ResT: Entropy-Aware Reshaping of Token-Level Policy Gradients for Robust Tool-Use RL

Zihan Lin[123*†] **Xiaohan Wang**[2*] **Jie Cao**[1] **Jiajun Chai**[2] **Guojun Yin**[2‡] **Wei Lin**[2] **Ran He**[1‡]

[1] MAIS & NLPR, Institute of Automation, CAS [2] Meituan [3] School of Artificial Intelligence, UCAS

01 MOTIVATION

The Problem with Uniform Token Credit

Standard GRPO assigns identical gradient weight to every token in a tool-use trajectory, but different segments have fundamentally different roles:

- Format tags — low entropy, purely structural
- Tool names — low entropy, semantically critical
- Parameters — medium entropy, value-sensitive
- Chain-of-thought — high entropy, open-ended reasoning

Uniform weighting wastes gradient on already-certain tokens and under-weights critical decision points.

02 THEORETICAL FOUNDATION

Lemma 2: Gradient Variance Bound

Token-level policy gradient variance is bounded by entropy:

$$\beta_t = \mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(y_t \mid y_{<t})\|^2\right] \cdot (1 - e^{-H_t})$$

Theorem 1: Variance Upper Bound

For entropy-aware reweighted gradient:

$$\mathrm{Var}(\hat{g}_i^{\mathrm{rw}}) \leq \mathbb{E}[\hat{A}^2] \cdot \sum_t \beta_t \cdot \tilde{w}_t^2$$

Theorem 2: Optimal Token Weights

Optimal weights are inversely proportional to β_t:

$$\tilde{w}_t^* \propto 1/\beta_t \implies \text{Low-entropy tokens get more weight}$$

Key insight: Tokens the model is already confident about (format, tool names) carry less gradient noise and should receive stronger reinforcement signals.

03 METHOD: ResT

A. Dynamic Reward Scoring

Multi-granularity verification combining:

- Format matching score S_format (structural correctness)
- Accuracy score S_acc (Jaccard similarity for names/params)
- Dynamic scaling via training progress variable v

$$R = \beta_a \cdot S_{\mathrm{acc}} \cdot (1 - \nu) + \beta_f \cdot S_{\mathrm{format}} \cdot \nu$$

B. Entropy-Aware Token Reweighting

Partition response into 4 semantic regions:

Format Tags	Tool Names	Param Values	Chain of Thought
Low H	*Low H*	*Med H*	*High H*

Weight each region inversely proportional to its mean entropy. Normalize by sequence mean to maintain gradient scale.

C. Curriculum Learning

Progressive focus shift across training:

- Early: prioritize format tokens (learn structure first)
- Mid: increase parameter weight (learn correct values)
- Late: boost CoT weight (refine reasoning quality)
- Tool name weight remains constant throughout

D. No KL Penalty

Replace KL divergence penalty with entropy reweighting + PPO clipping + curriculum scheduling. Avoids reference-model overhead while maintaining stability.

$$\mathcal{L}_{\mathrm{ResT}} = -\frac{1}{|G|} \sum_g \sum_i \frac{\omega_t}{T} \min\left(r_{i,t}\,\hat{A}_i,\ \mathrm{clip}(r_{i,t},\, 1-\varepsilon,\, 1+\varepsilon)\,\hat{A}_i\right)$$

04 MAIN RESULTS: BFCL

Multi-Turn Tool Use Benchmark

Model	GRPO	Dr.GRPO	DAPO	ResT
Qwen3-1.7B	14.50	15.25	14.50	**16.00**
Qwen3-4B	41.62	48.62	42.38	**50.38**
Qwen3-8B	36.00	38.12	38.25	**40.13**
Qwen3-14B	38.88	37.00	40.38	**44.25**

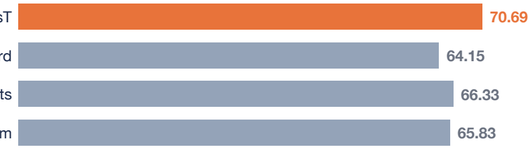
4B ResT (50.38%) surpasses GPT-4o (50.00%)
on multi-turn tool use — with 200x fewer parameters

05 API-BANK BENCHMARK

Model	GRPO	Dr.GRPO	ResT	Δ
1.7B	63.65	63.99	**64.99**	+1.34
4B	65.33	66.33	**68.68**	+3.35
8B	66.50	68.01	**70.69**	+4.19
14B	66.33	66.99	**69.35**	+3.02

06 ABLATION STUDY

API-Bank (Qwen3-8B) — each component is necessary:



KEY TAKEAWAY ResT reshapes token-level policy gradients using entropy-aware weighting, dynamic reward scoring, and curriculum learning — achieving SOTA on BFCL and API-Bank across all Qwen3 model sizes, with a 4B model surpassing GPT-4o on multi-turn tool use.

Contact linzihan24@mails.ucas.ac.cn