2025 RESEARCH STUDY

VL-RETHINKER:

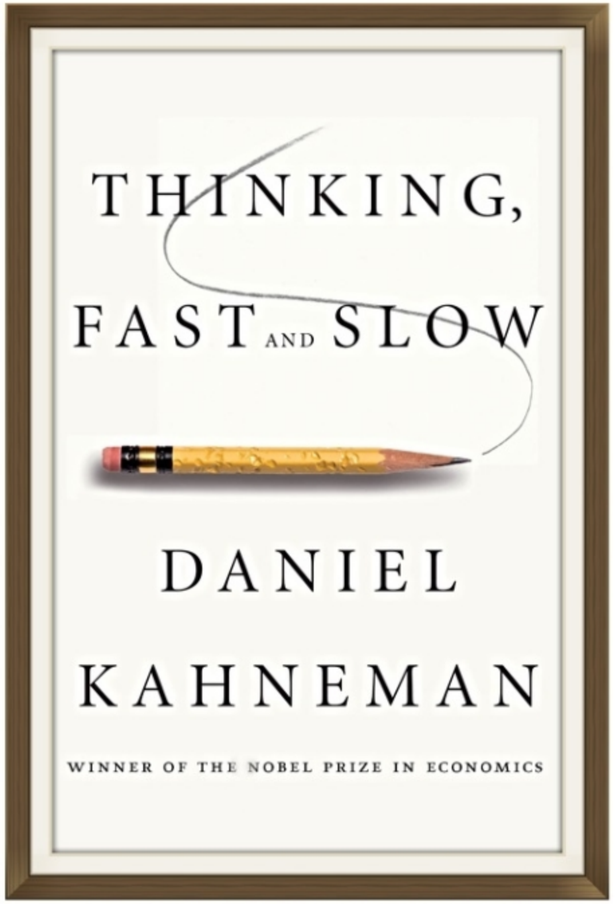
INCENTIVIZING SELF-REFLECTION OF VISION-LANGUAGE MODELS WITH REINFORCEMENT LEARNING

2025.08.06

PRESENTER : SKN-13 강지윤

생각에 관한 생각: 인간의 사고방식의 근본적 두가지 모드

FAST-THINKING VS SLOW-THINKING



FAST-THINKING
직관(무의식)

System 1 "Fast"

DEFINING CHARACTERISTICS
Unconscious
Effortless
Automatic

WITHOUT Self-Awareness or Control

"What You See Is All There Is"

ROLE
Assess the Situation
Deliver Updates

GPT-4o

SLOW-THINKING
논리적 추론(의식)

System 2 "Slow"

DEFINING CHARACTERISTICS
Deliberate and Conscious
Effortful
Controlled Mental Process

WITH Self-Awareness or Control

Logical and Skeptical

ROLE
Seeks New Information
Makes Decisions

GPT-o1
DeepSeek-R1
Kimi-1.5
Gemini -Thinking

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

PROBLEM STATE & OBSERVATIONS

SLOW-THINKING 멀티 모달과 수학 과제의 RL 학습 차이

(fast-thinking은 멀티모달과 text가 비슷함)

Math Focus Tasks	**강화학습 훈련 시, SLOW THINKING 유도 → 긴 REASONING TRACE 발생** CHAIN-OF-THOUGHT GENERATION, STEP-BY-STEP 추론
Multimodal Tasks	**비슷한 방식의 강화학습 훈련에도 SLOW THINKING이 거의 나타나지 않음** VLM은 즉답 경향, CHAIN-OF-THOUGHT 없이 바로 답

Problem State	Why don't multimodal models naturally develop longer chains of thought or reflective behaviors during RL training?

KEY QUESTION

How can we effectively incentivize
multimodal slow-thinking capabilities in Vision-Language Models?

멀티모달도 사람처럼 성찰을 하게 해보자...

How can we effectively incentivize
multimodal slow-thinking capabilities in Vision-Language Models?

REINFORCE LEARNING 을 DEVELOP 시키는 방향으로 !

SUMMARY

MAIN CONTRIBUTIONS

[CONTRIBUTIONS 1]

Direct RL approach for improving VLM reasoning

- ✓ 복잡한 SUPERVISED 학습 대신, 간단한 RL로도 효과적

[CONTRIBUTIONS 2]

Selective Sample Replay (SSR)

- ✓ GRPO 기반 RL을 더 안정적이고 효과적으로 만듦

[CONTRIBUTIONS 3]

Forced Rethinking strategy

- ✓ 모델이 자발적으로 자기 반성 (self-reflection) 을 하도록 유도

PRELIMINARIES

PROBLEM FORMULATION

INPUT

$x = (I, Q)$

Visual & Text

$x \in \mathcal{V} \times \mathcal{T}$

OUTPUT

textual response

$y \in \mathcal{Y}$

LEARNING OBJECTIVES

멀티모달 질의-정답 데이터 분포

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} [r(y, x)]$$

이진 보상 함수
1(정답) 또는 0(오답)

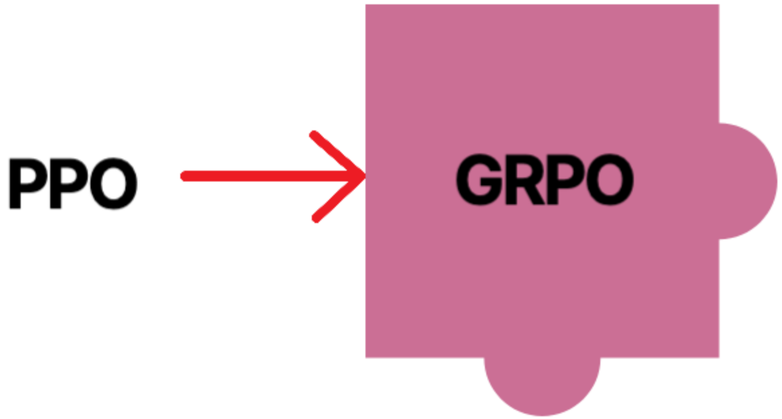
policy $\pi_\theta(y|x)$

입력 x가 주어졌을 때 응답 y를 낼 확률 분포(정책)

데이터 분포에서 뽑은 질의 x에 대해, 정책 π θ가 생성한 응답 y의 정답률(보상 기대값)

REINFORCE LEARNING

How can we effectively incentivize
multimodal slow-thinking capabilities in Vision-Language Models?



불확실한 REWARD

MULTIMODAL에서는 왜 PPO 대신 GRPO를 쓸까?

"A man is playing soccer."

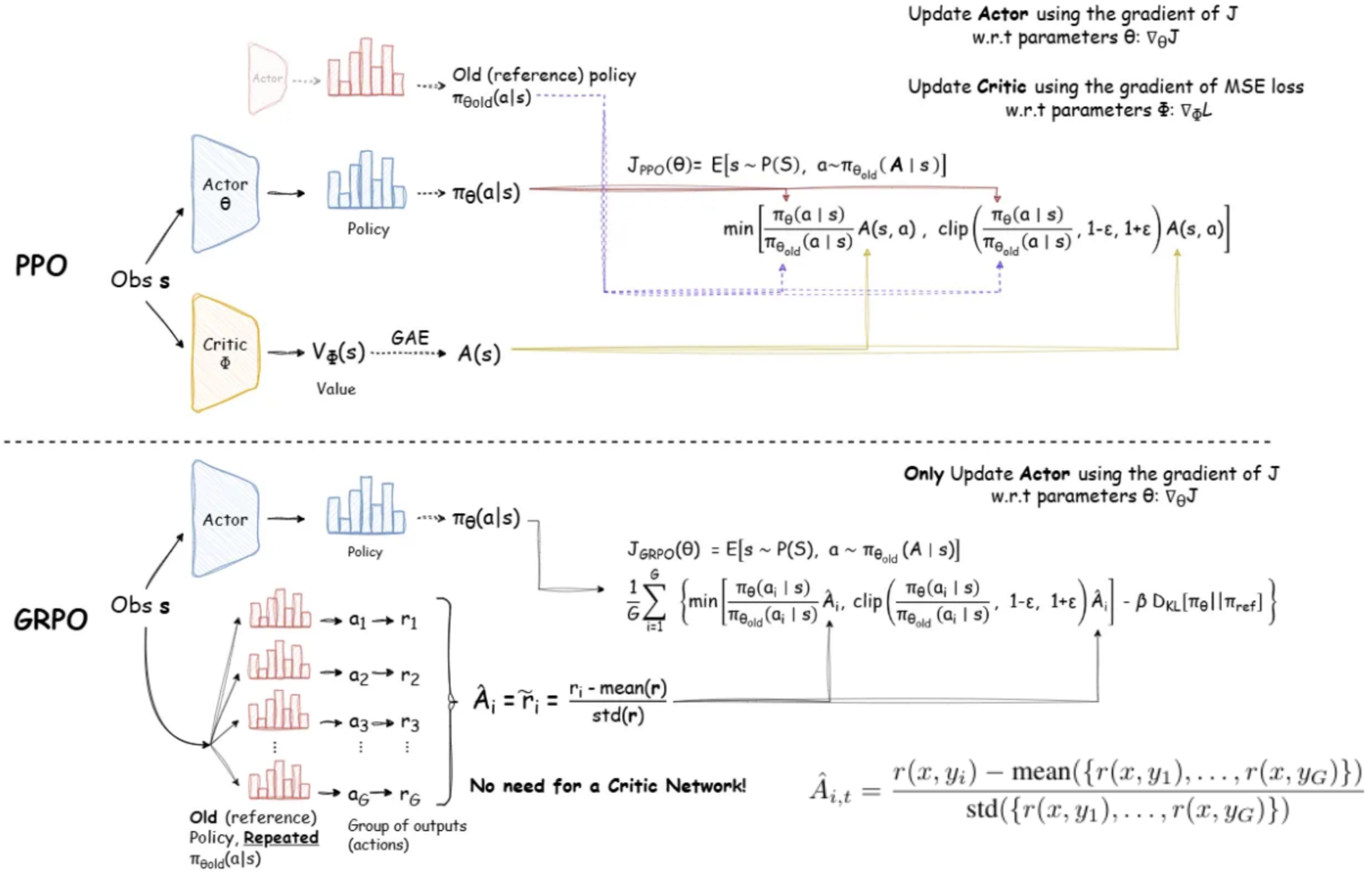
"A person kicks a ball on the field."

절대적인 정답이 없는 경우가 많은 multimodal reasoning

➡ 이런 환경에서 PPO는 신뢰할 수 없는 reward로 학습하게 되어 불안정한 학습 발생

TERMINOLOGY

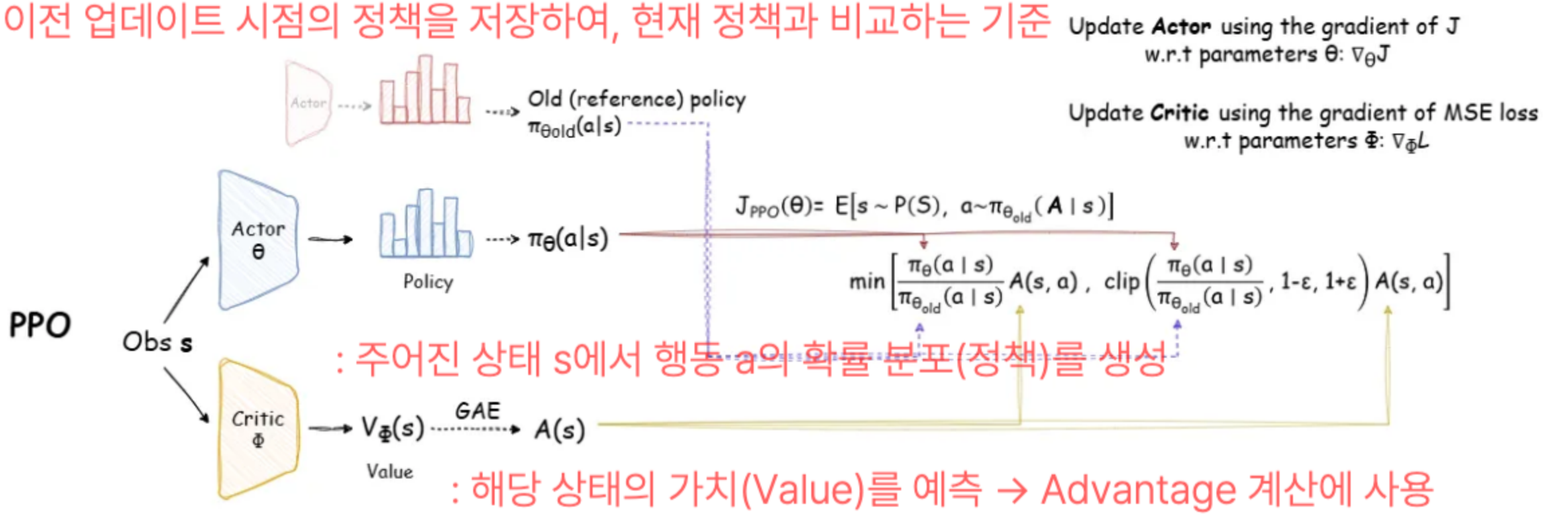
GROUP RELATIVE POLICY OPTIMIZATION(GRPO)



TERMINOLOGY

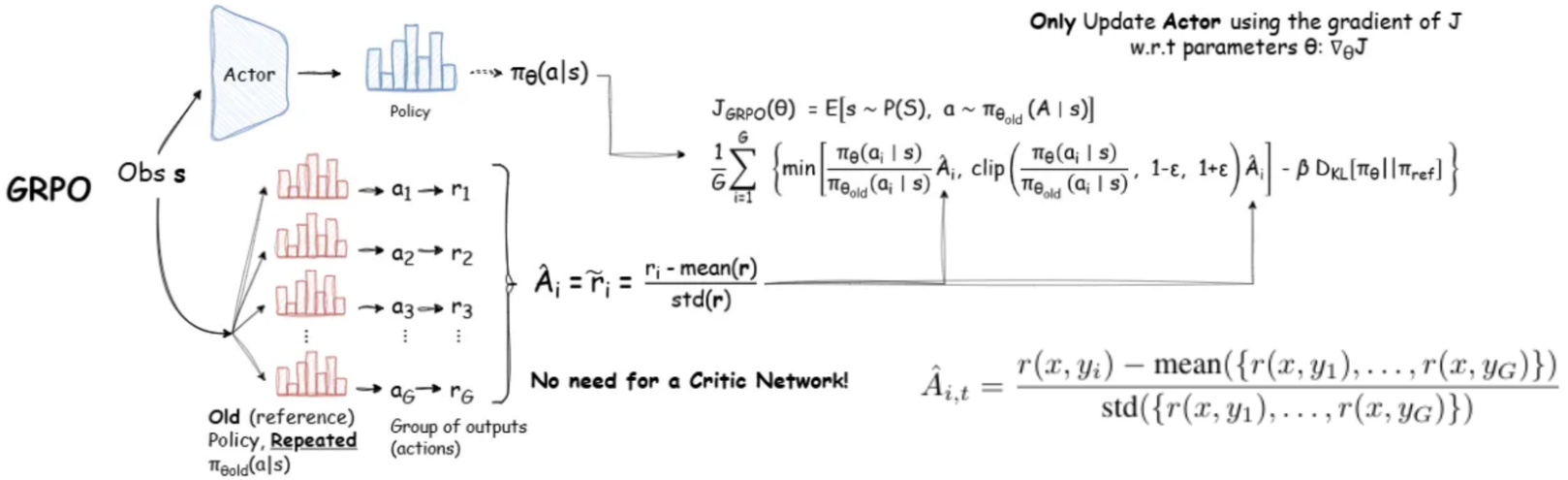
GROUP RELATIVE POLICY OPTIMIZATION(GRPO)

이전 업데이트 시점의 정책을 저장하여, 현재 정책과 비교하는 기준



: 주어진 상태 s에서 행동 a의 확률 분포(정책)를 생성

: 해당 상태의 가치(Value)를 예측 → Advantage 계산에 사용



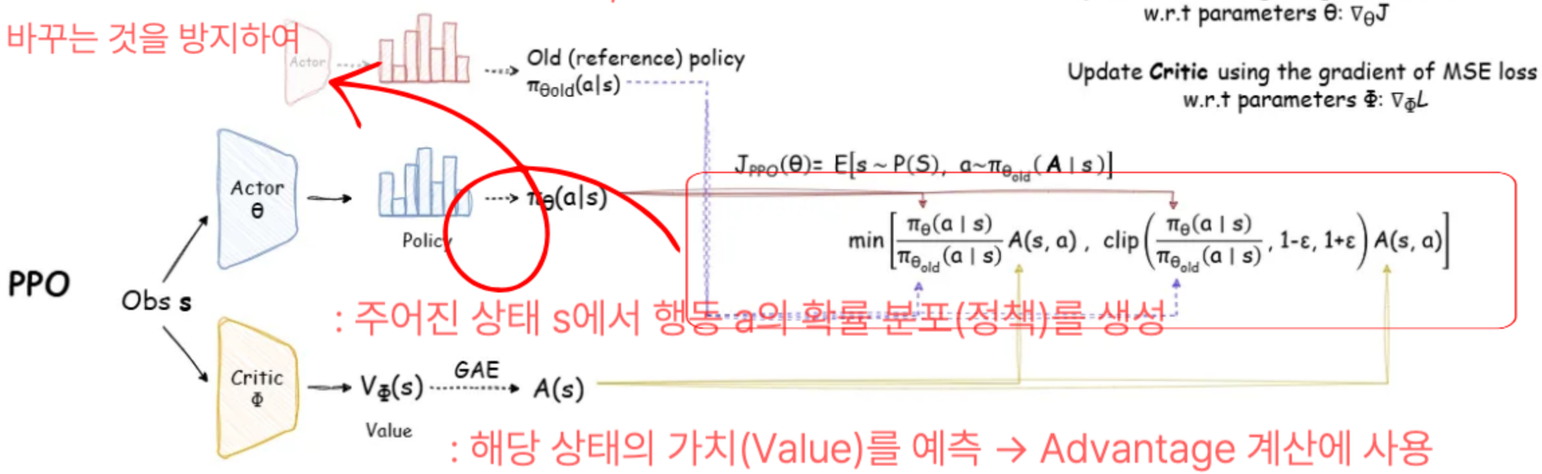
$$\hat{A}_{i,t} = \frac{r(x, y_i) - \text{mean}(\{r(x, y_1), \ldots, r(x, y_G)\})}{\text{std}(\{r(x, y_1), \ldots, r(x, y_G)\})}$$

TERMINOLOGY

GROUP RELATIVE POLICY OPTIMIZATION(GRPO)

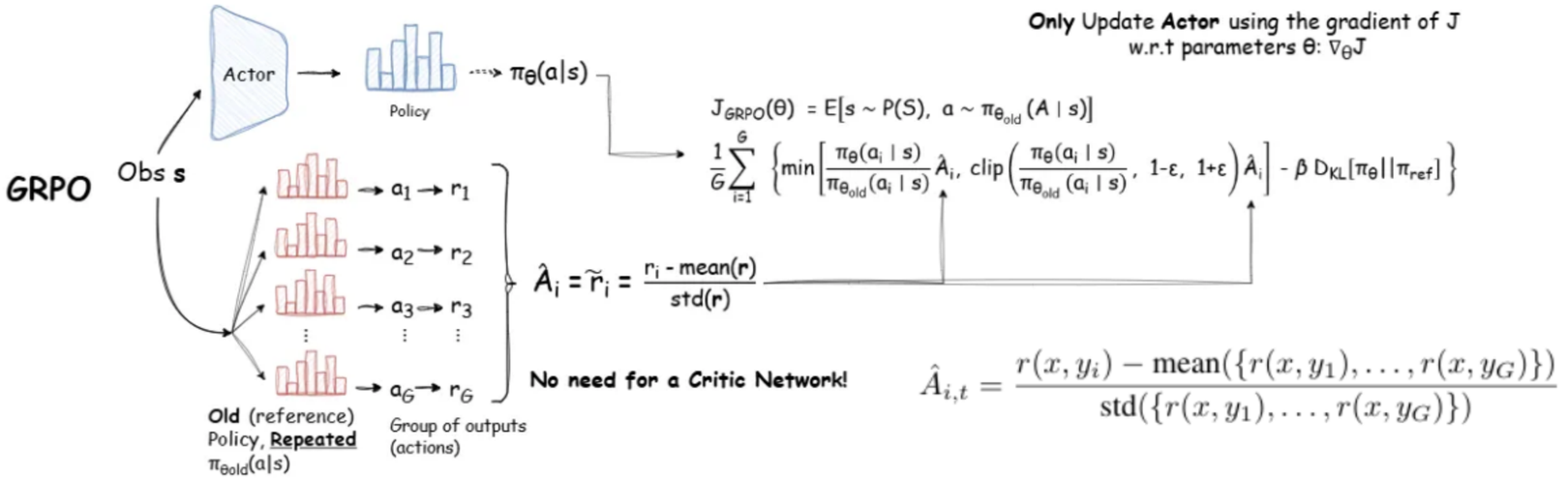
Clipped Surrogate Loss
:정책(policy)을 너무 급격하게 바꾸는 것을 방지하여
학습을 안정화하기 위한 목적

이전 업데이트 시점의 정책을 저장하여, 현재 정책과 비교하는 기준



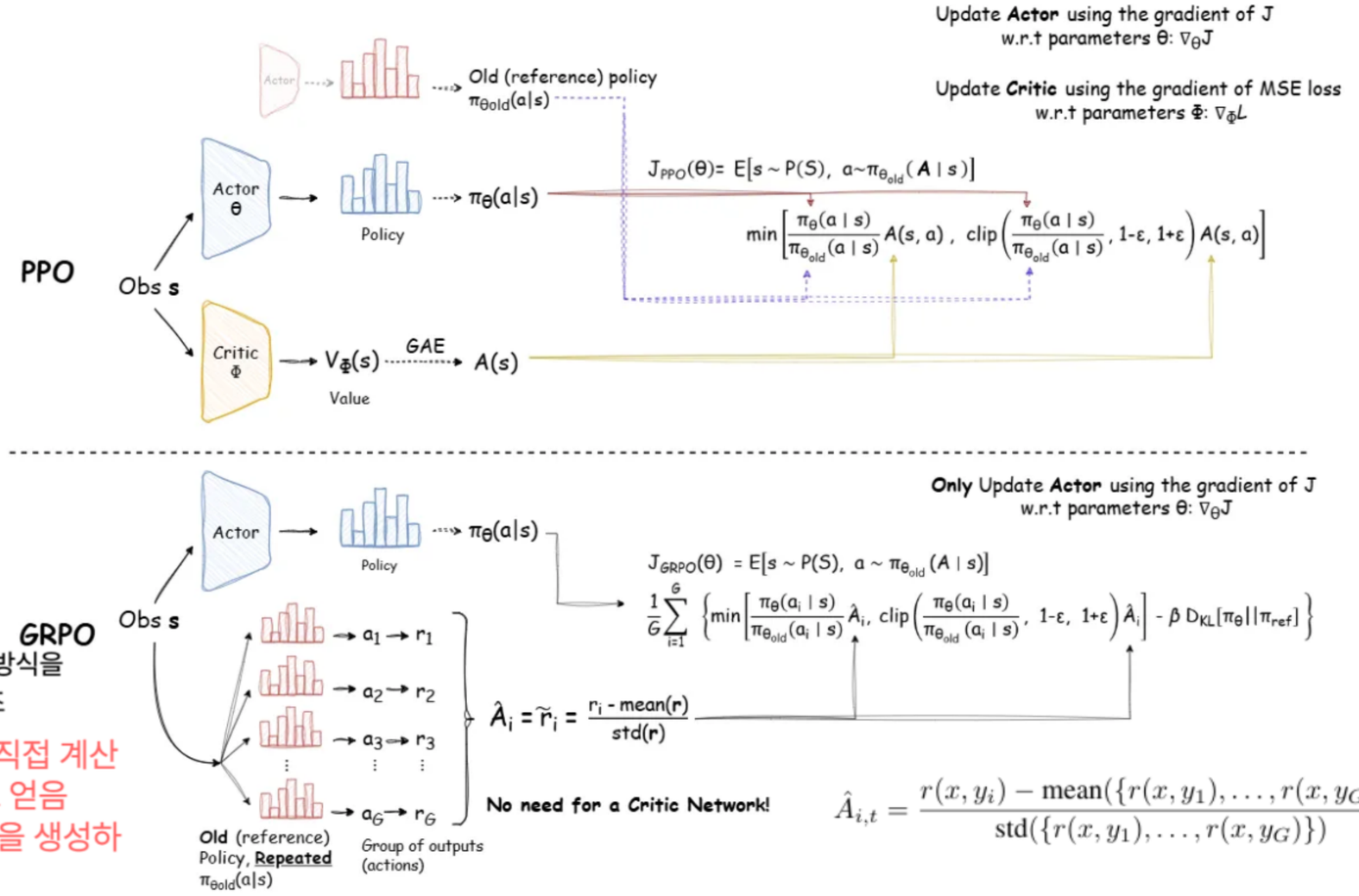
: 주어진 상태 s에서 행동 a의 확률 분포(정책)를 생성

: 해당 상태의 가치(Value)를 예측 → Advantage 계산에 사용



TERMINOLOGY

GROUP RELATIVE POLICY OPTIMIZATION(GRPO)



GRPO는 PPO의 철학을 따르지만, reward 처리 방식을 'group-level 상대 평가' 방식으로 바꾼 구조

- Critic 네트워크가 없음 → Advantage를 직접 계산
- Advantage는 그룹 내 상대 보상 정규화로 얻음
- 한 상태 s에서 G개의 응답(a_1, a_2, ..., a_G)을 생성하고, 각 보상 r 을비교

TERMINOLOGY

GROUP RELATIVE POLICY OPTIMIZATION(GRPO)

GRPO는 PPO의 철학을 따르지만, reward 처리 방식을
'group-level 상대 평가' 방식으로 바꾼 구조

$$\hat{A}_{i,t} = \frac{r(x, y_i) - \text{mean}(\{r(x, y_1), \ldots, r(x, y_G)\})}{\text{std}(\{r(x, y_1), \ldots, r(x, y_G)\})}$$

r(x,yi) x에 대한 i번째 응답에 부여된 보상에서
생성한 응답들에게 각각 부여된 보상에 대한 평균값으로 빼주면.
같은 입력 x에 대한 응답 y1,y2,...,yGy_1, y_2, ..., y_Gy1,y2,...,yG 들의 보상을 비교해서,
- 평균보다 높으면 +, 낮으면 -가 되는 상대적 advantage을 계산
- 정규화(std)를 통해 보상 스케일이 안정되며, vanishing gradient를 방지

GRPO의 문제점

VANISHING ADVANTAGES

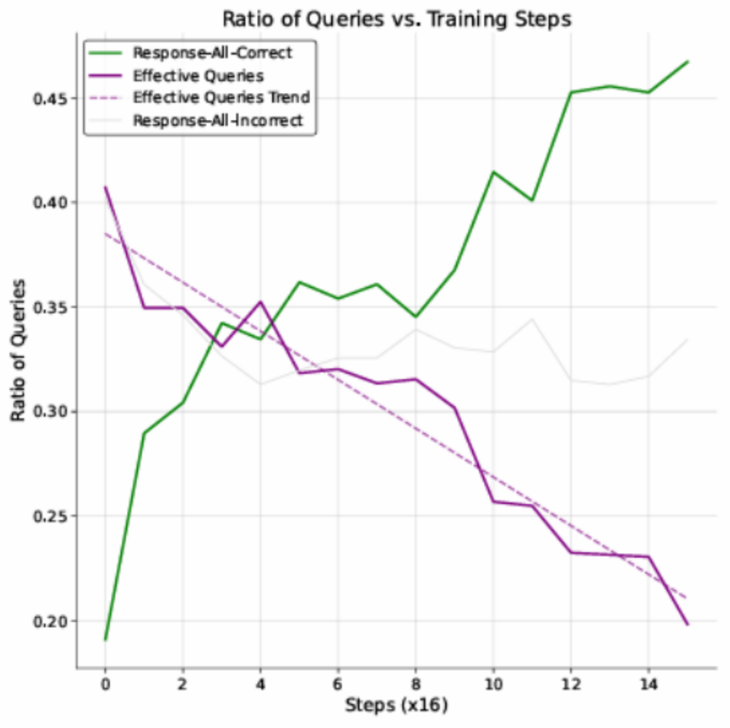


Figure 2: Illustration of the Vanishing Advantages problem. Training of 72B rapidly saturates, leading to a significant decrease of effective queries to only 20% within 256 steps.

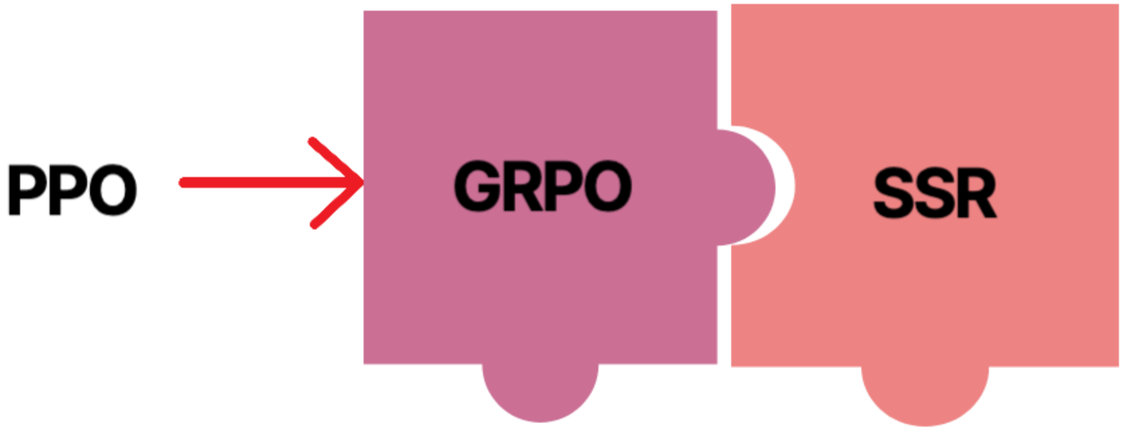
문제 상황: GRPO는 query group 내 응답들이 모두 정답이거나 모두 오답일 경우,
→ 보상이 동일하게 주어짐
→ 결과적으로 Advantage(이점)가 0이 됨
→ ⚠ Gradient가 0 → 학습이 멈춤

Advantage = 0 이면 정책(policy)이 업데이트되지 않음
결국 해당 query group은 학습에 기여하지 못함
시간이 지날수록 모델이 정답률이 높아져서

→ 유효한 쿼리 비율이 감소함

REINFORCE LEARNING

How can we effectively incentivize
multimodal slow-thinking capabilities in Vision-Language Models?



VANISHING ADVANTAGE 문제 해결을 위한

SSR(SELECTIVE SAMPLE REPLAY)

GRPO 학습 중에 의미 있는 학습 경험만 골라 다시 학습(replay)하는 방식

SSR enhances GRPO by integrating an experience replay mechanism that strategically samples high-value experiences from past iterations, similar to Prioritized Experience Replay [Schaul et al., 2015] in Temporal Difference learning.

샘플 j의 Advantage (이전 step에서 계산) 크기만 반영

$$P(\text{select } j) = \frac{|\hat{A}_j|^\alpha}{\sum_{k \in \mathcal{B}_{\text{replay}}} |\hat{A}_k|^\alpha}$$

α: 큰 값일수록 큰 Advantage에 더 집중
(α=0이면 전부 균등확률,
α>1이면 큰 값 편향 강화)

1. GRPO 훈련 중 얻은 과거 경험 (입력 , 응답 , 상대 advantage ^)를 저장

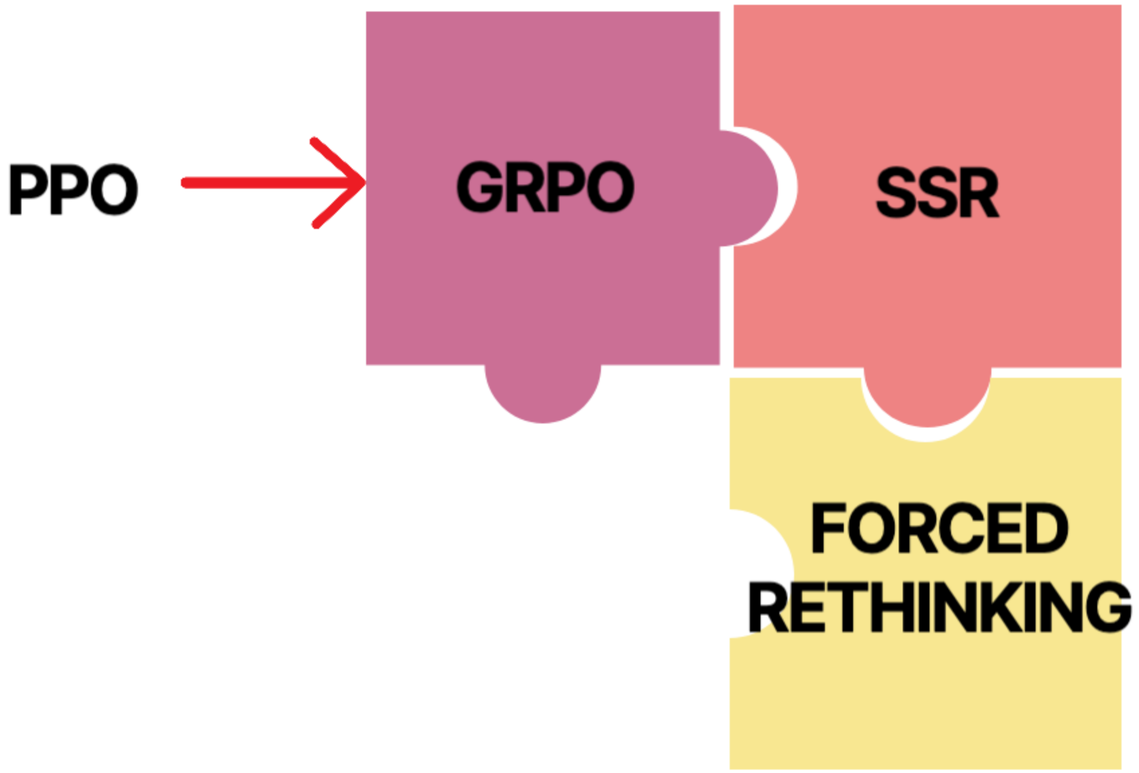
단, $|\hat{A}_i| > 0$ 인 샘플만 저장 (즉, 의미 있는 신호만 보존)

2. 매 훈련 step마다, 현재 batch에 과거 buffer에서 일부 샘플을 추가
→ rehearsal이라고 부름

3. 우선순위 샘플링: 우선순위는 의 크기 (절대값) 기준
즉, 매우 좋았거나 나빴던 응답을 우선적으로 다시 학습

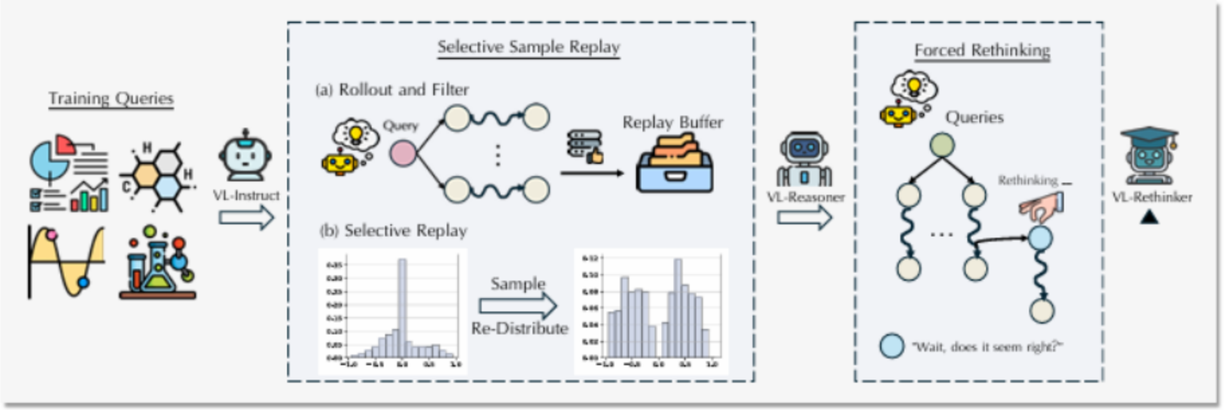
REINFORCE LEARNING

How can we effectively incentivize
multimodal slow-thinking capabilities in Vision-Language Models?



OPTIMIZATION STABILITY

FORCED RETHINKNING



GRPO + SSR 로도 최적화 안정성은 높아졌지만,
self-correction이나 self-reflection 같은 심층 추론 과정이 자연스럽지 못함.

How to solve? -> "force rethinking"
the model to engage in more extensive internal deliberation before producing a final answer

방식1. 프롬프트 내 방식 제공

입력 프롬프트에 직접 "regularly perform self-reflection on your ongoing reasoning"와 같은 힌트 포함.

이런 맥락적 신호(contextual cue)로 인해 모델이 rethinking sequence를 더 잘 생성

방식2. RL Rollout에서의 직접적 개입(core principle)

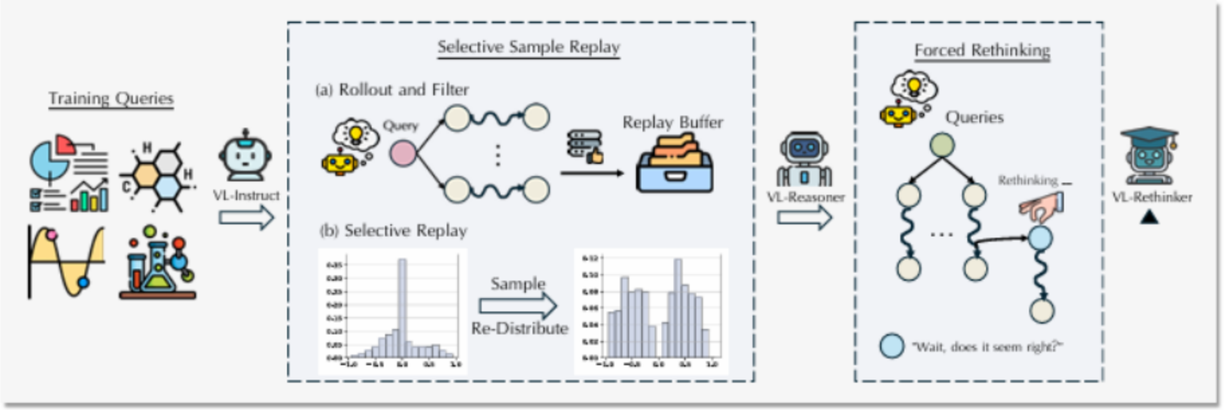
모델이 입력 x에 대해 첫 번째 답변 y1을 생성하면,
그 y1 뒤에 rethinking trigger를 붙임.
그리고 이 y1 + trigger를 다시 모델에 넣어, 추가 답변 y2를 생성하게 함.
최종적으로 y = y1 ⊕ trigger ⊕ y2 형태가 됨.

이 트리거(trigger)는 3가지 유형이 있음:

- self-verification
- self-correction
- self-questioning

OPTIMIZATION STABILITY

FORCE RETHINKNING



GRPO + SSR 로도 최적화 안정성은 높아졌지만,
self-correction이나 self-reflection 같은 심층 추론 과정이 자연스럽지 못함.

How to solve? -> "force rethinking"
the model to engage in more extensive internal deliberation before producing a final answer

Rethinking Instruction Prompt

```
{question}
Guidelines:

Please think step by step, and **regularly perform self-questioning, self
-verification, self-correction to check your ongoing reasoning**, using
connectives such as "Wait a moment", "Wait, does it seem right?", etc. Remember
to put your final answer within \\boxed{}.
```

During the Forced Rethinking training stage, we use the above prompt to encourage self-reflection, and use three types of rethinking textual triggers.

Rethinking Triggers

```
self_questioning = "\n\nWait, does it seem right?"
self_correction = "\n\nWait, there might be a mistake"
self_verification = "\n\nWait, let's double check"
```

방식2. RL Rollout에서의 직접적 개입(core principle)

모델이 입력 x에 대해 첫 번째 답변 y1을 생성하면,
그 y1 뒤에 rethinking trigger를 붙임.
그리고 이 y1 + trigger를 다시 모델에 넣어, 추가 답변 y2를 생성하게 함.
최종적으로 y = y1 ⊕ trigger ⊕ y2 형태가 됨.

이 트리거(trigger)는 3가지 유형이 있음:

- self-verification
- self-correction
- self-questioning

EXPERIMENTS

연구 문제

KEY QUESTIONS

Q1: *Method Effectiveness:* How does our approach enhance performance on comprehensive multi modal benchmarks compared to existing MLLMs?

Q2: *Ablation Studies:* How do the proposed Selective Sample Replay (SSR), Forced Rethinking, and curated data affect performance?

Q3: *Effectiveness of the learned rethinking behaviors:* Do the model learn to effectively and spontaneously perform deliberate thinking?

DESIGN THE EXPERIMENT

TRAINING DATA AND BENCHMARKS

DATASET

공개 데이터셋(Du et al., Yang et al., Meng et al.) +
웹에서 새로 수집한 데이터를 통합해 초기 "seed" 쿼리셋을 만듦.

데이터 품질을 높이기 위해 엄격한 기준 적용:

VLM이 명확히 답변 가능한 객관적 쿼리만 수집
문제가 있거나(불완전, 애매),
너무 쉽거나 검증이 어려운 샘플은 제거

언어 다양성과 지식 강화를 위해
기존 쿼리들을 다양한 표현으로 rephrasing하여 확장함.

결과적으로 38,870개 고품질 쿼리셋 생성.

DATA SUBSET 운영

훈련 중 seed 쿼리에 대해 RL 성능이 빨리 포화에 도달하는 현상 관찰
일부 쿼리는 모델이 항상 맞추거나, 항상 틀림 → 학습 의미가 떨어짐

이를 해결하기 위해 모델 크기별로 난이도·성능이 다른
쿼리 subset을 별도로 큐레이션:

7B 모델: 약 16,000개 쿼리
32B, 72B 모델: 약 20,000개 쿼리

모델 스케일(파라미터 크기)에 따라 데이터 난이도를 조절

DESIGN THE EXPERIMENT

IMPLEMENTAION

- OpenRLHF 프레임워크를 사용해 알고리즘을 구현.

OpenRLHF: RLHF(강화학습 기반 인간 피드백) 오픈소스 라이브러리 모델 훈련

- 각 쿼리셋(질문 데이터셋)마다 최대 3에폭(epoch) 학습 진행
- 최종 체크포인트는 검증용 validation set의 평균 reward(보상)를 기준으로 선정

- train paradigm: Near on-policy RL 적용

-> 행동 정책(behavior policy과 개선 정책(improvement policy를

1024개 쿼리(1에피소드)마다 동기화

- SSR(Selective Sample Replay) 버퍼

각 에피소드마다 지속됨 → 에피소드가 끝나면 초기화됨

- 샘플링 & 배치

쿼리 1개당 8개 응답(response) 생성

- 훈련 배치 크기(batch size): 512 쿼리-응답 쌍
- 각 쿼리에 대해 최대 2개의 정답 'rethinking trajectory'만 허용

EXPERIMENTS

MAIN RESULT - 72B

→ 과학, 인문, 사회, 예술 등 여러 학문 분야에 걸친 멀티모달·멀티도메인 추론 능력 평가

→ 현실 세계의 대규모·다양하고 희귀한(롱테일) 문제들로 구성된 종합 벤치마크

Model	Math-Related			Multi-Discipline			Real-World
	MathVista testmini	MathVerse testmini	MathVision test	MMMU-Pro overall	MMMU val	EMMA full	MEGA core
Proprietary Model							
OpenAI-o1	73.9	57.0	60.3	62.4	78.2	45.7	56.2
OpenAI-GPT-4o	60.0	41.2	30.6	51.9	69.1	32.7	52.7
Claude-3.5-Sonnet	67.7	47.8	33.5	51.5	68.3	35.1	52.3
Gemini-2.0-Flash	73.4	54.6	41.3	51.7	70.7	33.6	54.1
Open-Source Models							
Llama4-Scout-109B	70.7	-	-	<u>52.2</u>	69.4	24.6	31.8
InternVL-2.5-78B	72.3	51.7	34.9	48.6	61.8	27.1	44.1
QvQ-72B	71.4	48.6	35.9	51.5	70.3	32.0	8.8
LLava-OV-72B	67.5	39.1	30.1	31.0	56.8	23.8	29.7
Qwen-2.5-VL-32B	74.7	48.5	<u>38.4</u>	49.5	†59.4	31.1	13.3
Qwen-2.5-VL-72B	<u>74.8</u>	<u>57.2</u>	38.1	51.6	†67.0	<u>34.1</u>	<u>49.0</u>
VL-Rethinker-32B	78.8	56.9	40.5	50.6	65.6	37.9	19.9
VL-Rethinker-72B	**80.4**	**63.5**	**44.9**	**55.9**	**68.8**	**38.5**	**51.3**
Δ (Ours - Open SoTA)	+5.6	+6.3	+6.8	+3.7	-1.4	+4.4	+2.3

Table 1: Comparison between our 72B model and other state-of-the-art models. The notation of † indicates reproduced results using our evaluation protocols.

EXPERIMENTS

MAIN RESULT - 7B

Model	Math-Related			Multi-Discipline			Real-World
	MathVista testmini	MathVerse testmini	MathVision test	MMMU-Pro overall	MMMU val	EMMA full	MEGA core
General Vision-Language Models							
InternVL2-8B	58.3	-	17.4	29.0	51.2	19.8	26.0
InternVL2.5-8B	64.4	39.5	19.7	34.3	56.0	-	30.4
QwenVL2-7B	58.2	-	16.3	30.5	54.1	20.2	34.8
QwenVL2.5-7B	68.2	46.3	25.1	36.9	†54.3	21.5	35.0
Llava-OV-7B	63.2	26.2	-	24.1	48.8	18.3	22.9
Kimi-VL-16B	68.7	44.9	21.4	-	†55.7	-	-
Vision-Language Reasoning Models							
MM-Eureka-8B (Intern)	67.1	40.4	22.2	27.8	49.2	-	-
MM-Eureka-7B (Qwen)	73.0	50.3	26.9	-	-	-	-
R1-VL-7B	63.5	40.0	24.7	7.8	44.5	8.3	29.9
R1-Onevision-7B	64.1	46.4	29.9	21.6	-	20.8	27.1
OpenVLThinker-7B	70.2	47.9	25.3	37.3	52.5	26.6	12.0
VL-Rethinker-7B	**74.9**	**54.2**	**32.3**	**41.7**	**56.7**	**29.7**	**37.2**
Δ (Ours - Prev SoTA)	+4.7	+6.3	+2.4	+4.4	+0.7	+3.1	+2.2

Table 2: Comparison between our 7B model and other general and reasoning vision-language models. † means that the results are reproduced by us.

ABLATION STUDY

ABLATION ON SELECTIVE SAMPLE REPLAY(SSR)

SSR 유지, Forced Rethinking 제거

SSR 제거, Fiter만 적용
SSR & Filtering 제거, GRPO만 사용

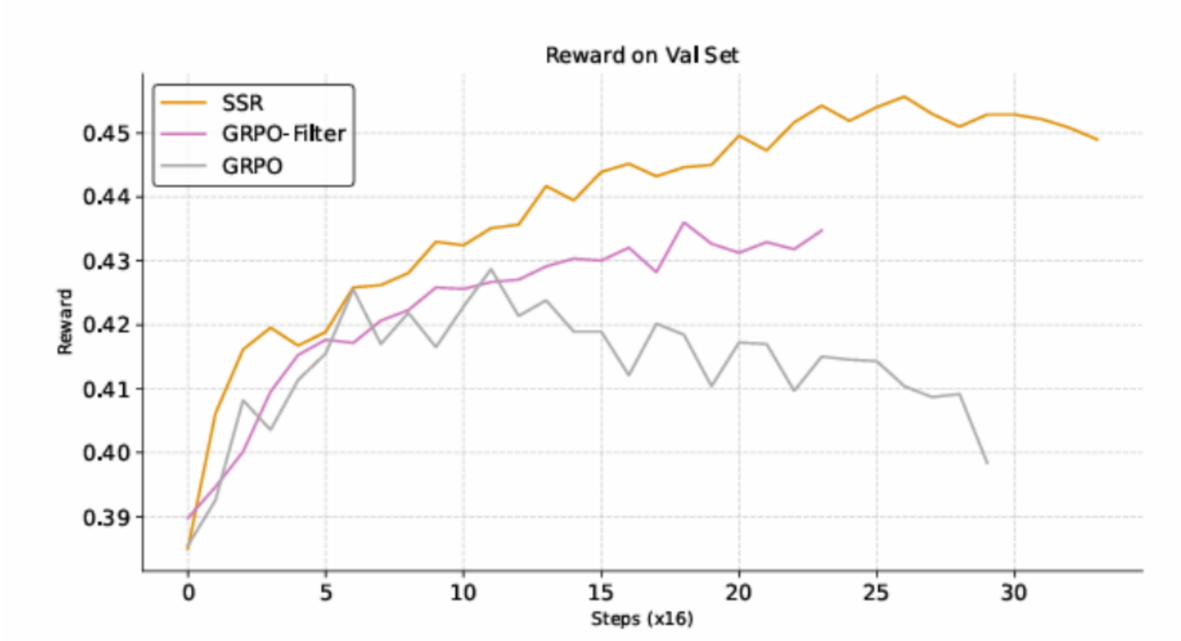
텍스트 일부 제외(13K만 사용)
과학 및 텍스트 제외 (11K만 사용)

Model	RL-Algo	Data	MathVision	MathVista	MathVerse	MMMU-Pro	EMMA
VL-Rethinker-7B	SSR	16K	32.3	74.9	54.2	41.7	29.7
w/o 'Forced-Rethinking'	SSR	16K	29.8	72.4	53.2	40.9	29.5
- no SSR	Filter	16K	28.5	72.0	50.0	40.0	26.9
- no SSR& Filter	GRPO	16K	26.0	70.9	51.4	38.8	26.2
- no Text	SSR	13K	29.1	73.5	53.5	41.1	28.7
- no Science&Text	SSR	11K	28.0	71.6	50.3	39.7	28.0

Table 3: Ablation Results to show the impact of SSR and Data Mix.

ABLATION STUDY

ABLATION ON SELECTIVE SAMPLE REPLAY(SSR)



- GRPO (기본)

학습 후반에 과적합(overfitting) 현상이 두드러짐
그 결과 성능이 오히려 하락
원인: 학습이 진행될수록 advantage가 0에 가까운 샘플 비중이 증가 → 학습 신호 약화 → 실질적인 batch 크기 감소 → 학습 불안정

- GRPO-Filter

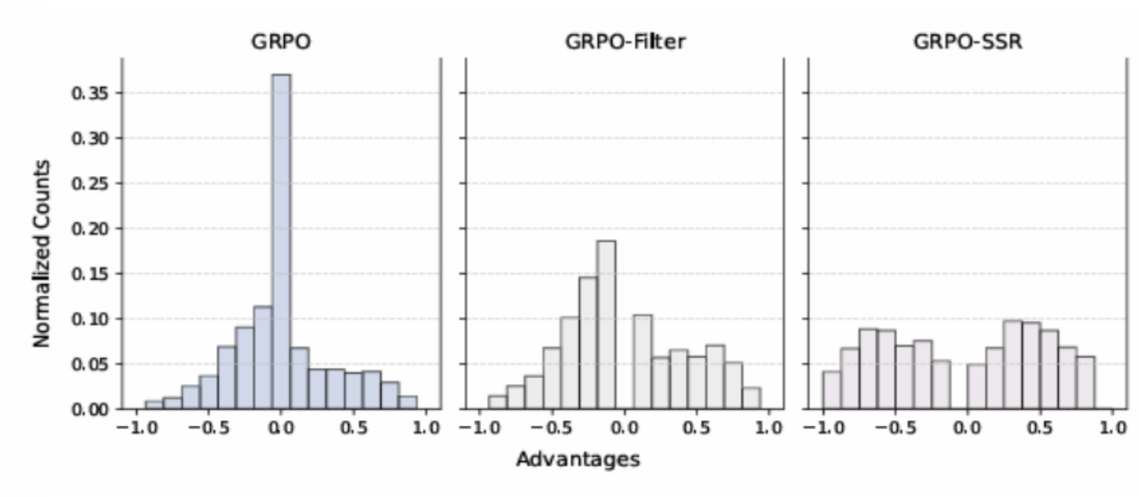
zero advantage 샘플을 제거하여 GRPO보다 안정적이지만, 성능 개선 폭은 제한적

- GRPO-SSR

학습 전반에서 가장 안정적
최종 수렴 성능도 최고
Filtering + Selective Replay의 결합 효과가 긍정적으로 작용

ABLATION STUDY

ABLATION ON SELECTIVE SAMPLE REPLAY(SSR)



- GRPO

advantage 분포가 0 부근에 집중된 왜곡된 형태
많은 샘플이 gradient를 거의 제공하지 못함(비효과적)

- GRPO-Filter

0 부근 극단적 피크는 완화됐지만, 여전히 중앙(0)에 강한 편향 존재
작은 advantage 샘플 다수가 남아 있음

- GRPO-SSR

advantage 분포를 0에서 멀리 분산시킴
절대 advantage가 큰 샘플(예: 어려운 문제 정답, 쉬운 문제 오답)에 더 높은 비중
이런 샘플은 결정 경계(decision boundary) 근처에 있어 모델 학습에 더 유익

ABLATION STUDY

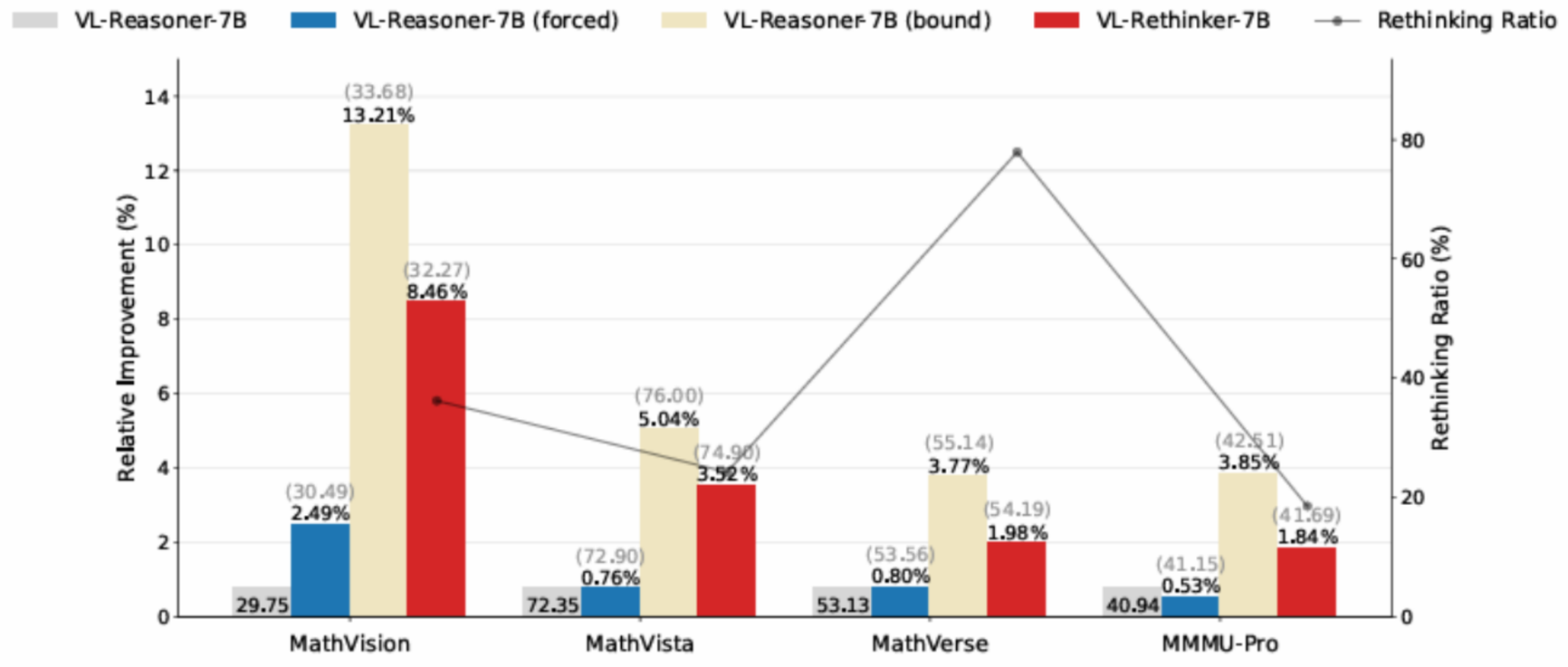
ANAYLSIS ON FORCED RETHINKING

w/o Forced Rethinking (VL-Reasoner)

Baseline 모델에 테스트 시 모든 질의에 강제로 재사고 단계 적용

테스트 시 오답인 경우에만 재사고 적용 → 최대 향상 가능치 추정

Forced Rethinking 기법으로 학습된 최종 모델



Relative Improvement (%): 베이스라인 대비 성능 향상 비율 (막대)
Rethinking Ratio (%): 테스트 시 자발적으로 재사고를 사용한 비율 (선)

CONCLUSION

“너 자신을 알라.”

“Know thyself.”
-Socrates