JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 12, pp. 51-60, December 2025
https://doi.org/10.9708/jksci.2025.30.12.051

SuperSmall-R1: A Lightweight Reinforcement Learning Model for Mathematical Reasoning

Jaegun Lee*, Janghoon Choi**

*M.S. Student, Graduate School of Data Science, Kyungpook National University, Daegu, Korea
**Associate Professor, Graduate School of Data Science, Kyungpook National University, Daegu, Korea

[Abstract]

In this study, we propose SuperSmall-R1, a lightweight reasoning model specialized for mathematical problem solving, built upon the compact text language model DeepSeek-R1-Distill-Qwen-1.5B. Unlike conventional approaches, the proposed model improves performance solely through reinforcement learning without supervised fine-tuning (SFT). Specifically, we introduce ZeroGRPO, a variant of the GRPO algorithm with the KL-divergence penalty removed, thereby increasing the freedom of policy exploration. In addition, instead of employing a complex penalty-based reward scheme, we adopt a simple yet effective reward function based only on format consistency and answer correctness. To address the issue of insufficient rewards when tackling difficult problems from the beginning, we further incorporate curriculum learning, where the problem difficulty is gradually increased. Experimental results on the Math-500 benchmark demonstrate that our approach outperforms not only the base model but also existing methods based on GRPO and Penalty GRPO, confirming that mathematical reasoning ability can be effectively enhanced even under resource-constrained environments.

▸ **Key words**: Reinforcement learning, Policy optimization, Curriculum learning, Reward design, Large Language Model

[요 약]

본 연구에서는 경량 텍스트 언어모델 DeepSeek-R1-Distill-Qwen-1.5B 를 기반으로 수학 문제 해결에 특화된 소형 추론모델 SuperSmall-R1을 제안한다. 제안된 모델은 별도의 지도 미세조정(SFT) 없이, 강화학습만으로 성능을 향상시킨다. 구체적으로, 기존의 GRPO 알고리즘에서 KL 발산 페널티를 제거한 ZeroGRPO를 제안하여 정책의 탐색 자유도를 높이고자 하였다. 또한 기존의 복잡한 penalty 기반 보상 체계 대신, 간결한 포맷 일치 및 정답 여부만으로 구성된 단순하면서도 효과적인 보상 함수를 적용하였다. 특히 초기부터 어려운 문제를 풀 경우 발생하는 보상 부족 문제를 해결하기 위해 난이도를 점진적으로 높이는 커리큘럼 학습(curriculum learning)을 도입하였다. Math-500 벤치마크를 활용한 실험 결과, 제안된 방법은 원본 베이스모델뿐 아니라 기존 GRPO, Penalty GRPO 알고리즘을 사용한 방법보다 높은 정확도를 보이며, 제한된 자원 환경에서도 수학 문제 해결 능력이 효과적으로 향상됨을 확인하였다.

▸ **주제어**: 강화학습, 정책 최적화, 커리큘럼 학습, 보상 디자인, 대형언어모델

• First Author: Jaegun Lee, Corresponding Author: Janghoon Choi
*Jaegun Lee (leejken530@knu.ac.kr), Graduate School of Data Science, Kyungpook National University
**Janghoon Choi (jhchoi09@knu.ac.kr), Graduate School of Data Science, Kyungpook National University
• Received: 2025. 10. 23, Revised: 2025. 11. 10, Accepted: 2025. 11. 24.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

대형 언어 모델(LLM)의 발전은 자연어 처리 분야에 혁명적인 변화를 가져왔으나, 수학적 추론과 같은 특정 영역에서는 여전히 근본적인 한계를 보이고 있다[1]. 특히 복잡한 기호 조작, 다단계 논리적 추론, 정확한 계산이 요구되는 수학 문제 해결에서 LLM은 인간 전문가 수준에 크게 미치지 못하고 있다[2]. 이러한 한계는 단순히 모델 크기를 늘리는 것만으로는 해결되지 않으며, 수학적 추론에 특화된 학습 방법론이 필요함을 시사한다. 또한 실제 응용 환경에서는 계산 자원의 제약으로 인해 대규모 모델 배포가 어려운 경우가 많다. 교육현장, 모바일 기기, 엣지 컴퓨팅 환경 등에서 활용 가능한 경량 수학 추론 모델의 개발은 실용적 측면에서 매우 중요하나, 이에 대한 연구는 상대적으로 부족하였다.

기존의 강화학습 기반 미세조정(Reinforcement Learning from Human Feedback, RLHF) 접근법들은 일반적인 대화 품질 향상에는 효과적이었으나, 수학 문제 해결이라는 특수한 도메인에 적용할 때 몇 가지 문제점을 노출한다. 첫째, PPO[3], DPO[4], GRPO[5] 등의 방법은 KL divergence 제약을 통해 베이스 모델로부터의 과도한 이탈을 방지하지만, 이는 동시에 모델이 새로운 문제 해결 전략을 탐색하는 것을 제한할 수 있다. 둘째, 수학 문제는 명확한 정답과 형식이 존재하는 특성상, 일반적인 텍스트 생성과는 다른 보상 체계가 필요하다. 셋째, 초기 학습 단계에서 어려운 문제에 대한 지속적인 실패는 모델 학습을 불안정하게 만들 수 있다. 본 연구에서는 이러한 문제들을 해결하기 위해 1.5B 파라미터 규모의 경량 모델 기반 수학 추론 시스템 **SuperSmall-R1**을 제안한다.

제안하는 접근법은 세 가지 핵심 기여점이 있다. 첫째, ZeroGRPO는 KL divergence 페널티를 제거해 모델의 탐색 공간을 넓히고 수학적 추론에 필요한 창의적 문제 해결 능력을 강화한다. 둘째, 복잡한 중간 평가 단계를 없애고 최종 답안의 정확성과 형식 준수만을 기준으로 평가하는 단순한 이진 보상 체계를 도입해 학습 안정성을 높인다. 마지막으로, 커리큘럼 학습 방식을 적용해 문제 난이도를 점진적으로 높이며 모델이 기초 수학 개념에서 시작해 점차 고급 문제 해결 능력으로 발전하도록 유도한다.

본 논문의 구성은 다음과 같다. II장에서는 관련 연구와 강화학습 배경 지식을 살펴보고 , III장에서는 제안하는 ZeroGRPO, 보상 함수, 커리큘럼 학습 전략을 상세히 설명한다. IV장에서는 실험을 통해 제안 방법의 성능을 검증하며 , V장에서 결론을 맺는다.

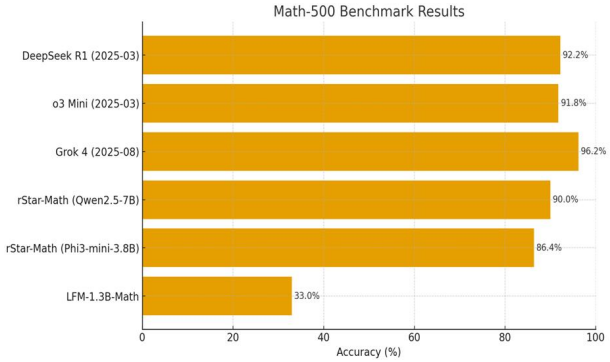

Fig. 1. Math-500 benchmark accuracy results.

II. Preliminaries

1. Related works

1.1 Mathematical Reasoning in Large Language Models

GPT-4[6], Claude[7], Gemini[8] 등의 최신 LLM 은 일반적인 언어 이해와 생성에서 인상적인 성능을 보이지만, 수학적 추론에서는 여러 한계를 드러낸다. Math-500 벤치마크에서는 최근 모델들이 90% 이상의 정확도를 기록하지만, 이는 상대적으로 제한된 범위의 문제에 국한된다. Hendrycks et al.[2]은 MATH 데이터셋을 통해 최고 성능의 모델들도 고등학교 수준의 수학 문제에서 50% 미만의 정확도를 보임을 입증하였다. 이는 LLM들이 언어 이해와 생성에서는 인상적인 성능을 내더라도, 복잡한 수학적 추론에서는 여전히 한계를 갖고 있음을 시사한다.

1.2 Math-specialized Language Models

수학적 추론 능력 향상을 위한 전문 모델들이 개발되어 왔다. Minerva[9]는 540B 파라미터 규모의 모델로 수학 및 과학 문제에 특화되었으며, MathCoder[10]는 코드 생성을 통한 수학 문제 해결을 시도하였다. 그러나 이러한 모델들은 대부분 수십억에서 수천억 개의 파라미터를 요구하여, 실제 배포 환경에서의 활용이 제한적이다. 또한 대규모 수학 데이터셋에 대한 지도 학습에 의존하여 데이터 수집과 라벨링에 상당한 비용이 소요된다.

1.3 Reinforcement Learning-based Fine-tuning

RLHF는 인간의 선호도를 반영하여 모델을 개선하는 효과적인 방법으로 입증되었다. PPO[3]는 정책 그래디언트 방법의 안정성을 개선하였고, DPO[4]는 보상 모델 없이 직접적인 선호도 최적화를 가능하게 하였다. GRPO[5]는 그룹 단위의 상대적 평가를 통해 학습 효율성을 높였다.

그러나 이러한 방법들은 주로 일반적인 대화 품질 향상에 초점을 맞추고 있으며, 수학 문제와 같이 명확한 정답이 존재하고 긴 추론 과정이 필요한 작업에는 최적화되어 있지 않다. 특히 KL divergence 제약은 베이스 모델의 행동을 과도하게 보존하려 하여, 새로운 문제 해결 전략의 발견을 저해할 수 있다.

1.4 Lightweight Model Research

최근 Phi[11], TinyLlama[12] 등의 연구는 작은 모델도 적절한 학습 방법을 통해 경쟁력 있는 성능을 달성할 수 있음을 보여주었다. 그러나 이러한 연구들은 주로 일반적인 언어 능력에 초점을 맞추고 있으며, 수학적 추론이라는 특수한 영역에 대한 경량 모델 연구는 여전히 부족하다.

1.5 Distinct Contributions of This Study

본 연구는 기존 연구와 다음과 같은 측면에서 차별화된다. 1.5B 파라미터의 초경량 모델을 사용해 자원 효율적이며 실용적인 배포가 가능하다. 또한 별도의 지도학습(SFT) 없이 강화학습만으로 성능을 향상시킴으로써 데이터 의존도를 낮췄다. 수학 문제의 구조적 특성을 반영한 보상 체계와 학습 전략을 설계해 도메인 특화 최적화를 달성했으며, KL 제약을 제거하여 모델이 보다 자유롭게 탐색하고 창의적인 문제 해결 경로를 모색할 수 있게 했다. 이로써 제한된 자원 환경에서도 효과적인 수학 추론이 가능한 실용적 대안을 제시한다.

2. Reinforcement Learning Preliminaries

강화학습(Reinforcement Learning, RL)은 에이전트가 환경과 상호작용하며 보상을 최대화하는 방향으로 정책(policy)을 학습하는 기계학습 패러다임이다. RL에서는 보통 상태(state) $s \in S$, 행동(action) $a \in A$, 보상 함수(reward function) $r(s,a)$, 그리고 정책 $\pi_\theta(a|s)$로 구성된다. 목표는 누적 보상(expected return)을 최대화하는 최적 정책 π^*를 찾는 것이다. 이를 수식으로 표현하면 다음과 같다.

$$J(\theta) = \mathrm{E}_{a \sim \pi_\theta(\cdot|s)}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)] \quad (1)$$

여기서 $\gamma \in (0,1]$는 할인율(discount factor)을 의미한다.

2.1 GRPO (Group Relative Policy Optimization)

GRPO(Group Relative Policy Optimization)는 최근 제안된 강화학습 알고리즘으로, 기존 PPO(Proximal Policy Optimization)의 변형이다. PPO는 정책 업데이트의 안정성을 위해 KL-divergence 기반의 clipping을 사용하지만, GRPO는 여러 샘플(group) 간의 상대적 성능 비교를 통해 효율적으로 학습한다.

구체적으로, 주어진 입력 x 에 대해 정책 π_θ에서 G개의 응답 $\{y_1, \cdots, y_G\}$를 샘플링한다. 이후 각 응답에 대해 보상 r_i를 계산하고, 상대적 순위를 기반으로 advantage A_i를 정의한다. 정책의 목적함수는 다음과 같이 주어진다.

$$L_{GRPO}(\theta) = \mathrm{E}_{x \sim D, y_i \sim \pi_\theta}$$
$$[\frac{1}{G}\sum_{i=1}^{G} \log \pi_\theta(y_i|x) \cdot A_i - \beta KL(\pi_\theta(\cdot|x) \| \pi_{ref}(\cdot|x))] \quad (2)$$

여기서 π_{ref}는 기준 정책(reference model, β는 KL-divergence 항의 중요도를 조절하는 계수이다.

2.2 KL Divergence

KL 발산(Kullback-Leibler divergence)은 두 확률 분포 간의 차이를 측정하는 지표이다. RLHF(RL from Human Feedback)에서 자주 사용되며, 새로운 정책이 기존 reference policy에서 과도하게 벗어나지 않도록 제약 역할을 한다.

$$D_{KL}(\pi_\theta \| \pi_{ref}) = \sum_y \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \quad (3)$$

그러나 수학 문제와 같이 새로운 추론 전략의 탐색이 필요한 경우, KL 제약은 모델의 자유도를 지나치게 억제하여 성능 향상을 방해할 수 있다.

2.3 Reward Functions in Mathematical Reasoning

일반적인 RLHF에서는 사람이 평가한 선호도나 대화 품질을 보상으로 사용한다. 하지만 수학 문제는 정답 여부가 명확히 정의되므로, 불필요하게 복잡한 보상 체계 대신 단순하고 직접적인 보상을 정의할 수 있다. 기존 연구에서는 보통 다음과 같은 보상을 사용한다.

$$R = \begin{cases} 1, & \text{if } y = \hat{y} \\ -1, & \text{otherwise} \end{cases} \quad (4)$$

모델의 응답이 정답일 경우 +1, 오답일 경우 −1 로 보상과 처벌을 확실하게 하였다. 하지만 초소형 베이스모델의 경우 초기 정답률이 매우 낮기 때문에 오답에 대한 처벌이 누적되면서 평균 보상이 0 이하로 급격히 떨어지는 문제가 발생한다.

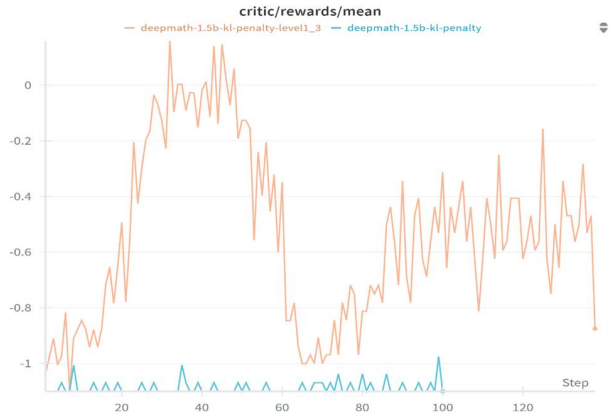

Fig. 2. Critic reward mean distribution.

Fig 2의 예시를 통해 살펴보면, 음수의 penalty로 인해서 reward 가 0 이상이 된 경우는 극히 적었으며, 학습 과정에서 다시 하락하는 현상이 있다. 이에 따라 학습이 점차 정체되거나 성능이 오히려 저하되는 현상이 나타난다. 따라서 단순한 이진 보상 체계는 자원 제약 환경에서의 초경량 모델 성능 향상에는 적합하지 않음을 확인하였다.

III. The Proposed Scheme

1. Model Architecture

본 연구에서는 DeepSeek-R1-Distill-Qwen-1.5B를 baseline 모델로 채택하였다. 이 모델은 1.5B 파라미터를 가진 경량 텍스트 전용 언어 모델로, 멀티모달 기능은 포함하지 않는다. 이러한 소규모 모델을 선택한 이유는, 해당 모델의 제한된 컴퓨팅 자원 환경에서의 효율적인 학습 및 추론 성능을 최대한 활용하기 위함이다, 또 이러한 특성은 추후 실제 응용 환경에서의 배포 가능성을 높일 수 있어 이점을 가진다.

2. Training Pipeline

2.1 Curriculum Learning

훈련에는 DeepMath-103K[13] 데이터셋을 사용하였다. 해당 데이터셋은 103,000개의 고품질 수학 문제로 구성되어 있으며, 모든 문제는 검증 가능한 최종 답안을 포함한다.본 연구에서는 문제를 난이도에 따라 Level 1-3 (기초), Level 4 (중급), Level 5 (고급)으로 분류하였다. 학습은 낮은 난이도의 문제에서 시작하여 모델 성능이 향상됨에 따라 점진적으로 더 어려운 문제를 도입하는 방식으로 진행하였다. 이는 1.5B 규모의 소형 모델이 단계적으로 수학적 추론 능력을 습득할 수 있도록 돕는다.또한 Table1은 각 난이도별 대표 예시 문제를 나타내고 있다.

2.2 ZeroGRPO(Zero KL-divergence GRPO)

1.5B 규모의 소형 모델에서는 KL 제약이 탐색 공간 축소, 학습 속도 저하, 국소 최적해(local optimum) 문제를 유발하여 policy의 학습이 올바르게 이루어지지 않았다. 따라서 본 연구에서는 (2)의 수식에서 KL 항을 제거한 ZeroGRPO를 제안하며, 목적함수 L는 다음과 같이 단순화된다.

$$L_{Zero}(\theta) = \mathrm{E}_{x \sim D, y \sim \pi_\theta}[\frac{1}{G}\sum_{i=1}^{G}\log\pi_\theta(y_i|x) \cdot A_i] \quad (5)$$

이를 통해 모델이 기존의 baseline 모델의 정책에 구속되지 않고 더 제약 없이 최적 정책을 탐색할 수 있도록 하였다. 특히 수학 문제 해결과 같이 정확한 답을 요구하는 작업에서는 baseline 모델의 일반적인 언어 생성 패턴을 벗어나 문제 해결에 특화된 새로운 행동을 기반으로 학습하는 것이 중요하며, 본 연구에서 제안하는 ZeroGRPO 가 이를 효과적으로 지원할 수 있도록 설계하였다. Fig. 4에서는 기존 방법론과 제안하는 방법론을 비교하여 시각화하였다.

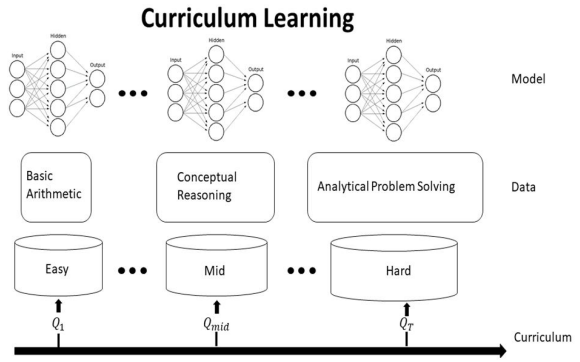

Fig. 3. Curriculum Learning Scheme. The model learns progressively from easy to hard datasets. Example: easy → mid → hard

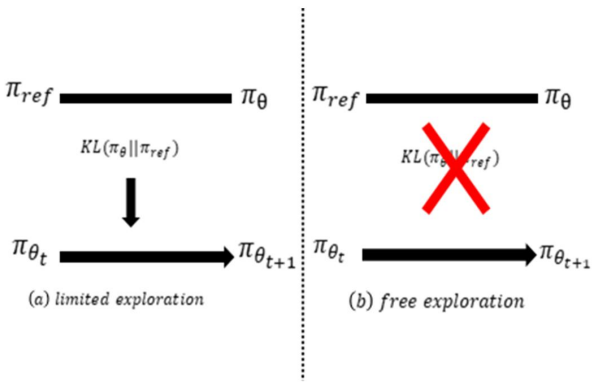

Fig. 4. Exploration Difference.Comparison between standard GRPO and the proposed ZeroGRPO.The KL-divergence constraint in GRPO restricts exploration (a),while ZeroGRPO removes the KL term, allowing unconstrained policy updates (b).

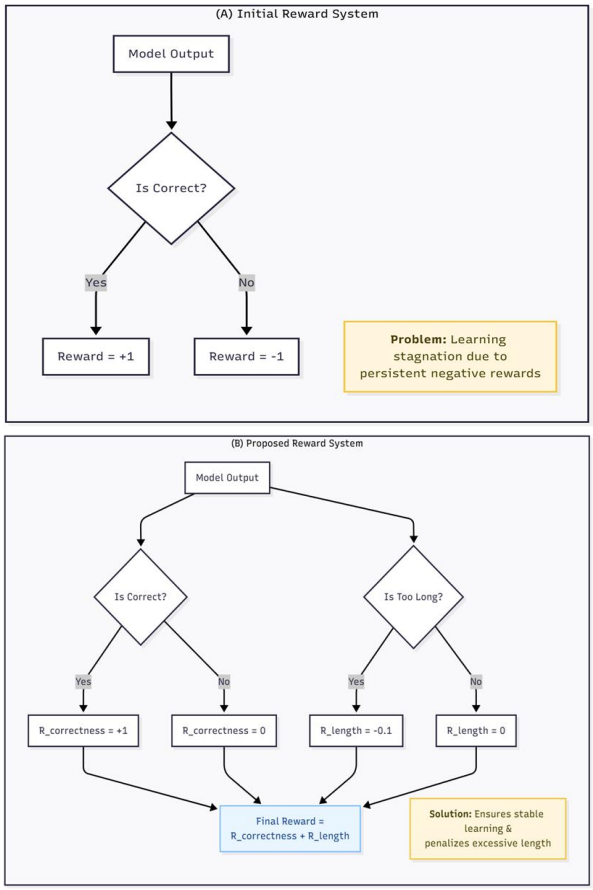

Fig. 5. Reward design comparison between the initial and proposed reward mechanisms. The proposed reward design can stabilize policy update and encourage further exploration.

2.3 Reward Design

초기 실험에서 기존의 strict bipolar 보상 체계(정답: +1, 오답: −1)를 적용했을 때, 1.5B 모델은 학습 초기에 오답이 빈번하게 발생하면서 지속적인 음의 보상을 누적 받아 학습이 정체되는 문제가 나타났다. 이를 완화하기 위해 본 연구에서는 비음성(non-negative) 기반의 단순화된 보상 함수를 설계하였다. 정답 여부에 따라 +1 또는 0을 부여하는 정답 보상은 모델이 올바른 출력을 생성할 때 명확한 강화 신호를 받게 하며, 오답일 경우 −1의 처벌 대신 0으로 처리하여 학습 초반의 불안정한 탐색 구간에서도 손실 폭을 줄인다. 이때 정답 여부에 대한 보상은 다음과 같이 정의된다.

$$R_{correctness} = \begin{cases} 1, & \text{if } y = \hat{y} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

또한 모델이 불필요하게 긴 출력을 생성하는 문제를 방지하기 위해 길이 패널티를 추가하였다. 전체 최대 허용 길이$L_{\max}$의 95%를 초과하는 경우, 아래와 같이 소규모의 패널티(−0.1)를 부여하여 장문 생성을 억제하였다.

$$R_{\text{length}} = \begin{cases} -0.1, \text{if } \text{len(y)} > 0.95^{*}\text{L}_{\max} \\ 0, \text{ otherwise} \end{cases} \quad (7)$$

최종 보상은 $R = R_{correctness} + R_{\text{length}}$로 구성되며, 정답을 맞춘 경우에는 +1의 보상이 그대로 유지되지만 불필요하게 긴 출력에는 미세한 감점을 주어 정확성과 간결성의 균형을 학습하도록 유도한다. 이와 같은 단순화된 비음성 보상 구조는 초기 학습 안정성을 크게 향상시키며, 특히 소형 모델의 경우 음의 보상 누적 문제를 근본적으로 방지해 빠르고 안정적인 수렴을 가능하게 한다. 이러한 보상 체계는 모델이 초기 학습 단계에서도 안정적으로 학습할 수 있도록 하며 특히 소형 모델에서 효과적인 것으로 나타났다. 본 논문에서 제안하는 보상 구성은 Fig. 5에서 시각화하여 비교하였다.

IV. Experiments

4.1 Experimental Setup

4.1.1 Benchmark Datasets

모델 평가를 위해 Math-500 벤치마크를 구성하였다. 이는 기존 MATH 데이터셋에서 난이도별로 문제를 선별한 것으로, 5 개의 난이도 레벨에 각 100 문제씩 총 500 문제로 구성되어 있으며 난이도별 특성과 예시문제를 Table 1과 Table 2 에서 정리하였다.

Table 1. Features by Difficulty Level

Level	Features
1-3	Focused on basic calculations, solvable with short reasoning, efficient learning and generalization effects
4-5	Multi-level logic, complex concept problems, and requires long Chain-of-Thought

Table 2. Examples by difficulty level used in the curriculum learning of this study

Level	Example
Level 1	Calculate the value of the expression: $6\times7+4\times7+6\times3+4\times3$
Level 2	Determine whether the logical statement $\forall x\, P(x) \mapsto \exists x\, P(x)$ is true
Level 3	If two distinct numbers are selected at random from the first seven prime numbers, what is the probability that their sum is an even number?
Level 4	What is the product of all real numbers that are doubled when added to their reciprocals?
Level 5	Determine the points in the complex plane where the function $g(z) = \|z\|^4$ is differentiable and analytic

Table 3. Model Configurations and Training Methods

Model	#Params	Training Method
DeepSeek-R1-Distill-Qwen-1.5B	1.5B	Baseline Model (No training)
Math1.5B-level1-3-kl-penalty	1.5B	Standard GRPO (KL reward penalty included)
Math1.5B-level1-3-kl-nopenalty	1.5B	Standard GRPO (KL loss included, reward penalty removed)
Math1.5B-level1-3-no-kl (Ours)	1.5B	ZeroGRPO + Proposed reward function

Table 4. Experiment Settings

Hyperparameter	Value
Leaning rate	5e-6
Batch size	16
Temperature	0.6
Max. prompt length	1024
Max. response length	4096
Total training steps	100 steps per epoch

Table 5. Training Runtime

Run name	Runtime
DeepSeek-R1- Distill-Qwen-1.5B	0초(0분)
Math1.5B-level1-3-kl-penalty	3502초(58.4분)
Math1.5B-level1-3-kl-nopenalty	3480초(58.0분)
Math1.5B-level1-3-no-kl	5431초(90.5분)
Math1.5B-level4-no-kl	7812초(130.2분)
Math1.5B-level5-no-kl(Ours)	8556초(142.6분)
Math1.5B-level6-no-kl	8280초(138.0분)

Math-500 은 산술(Arithmetic), 대수(Algebra), 기하(Geometry), 확률/통계(Probability/Statistics) 등 다양한 수학 영역을 포괄하며, 각 문제는 문제 텍스트와 정답으로 구성된다.

4.1.2 Baseline Models

제안 방법의 효과를 검증하기 위해 Table 2와 같은 baseline 모델들과 비교 실험을 수행하였다. Baseline 모델을 기반으로, 모델별로 기존 GRPO와 같은 보상 및 KL-divergence 구성, 보상 체계 개선, KL-divergence 제거 등의 차이를 두어 본 방법론의 효과를 검증하였다.

4.1.3 Training Environment and Hyperparameters

모든 실험은 Ubuntu 20.04 기반의 소프트웨어 환경과 NVIDIA RTX A6000 48GB x4를 장착한 하드웨어 시스템에서 구동되었으며, 실험을 위한 하이퍼파라미터 설정은 Table 4.의 설정에 따라 실험을 수행하였으며, 각 모델의 학습시간은 Table 5에 정리되어 있다.

4.2 Main Experimental Results

4.2.1 Performance Comparison across GRPO Variants

본 연구에서는 제안하는 ZeroGRPO 의 효과를 검증하기 위해 다양한 GRPO 변형을 비교 실험하였다. 모든 모델은 동일한 커리큘럼(Level 1-3)으로 학습되었으며, KL 페널티 유무가 미치는 영향을 중점적으로 분석하였다.

KL 페널티 제거의 효과로, KL 제약을 완전히 없앤 ZeroGRPO는 기존 GRPO 대비 2.5배 이상의 성능 향상을 보였다. 반면, 부분적으로만 KL 항을 제거한 kl-nopenalty 설정은 소폭의 개선에 그쳤으며, 완전한 KL 제거가 훨씬 더 효과적임을 확인하였다.

이러한 결과는 KL 제약이 모델의 탐색 범위를 과도하게 제한하여 새로운 문제 해결 경로를 학습하는 데 방해가 됨을 의미한다. ZeroGRPO는 이러한 제약을 제거함으로써 보다 다양한 추론 패턴을 시도하고, 복잡한 수학적 구조를 스스로 발견할 수 있었다. 특히 고난도 문제(Level 5) 구간에서 창의적 접근 방식이 두드러지게 향상되어, 모델이 단순한 패턴 복제에서 벗어나 일반화된 수학적 사고를 학습했음을 보여준다.

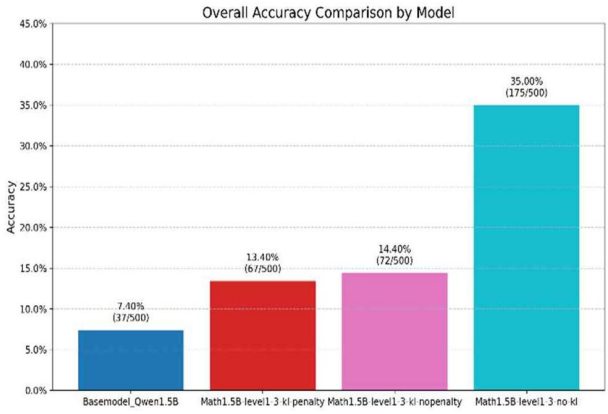

Fig. 6. Overall Performance Comparison of GRPO Variants

Table 6. Model Performance Comparison

Model	Accuracy	# of correct answers	Performance improvement
Basemodel_Qwen1.5B	7.4%	37/500	
Math1.5B-level1-3-kl- penalty	13.4%	67/500	+81.1%
Math1.5B-level1-3-kl- nopenalty	14.4%	72/500	+94.6%
Math1.5B-level1-3-no-kl (Ours)	**35.0%**	**175/500**	**+373.0%**

4.2.2 Difficulty-level Performance Analysis

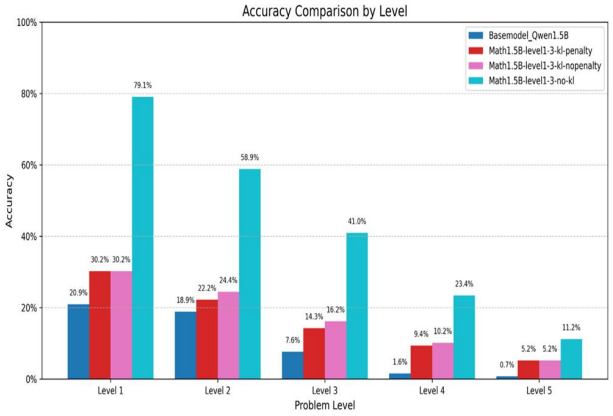

Fig. 7. Accuracy Comparison by Level

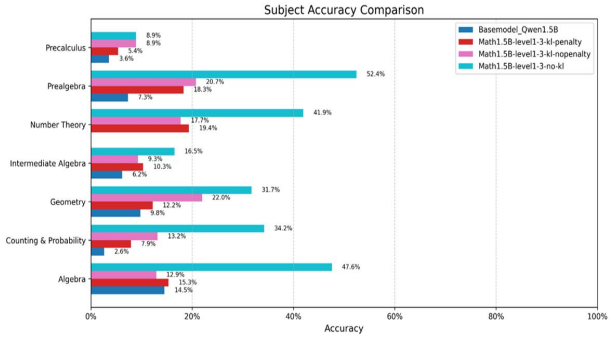

Fig. 8. Subject Accuracy Comparison

난이도별 문제 풀이 성능의 비교를 위한 실험 또한 수행하였다. Fig. 7은 난이도 수준별 모델의 성능을 보여준다. 해당 실험에서, 난이도 1-2 수준에서는 ZeroGRPO가 뚜렷한 우위를 보였으며(난이도 1: 79.1%, 난이도 2: 58.9%), 난이도 3에서도 ZeroGRPO(41.0%)와 기존 GRPO(약 15%) 간에 큰 성능 격차가 유지되었다. 난이도 4-5에서는 모든 모델의 성능이 전반적으로 낮았으나, ZeroGRPO는 상대적 우위를 지속적으로 보였다. 이러한 결과는 KL 제약을 제거함으로써, 특히 학습 데이터에 포함된 난이도 구간(난이도 1-3)에서 모델의 표현력이 크게 향상되었다.

4.2.3 Domain-specific Performance Analysis in Mathematics

수학 문제 영역별 문제 풀이 성능을 위한 비교 실험을 수행하였다. Fig. 8와 같이 주제별 성능을 분석한 결과, ZeroGRPO의 우수성이 모든 수학 영역에서 일관되게 확인되었다. 영역별 성능 비교에서 ZeroGRPO는 기존 GRPO 대비 뚜렷한 향상을 보였다. 구체적으로, Prealgebra 영역에서 52.4% 대 약 19%로 2.7배 향상되었고, Algebra에서는 47.6% 대 약 14%로 3.4배 향상되었다. Number Theory에서는 41.9% 대 약 18%로 2.3배, Counting & Probability에서는 34.2% 대 약 10%로 3.4배, Geometry에서는 31.7% 대 약 16%로 2.0배의 성능 향상을 각각 보였다. 이러한 결과는 ZeroGRPO가 특정 주제에 국한되지 않고, 다양한 수학 영역 전반에서 안정적이고 일관된 성능 개선을 이끌어내었다.

4.3 Analysis and Discussion

4.3.1 Negative Impact of KL Constraints

실험 결과는 KL divergence 제약 항이 소형 모델의 수학적 추론 능력 향상을 크게 제한한다는 점을 보여준다. 기존 GRPO의 KL 페널티는 다음과 같은 부정적 영향을 미쳤다. 첫째, 모델이 참조 정책(reference policy)으로부터 크게 벗어나지 못하도록 제한하였다. 둘째, 복잡한 수학적 추론에 필요한 새로운 문제 해결 전략의 학습을 방해하였다. 셋째, 특히 다단계 추론이 요구되는 문제에서 성능 저하를 초래하였다. 이러한 결과는 KL 제약이 소형 모델의 표현력과 적응력을 저해하는 핵심 요인이다.

본 연구의 working hypothesis는, 소형 모델에서는 KL 항이 탐색(exploration)을 과도하게 억제하여 정책 개선에 필요한 분포 이동(distribution shift)을 수행하지 못하게 만든다는 점이다. 기존 GRPO의 업데이트는 다음과 같은 Regularized Policy Optimization 문제로 해석된다.

$$\theta_{t+1} = \arg\max_{\theta} \left[\mathrm{E}_{x \sim \pi_{\theta_t}} [r(x)] - \lambda \mathrm{KL}(\pi_\theta \| \pi_{\theta_t}) \right]$$

여기서 $r(x)$ 는 reward model이 계산한 보상,λ 는 KL penalty 계수이다. 위 식은 아래와 같은 제약 최적화 문제와 동일하다.

$$\max_{\theta} \mathrm{E}[r(x)] \text{ s.t. } \mathrm{KL}(\pi_\theta \| \pi_{\theta_t}) \leq \epsilon$$

즉, KL 항은 정책이 참조 정책(reference policy)에서 벗어날 수 있는 허용 이동량 자체를 제한한다. 대형 모델은 파라미터 공간이 넓어 이 제약 내에서도 reasoning trajectory를 새로 탐색할 수 있지만, 소형 모델은 capacity가 낮기 때문에 KL 제약이 곧바로 "탐색 억제"로 작동한다. 실제로 정책 경사는 아래와 같이 KL 항에 의해 축소된다.

$$\nabla_\theta J_{KL} = \mathrm{E}\left[\nabla_\theta \log \pi_\theta(x) r(x) - \lambda \nabla_\theta \mathrm{KL}(\pi_\theta \| \pi_{\theta_t}) \right]$$

따라서 보상 신호가 충분히 크더라도

$$\| \theta_{t+1} - \theta_t \| \propto \frac{1}{\lambda}$$

와 같이 업데이트 크기가 강하게 제한되며,소형 모델이 새로운 reasoning pattern을 발견하는 데 필요한 분포 이동

량이 확보되지 않는다.이 점에서 KL 제거는 작은 모델에게 더 넓은 탐색 공간을 허용하고,본 연구의 실험 결과는 이러한 working hypothesis와 일관된다.

다만 KL 항을 제거할 경우 $\Delta\mathrm{KL} = \mathrm{KL}(\pi_{\theta_{t+1}} \| \pi_{\theta_t})$이 과도하게 증가하여 정책의 불안정성을 유발할 가능성이 있다. 따라서 후속 연구에서는 KL 제거를 기본으로 하되,adaptive temperature, trust-region 기반 soft-regularization 또는 주기적 re-centering 등의 안정화 기법을 결합하는 방향이 필요하다.

4.3.2 Superiority of the Proposed Method

ZeroGRPO의 373% 성능 향상은 세 가지 주요 요인에 기인한다. 첫째, KL 제약을 제거함으로써 모델의 탐색 자유도가 증가하여 보다 다양한 문제 해결 전략을 탐색할 수 있었다. 둘째, 수학 문제의 특성에 맞추어 설계된 명확한 보상 신호 최적화가 학습 효율을 극대화하였다. 셋째, 난이도 1-3에 집중한 커리큘럼 학습과 ZeroGRPO의 결합이 강력한 시너지 효과를 발휘하였다. 이러한 요인들이 복합적으로 작용하여 ZeroGRPO는 기존 GRPO 대비 압도적인 성능 향상을 달성하였다.

4.4 Effectiveness of Curriculum Learning

본 연구에서는 난이도별 순차 학습(Level 1-3 → Level 4 → Level 5)이 모델의 학습 효율과 추론 능력에 미치는 영향을 심층적으로 분석하였다. 낮은 난이도의 문제를 통해 기본적인 수학 연산과 논리적 사고 구조를 먼저 학습한 후, 점차 복잡도가 높은 문제로 확장함으로써 모델이 자연스럽게 추론 단계를 내재화하도록 유도하였다.

4.4.1 Curriculum Expansion Experiments

Fig 9과 Fig 10는 점진적으로 높은 난이도를 포함하여 학습한 모델들의 성능을 보여준다.

구체적으로, Math1.5B-level1-3-no-kl은 난이도 1-3만으로 학습하여 35.0%의 정확도를 달성하였다. Math1.5B-level4-no-kl은 학습 범위를 난이도 4까지 확장하면서 46.6%의 정확도를 기록하였으며, Math1.5B-level5-no-kl은 난이도 5까지 확장 학습하였으나 정확도는 44.2%로 소폭 감소하였다. 마지막으로, Math1.5B-level5_5-no-kl은 난이도 5에 집중적으로 추가 학습을 진행하여 44.4%의 정확도를 보였다. 이러한 결과는 학습 난이도의 확장이 일정 수준까지는 성능 향상에 기여하지만, 난이도 5 이상의 학습에서는 추가적인 성능 개선이 제한적이다.

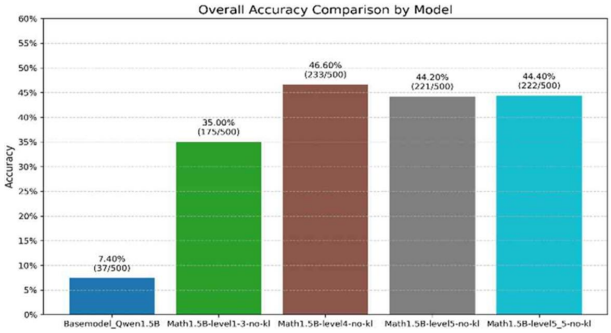

Fig. 9. Overall Accuracy Comparison by Model

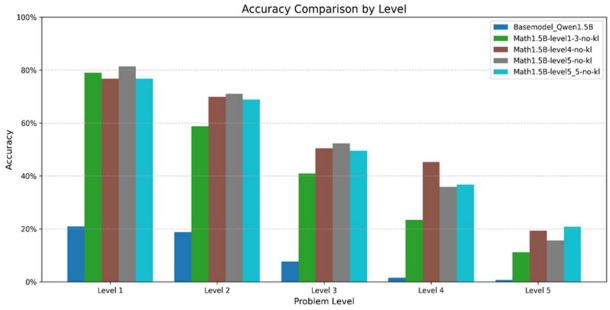

Fig. 10. Accuracy Comparison by Level

4.4.2 Key Findings

본 연구의 실험 결과를 통해 도출한 특성은 아래와 같다.

(1) 최적 커리큘럼 범위는 Level 4까지였다.

모델은 학습 범위를 Level 4로 확장했을 때 최고 성능(46.6%)을 달성하였다. 이는 Level 1-3까지만 학습한 경우보다 약 33% 높은 결과로, 난이도 4 수준의 문제를 추가 학습하는 것이 일반화 성능 향상에 실질적인 기여를 한다는 점을 보여준다.

(2) Level 5 확장 시 Catastrophic Forgetting이 발생했다.

Level 5까지 학습을 확장하면 전체 정확도가 오히려 하락하였다. 특히 Fig. 10에서 확인되듯, 난이도 1-2의 쉬운 문제에서도 정확도가 감소하였다. 이는 새로운 고난도 문제 학습이 기존에 습득한 저난이도 패턴을 덮어쓰는 catastrophic forgetting 현상으로 해석된다.

(3) 모델 용량 한계로 인한 성능 포화가 확인되었다.

1.5B 파라미터 모델은 Level 5 이상의 문제를 학습해도 성능 개선이 거의 없었다. 이는 소형 모델의 표현 용량이 이미 포화 상태에 도달했음을 의미하며, 따라서 Level 6 이상의 확장은 비효율적이라 판단된다.

(4) 난이도별 성능 변화는 특정 구간에서만 향상되었다.

난이도 1-3 에서는 모든 커리큘럼 설정에서 안정적인 성능을 유지했다. 반면 난이도 4 구간에서는 Level 4 까지 학습한 모델만이 성능이 두 배 이상 향상되었으며(23.4% → 46%), 난이도 5에서는 모든 모델이 20% 미만의 낮은

성능을 보였다. 이는 커리큘럼 확장의 효과가 난이도별로 국한된다.

종합적으로, 소형 모델의 커리큘럼 러닝에서는 적정 난이도 범위의 선택이 결정적이었다. 쉬운 문제만 학습하면 일반화가 제한되고, 지나치게 어려운 문제를 포함하면 오히려 성능이 저하될 수 있다. 이러한 결과는 소형 모델의 커리큘럼 러닝에서 적절한 난이도 범위 선택이 중요함을 보여준다. 너무 쉬운 문제만으로는 일반화가 제한되지만, 너무 어려운 문제까지 포함하면 오히려 전반적인 성능이 저하될 수 있다.

4.5 Limitation

본 연구는 단일 러닝 결과만을 기반으로 하여 모델 초기화(seed) 변화에 따른 성능 분산 분석을 수행하지 않았다. 또한 정확도에 대한 신뢰 구간이나 분산지표를 제시하지 못해 결과의 통계적 안정성을 확인하기 어렵다. 이러한 점은 후속 연구에서 반복 실험과 통계적 검증을 통해 보완되어야 한다.

V. Conclusion

본 연구에서는 1.5B 파라미터의 경량 모델만으로도 효과적인 수학 추론이 가능함을 입증하는 SuperSmall- R1을 제안한다. 제안된 방법은 세 가지 핵심 혁신을 통해 기존 접근법의 한계를 극복하였다.

첫째, ZeroGRPO 를 통해 KL divergence 제약을 완전히 제거함으로써 소형 모델이 갖는 제한된 표현 능력내에서도 최대한의 탐색 자유도를 확보하였다. 실험결과, KL 제약이 있는 기존 GRPO 대비 2.5 배 이상의 성능 향상을 달성하였으며, 이는 수학적 추론과 같은 특수 도메인에서는 베이스 모델로부터의 이탈을 허용하는 것이 오히려 유리하다.

둘째, 복잡한 중간 단계 평가 대신 최종 답안의 정확성과 형식 준수만을 평가하는 단순화된 보상 체계를 도입하여 학습의 안정성을 크게 향상시켰다. 이는 특히 초기 학습 단계에서 지속적인 실패로 인한 학습 불안정성을 방지하는 데 효과적이었다.

셋째, 난이도별 커리큘럼 학습을 통해 모델이 기초적인 수학 개념부터 체계적으로 학습할 수 있도록 하였다. 실험결과 Level 4 까지의 점진적 확장이 최적의 성능을 보였으며, 이는 소형 모델의 제한된 용량을 고려한 적절한 난이도 범위 선택의 중요성을 보여준다. 결론적으로, 본 연구는 제한된 자원 환경에서도 효과적인 수학 추론이 가능한 실용적 솔루션을 제시하였다. 특히 KL 제약 제거, 단순화된 보상 체계, 커리큘럼 학습의 조합이 소형 모델의 성능을 극대화하는데 핵심적임을 입증하였다. 이는 향후 경량 AI 시스템 개발에 중요한 시사점을 제공하며, 교육 기술, 개인화 학습 도구, 모바일 AI 애플리케이션 등 다양한 분야에서의 활용이 기대된다.

ACKNOWLEDGEMENT

This research was supported by the Republic of Korea Government (Ministry of Science and ICT) through the research fund of the National Research Foundation of Korea (NRF) and Information & Communications Technology Planning & Evaluation (IITP) under grants NRF-2021R1C1C2095450, RS-2023- 00242528, RS-2024-00437756.

REFERENCES

[1] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models,” in Proc. ICLR, 2025.

[2] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, et al., “Measuring Mathematical Problem Solving with the MATH Dataset,” in Proc. NeurIPS, Track on Datasets and Benchmarks, 2021.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov,“Proximal Policy Optimization Algorithms,”arXiv preprint arXiv:1707.06347, 2017 doi: 10.48550/arXiv.1707.06347

[4] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn,“Direct Preference Optimization: Your Language Model is Secretly a Reward Model,” in Proc. NeurIPS, 2023, Oral Presentation.

[5] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, et al., “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models,” arXiv preprint arXiv:2402.03300, 2024. doi: 10.48550/arXiv.2402.03300

[6] OpenAI,“GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023. doi: 10.48550/arXiv.2303.08774

[7] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, et al.,“Constitutional AI: Harmlessness from AI Feedback,” arXiv preprint arXiv:2212.08073, 2022. doi: 10.48550/arXiv.2212.08073

[8] R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, et al., “Gemini: A Family of Highly Capable Multimodal Models,”arXiv preprint arXiv:2312.11805, 2023. doi: 10.48550/arXiv.2312.11805

[9] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, et al., “Solving Quantitative Reasoning Problems with Language Models,” in Proc. NeurIPS, 2022.

[10] K. Wang, H. Ren, A. Zhou, Z. Lu, S. Luo, et al., “MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning,”in Proc. ICLR 2024, 2024.

[11] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, et al.,“Textbooks Are All You Need,” submitted to Proc. ICLR, 2024.

[12] Zhang, Peiyuan; Zeng, Guangtao; Wang, Tianduo; and Wei Lu. 2024. “TinyLlama: An Open-Source Small Language Model.” arXiv preprint arXiv:2401.02385. doi:10.48550/arXiv.2401.02385

[13] He et al.,“DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning,”submitted to Proc. ICLR, 2026

Author

Jaegun Lee received his B.S. degree from Gyeongguk National University (GNU), Andong, Korea, in 2022, and subsequently completed a B.S. degree in Statistics from Korea National Open University (KNOU) in 2024.

He is currently pursuing his M.S. degree in Data Science at the Graduate School of Data Science, Kyungpook National University (KNU), Korea. His research interests include reinforcement-learning-based computer vision, diffusion-RL integration for text-to-image generation, and semantic segmentation.

Janghoon Choi received the B.S. degree in electrical and computer engineering, and Ph.D. degree in electrical engineering and computer science from Seoul National University, Korea, in 2013 and 2021, respectively. He joined the faculty of the Graduate School of Data Science at Kyungpook National University, Daegu, Korea, in Sep. 2022. He is currently an Associate Professor in the Graduate School of Data Science at Kyungpook National University. He is interested in computer vision problems including visual tracking, video understanding and image restoration.