (2024) ▲☆	4.31	7.08	3.94	7.66	6.05	5.98	7.99	6.14
DMPO Shi et al. (2024) ▲	4.55	7.34	4.05	7.87	6.32	6.21	8.22	6.37
SDPO Kong et al. (2025) ▲●	4.69	**7.54**	4.21	8.04	6.58	6.50	8.32	6.55
T^{3}PO (Ours) ✪	**4.96**	7.48	**4.36**	**8.46**	**7.00**	**6.81**	**8.50**	**6.80**
Qwen2.5-7B-Instruct								
Base	4.63	7.31	4.36	7.91	5.43	5.11	7.75	6.07
DPO Rafailov et al. (2023) ▲	5.51	8.05	5.21	8.86	7.11	6.83	8.86	7.20
ETO Song et al. (2024) ▲☆	5.40	7.81	5.08	8.57	6.91	6.66	8.64	7.01
DMPO Shi et al. (2024) ▲	5.73	8.09	5.17	8.73	7.25	7.43	8.34	7.25
SDPO Kong et al. (2025) ▲●	6.18	8.25	**5.29**	9.14	7.65	7.70	**9.17**	7.63
T^{3}PO (Ours) ✪	**6.26**	**8.56**	5.23	**9.22**	**8.35**	**7.78**	9.11	**7.79**
Qwen2.5-32B-Instruct								
Base	5.48	8.13	4.44	9.21	6.53	5.89	8.33	6.86
DPO Rafailov et al. (2023) ▲	7.13	8.91	5.72	9.70	7.89	7.16	9.19	7.96
ETO Song et al. (2024) ▲☆	7.24	9.11	5.67	9.14	8.68	7.46	9.47	8.11
DMPO Shi et al. (2024) ▲	7.72	9.08	6.09	9.35	8.36	7.42	9.23	8.18
SDPO Kong et al. (2025) ▲●	7.90	8.85	6.58	9.72	8.65	7.64	9.50	8.40
T^{3}PO (Ours) ✪	**8.01**	**9.16**	**6.62**	**9.92**	**8.82**	**7.79**	**9.51**	**8.55**

Proofs in Appendix E.2 and E.3.

2.2 Self-Evaluated Turn Weighting

We introduce continuous turn weighting using frozen reference model for multi-perspective scoring. For each turn h, we collect scores across M dimensions with G repetitions per dimension. Kernel-weighted averaging suppresses outliers based on deviation from median:

$$w_i = \exp\left(-\frac{(s_{h,i} - \text{med}(\{s_{h,j}\}))^2}{2\sigma^2}\right). \tag{4}$$

Normalized turn weights $\omega_h = \bar{s}_h / \sum_{h'} \bar{s}_{h'}$ apply symmetrically to τ^w and τ^l. Full T^{3}PO objective:

$$\mathcal{L}_{\text{T}^3\text{PO}} = -\mathbb{E}\left[\log\sigma\left(\beta \sum_{h=0}^{H-1} \omega_h \cdot \Delta_h\right)\right], \quad \Delta_h = \sum_k \gamma^k \log \frac{\pi_\theta(y^w_{h,k}|\cdot)}{\pi_{\text{ref}}(y^w_{h,k}|\cdot)} - \sum_k \gamma^k \log \frac{\pi_\theta(y^l_{h,k}|\cdot)}{\pi_{\text{ref}}(y^l_{h,k}|\cdot)}. \tag{5}$$

Motivation, scoring templates, and implementation details in Appendix D and F.

3 Experiments

3.1 Experimental Setup

Datasets. We construct 1,296 preference pairs from MT-Bench-101 (Bai et al., 2024) using a 7:3 train-test split (details in Appendix G). Evaluation is conducted on two independent benchmarks: (1) the **held-out subset of MT-Bench-101** for in-distribution performance across 7 fine-grained

Table 2: Ablation study on Qwen2.5-7B-Instruct. We evaluate the independent contributions of token-level discounting (γ) and turn-level weighting (ω). Results demonstrate that both mechanisms provide complementary improvements, with their combination achieving the best performance across all task categories.

Method	GR	AR	MR	CC	TS	CM	SC	Avg.
T^{3}PO (Full)	**6.26**	**8.56**	**5.23**	**9.22**	**8.35**	**7.78**	**9.11**	**7.79**
w/o Token Discount (γ)	5.83	8.32	5.01	8.98	8.03	7.53	8.86	7.51
w/o Turn Weighting (ω)	5.88	8.37	5.07	9.03	8.10	7.59	8.92	7.56
w/o Both	5.75	8.24	4.95	8.91	7.94	7.47	8.79	7.43

abilities; (2) **MT-Bench** (Zheng et al., 2023) (80 questions, 8 categories) for cross-dataset generalization—despite the similar name, MT-Bench is completely independent from MT-Bench-101 with no data overlap.

Training Configuration. We train Qwen2.5-Instruct models (3B/7B/32B) using AdamW optimizer for 3 epochs with learning rate 5e-7. Key hyperparameters: $\beta = 0.1$, $\gamma = 0.98$, $M = 4$ scoring perspectives for turn evaluation, $G = 5$ repetitions per perspective to reduce variance. All baselines use identical settings. Full training details are in Appendix H.

Evaluation Benchmarks and Metrics. Following Bai et al. (2024), we evaluate on MT-Bench-101 using LLM-as-judge (e.g., GPT-4.1, GPT-4o, DeepSeek-V3) across seven dimensions including general reasoning, mathematical reasoning, context memory, and self-correction, with minimum-score aggregation. Generation uses temperature 0.7, top-p 0.9, max tokens 4096. Results are averaged over five runs.

Baselines. We compare against: **Base** (Qwen2.5-Instruct without preference optimization); **DPO** (Rafailov et al., 2023) (turn-level optimization); **ETO** (Song et al., 2024) (session-level over entire trajectories); **DMPO** (Shi et al., 2024) (SAOM framework for session-level); and **SDPO** (Kong et al., 2025) (segment-level with heuristic selection).

3.2 Main Results

Table 1 presents evaluation results on MT-Bench-101 across Qwen2.5-3B, 7B, and 32B scales, where T^{3}PO achieves the highest average scores compared to all baselines. The improvements stem from our dual-granularity design addressing signal dilution in uniform weighting: temporal token discounting sharpens optimization on early tokens determining semantic trajectories, while self-evaluated turn weighting allocates more effort to high-signal turns, both yielding robust cross-scale improvements. Cross-evaluator validation using GPT-4o and DeepSeek-V3 (Appendix L) confirms consistent superiority across all judges, eliminating judge-specific biases. Cross-dataset evaluation on the independent MT-Bench (Appendix I) shows identical performance ranking, confirming generalization beyond training distributions.

3.3 Ablation Study

Table 2 isolates the contributions of our dual-granularity components on Qwen2.5-7B-Instruct. Removing Token-level Discounting yields larger performance drops, while excluding Turn-level Weighting also degrades results. Removing both components produces the weakest performance. These patterns suggest the two mechanisms target complementary aspects of multi-turn preference learning, with their combination yielding additive benefits. Hyperparameter sensitivity analysis and case studies validating our design choices are provided in Appendix N and O.

4 Conclusion

We presented T^{3}PO, a dual-granularity framework for multi-turn preference optimization through token-level temporal discounting (prioritizing early high-impact tokens with provable partition cancel-

lation) and turn-level self-evaluated weighting (eliminating external dependencies via reference model scoring). Experiments across Qwen2.5 models (3B/7B/32B) demonstrate consistent improvements over baselines, with ablations confirming complementary contributions from both mechanisms.

REFERENCES

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Xin-Qiang Cai and Masashi Sugiyama. Vi-curl: Stabilizing verifier-independent rl reasoning via confidence-guided variance reduction. *arXiv preprint arXiv:2602.12579*, 2026.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *arXiv preprint arXiv:2510.00915*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Ruiyi Ding, Yongxuan Lv, Xianhui Meng, Jiahe Song, Chao Wang, Chen Jiang, and Yuan Cheng. Prpo: Aligning process reward with outcome reward in policy optimization, 2026. URL https://arxiv.org/abs/2601.07182.

Yangyi Fang, Jiaye Lin, Xiaoliang Fu, Cong Qin, Haolin Shi, Chaowen Hu, Lu Pan, Ke Zeng, and Xunliang Cai. How to allocate, how to learn? dynamic rollout allocation and advantage modulation for policy optimization. *arXiv preprint arXiv:2602.19208*, 2026a.

Yangyi Fang, Jiaye Lin, Xiaoliang Fu, Cong Qin, Haolin Shi, Chang Liu, and Peilin Zhao. Proximity-based multi-turn optimization: Practical credit assignment for llm agent training. *arXiv preprint arXiv:2602.19225*, 2026b.

Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89(1):123–156, 2012.

Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pp. 2014–2025. PMLR, 2023.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. Adaptive sample scheduling for direct preference optimization. *arXiv preprint arXiv:2506.17252*, 2025.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuefeng Xiao, Hongyan Xie, Huaqiu Li, Songshi Liang, Zhongxiang Dai, Fuzhen Zhuang, Jianxin Li, Yikun Ban, and Deqing Wang. Real-time aligned reward model beyond semantics. 2026. URL https://api.semanticscholar.org/CorpusID:285240754.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. SDPO: Segment-level direct preference optimization for social agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12409–12423, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.607. URL https://aclanthology.org/2025.acl-long.607/.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. Rl with kl penalties is better viewed as bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, 2022.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

Dengcan Liu, Fengkai Yang, Xiaohan Wang, Shurui Yan, Jiajun Chai, Jiahao Li, Yikun Ban, Zhendong Mao, Wei Lin, and Guojun Yin. Cdrrm: Contrast-driven rubric generation for reliable and interpretable reward modeling, 2026. URL https://arxiv.org/abs/2603.08035.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. Direct multi-turn preference optimization for language agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2312–2324, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.138. URL https://aclanthology.org/2024.emnlp-main.138/.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.409. URL https://aclanthology.org/2024.acl-long.409.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Yuqiao Tan, Minzheng Wang, Shizhu He, Huanxuan Liao, Chengfeng Zhao, Qiunan Lu, Tian Liang, Jun Zhao, and Kang Liu. Bottom-up policy optimization: Your language model policy secretly contains internal policies. *arXiv preprint arXiv:2512.19673*, 2025. URL https://arxiv.org/abs/2512.19673.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your group-relative advantage is biased, 2026. URL https://arxiv.org/abs/2601.08521.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. *arXiv preprint arXiv:2408.13518*, 2024.

Zhiqi Yu, Zhangquan Chen, Mengting Liu, Heye Zhang, and Liangqiong Qu. Unveiling implicit advantage symmetry: Why grpo struggles with exploration and difficulty adaptation. *arXiv preprint arXiv:2602.05548*, 2026.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yixiao Zhou, Yang Li, Dongzhou Cheng, Hehe Fan, and Yu Cheng. Look inward to explore outward: Learning temperature policy from llm internal states via hierarchical rl. *arXiv preprint arXiv:2602.13035*, 2026.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Related Works

A.1 Direct Preference Optimization

LLM alignment has evolved from two-stage RLHF (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022) (training Bradley-Terry reward models (Bradley & Terry, 1952) followed by PPO (Schulman et al., 2017) under KL constraints (Korbak et al., 2022)) toward streamlined approaches. DPO (Rafailov et al., 2023) eliminates explicit reward modeling via closed-form Bradley-Terry optimization directly on preference data, inspiring finer-grained extensions: Token-level DPO (Zeng et al., 2024) integrates forward KL constraints, Step-DPO (Lai et al., 2024) operates on reasoning steps, and SePO (Yang et al., 2024) introduces token-level reward estimation. However, these methods focus on single-turn settings and fail to address temporal dynamics in multi-turn interactions.

A.2 Multi-Turn Preference Optimization

Extending preference optimization to multi-turn scenarios introduces fundamental challenges in capturing heterogeneous turn contributions. Pioneering work (Fürnkranz et al., 2012; Christiano et al., 2017; Hejna III & Sadigh, 2023) adopts two-stage approaches, inheriting RLHF complexity. Recent single-stage methods address multi-turn structure at varying granularities: ETO (Song et al., 2024) applies DPO loss to complete trajectories with uniform turn weighting; DMPO (Shi et al., 2024) introduces SAOM framework with length normalization; SDPO (Kong et al., 2025) proposes segment-level optimization via binary selection but requires external models. T^3PO operates at token and turn levels through principled temporal discounting with theoretical guarantees and self-evaluated turn weighting without external dependencies.

A.3 Broader Perspectives on RL-based Alignment

Beyond preference optimization, recent works have explored fine-grained policy optimization and credit assignment mechanisms for LLM alignment across multiple dimensions. In multi-turn and long-horizon settings, prior studies improve credit assignment via localized propagation (Fang et al., 2026b) and adaptive sampling strategies (Fang et al., 2026a), while revealing limitations of existing advantage estimation methods, including bias in group-relative objectives (Yang et al., 2026) and challenges in exploration and difficulty adaptation (Yu et al., 2026). Complementary efforts focus on reward modeling and alignment stability, enhancing robustness through adaptive sampling (Huang et al., 2025), real-time reward alignment beyond semantic signals (Huang et al., 2026), and interpretable rubric-based reward construction (Liu et al., 2026), as well as aligning process-level and outcome-level supervision (Ding et al., 2026). Furthermore, verifier-based and reasoning-oriented RL frameworks address noisy or imperfect feedback (Cai et al., 2025; Cai & Sugiyama, 2026), while hierarchical and internal-policy perspectives (Tan et al., 2025; Zhou et al., 2026) suggest that LLM policies may implicitly encode multi-level decision processes.

B Extended Occupancy Measure Formulation

Multi-turn dialogue trajectories $\tau = (s_0, a_0, \ldots, s_H, a_H)$ admit occupancy measure:

$$d^\pi(\tau) = \mu_0(s_0) \prod_{h=0}^{H} \left[\pi(a_h|s_h) \cdot \mathcal{T}(s_{h+1}|s_h, a_h)\right], \tag{6}$$

where $\mu_0(s_0)$ is the initial distribution and $\mathcal{T}$ represents environment dynamics (user responses). Under KL-constrained optimization:

$$\max_d \mathbb{E}_{\tau\sim d}\left[\sum_{h=0}^{H} r(s_h, a_h)\right] - \beta D_{\text{KL}}[d(\tau)\|d^{\pi_{\text{ref}}}(\tau)], \tag{7}$$

the optimal distribution is:

$$d^*(\tau) = \frac{1}{Z} d^{\pi_{\text{ref}}}(\tau) \exp\left(\frac{1}{\beta}\sum_{h=0}^{H} r(s_h, a_h)\right), \tag{8}$$

where $Z(s_0)$ depends only on initial state.

Segment-Level Formulation. For index set $\mathcal{I}$ with $L = |\mathcal{I}|$, the segment-level objective under equal-length constraint $L^w = L^l = L$ is:

$$\mathcal{L}(\pi_\theta) = -\mathbb{E}_{(\tau^w,\tau^l)\sim\mathcal{D}}\left[\log\sigma\left(\frac{\beta}{L}\sum_{h\in\mathcal{I}^w}\log\frac{\pi_\theta(a_h^w|s_h^w)}{\pi_{\text{ref}}(a_h^w|s_h^w)} - \frac{\beta}{L}\sum_{h\in\mathcal{I}^l}\log\frac{\pi_\theta(a_h^l|s_h^l)}{\pi_{\text{ref}}(a_h^l|s_h^l)}\right)\right]. \tag{9}$$

Fine-Grained Extensions. This formulation assumes uniform contributions at token and turn levels. Empirical analysis (Appendix O) reveals position-dependent KL decay and heterogeneous turn signals, motivating token discounting (Section 2.1) and turn weighting (Section 2.2) while preserving partition cancellation (Theorem 1).

C Token Discounting: Detailed Analysis

Decomposition $\log\pi(a_h|s_h) = \sum_k \log\pi(y_{h,k}|\cdot)$ assumes uniform token contributions. KL divergence exhibits monotonic decay within turns (Appendix O), motivating discount factor $\gamma \in (0, 1)$. The discounted formulation applies exponential decay to token positions, prioritizing early high-signal tokens. Theorem 1 establishes partition cancellation; Theorem 2 characterizes the approximation-optimization trade-off.

D Turn Weighting: Detailed Motivation

Multi-turn dialogues exhibit heterogeneous turn-level signal distribution. Standard uniform weighting misallocates effort; binary selection discards context. We introduce continuous turn weighting via frozen reference model self-evaluation, eliminating external dependencies.

D.1 Multi-Perspective Scoring Implementation

For turn h in pair (τ^w, τ^l), we assess signal strength across multiple dimensions via discrete scoring, repeated multiple times per dimension. Kernel-weighted averaging suppresses outliers based on deviation from median, where bandwidth is set via median absolute deviation. Normalized weights apply symmetrically to τ^w and τ^l, preserving partition cancellation (Theorem 1). The full objective integrates token discounting and turn weighting as specified in Eq. 5.

E THEORETICAL FOUNDATION

E.1 EXTENSION TO MULTI-TURN DIALOGUES

For trajectories $\tau = \{(s_0, a_0), \ldots, (s_{H-1}, a_{H-1})\}$ where s_h denotes dialogue history and a_h denotes response at turn h, the occupancy measure decomposes as:

$$d^\pi(\tau) = P(s_0) \prod_{h=0}^{H-1} \left[\pi(a_h|s_h) \cdot P(s_{h+1}|s_h, a_h)\right]. \tag{10}$$

From Eq. 8, trajectory reward is:

$$r(\tau) = \beta \sum_{h=0}^{H-1} \log \frac{d^*(s_h, a_h)}{d_{\text{ref}}(s_h, a_h)} + H \cdot \beta \log Z. \tag{11}$$

For preference pairs (τ^w, τ^l), the reward difference becomes:

$$r(\tau^w) - r(\tau^l) = \beta \sum_{h=0}^{H^w-1} \log \frac{d^*(s_h^w, a_h^w)}{d_{\text{ref}}(s_h^w, a_h^w)} - \beta \sum_{h=0}^{H^l-1} \log \frac{d^*(s_h^l, a_h^l)}{d_{\text{ref}}(s_h^l, a_h^l)} + (H^w - H^l)\beta \log Z. \tag{12}$$

Under equal-turn constraint $H^w = H^l = H$, the residual term vanishes and Z cancels:

$$r(\tau^w) - r(\tau^l) = \beta \sum_{h=0}^{H-1} \Big[\log \frac{d^*(s_h^w, a_h^w)}{d_{\text{ref}}(s_h^w, a_h^w)} - \log \frac{d^*(s_h^l, a_h^l)}{d_{\text{ref}}(s_h^l, a_h^l)}\Big]. \tag{13}$$

Token-level decomposition $\pi(a_h|s_h) = \prod_{k=0}^{T_h-1} \pi(y_{h,k}|s_h, y_{h,<k})$ where $y_{h,<k} = \{y_{h,0}, \ldots, y_{h,k-1}\}$ yields:

$$r(\tau^w) - r(\tau^l) = \beta \sum_{h=0}^{H-1} \Big[\sum_{k=0}^{T_h^w-1} \log \frac{\pi_\theta(y_{h,k}^w|\cdot)}{\pi_{\text{ref}}(y_{h,k}^w|\cdot)} - \sum_{k=0}^{T_h^l-1} \log \frac{\pi_\theta(y_{h,k}^l|\cdot)}{\pi_{\text{ref}}(y_{h,k}^l|\cdot)}\Big]. \tag{14}$$

Partition cancellation occurs at turn level, independent of token-level length variations within each turn.

Distinction from DMPO. DMPO (Shi et al., 2024) applies turn-level discounting across dialogue turns to model temporal decay of turn importance; our token-level discounting operates within each turn's response, capturing intra-turn signal concentration validated by KL analysis (Figure 6(a)). DMPO's turn-level discount addresses which turns matter; our token-level discount addresses which tokens within each turn matter. We prove partition cancellation (Theorem 1) and characterize approximation-optimization trade-off (Theorem 2); DMPO provides no such analysis.

E.2 PROOF OF PARTITION FUNCTION CANCELLATION (THEOREM 1)

Under Maximum Entropy RL, optimal policy $\pi^*(a|s) = \exp((Q^*(s,a) - V^*(s))/\beta)$ satisfies Bellman equation with token-level discount $\gamma \in (0, 1]$:

$$Q^*(s_k, a_k) = r(s_k, a_k) + \beta \log \pi_{\text{ref}}(a_k|s_k) + \gamma V^*(s_{k+1}). \tag{15}$$

For sequence $\tau = \{(s_0, a_0), \ldots, (s_{L-1}, a_{L-1})\}$, recursive expansion yields:

$$\begin{aligned}
&\sum_{k=0}^{L-1} \gamma^k r(s_k, a_k) \\
&= \sum_{k=0}^{L-1} \Big[\gamma^k Q^*(s_k, a_k) - \gamma^k \beta \log \pi_{\text{ref}}(a_k|s_k) - \gamma^{k+1} V^*(s_{k+1})\Big] \\
&= Q^*(s_0, a_0) - \beta \log \pi_{\text{ref}}(a_0|s_0) + \sum_{k=1}^{L-1} \gamma^k \Big[Q^*(s_k, a_k) - V^*(s_k) - \beta \log \pi_{\text{ref}}(a_k|s_k)\Big].
\end{aligned}$$

Substituting optimal policy relation $Q^*(s_k, a_k) - V^*(s_k) = \beta \log \pi^*(a_k|s_k)$ yields:

$$\sum_{k=0}^{L-1} \gamma^k r(s_k, a_k) = V^*(s_0) + \sum_{k=0}^{L-1} \gamma^k \beta \log \frac{\pi^*(a_k|s_k)}{\pi_{\text{ref}}(a_k|s_k)}. \tag{16}$$

Define composite trajectory reward $R(\tau) = \sum_{h=0}^{H} \sum_{k=0}^{T_h-1} \gamma^k \log \frac{\pi_\theta(y_{h,k}|\cdot)}{\pi_{\text{ref}}(y_{h,k}|\cdot)}$. Under SAOM framework:

$$d^*(\tau) = \frac{1}{Z} d^{\pi_{\text{ref}}}(\tau) \exp\left(\frac{1}{\beta} R(\tau)\right), \tag{17}$$

where $Z(s_0)$ absorbs local baseline terms $V^*(s_0)$ from Eq. equation 16.

For equal-length segments $L^w = L^l = L$ from the same prompt, length-normalized baselines $\frac{V^*(s_0)}{L}$ cancel identically. Taking log-ratio difference eliminates partition function:

$$\log \frac{d^*(\tau^w)}{d^{\pi_{\text{ref}}}(\tau^w)} - \log \frac{d^*(\tau^l)}{d^{\pi_{\text{ref}}}(\tau^l)} = \frac{1}{\beta} R(\tau^w) - \frac{1}{\beta} R(\tau^l), \tag{18}$$

yielding final objective where Z is perfectly eliminated:

$$\mathcal{L}_{\text{T}^3\text{PO}} \propto -\mathbb{E}\Big[\log \sigma\Big(\beta \sum_{h=0}^{H} \big[\sum_{k=0}^{T_h^w-1} \gamma^k \log \frac{\pi_\theta(y_{h,k}^w|\cdot)}{\pi_{\text{ref}}(y_{h,k}^w|\cdot)} - \sum_{k=0}^{T_h^l-1} \gamma^k \log \frac{\pi_\theta(y_{h,k}^l|\cdot)}{\pi_{\text{ref}}(y_{h,k}^l|\cdot)} \big]\Big)\Big]. \tag{19}$$

□

E.3 Proof of Approximation-Optimization Trade-off (Theorem 2)

Define value function $V_\gamma^\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{L_{\max}-1} \gamma^k r(s_k, a_k)|s_0 = s]$, state visitation $d^\pi(s) = \sum_{k=0}^{L_{\max}-1} \Pr(s_k = s|\pi, s_0)$, total variation $d_{\text{TV}}(\pi, \pi') = \frac{1}{2} \sum_a |\pi(a|s) - \pi'(a|s)|$, and expected divergence $\overline{d}_{\text{TV}}(\pi, \pi^*) = \mathbb{E}_{s\sim d^{\pi^*}}[d_{\text{TV}}(\pi(\cdot|s), \pi^*(\cdot|s))]$.

Evaluating policy π (optimized with discount γ) against optimal π^* under undiscounted metric, the performance gap decomposes as:

$$\mathcal{E}(\pi) := V_1^{\pi^*}(s_0) - V_1^\pi(s_0) = \underbrace{\left[V_1^{\pi^*} - V_\gamma^{\pi^*}\right]}_{\text{Bias 1}} + \underbrace{\left[V_\gamma^{\pi^*} - V_\gamma^\pi\right]}_{\text{Opt. Error}} + \underbrace{\left[V_\gamma^\pi - V_1^\pi\right]}_{\text{Bias 2}}. \tag{20}$$

For any policy π' with $|r(s, a)| \le R_{\max}$:

$$\begin{aligned} \left|V_1^{\pi'}(s_0) - V_\gamma^{\pi'}(s_0)\right| &= \left|\mathbb{E}_{\tau\sim\pi'}\left[\sum_{k=0}^{L_{\max}-1} (1-\gamma^k) r(s_k, a_k)\right]\right| \\ &\le R_{\max} \sum_{k=0}^{L_{\max}-1} (1-\gamma^k) = R_{\max}\left[L_{\max} - \frac{1-\gamma^{L_{\max}}}{1-\gamma}\right]. \end{aligned} \tag{21}$$

Applying to both $\pi' = \pi^*$ and $\pi' = \pi$ yields:

$$\text{Bias 1} + \text{Bias 2} \le 2R_{\max}\left[L_{\max} - \frac{1-\gamma^{L_{\max}}}{1-\gamma}\right]. \tag{22}$$

By Performance Difference Lemma for finite-horizon MDPs (Kakade & Langford, 2002):

$$V_\gamma^{\pi^*}(s_0) - V_\gamma^\pi(s_0) = \frac{1-\gamma^{L_{\max}}}{1-\gamma} \mathbb{E}_{s\sim d^{\pi^*}}\Big[\sum_a (\pi^*(a|s) - \pi(a|s)) Q_\gamma^\pi(s, a)\Big]. \tag{23}$$

Since $Q_\gamma^\pi(s, a) \le \frac{1-\gamma^{L_{\max}}}{1-\gamma} R_{\max}$ and $\sum_a |\pi^*(a|s) - \pi(a|s)| = 2d_{\text{TV}}(\pi^*, \pi)$:

$$V_\gamma^{\pi^*}(s_0) - V_\gamma^\pi(s_0) \le 2R_{\max} \cdot \overline{d}_{\text{TV}}(\pi, \pi^*) \cdot \frac{(1-\gamma^{L_{\max}})^2}{(1-\gamma)^2}. \tag{24}$$

Combining Eq. equation 22 and optimization error bound yields:

$$\mathcal{E}(\pi_\gamma) \le 2R_{\max}\left[L_{\max} - \frac{1-\gamma^{L_{\max}}}{1-\gamma}\right] + 2R_{\max} \cdot \overline{d}_{\text{TV}}(\pi_\gamma, \pi^*) \cdot \frac{(1-\gamma^{L_{\max}})^2}{(1-\gamma)^2}. \tag{25}$$

As $\gamma \to 0$, approximation bias dominates; as $\gamma \to 1$, optimization error explodes. Optimal $\gamma \in (0, 1)$ balances these effects. □

F SELF-EVALUATION PROMPTS

We employ multi-perspective scoring across dimensions, each repeated multiple times per dimension.

F.1 PERSPECTIVE 1: SEMANTIC COHERENCE

Semantic Coherence Prompt

Task: Evaluate discriminative signal strength based on semantic coherence.
Context: - Dialogue History: {dialogue_history} - User Query at Turn {turn_id}: {user_query} - Response A: {response_chosen} - Response B: {response_rejected}
Criterion: Quality difference in semantic coherence (logical flow, relevance, structural clarity)?
Scoring: 5: Very strong—one coherent, other incoherent/off-topic; 4: Strong—noticeable gap affecting understanding; 3: Moderate—minor differences; 2: Weak—subtle differences; 1: No signal—similar coherence.
Output: Single integer 1-5.

F.2 PERSPECTIVE 2: ERROR SEVERITY

Error Severity Prompt

Task: Evaluate discriminative signal strength based on error severity.
Context: [Same as above]
Criterion: Difference in error severity (factual, logical, grammatical)?
Scoring: 5: Very strong—severe error fundamentally undermining answer; 4: Strong—significant error affecting correctness; 3: Moderate—noticeable but not fatal error; 2: Weak—minor error with limited impact; 1: No signal—similar error levels.
Output: Single integer 1-5.

F.3 PERSPECTIVE 3: LOGICAL CONSISTENCY

Logical Consistency Prompt

Task: Evaluate discriminative signal strength based on logical consistency.
Context: [Same as above]
Criterion: Quality difference in logical consistency (coherence with history, avoiding self-contradiction)?
Scoring: 5: Very strong—one contradicts prior statements; 4: Strong—noticeable inconsistency affecting trust; 3: Moderate—minor inconsistency; 2: Weak—subtle differences; 1: No signal—similar consistency.
Output: Single integer 1-5.

F.4 PERSPECTIVE 4: FACTUAL ACCURACY

Factual Accuracy Prompt

Task: Evaluate discriminative signal strength based on factual accuracy.
Context: [Same as above]
Criterion: Difference in factual correctness?
Scoring: 5: Very strong—one correct, other incorrect; 4: Strong—significant discrepancy affecting reliability; 3: Moderate—minor factual difference; 2: Weak—subtle differences; 1: No signal—similar accuracy.
Output: Single integer 1-5.

G Training Data and Evaluation Protocol

G.1 Preference Pair Construction

We construct 1,296 preference pairs from MT-Bench-101 (Bai et al., 2024) (1,388 dialogues, 7 abilities) via 7:3 train-test split with stratified sampling. Teacher models (GPT-4.1, DeepSeek-V3) generate positive τ^w (high-quality instructions) and negative τ^l (degraded constraints). Equal-turn constraint $L^w = L^l$ (Theorem 1) is ensured through matched sampling: for each τ^w with L turns, we select τ^l candidates with exactly L turns. MT-Bench-101 prompts have fixed multi-turn structures, so responses to the same prompt inherently share identical turn counts. Training set: 1,296 pairs (2-8 turns), manually verified by three annotators.

G.2 Evaluation Benchmarks

Evaluation uses held-out MT-Bench-101 (30%, in-distribution) and independent MT-Bench (Zheng et al., 2023) (80 questions, 8 categories) for cross-dataset generalization. Despite similar naming, MT-Bench and MT-Bench-101 have zero data overlap in prompts, content, task design, or evaluation methodology.

H Implementation Details

For each turn h in pair (τ^w, τ^l), we query reference model π_{ref} with multiple perspectives, repeating each multiple times. Scores per turn are aggregated via kernel-weighted averaging (Eq. 4), then normalized to obtain turn weights ω_h (Section 2.2).

I Cross-Dataset Generalization Results on MT-Bench

T^3PO consistently outperforms all baselines across model scales. Performance ranking remains identical to MT-Bench-101 results (T^3PO > SDPO > DMPO > DPO > ETO), demonstrating that dual-granularity mechanisms capture fundamental preference structures transferring across datasets.

J Human Annotation for Validation

Correlation analysis in Figure 6(b) includes author assessments. Three co-authors independently rated turn importance on 1-5 scale following standardized instructions. Annotators participated voluntarily without compensation and provided informed consent for anonymized aggregation. Evaluation involved synthetic model-generated dialogues without sensitive content. This internal validation of non-sensitive synthetic data was exempt from ethics review under institutional minimal-risk guidelines.

K Scope of Comparison

Token-level methods (Token-DPO (Zeng et al., 2024), Step-DPO (Lai et al., 2024), SePO (Yang et al., 2024)) are excluded as they lack multi-turn theoretical foundations. These methods provide no partition cancellation analysis for multi-turn settings—our Theorem 1 proves cancellation requires turn-level equal-length constraint $H^w = H^l$, not token-level properties. Their theoretical foundations do not extend to multi-turn trajectories $\tau = \{(h_0, y_0), \ldots, (h_{H-1}, y_{H-1})\}$ where turn structure must be explicitly modeled. Applying single-turn methods to multi-turn benchmarks conflates problem formulation mismatch with methodological differences, yielding uninterpretable results. Our token discounting operates within the proven turn-level framework, inheriting cancellation guarantees while adding intra-turn positional modeling—conceptually distinct from token-level methods lacking multi-turn foundations.

Table 3: Detailed Training Configuration

Configuration	Value
Model & Architecture	
Base Model	Qwen2.5-Instruct
Model Scales	3B / 7B / 32B
Fine-tuning Type	Full Parameter
Optimization	
Optimizer	AdamW
Learning Rate	5e-7
LR Scheduler	Cosine Decay
Warmup Ratio	0.0
Training Epochs	3
Per-device Batch Size	1
Gradient Accumulation Steps	4
Gradient Clipping	Auto
Loss & Hyperparameters	
β	0.1
Token Discount Factor γ	0.98
Scoring Perspectives M	4
Repetitions per Perspective G	5
Data Processing	
Maximum Sequence Length	4096
Preprocessing Workers	16
DeepSpeed ZeRO Configuration	
ZeRO Stage	Stage 3
Optimizer Offload Device	CPU
Parameter Offload Device	CPU
Pin Memory	True
Overlap Communication	True
Contiguous Gradients	True
Sub Group Size	1e9
Max Live Parameters	1e9
Max Reuse Distance	1e9
Reduce Bucket Size	Auto
Stage3 Prefetch Bucket Size	Auto
Gather 16bit Weights on Save	True
System & Infrastructure	
Precision	BF16
Hardware	2×A100-80GB GPUs
Random Seed	11
DDP Timeout	180000s

L Robustness to Evaluation Protocol

To validate robustness beyond judge-specific biases, we conduct cross-evaluator validation using GPT-4o (advanced GPT-4 iteration) and DeepSeek-V3 (open-source model with different training paradigm).

Tables 6 and 7 present results across all scales (3B/7B/32B) and task categories. T^3PO maintains superior performance across all three evaluators, with performance ranking stable across judges (T^3PO consistently first, followed by SDPO), demonstrating that dual-granularity mechanisms capture fundamental dialogue quality aspects generalizing across diverse evaluation frameworks.

Table 4: Software Dependencies and Package Versions

Package	Version
Core Deep Learning	
transformers	4.43.0 – 4.46.2
accelerate	$\geq$ 0.30.1
deepspeed	0.14.0
peft	$\geq$ 0.11.1
trl	0.8.6 – 0.12.0
Data & Utilities	
datasets	$\geq$ 2.16.0
pandas	$\geq$ 2.0.0
numpy	$<$ 2.0.0
pyarrow	$\leq$ 20.0.0
scipy	latest
matplotlib	$\geq$ 3.7.0
Model & Interface	
sentencepiece, tiktoken, protobuf, einops	latest
gradio, fastapi, uvicorn	latest

Table 5: Cross-dataset generalization on MT-Bench. Models trained on MT-Bench-101 evaluated on independent MT-Bench. Scores 1-10 (GPT-4.1 judge). **Bold**: best per scale.

Method	3B	7B	32B
DPO	6.05	7.21	8.16
ETO	5.84	6.95	8.29
DMPO	6.28	7.39	8.46
SDPO	6.54	7.69	8.71
T^{3}PO (Ours)	**6.85**	**7.99**	**8.98**
vs. SDPO	*+0.31*	*+0.30*	*+0.27*

M Training Efficiency

Table 8 reports per-step training speeds. T^{3}PO achieves efficiency comparable to baselines: turn weights are precomputed offline during preprocessing and cached; token-level discounting involves lightweight element-wise operations adding negligible cost.

Self-Evaluation Overhead. Turn weighting requires multiple inference calls per turn using frozen reference model. All turn weights are computed offline during preprocessing—calculated once, cached alongside preference pairs, loaded directly during training without re-evaluation. Operating in inference-only mode without gradient computation, this one-time cost is negligible when amortized across multi-epoch training. Per-step speeds in Table 8 reflect this: no online evaluation during training.

N Hyperparameter Analysis

N.1 Token-level Discount Factor

Figure 3 shows discount factor impact on Qwen2.5-3B/7B. Performance exhibits stability across moderate discount values. Lower values degrade reasoning through excessive suppression of later tokens; uniform weighting underperforms in context-sensitive dimensions through diluted focus. We select a moderate value balancing early-token prioritization and coverage completeness.

Table 6: Performance evaluated by GPT-4o, supplementing Table 1. Settings identical except judge model. ▲: Uniform token, ☆: Uniform turn, ●: Binary selection, ✪: Fine-grained modeling.

Method	GR	AR	MR	CC	TS	CM	SC	Avg.
Qwen2.5-3B-Instruct								
Base	3.15	6.02	3.31	6.79	4.18	4.12	7.03	4.94
DPO ▲	4.38	7.19	3.96	7.73	6.17	6.03	8.10	6.22
ETO ▲☆	4.27	7.02	3.88	7.60	5.99	5.92	7.93	6.09
DMPO ▲	4.51	7.28	3.99	7.81	6.26	6.15	8.16	6.31
SDPO ▲●	4.65	7.48	4.15	7.98	6.52	6.44	8.26	6.50
T^{3}PO ✪	**4.92**	**7.56**	**4.30**	**8.39**	**6.76**	**6.62**	**8.45**	**6.71**
Qwen2.5-7B-Instruct								
Base	4.57	7.25	4.30	7.85	5.37	5.05	7.69	6.01
DPO ▲	5.45	7.99	5.15	8.80	7.05	6.77	8.80	7.14
ETO ▲☆	5.34	7.75	5.02	8.51	6.85	6.60	8.58	6.95
DMPO ▲	5.67	8.03	5.11	8.67	7.19	7.37	8.28	7.19
SDPO ▲●	6.12	8.19	**5.23**	9.08	7.59	7.64	**8.92**	7.54
T^{3}PO ✪	**6.24**	**8.24**	5.17	**9.16**	**7.79**	**7.72**	8.89	**7.60**
Qwen2.5-32B-Instruct								
Base	5.42	8.07	4.38	9.15	6.47	5.83	8.27	6.80
DPO ▲	7.07	8.85	5.66	9.64	7.83	7.10	9.13	7.90
ETO ▲☆	7.18	9.05	5.61	9.08	8.62	7.40	9.41	8.05
DMPO ▲	7.66	9.02	6.03	9.29	8.30	7.36	9.17	8.12
SDPO ▲●	7.84	8.79	6.52	9.66	8.59	7.58	9.44	8.35
T^{3}PO ✪	**7.95**	**9.10**	**6.56**	**9.86**	**8.76**	**7.73**	**9.45**	**8.49**

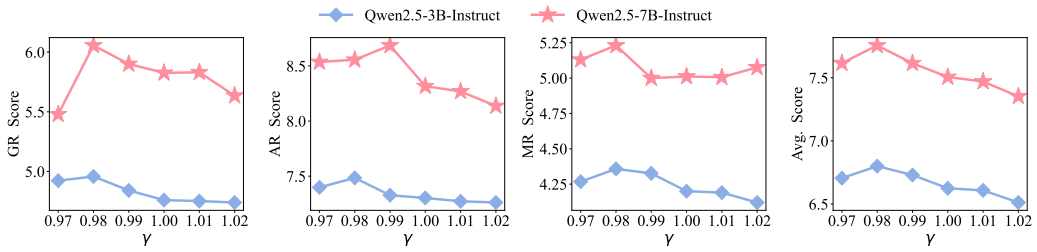


Figure 3: Performance sensitivity to token discount factor across task categories.

N.2 Turn Scoring Configuration

Figure 4 examines repetition count and aggregation strategy. Left: repetition exhibits stability-efficiency trade-off (low values cause variance, excessive repetition yields diminishing returns). Right: kernel weighting (Eq. 4) consistently outperforms mean aggregation by reducing outlier influence.

N.3 Impact of Dialogue Length

Figure 5 shows performance when trained on dialogues up to 2, 4, 6, 8 turns. Performance improves with longer training dialogues, saturating beyond 6 turns across all scales (3B/7B/32B), demonstrating effective leverage of extended contexts without requiring extensive long-dialogue data.

Table 7: Performance evaluated by DeepSeek-V3. Format identical to Table 1.

Method	GR	AR	MR	CC	TS	CM	SC	Avg.
Qwen2.5-3B-Instruct								
Base	2.85	5.76	3.02	6.51	3.89	3.84	6.75	4.66
DPO ▲	4.08	6.92	3.68	7.45	5.91	5.78	7.84	5.95
ETO ▲☆	3.98	6.76	3.61	7.32	5.73	5.67	7.67	5.82
DMPO ▲	4.21	7.01	3.72	7.53	5.99	5.90	7.91	6.04
SDPO ▲●	4.36	7.22	3.89	7.71	6.25	6.18	8.03	6.23
T^{3}PO ✪	**4.62**	**7.25**	**4.02**	**8.15**	**6.68**	**6.49**	**8.19**	**6.49**
Qwen2.5-7B-Instruct								
Base	4.28	6.98	4.02	7.56	5.08	4.78	7.42	5.73
DPO ▲	5.18	7.71	4.86	8.52	6.79	6.49	8.54	6.87
ETO ▲☆	5.07	7.48	4.73	8.23	6.57	6.32	8.29	6.67
DMPO ▲	5.39	7.75	4.82	8.40	6.92	7.11	8.03	6.92
SDPO ▲●	5.86	7.91	**4.95**	8.79	7.33	7.38	8.78	7.29
T^{3}PO ✪	**5.94**	**8.21**	4.88	**8.88**	**7.97**	**7.45**	**8.86**	**7.46**
Qwen2.5-32B-Instruct								
Base	5.12	7.79	4.09	8.86	6.19	5.56	8.01	6.52
DPO ▲	6.78	8.57	5.35	9.34	7.54	6.83	8.87	7.61
ETO ▲☆	6.89	8.74	5.28	8.78	8.34	7.13	9.14	7.76
DMPO ▲	7.38	8.77	5.72	8.99	8.01	7.09	8.91	7.84
SDPO ▲●	7.54	8.51	6.23	9.37	8.31	7.29	9.18	8.06
T^{3}PO ✪	**7.66**	**8.83**	**6.27**	**9.59**	**8.49**	**7.46**	**9.19**	**8.21**

Table 8: Training speed (steps/second). T^{3}PO maintains efficiency via offline turn weight precomputation.

Method	3B	7B	32B
DPO	0.063	0.038	0.006
ETO	0.070	0.039	0.007
DMPO	0.058	0.035	0.006
SDPO	0.067	0.042	0.009
T^{3}PO	**0.064**	**0.041**	**0.009**

O Case Study: Mechanism Validation

O.1 Token-Level Discounting Validation

Figure 6(a) visualizes KL divergence across relative token positions within turns. Baseline shows gradual decay; our method exhibits sharper early-position reduction, confirming token-level discounting effectively prioritizes high-impact tokens.

O.2 Turn-Level Evaluation Consistency

We compute Spearman correlations between human experts, self-evaluation using frozen reference model, and external judges (GPT-4, DeepSeek-V3). Figure 6(b) shows high pairwise correlations, with self-evaluation achieving comparable correlation with human consensus to external models, confirming reliable turn identification without external dependencies.

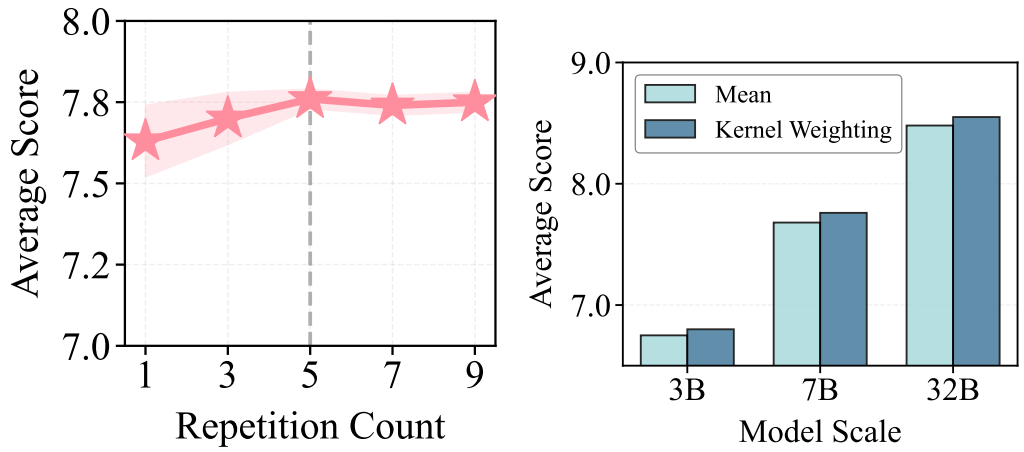


Figure 4: Turn scoring configuration. Left: Repetition count stability-efficiency trade-off. Right: Kernel weighting outperforms mean aggregation.

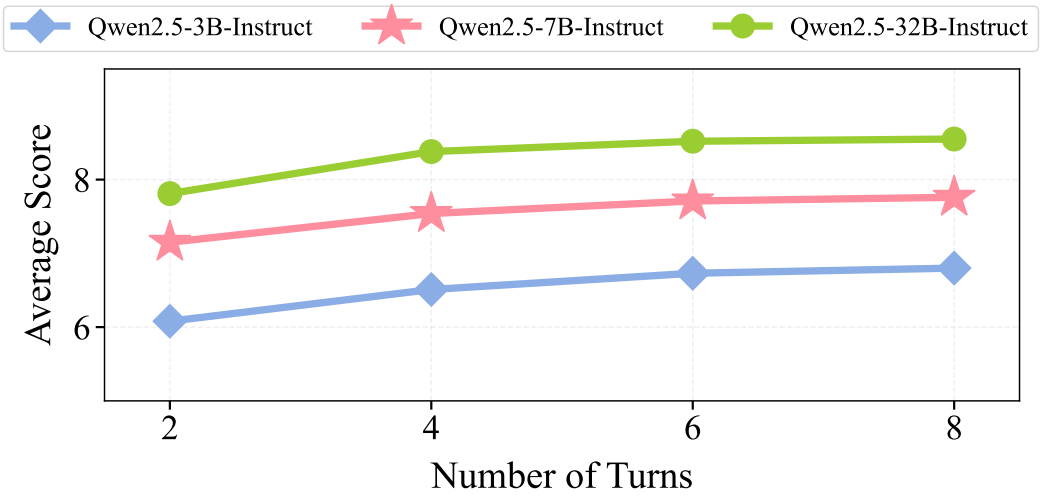
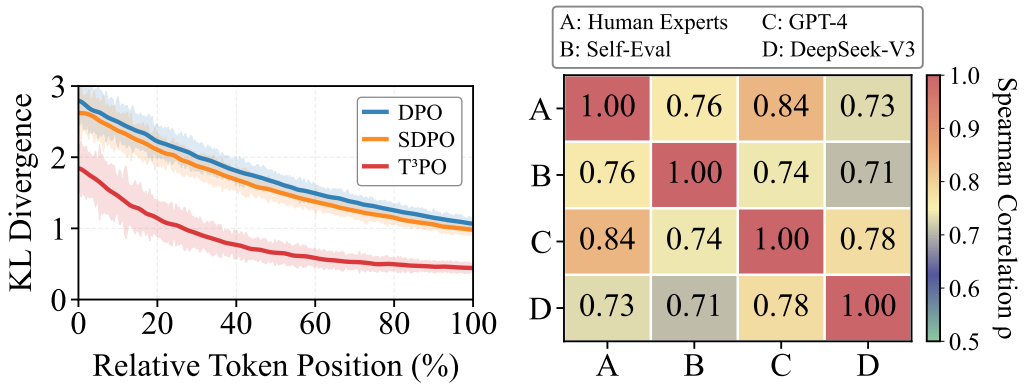


Figure 5: Performance when trained on dialogues up to 2, 4, 6, 8 turns across three scales.



(a) Token Discounting (b) Turn Weighting

Figure 6: Dual-granularity validation. (a) KL divergence shows T^3PO's sharper early-position reduction. (b) High cross-evaluator correlations (A: Human, B: Self-Eval, C: GPT-4, D: DeepSeek-V3).