ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Model Sees but Does Not Learn: Eliminating Error Propagation in Reasoning Distillation

Jaeeun Jang[1], Hansle Lee[1], Wonjun Cho[1] and Sangmin Kim[1]

1- Hanwha Systems - AI Research Team
Pangyoyeok-ro, Seongnam-si, Gyeonggi-do 13524 - Republic of Korea

Abstract. Small LLMs struggle to acquire robust reasoning through RLVR due to instability and reward sparsity, and standard SFT, which directly imitates teacher-generated reasoning traces, inherits erroneous intermediate steps and collapses reasoning diversity. We introduce VGGM, a selective-learning objective that applies gradients only to verified correction segments, preventing error memorization and mitigating excessive entropy collapse. VGGM yields more stable optimization, stronger metacognitive correction behavior, and substantially higher data efficiency. Across GSM8K, MATH, and AMC'23, VGGM consistently outperforms standard SFT and, when combined with GRPO, achieves performance approaching DeepSeek-R1-distillation models while using 40× less supervised data.

1 Introduction and Related Work

Large Language Models (LLMs) have recently shown strong advances in mathematical and logical reasoning, and two primary paradigms have emerged for eliciting such reasoning abilities. The first is Reinforcement Learning with Verifiable Rewards (RLVR), introduced by DeepSeek-R1 [1]. RLVR directly optimizes reasoning quality using structured outcome-based rewards and has shown strong performance in large models. However, recent empirical observations highlight its limitations in small models: RLVR training is inherently unstable, sensitive to exploration noise, and often fails to converge to reliable reasoning behaviors due to the difficulty of long-horizon credit assignment and reward sparsity. A second, widely adopted approach for enabling reasoning in small models is supervised fine-tuning (SFT) on reasoning trajectories generated by a strong teacher model. Works such as S1 [2] and LIMO [3] show that SFT alone can yield surprisingly strong performance, even rivaling much larger models. Consequently, SFT has become the de facto standard for distilling reasoning into compact student models due to its simplicity, stability, and data efficiency. However, this approach suffers from two fundamental limitations that hinder generalizable reasoning.

First, SFT treats every step in a reasoning trace as equally reliable, even when early steps are misleading or later corrected. In practice, this means the model also learns noisy intermediate reasoning that should not contribute to its final solution. Moreover, this uniform supervision blurs the distinction between mistakes and their resolution, preventing the model from learning how to revise or regulate its own reasoning. Second, SFT fundamentally collapses reasoning diversity. The cross-entropy objective concentrates probability mass on a single next-token target, making the model overly deterministic and optimizing it for

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

`pass@1` rather than `pass@N` [4]. This collapse is especially harmful at key forking tokens—high-entropy positions where multiple valid reasoning paths branch—where SFT sharply reduces entropy and suppresses alternative trajectories. As a result, the model loses the flexibility to explore diverse reasoning chains, repeatedly reproduces the same incorrect solution under multi-sample decoding, and generalizes poorly to problems requiring path-level variability [5].

2 Verification-Guided Gradient Masking

Standard SFT assumes that longer reasoning traces naturally improve performance, indiscriminately training all tokens via next-token prediction [2, 3]. However, this oversimplified assumption ultimately limits the development of robust reasoning capabilities, a shortcoming also observed in DeepSeek-R1-Distill models trained on 800k samples generated by DeepSeek-R1 [1], as we discuss below.

The Error Propagation Problem. Reasoning sequences often include intermediate mistakes later corrected through verification, e.g., "The total apples are 120. Therefore, remaining apples are $120 \times \frac{1}{4} = 40$ (incorrect). Wait, let me recalculate: $120 \times \frac{1}{4} = 30$ (correct)". However, standard SFT assigns equal weight to all tokens, $\mathcal{L}_{\text{traditional}} = -\sum_{t=1}^{T} \log P_\theta(y_t|x, y_{<t})$, learning both incorrect and corrected steps indiscriminately. This leads to the internalization of flawed reasoning, causing error accumulation and reduced self-correction ability.

Beyond memorizing erroneous steps, indiscriminate SFT introduces gradient noise analogous to training with noisy labels [6]: incorrect intermediate calculations push parameters toward wrong solution regions, while later corrective steps pull them back in the opposite direction. This tug-of-war inflates gradient variance and destabilizes convergence, causing oscillatory loss behavior.

The Metacognitive Capability Gap. Beyond error propagation, existing methods fail to develop genuine metacognitive capabilities—the capacity to monitor and regulate their own reasoning. The essence of System-2 reasoning lies not in sequential inference but in self-verification and error correction that enable reflection. However, models trained with standard SFT merely imitate verification cues (e.g., "Wait!", "Let me reconsider") without true self-evaluation [7], reflecting surface-level mimicry rather than genuine metacognitive reasoning.

Verification-Guided Gradient Masking (VGGM). We introduce VGGM to address both issues through selective learning, enabling models to "learn only from correct segments" while "genuinely internalizing verification rather than merely mimicking patterns". Our objective applies gradients conditionally:

$$\mathcal{L}_{\text{adaptive}} = -\sum_{i=1}^{N} \sum_{t=t_{start}}^{T} \log P_\theta(y_{i,t}|x_i, y_{i,<t}), \quad t_{start} = \begin{cases} 1 & \text{if } v_i = 0 \\ T_{last} & \text{if } v_i = 1 \end{cases} \tag{1}$$

where T_{last} denotes the start of the final verification segment. For samples containing verification ($v_i = 1$), gradients are applied only from T_{last} onward to avoid learning erroneous intermediate steps, enabling the model to internalize

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

genuine correction behavior. We employ the OpenR1-Math-220k [8], leveraging DeepSeek-R1's reasoning, yielding 29,848 samples. Using Claude-3.5-Sonnet, we automatically identify final verification elements and insert standardized "Wait!" triggers at T_{last} for consistent metacognitive learning signals. Since about 90% of samples contain verification ($v_i = 1$), relying solely on this setting would bias training toward verification-conditioned reasoning. To maintain balance, we sample 12k instances with $v_i = 1$, 8k with $v_i = 0$, and reserve 9,848 for GRPO [1], enabling the model to learn both direct problem-solving and self-verification.

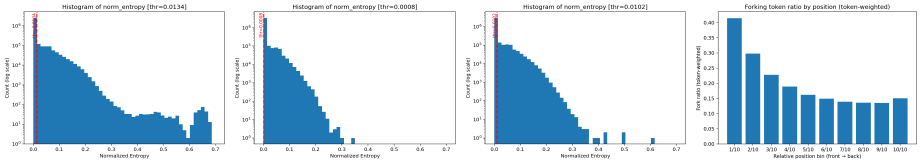
Our method is conceptually analogous to the GCE loss for noisy-label training, which attenuates gradients from unreliable supervision [6]. The GCE work pointed out that cross-entropy is sensitive to noisy labels and tends to memorize error patterns. Just as GCE prevents models from memorizing erroneous supervision at the label level, our VGGM prevents models from internalizing erroneous intermediate steps at the token level. Both approaches converge on the same principle: robustness arises not from learning everything indiscriminately, but from selectively emphasizing clean supervision while suppressing noise, making VGGM a reasoning-specialized extension of GCE for sequence-level training.

Complementing this mitigation of error propagation, VGGM also addresses the metacognitive capability gap through its handling of pre-*Wait!* prefixes. Although the full sequence still passes through the attention mechanism in the forward pass—meaning that pre-*Wait!* tokens can indirectly influence hidden states and propagate gradients into the model parameters—we set the loss to zero for all such pre-*Wait!* tokens, so these erroneous prefixes serve only as contextual conditioning rather than misleading supervision. In effect, the model *sees but does not learn from* early mistakes: credit assignment is restricted solely to the verified correction segment, enabling the model to learn how to revise an errorful trajectory instead of imitating it. This contrasts with truncation, which removes the error context and prevents learning the realistic error → detection → correction pattern that is essential for genuine metacognitive reasoning.

In addition, VGGM preserves broader token-level uncertainty by updating only verified suffixes. Since gradients are applied exclusively to the final correction segment, earlier tokens, where uncertainty is naturally higher due to branching reasoning paths, remain unsharpened by cross-entropy updates.

3 Experiments

3.1 Entropy Preservation Analysis



(a) Base (b) standard SFT (c) VGGM (d) Fork-Token Ratio

Fig. 1: Token-level entropy histograms and fork-token positional distribution.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

We evaluate entropy behavior on the MATH dataset by generating 8 rollouts per question and computing token entropy as: $H_t = -\sum_{j=1}^{V} p_{t,j} \log p_{t,j}$, where $p_t = \text{Softmax}(z_t/T)$. Columns (a)-(c) in Figure 1 show token entropy distributions for the base model, standard SFT, and VGGM, and column (d) shows the positional distribution of high-entropy fork-tokens across reasoning steps [5].

Standard SFT shows clear entropy collapse, sharpening logits across the full sequence—even at steps where uncertainty should remain to support alternative reasoning paths. In contrast, VGGM maintains substantially higher entropy in the early reasoning phase by limiting gradient updates to verified suffixes, preventing premature commitment to a single trajectory. Consistent with this pattern, the positional distribution of forking tokens shows that high-entropy branching decisions primarily occur early in the reasoning trace, where VGGM preserves exploratory diversity while standard SFT rapidly collapses uncertainty.

3.2 Effect of VGGM on Reasoning Performance

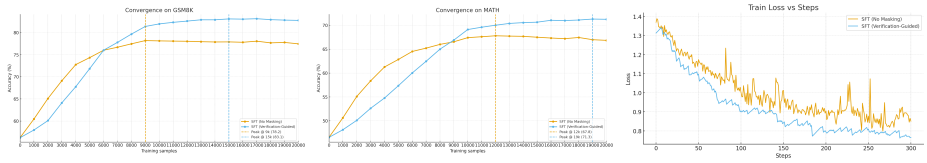
We evaluate how VGGM improves reasoning quality in Qwen2.5-7B-Base [9] across three mathematical benchmarks: GSM8K [10], MATH [11], and AMC'23.

Table 1 shows that VGGM improves supervised fine-tuning by blocking gradients from incorrect segments, yielding a cleaner and more stable reasoning foundation than standard SFT. Moreover, VGGM shows far less entropy collapse, preserving more diverse reasoning paths before RL. This retained diversity gives GRPO a broader exploration space, enabling substantially larger RL-driven gains [12]. As a result, VGGM benefits far more from GRPO than standard SFT, ultimately achieving the highest performance across all benchmarks.

Table 1: Benchmark performance.

Method	GSM8K	MATH	AMC
Base (No SFT)	66.2	46.5	28.2
Base + GRPO	74.2	50.5	30.1
SFT (No Masking)	78.2	67.8	52.3
SFT (No Masking) + GRPO	80.7	69.0	54.1
SFT (VGGM)	83.1	71.3	58.7
SFT (VGGM) + GRPO	**87.3**	**74.9**	**62.7**

3.3 Stabilizing Training Through VGGM



(a) GSM8K Convergence (b) MATH Convergence (c) Training Loss Stability

Fig. 2: Convergence and loss stability comparison: No-Masking vs. VGGM

Without masking, accuracy peaks early (**9k**/**12k** on GSM8K/MATH) and then declines, as fitting erroneous reasoning steps pushes optimization away from globally optimal regions and traps training in *suboptimal local minima*. In contrast, VGGM continues to improve until **15k**/**19k** samples, showing more stable convergence and better generalization (Figure 2). Although non-masked SFT learns faster initially, VGGM delivers steadier progress and stronger systematic

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

reasoning. Loss curves reflect the same pattern: non-masked SFT exhibits frequent spikes from conflicting gradients, whereas VGGM shows a smooth downward trajectory, confirming stabilized optimization and consistent convergence.

3.4 Data-Efficient Comparison with DeepSeek-R1-Distill models

We follow the same evaluation protocol as in our main experiments on `GSM8K`, `MATH`, and `AMC'23`, using `Llama3.1-8B-Base` and `Qwen2.5-7B-Base` to enable direct comparison with publicly available DeepSeek-R1 distillation checkpoints. While DeepSeek-R1 distill models are trained on **800k samples generated by the DeepSeek-R1 teacher** [1], our **OpenR1-Math-220** dataset contains only **20k** examples (40× smaller) but is produced by the same DeepSeek-R1 teacher with identical reasoning style and verification structure [8]. This ensures matched data quality and distribution, so performance differences primarily reflect **data scale**, enabling a fair assessment of verification-guided data efficiency.

Table 2: Comparison of VGGM-trained models with DeepSeek-R1 distillation models.

Model / Setting	Student Model	GSM8K	MATH	AMC
Llama3.1-8B	Llama3.1-8B-Base	45.1	19.8	12.3
Llama3.1-8B + SFT (No Masking, 20k)	Llama3.1-8B-Base	74.1	57.9	45.2
Llama3.1-8B + SFT (VGGM, 20k)	Llama3.1-8B-Base	**79.1**	**62.9**	**50.4**
DeepSeek-R1-Distill-Llama (No Masking, 800k)	Llama3.1-8B-Base	83.4	68.4	57.2
Qwen2.5-7B + SFT (VGGM, 20k)	Qwen2.5-7B-Base	**83.1**	**71.3**	**58.7**
DeepSeek-R1-Distill-Qwen (No Masking, 800k)	Qwen2.5-Math-7B-Base	89.2	76.7	65.8

VGGM nearly matches DeepSeek-R1-Distill models despite using far less data, demonstrating its **data-efficient learning**. (1) Whereas DeepSeek-R1-Distill-Llama is trained on 800k samples (83.4/68.4/57.2), standard SFT with only 20k examples reaches just 74.1/57.9/45.2, highlighting its strong reliance on data scale. (2) With the same 20k samples, VGGM achieves 79.1/62.9/50.4, substantially closing this gap by using verified reasoning signals. (3) On Qwen2.5, VGGM (83.1/71.3/58.7) approaches DeepSeek-R1-Distill-Qwen (89.2/76.7/65.8), and its training curves continue rising, suggesting potential parity with large-scale distillation models even without massive teacher-generated reasoning traces.

3.5 Synergy Between VGGM and Test-time Budget Forcing

We analyze how VGGM interacts with **test-time budget forcing**, a reasoning-length control method from the s1 framework [2] that inserts "Wait!" tokens to encourage models to think longer. Under standard SFT, a single *Wait!* token slightly reduced GSM8K accuracy (−0.9%), yielded only small gains on MATH (+0.6%), and produced a minor improvement on AMC'23 (+0.2%). In contrast, VGGM delivered consistent gains of **+1.1%** on GSM8K, **+2.3%** on MATH, and **+1.8%** on AMC'23. These results show that budget forcing is effective only when the model has a **semantically grounded sense of verification**: without VGGM, the *Wait!* token adds superficial verbosity, whereas with VGGM it reliably triggers genuine *re-evaluation and correction*, indicating strong synergy between verification-aware training and budget-controlled inference.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Conclusion

Selective supervision is more effective than full-sequence imitation for training robust reasoning in small LLMs. By masking gradients on erroneous prefixes and learning only from verified corrections, VGGM stabilizes optimization, prevents error propagation, and preserves token-level entropy at key branching points, maintaining reasoning diversity. This yields a stronger initialization for RL and leads to larger gains from GRPO. Despite using far less data, VGGM approaches the performance of large-scale DeepSeek-R1 distillation, demonstrating data efficiency and viability as an alternative to massive teacher-generated distillation.

Acknowledgements

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) grant funded by the Defense Acquisition Program Administration (DAPA), under Grant KRIT-CT-23-021.

References

[1] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[2] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[3] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[4] Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. In *Workshop on Reasoning and Planning for Large Language Models*.

[5] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[6] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[7] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.

[8] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, 2025.

[9] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

[10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[12] Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*, 2025.