Stabilizing On-Policy Distillation for MLLM Reasoning with Global Normalization

Anonymous ACL submission

Abstract

On-policy distillation (OPD) has recently emerged as an important post-training paradigm. By using a stronger teacher model to provide dense, fine-grained supervision for sampled trajectories, OPD offers a clear advantage over reinforcement learning with verifiable rewards (RLVR), which typically depends on sparse binary or outcome-based environmental feedback. However, naive token-level distillation can suffer from gradient instability, due to magnitude misalignment in outlier states. To address this issue, we propose Globally Normalized Distillation Policy Optimization (GNDPO), a practical method that stabilizes optimization by transforming raw KL scores into batch-level relative advantages. This normalization effectively mitigates gradient explosions while retaining the benefits of token-level guidance. Experimental results show that GNDPO substantially improves training robustness and downstream performance across multimodal reasoning tasks.

1 Introduction

Post-training is a critical stage for injecting fine-grained reasoning abilities and domain-specific knowledge into large language models (LLMs) and multimodal large language models (MLLMs). A typical post-training pipeline consists of two stages: supervised fine-tuning (SFT) and reinforcement learning (RL) (Guo et al., 2025; Zheng et al., 2025; Hu et al., 2025; Qin et al., 2025; Song and Zheng, 2026). Although SFT provides direct linguistic supervision, it suffers from exposure bias at inference time. In contrast, on-policy RL methods alleviate distribution shift through online sampling, but they are often constrained by sequence-level sparse rewards. Such terminal-only feedback provides limited dense supervision for long-chain reasoning tasks, resulting in inefficient credit assignment and unstable training dynamics (Team et al., 2026).

Recently, on-policy distillation (OPD) has emerged as a promising alternative that bridges supervised fine-tuning and reinforcement learning by using the teacher model's logits as dense token-level supervision (Lu and Thinking Machines Lab, 2025; DeepSeek-AI, 2026; Xiao et al., 2026). Despite its potential, existing OPD methods often suffer from substantial gradient instability. We observe that the raw KL divergence between the student and teacher models can fluctuate sharply when the student explores out-of-distribution (OOD) states. In particular, when the student generates tokens that the teacher assigns very low probability, the resulting log-ratio can induce large gradient spikes, disrupting the student's pretrained representations and leading to unstable or divergent training.

To address this issue, we introduce Globally Normalized Distillation Policy Optimization (GNDPO), a robust training framework that reformulates distillation as a relative optimization problem. Instead of directly optimizing raw teacher feedback, GNDPO converts token-level distillation rewards into relative advantages within each training batch. This global normalization mechanism dynamically rescales gradients and suppresses anomalous spikes, preventing optimization from being dominated by localized high-magnitude mismatches. As a result, GNDPO guides the student toward teacher-aligned distributions in a stable and controllable manner, providing a practical solution for large-scale model post-training.

To summarize, our contributions are threefold:

- We propose a globally normalized distillation mechanism that transforms raw teacher feedback into relative advantages, thereby stabilizing optimization in practical training settings.
- We conduct extensive experiments across diverse multimodal reasoning benchmarks, showing that GNDPO consistently improves

generation quality and training robustness over state-of-the-art RL and standard distillation baselines.

2 Preliminary

2.1 On-policy RL

Recent advancements of on-policy reinforcement learning methods such as GRPO (Guo et al., 2025) and GSPO (Zheng et al., 2025), has been increasingly favored in reinforcement learning for its efficiency and simplicity by replacing the traditional value model with a group-relative advantage estimation mechanism.

Formally, consider a given question q, where the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_i\}_{i=1}^{G}$ and computes their rewards $\{r_i\}_{i=1}^{G}$. The group-relative advantage for the j-th response, A_j, is subsequently derived by normalizing the aggregated rewards across the entire group:

$$A_i = \frac{r_i - \text{mean}\{r_1, \ldots, r_G\}}{\text{std}\{r_1, \ldots, r_G} \tag{1}$$

Based on this advantage estimation, the GSPO optimization objective is formulated as:

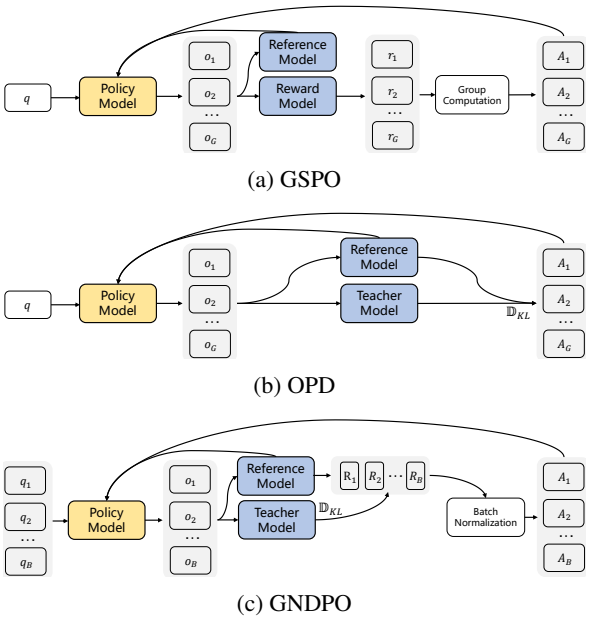
$$\begin{aligned}\mathcal{J}_{\text{GSPO}}(\theta) =& \mathbb{E}_{(q,o_i)\sim D, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Bigg[\frac{1}{G} \\ & \sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(s_{i,t}(\theta)A_{i,t}, \\ & \text{clip}(s_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_{i,t}\Big)\Bigg]\end{aligned} \tag{2}$$

where the probability ratio is defined as $\text{sg}[s_i(\theta)] \cdot \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\text{sg}[\pi_\theta(o_{i,t}|q,o_{i,<t})]}$ and ϵ represents the clipping threshold. For simplicity, the KL-divergence regularization term is omitted in this formulation.

2.2 On-Policy Distillation

On-policy distillation generates trajectories with the student model and evaluates them at token-level granularity using a stronger teacher model. Unlike GRPO-style methods, which rely on sparse reward signals, on-policy distillation provides dense token-level supervision from the teacher model throughout the generated trajectory.

Formally, given a question q, the student policy $\pi_{\theta_{\text{old}}}$ first generates a response o_i and computes the token-level logits. The teacher model then evaluates the same response and produces its corresponding logits. We use the reverse KL divergence between the student and teacher distributions as the reward signal for each token:



(a) GSPO

(b) OPD

(c) GNDPO

Figure 1: The comparison of GSPO, OPD and our proposed GNDPO method. GSPO relies on sparse, sequence-level rewards, which makes detailed credit assignment difficult. OPD provides dense, token-level guidance using raw KL divergence, but it is prone to gradient explosion. Our GNDPO solves this problem by normalizing the rewards within each batch, ensuring both high signal density and stable training dynamics.

$$\hat{A}_{i,t} = -(\log \pi_\theta(o_i^t|q, o_j^{<t}) - \log \pi_{\text{teacher}}(o_i^t|q, o_j^{<t})) \tag{3}$$

We define the token-level reward as the negative reverse KL divergence between the student and teacher distributions, encouraging the student policy to match the teacher. This reward reaches its maximum when the student distribution exactly recovers the teacher distribution. The resulting rewards are then used in an importance-sampling-based RL objective to update the student model.

3 Method

3.1 Motivation

Although token-level dense rewards provide richer supervision than sequence-level sparse rewards, such as those used in GSPO, they can introduce substantial gradient instability during on-policy exploration.

In particular, when the student policy visits out-of-distribution (OOD) contexts or generates low-frequency tokens to which the teacher assigns near-zero probability, $\pi_{\text{teacher}}(o_j^t \mid q, o_j^{<t}) \to 0$, the log-ratio term $\log \frac{\pi_\theta(o_j^t|q,o_j^{<t})}{\pi_{\text{teacher}}(o_j^t|q,o_j^{<t})}$ can become extremely

large. As a result, the raw distillation reward $\hat{A}_{i,t}$ may decrease sharply, producing disproportionately large gradient updates for a small number of tokens:$\nabla_\theta \mathcal{L}(\theta) \propto \nabla_\theta \log \pi_\theta(o_j^t \mid q, o_j^{<t}) \cdot \hat{A}_{i,t}$ Directly optimizing these raw KL-based scores can cause the policy gradient to be dominated by a small subset of high-entropy or highly mismatched tokens. This magnitude misalignment across contexts disrupts stable policy updates and can lead to training divergence.

3.2 Globally Normalized Distillation Policy Optimization

To mitigate the gradient explosion caused by localized logit discrepancies, we propose Globally Normalized Distillation Policy Optimization (GNDPO). Instead of assigning the raw KL divergence directly as the optimization landscape, we calibrate the token-level rewards into a relative advantage within each training batch $\mathcal{B}$.

For each token $o_{i,t}$ in the batch, its normalized distillation advantage $\hat{A}_{i,t}^{\text{norm}}$ is formulated as:

$$\hat{A}_{i,t}^{\text{norm}} = \frac{\hat{A}_{i,t} - \text{mean}(A|A \in \mathcal{B})}{\text{std}(A|A \in \mathcal{B}) + \epsilon} \tag{4}$$

This global normalization mechanism offers two critical benefits:Adaptive Gradient Rescaling:

1. When catastrophic logit mismatches occur, the extreme negative rewards expand the batch standard deviation $\sigma_{\mathcal{B}}$. This automatically scales down the effective gradient step size, preventing anomalous tokens from rupturing the shared representations.

2. Context-Agnostic Scale Calibration: By shifting the baseline via $\mu_{\mathcal{B}}$, the objective transitions from tracking absolute distribution matching to optimizing relative token quality. This balances the optimization contribution across heterogeneous contexts of varying generation difficulties.

4 Experiments

4.1 Experimental Setup

Models and Datasets. We conduct experiments with the InternVL3.5 model family (Wang et al., 2025) at three scales: InternVL3.5-1B-Instruct, InternVL3.5-2B-Instruct, and InternVL3.5-4B-Instruct. We use InternVL3.5-8B as the teacher model for OPD and GNDPO. For training data, we use the multimodal mathematical reasoning dataset Geometry3K (Lu et al., 2021). We evaluate the model on various challenging multimodal reasoning benchmarks: MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), MathVision (Wang et al., 2024), MathVerse (Zhang et al., 2024), DynaMath (Zou et al., 2025), Wemath (Qiao et al., 2025), LogicVista (Xiao et al., 2024).

Baselines. We compare our method with following baselines: (1) GSPO, group sequence relative policy optimization with binary outcome rewards verified against ground-truth answers. (2) GSPO-OPD, group sequence relative policy optimization with on policy distill rewards by the teacher model.

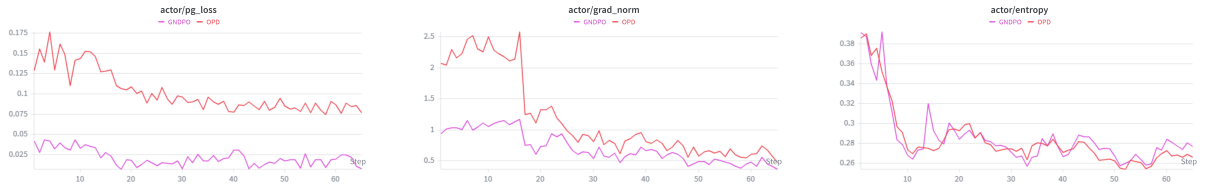
Implementation details. We implement our method based on the VERL (Sheng et al., 2025) framework. For GSPO, we sample 16 responses per problem. During training, the model's maximum context length is set to 24576, with a maximum prompt length of 8192 and a maximum response length of 16384. During evaluation, the model's maximum context length is set to 32768. We use Adam optimizer with a constant learning rate of 1e-6. The training batch size is set to 32, and we sample 16 responses per problem with a sampling temperature of 1.0. All experiments are conducted on 8 NVIDIA H20 100GB GPUs. More experimental details are in Appendix A. We will release our codes upon paper acceptance.

4.2 Main Results

Table 1 reports the evaluation results across seven multimodal reasoning benchmarks. GNDPO consistently outperforms the baseline methods in average accuracy across scales. Specifically, for InternVL3.5-4B, GNDPO leverages stabilized token-level advantages to achieve notable gains on complex reasoning datasets, including MMMU (+1.1%), MathVision (+1.1%), and WeMath (+1.6%), where precise reasoning chains are crucial. For InternVL3.5-2B, our method substantially boosts performance on MathVista (+1.3%), MathVision (+0.5%), and MathVerse (+1.7%). Concurrently, for the parameter-constrained InternVL3.5-1B, GNDPO drives distinct improvements on MathVista (+0.7%) and MMMU (+0.5%), proving that robust reward calibration is vital even for smaller models. While individual performance gains vary due to the baseline models' differing domain capacities, the consistent aggregate improvements underscore GNDPO's effectiveness in achieving stable behavioral alignment.

	MathVista MINI	MMMU VAL	MathVision	MathVerse (vision-only)	DynaMath (worst case)	WeMath	LogicVista	Overall
InternVL3.5-8B	78.4	73.4	56.8	61.5	37.7	57.0	57.3	60.3
InternVL3.5-4B								
Base (Instruct)	71.4	64.3	40.5	50.0	30.7	35.6	53.5	49.4
+GSPO	71.4	67.2	48.7	52.7	31.1	42.6	53.5	52.5
+OPD	**73.3**	67.2	50.5	52.9	**32.7**	47.2	53.9	53.9
+GNDPO	73.0	**68.3**	**51.6**	**53.2**	32.5	**48.8**	**54.4**	**54.5**
InternVL3.5-2B								
Base (Instruct)	60.8	53.0	27.0	39.6	19.8	28.1	41.2	38.5
+GSPO	62.3	54.7	35.9	43.2	19.8	28.2	39.6	40.5
+OPD	64.7	57.4	37.9	42.0	**24.4**	**34.9**	39.6	42.9
+GNDPO	**66.0**	**57.7**	**38.4**	**43.7**	23.8	34.8	**41.6**	**43.7**
InternVL3.5-1B								
Base (Instruct)	37.2	48.6	15.8	27.0	8.4	13.9	29.1	25.7
+GSPO	52.2	39.9	21.6	28.2	8.4	14.0	26.6	27.3
+OPD	55.9	47.1	24.0	28.3	11.8	23.1	30.9	31.6
+GNDPO	**56.6**	**47.6**	**24.5**	**28.4**	**12.0**	**23.7**	**31.3**	**32.0**

Table 1: Multimodal reasoning performance across different models with various optimization methods. **Bold** indicates the best result. "Overall" represents the average score across the seven selected benchmarks.



(a) Training loss dynamics comparison.

(b) Training grad_norm dynamics comparison.

(c) Training entropy dynamics comparison.

Figure 2: Training dynamics of OPD method and GNDPO method.

4.3 Training Dynamics

Figure 2 illustrates the training dynamics over 65 optimization steps. As shown in Figure 2(a), Standard OPD exhibits severe volatility due to raw, uncalibrated reward fluctuations, whereas GNDPO maintains a significantly lower and smoother loss profile throughout training. Figure 2(b) reveals that standard OPD suffers from a catastrophic gradient spike (ℓ_2 norm > 2.5) around step 15 caused by localized logit mismatches, while GNDPO successfully suppresses these anomalous spikes via global batch normalization. Figure 2(c) shows that GNDPO stabilizes the gradient scale without sacrificing the model's exploratory capacity.

5 Conclusion

In this paper, we introduced Globally Normalized Distillation Policy Optimization (GNDPO) to address the issue of gradient instability in on-policy token-level distillation. By reformulating raw, context-dependent teacher feedback into relative batch-level advantages, GNDPO successfully suppresses catastrophic gradient spikes caused by localized logit discrepancies. Extensive experiments across multiple multimodal reasoning benchmarks demonstrate that our framework consistently outperforms standard OPD and RLVR baselines in both average accuracy and training robustness.

Limitations

Despite its empirical success and enhanced stability, GNDPO exhibits a few limitations: first, since our global normalization relies on empirical statistics calculated over a training mini-batch, its stabilization capability may slightly degrade under extremely small batch sizes due to increased statistical variance; second, as an on-policy paradigm, it inherits the standard computational overhead of continuous token generation during rollouts, meaning the training wall-clock time remains bounded by model inference speeds; finally, while we successfully verified its efficacy on core mathematical and multimodal reasoning tasks, GNDPO's behavioral dynamics in open-ended creative text generation or multi-turn conversational alignment remain to be extensively explored.

References

DeepSeek-AI. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. Thinking Machines Lab: Connectionism. Accessed: 2026-05-19.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

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

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, and 1 others. 2026. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pages 48337–48383.

Parameter	Value
Learning rate	1×10^{-6}
Train batch size	32
PPO mini-batch size	32
Number of rollouts	16
Sampling temperature	1.0
Max prompt length	8192
Max response length	16384
Training steps	65
KL coefficient β	-

Table 2: Training hyper-parameters.

Parameter	Value
Max new tokens	32768
Temperature	0.6
Top-p	0.95
Top-k	-1

Table 3: Evaluation hyper-parameters.

A Appendix

A.1 Experimental Details

We provide the training and evaluation configurations in Table 2 and Table 3. Both GSPO, OPD and GNDPO methods use the same base hyperparameters where applicable to ensure fair comparison. We use the vlmevalkit (Duan et al., 2024) framework to evaluate the model performance.

A.2 Prompt

The reasoning prompt used in the model is shown as follows:

Prompt used for Answering Questions

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in \\boxed{}