DiPO Disentangled Perplexity Policy Optimization for Fine-grained Exploration-Exploitation Trade-Off

Xiaofan Li[1,2,†], Ming Yang[2,†], Zhiyuan Ma[2], Shichao Ma[2], Jintao Du[2], Yu Cheng[2], Weiqiang Wang[2], Zhizhong Zhang[1,‡], Xin Tan[1], Yanyun Qu[4], Lizhuang Ma[1,‡], Yuan Xie[1,3]

[1] East China Normal University [2] Ant Group [3] Shanghai Innovation Institute 4 Xiamen University

Fine-grained control of exploration and exploitation

Disentangle by correctness and perplexity; restore useful gradients in extreme groups.

01 Motivation: two failures of RLVR

Missing gradients. In GRPO-style training, all-zero (hard) and all-one (easy) groups have zero relative advantages. Hard groups lose exploration signals; easy groups lose exploitation signals.

Misaligned behavior. Incorrect responses can have low PPL, while correct responses can have high PPL. Uniformly encouraging exploration or exploitation ignores these mismatches.

02 DiPO: disentangle the space, reallocate the rewards

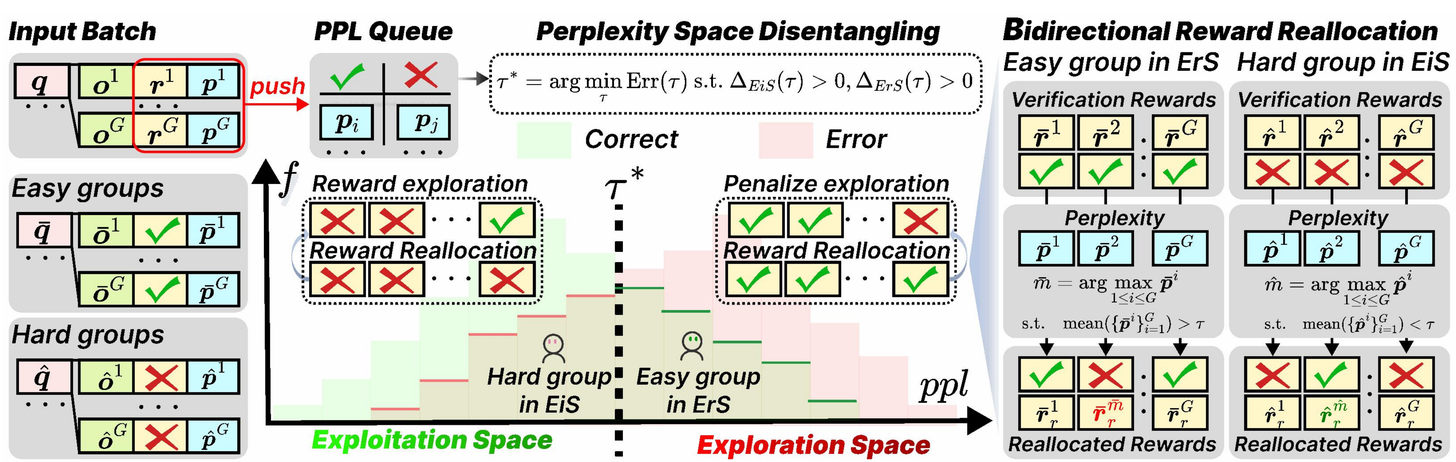


Figure 2. PPL queue → Perplexity Space Disentangling (PSD) → Bidirectional Reward Reallocation (BRR).

PSD | Select where to intervene

Cache two recent batches. Minimize correctness-classification error under 95% confidence-based advantage checks. Select hard groups with low mean PPL and easy groups with high mean PPL.

BRR | Change one reward per group

For the highest-PPL response, set reward 0 → 1 in selected hard groups and 1 → 0 in selected easy groups. Preserve the verification objective; add the BRR objective with α = 0.1.

J(DiPO) = J(DAPO, R) + α J(DAPO, Rr)

Code: github.com/FuNz-0/DiPO_RL

03 Experimental results

Mathematical reasoning

Six benchmarks · Accuracy / mean@8 (%)

Method	AIME24	AIME25	MATH	AMC	OLY	MIN	AVG
Qwen3-4B-Base							
Base model	9.58	3.75	48.48	31.48	25.52	24.82	23.94
GRPO	26.67	23.33	85.83	60.24	53.06	44.39	48.92
DAPO	26.25	23.75	86.43	61.90	53.88	44.34	49.43
DAPO w/ EL	26.67	**24.58**	86.78	62.95	**54.53**	44.53	50.01
CDE	26.67	24.17	85.93	62.35	52.25	43.11	49.08
DiPO (ours)	**29.17**	**24.58**	**87.00**	**63.70**	54.09	**44.76**	**50.55**
Qwen3-8B-Base							
Base model	8.33	9.17	66.78	39.46	33.30	31.75	31.47
GRPO	31.67	24.58	89.08	69.28	56.20	**48.62**	53.24
DAPO	30.08	25.83	89.43	69.12	56.90	48.02	53.23
DAPO w/ EL	33.75	25.42	**89.58**	69.87	57.21	47.56	53.90
CDE	31.67	26.25	89.35	68.07	57.11	47.75	53.37
DiPO (ours)	**35.00**	**27.50**	89.55	**71.23**	**57.73**	47.75	**54.79**
Qwen2.5-7B							
Base model	7.08	2.08	41.53	22.74	19.28	17.19	18.32
GRPO	20.42	15.42	79.15	58.43	42.42	36.95	42.13
DAPO	20.42	**16.67**	79.08	59.94	42.70	37.55	42.73
DAPO w/ EL	20.00	14.58	79.85	58.73	43.05	**39.65**	42.64
CDE	20.00	15.00	79.00	55.87	42.94	35.94	41.46
DiPO (ours)	**22.92**	**16.67**	**80.35**	**60.09**	**43.72**	37.59	**43.56**

Best AVG at all 3 scales; +1.56 pp on Qwen3-8B.

Function calling | BFCL v3

Method	Non-Live Acc	Live Acc	Multi Turn Acc	Relevance Detection	Irrelevance Detection
Qwen2.5-3B-Instruct					
Base model	42.52	53.96	1.00	64.71	56.01
SFT400	69.29	41.40	0.00	94.44	8.11
SFT400+PPO	78.29	58.76	5.12	94.12	54.70
SFT400+GRPO	76.21	64.15	1.75	88.89	71.76
PPO, Cold Start	82.42	67.78	4.88	94.12	73.87
ToolRL+GRPO	81.58	**73.78**	3.75	88.24	**84.85**
ToolRL+DAPO	82.19	69.43	8.00	81.25	57.60
ToolRL+DiPO	**83.42**	73.06	**8.62**	**100.00**	54.16
Qwen2.5-7B-Instruct					
Base model	66.02	53.51	4.25	76.47	62.66
SFT400	69.29	41.40	0.00	94.44	8.11
SFT400+PPO	83.90	51.84	0.25	**100.00**	29.66
SFT400+GRPO	80.69	46.51	0.25	**100.00**	14.19
PPO, Cold Start	79.33	63.17	0.38	88.89	52.92
ToolRL+GRPO	86.17	74.90	18.12	83.33	**76.68**
ToolRL+DAPO	**87.10**	76.31	19.75	87.50	67.25
ToolRL+DiPO	86.21	**76.83**	**24.50**	87.50	69.57