

Vision-Zero

Scalable VLM Self-Improvement via Strategic Gamified Self-Play

A Review & Implementation Overview



2025.10.11

장준한

Qinsi Wang¹, Bo Liu², Tianyi Zhou³, Jing Shi⁴, Yueqian Lin¹, Yiran Chen¹, Hai Helen Li¹, Kun Wan^{4*}, Wentian Zhao^{4*}

¹Duke University, ²National University of Singapore, ³University of Maryland, ⁴Adobe Inc.

<https://arxiv.org/abs/2509.25541>

<https://huggingface.co/papers/2509.25541>

<https://github.com/wangqinsi1/Vision-Zero>

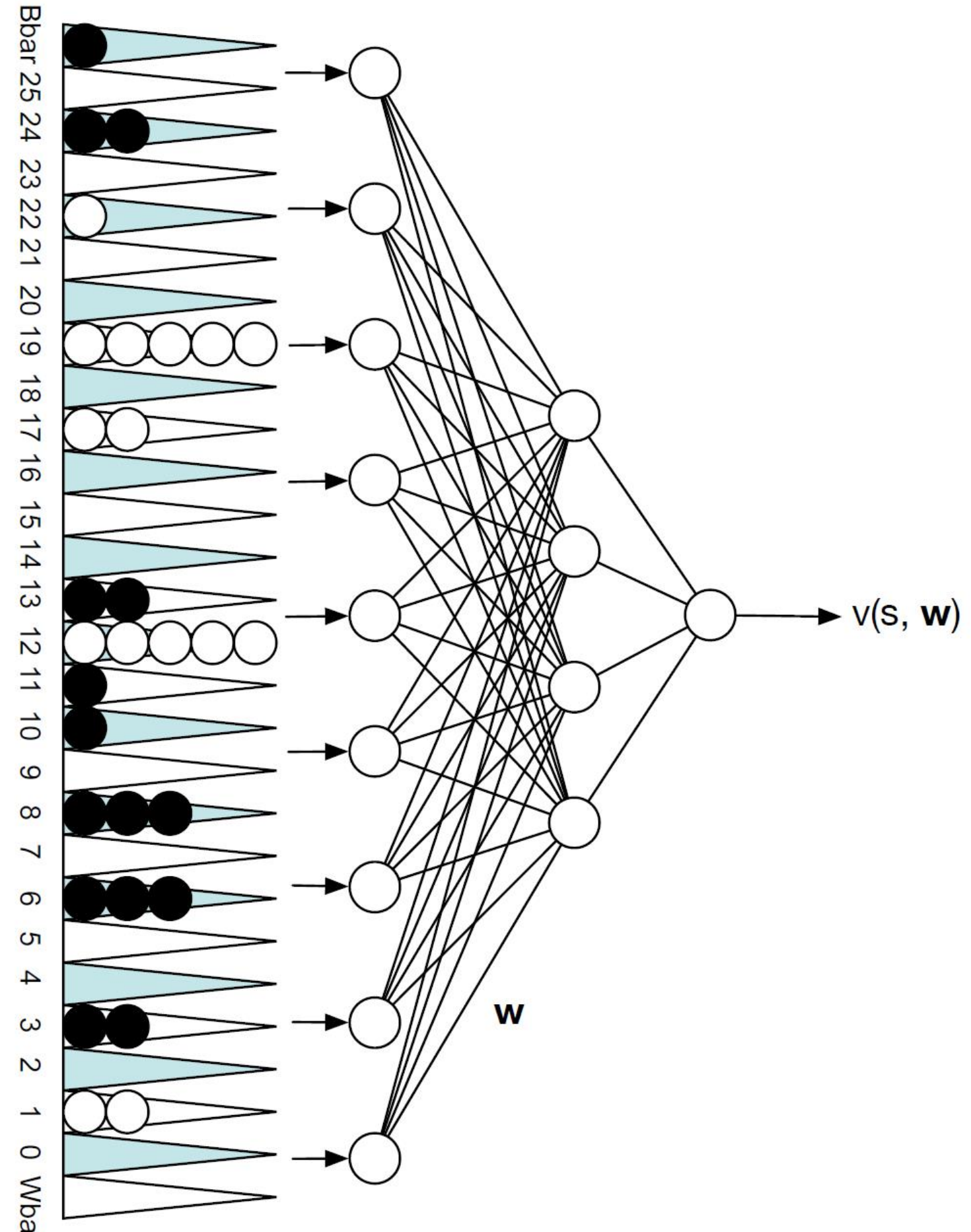
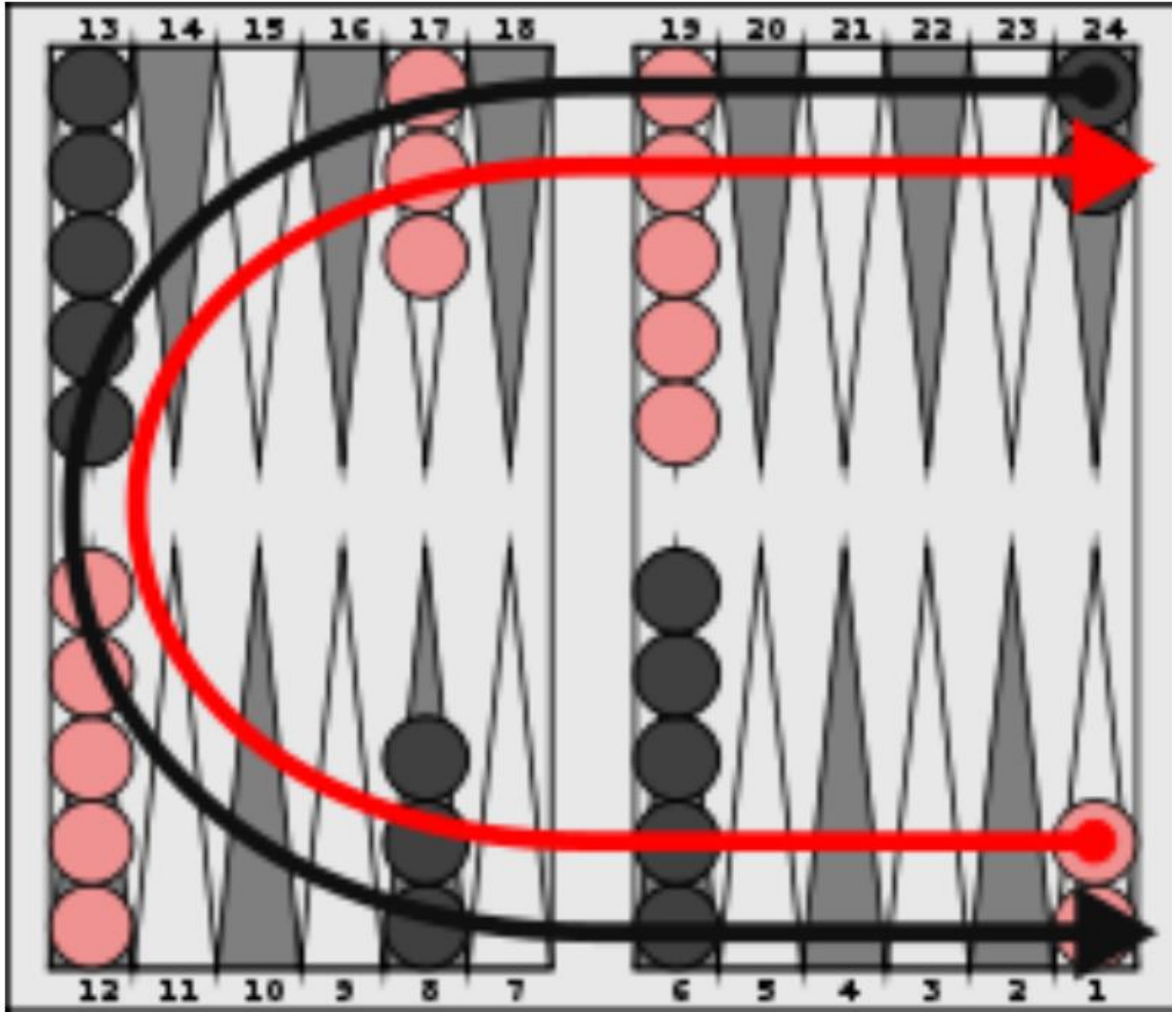
1. 서론 (Introduction)

- 최근 비전-언어 모델(VLM)은 다양한 멀티모달 과제에서 뛰어난 성과를 내고 있음
- 그러나 여전히 인간이 큐레이션한 데이터에 크게 의존한다는 구조적 한계를 지님
- 지도 학습, RLHF, RLVR 등 기존 접근 방식은 막대한 주석 비용과 데이터 부족 문제를 안고 있음
- 무엇보다 모델이 인간이 제공한 감독(supervision)을 넘어서 독자적으로 전략을 발견하기 어렵다는 '지식 상한선(Knowledge Ceiling)'을 만듦

1. 서론 (Introduction)

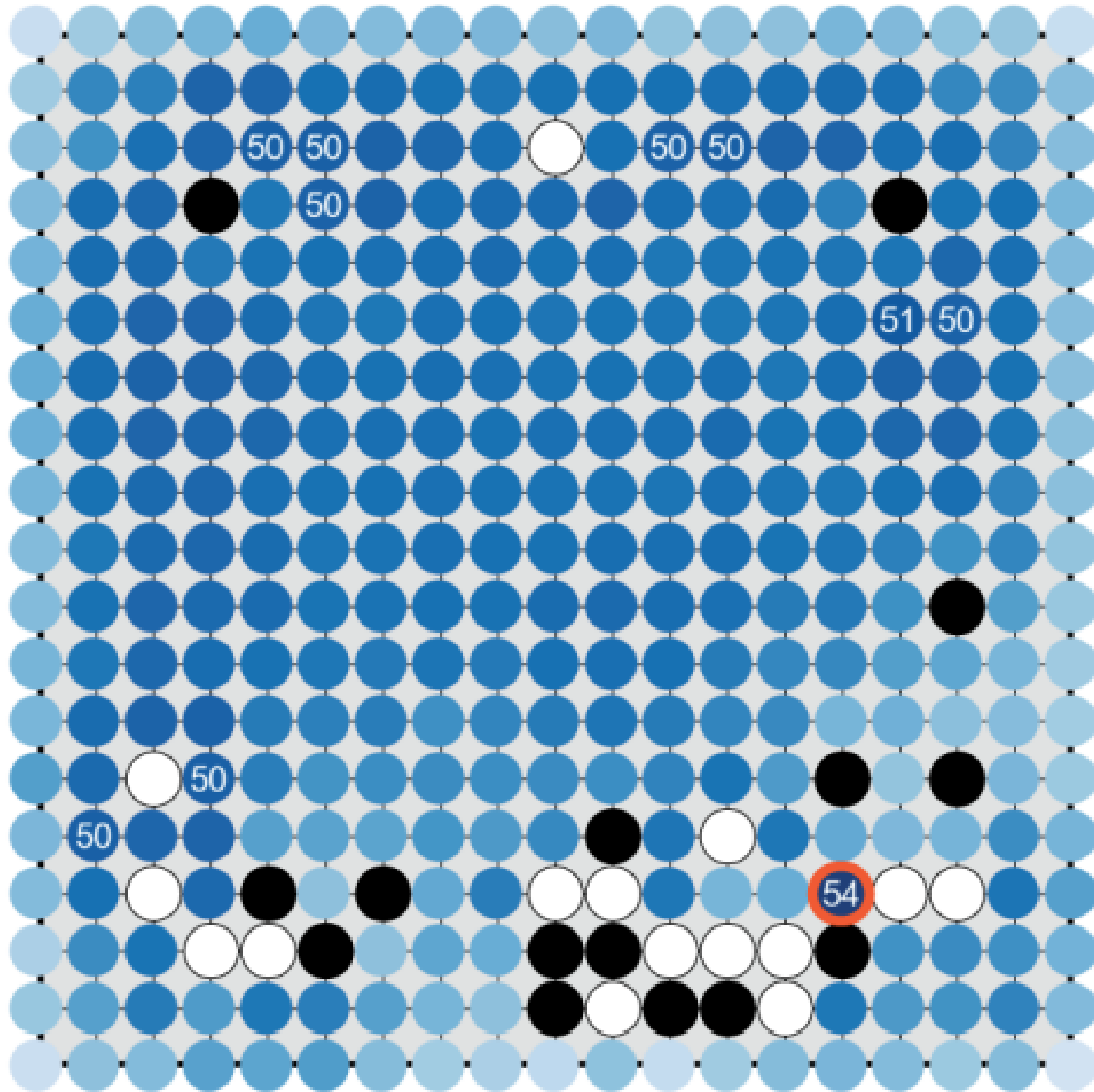
- 이러한 한계를 극복할 수 있는 유력한 해법은 **Self-Play**
 - Self-Play는 모델이 자기 복제본과 경쟁하는 과정을 통해 자동으로 피드백을 받음
 - 환경이 지속적으로 도전적인 상태를 유지하도록 하여 성능을 끊임없이 향상시킴
 - 실제로 TD-Gammon, AlphaGo, OpenAI Five와 같은 사례에서 인간 전문가를 능가하는 성취를 보춤

1. 서론 (Introduction)

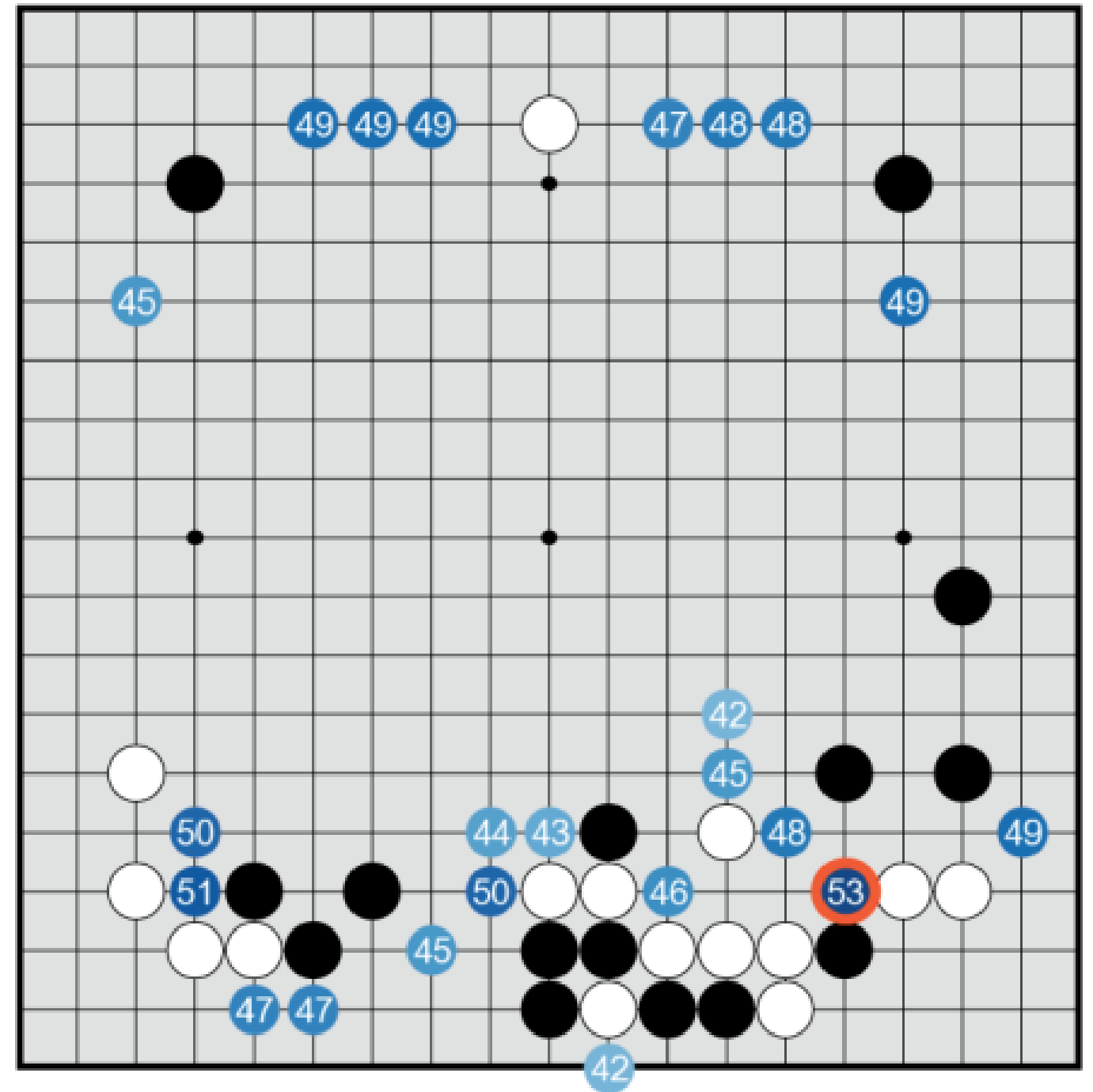


1. 서론 (Introduction)

- 이러한 한계를 극복할 수 있는 유력한 해법은 자기 대전(Self-



메|커|니|즘|임|이|입|증|되|





|| / | L | □ □ ∪ | H O □

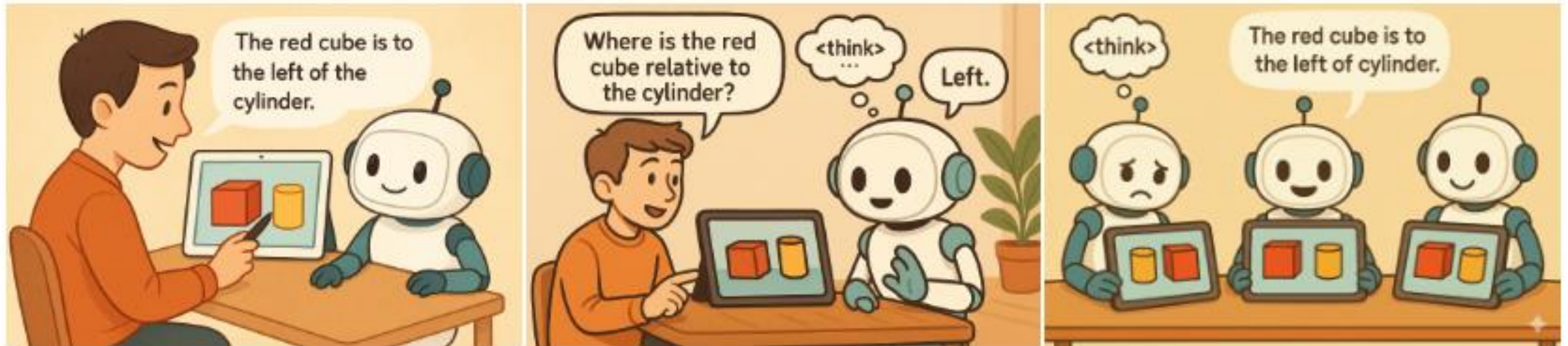
1. 서론 (Introduction)

- 최근에는 이러한 Self-Play 개념이 **LLM 학습**으로 확장되기 시작
- SPIRAL, Absolute Zero와 같은 연구들은 언어 기반 게임환경을 통해 모델이 명확한 규칙 속에서 경쟁하며 점진적으로 추론 능력을 높이도록 설계 됨
- 그러나 이러한 시도가 **VLM으로 확장된 사례는 거의 없으며**, 멀티모달 데이터의 높은 수집 비용을 고려할 때 이는 시급히 해결해야 할 과제

1. 서론 (Introduction)

- 이를 해결하기 위해 논문에서는 **Vision-Zero**를 제안
 - Vision-Zero는 미세한 시각적 차이를 가진 이미지 쌍을 활용해 “Who Is the Spy?”와 같은 전략적 Self-Play 게임 환경을 구성함
 - 인간 개입 없는 Zero-Human-in-the-Loop 학습을 가능하게 함
 - 또한 Iterative Self-Play Policy Optimization (Iterative-SPO)을 통해 Self-Play와 검증 가능한 보상을 결합하여 학습 안정성과 성능 향상을 모두 확보

1. 서론 (Introduction)



(a) Supervised Learning

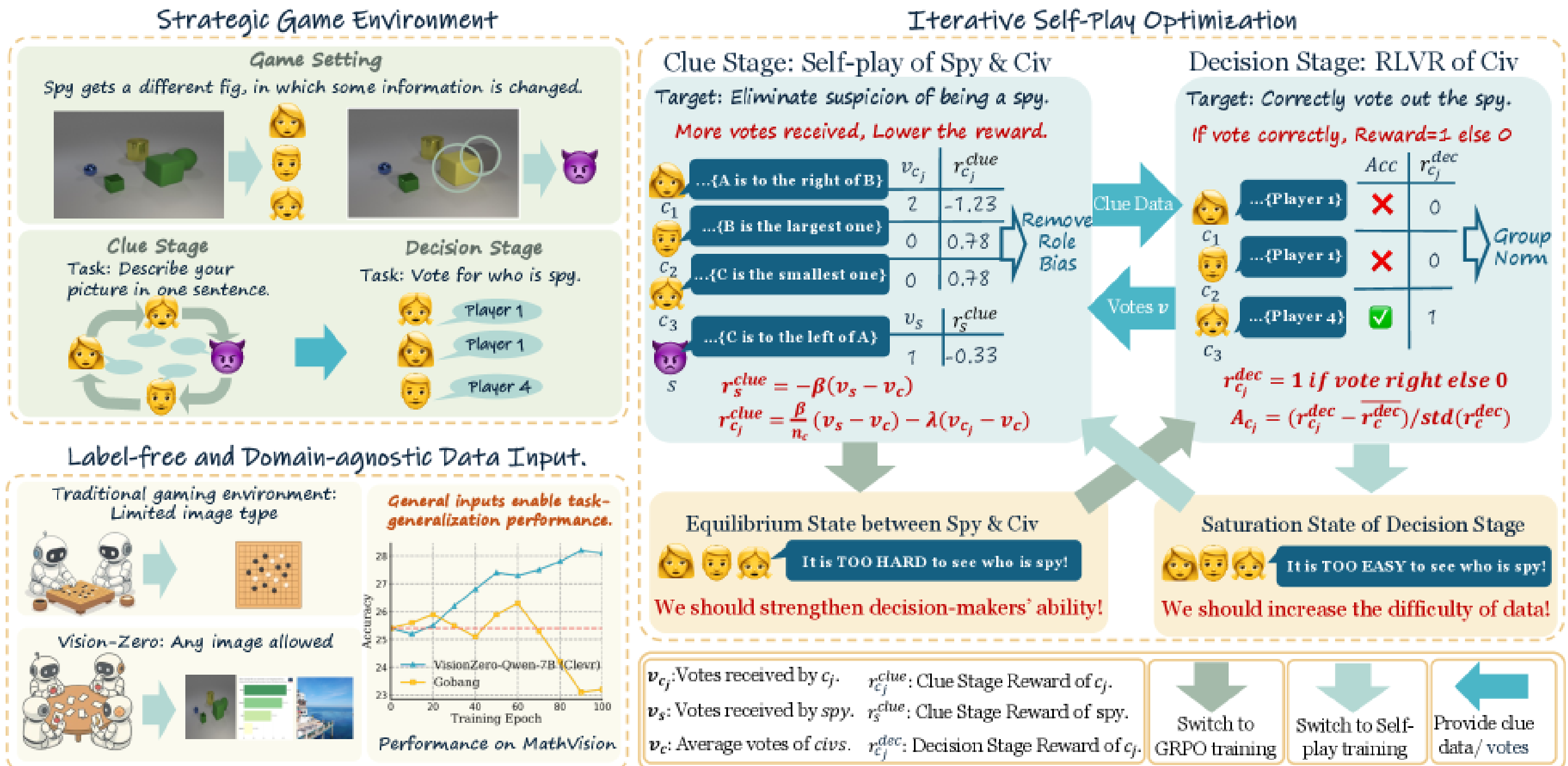
(b) Reinforcement Learning

(c) Vision-Zero

- A. 지도 학습은 인간이 큐레이션한 추론 궤적에 의존
- B. 강화 학습은 검증된 보상을 통해 자율 학습을 가능하게 하지만, 여전히 전문가가 설계한 질문-답변 쌍에 의존
- C. 비전 제로는 시각적 차이를 지닌 이미지 쌍을 활용한 자가 플레이 게임을 통해 인간 경험과 무관하게 학습 데이터를 생성하며, VLM이 확장 가능한 자기 계발을 이루도록 함

2. Vision-Zero: 일반화 가능한 게임화 교육 프레임워크

- Vision-Zero는 그림 3과 같이 일반적이고 확장 가능하며
고성능의 Self-Play 기반 VLM 학습 프레임워크



2. Vision-Zero: 일반화 가능한 게임화 교육 프레임워크

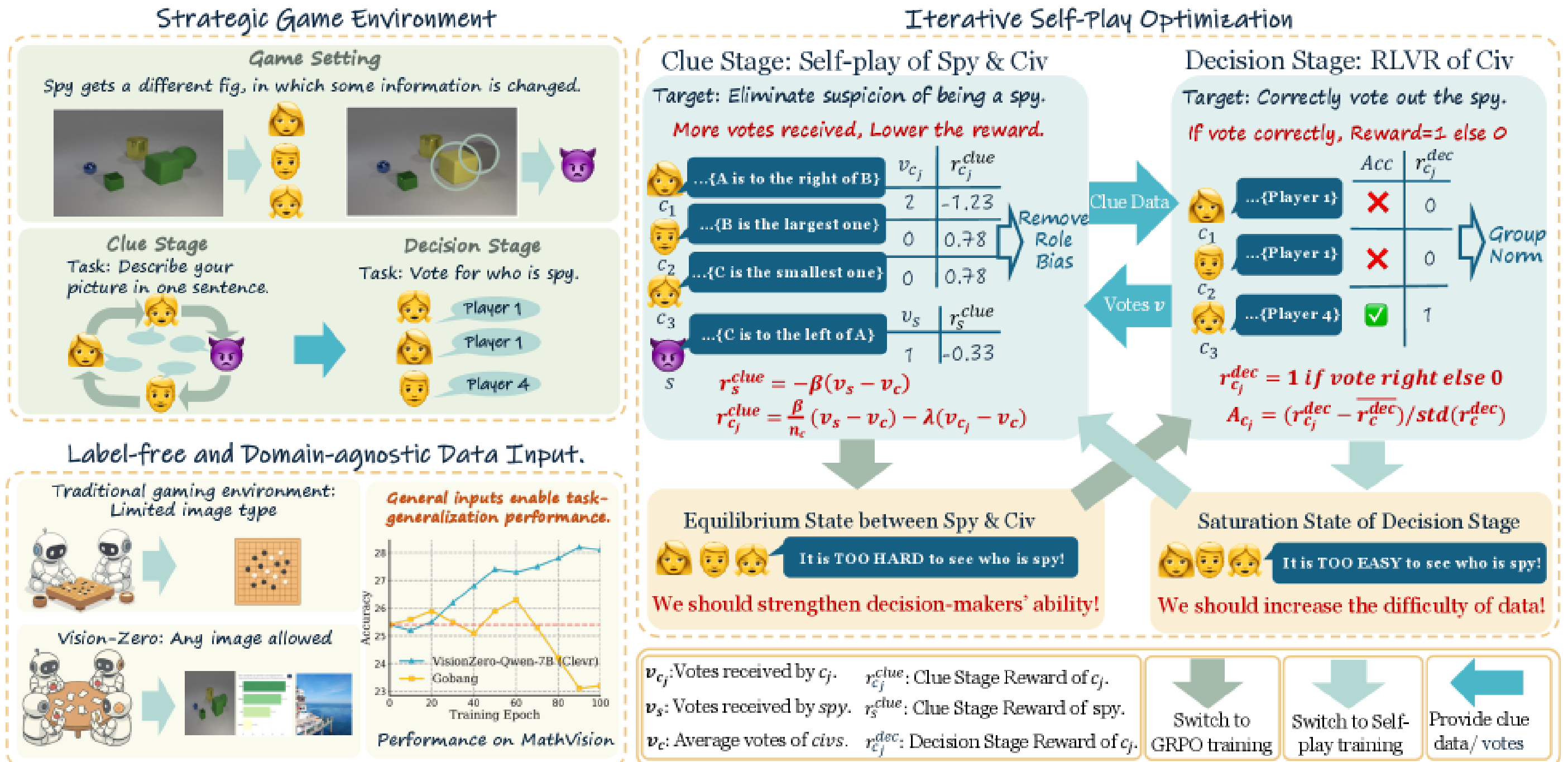
- Vision-Zero는 그림 3과 같이 **일반적이고 확장 가능하며 고성능의 Self-Play 기반 VLM 학습 프레임워크를 제안**
 1. **환경과 데이터 구성(2.1)** – 시각적 게임 환경을 정의하고 학습 데이터를 자동 생성
 2. **Iterative-SPO(2.2)** – Self-Play와 RLVR을 번갈아 적용해 지속적인 성능 향상을 이끔
 3. **비교 분석(2.3)** – 기존 인간 참여 학습과 대비해 Vision-Zero의 이점을 종합적으로 분석

2.1 환경 및 데이터

- Vision-Zero는 사회적 추론 게임 Who is the Spy에서 영감을 받은 전략적 게임 환경을 통해 모델이 상호작용하며 추론하도록 함
 - 여러 플레이어: N_c 명의 민간인 + 1명의 스파이
 - 스파이 이미지: 민간인 이미지와 미묘한 차이(객체의 추가/삭제/변형)
 - 단서 단계: 각자 이미지를 기반으로 단서를 제공. 스파이는 오도, 민간인은 정보 노출을 최소화
 - 결정 단계: 민간인은 단서와 이미지를 분석해 스파이를 식별

2.1 환경 및 데이터

- Vision-Zero는 전략적 게임 환경, 레이블 없는 입력, 반복적 SPO의 세 가지 축으로 구성 됨

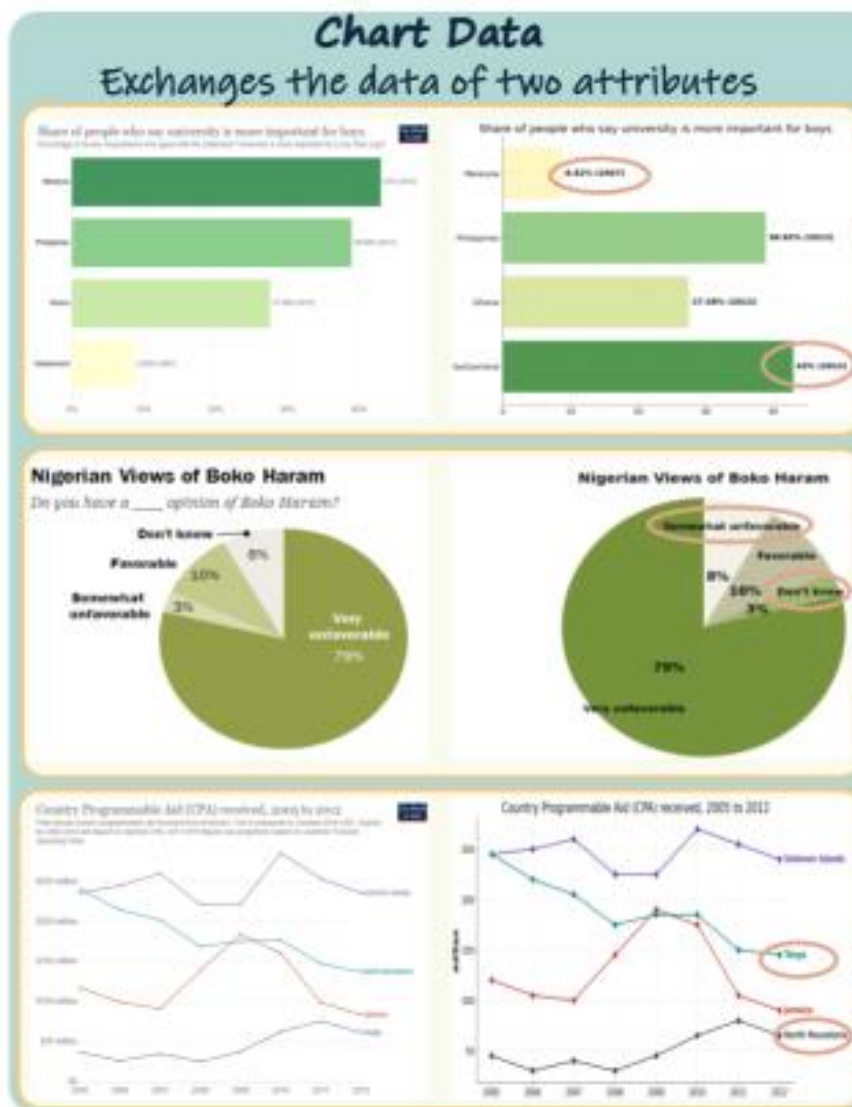
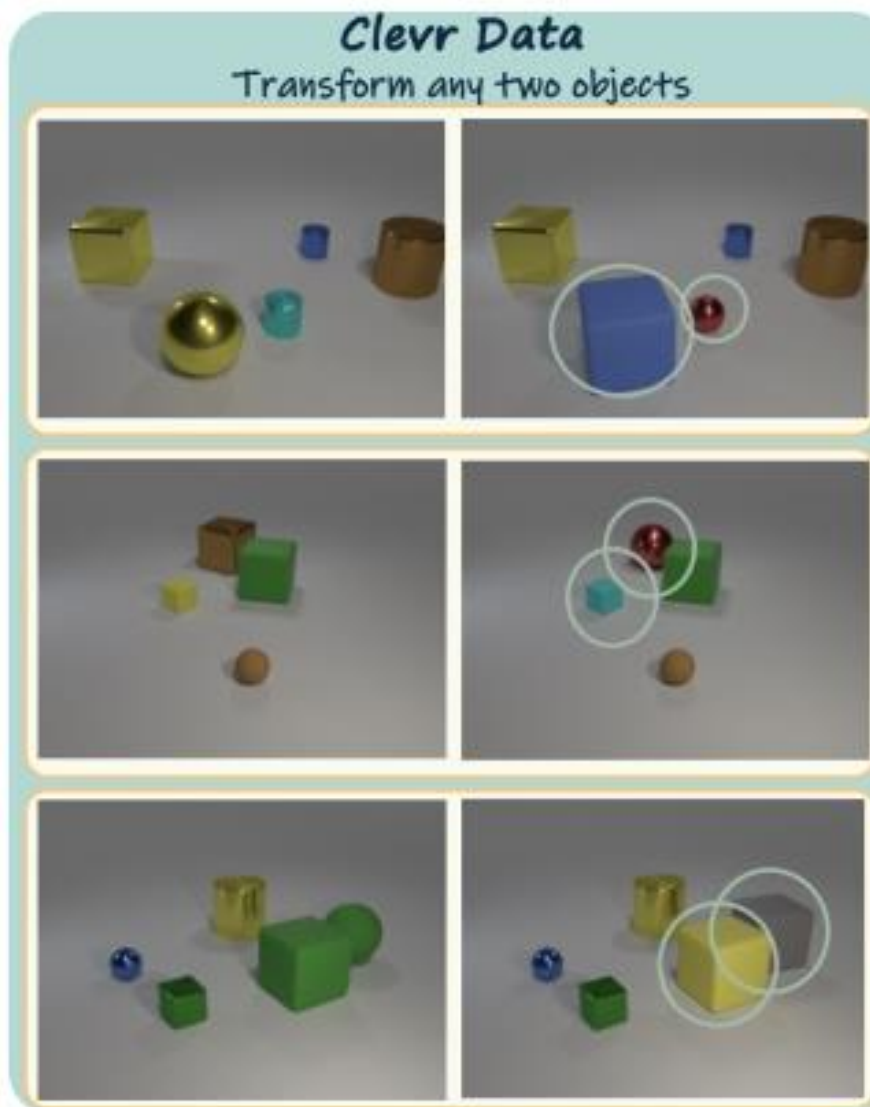


2.1 환경 및 데이터


- Vision-Zero는 전략적 게임 환경, 레이블 없는 입력, 반복적 SPO의 세 가지 축으로 구성 됨
 - 전략적 게임 환경: 역할별 전략적 행동을 요구하여 복합적 능력 발현
 - 레이블 없는 도메인 독립 입력: 임의의 이미지 입력을 지원해 다양성과 일반화 확보
 - Iterative-SPO:
 - 단서 단계 – 수신 투표 수에 반비례한 제로섬 보상(Self-Play)
 - 결정 단계 – 투표 정확성 기반 RLVR + 그룹 정규화

2.1 환경 및 데이터

- 핵심 메시지: Vision-Zero는 라벨이 없는 이미지 쌍만으로도 다양한 도메인에서 학습 가능
 - N_c : 민간인에게 원본 이미지
 - N_s : 스파이에게 수정된 이미지

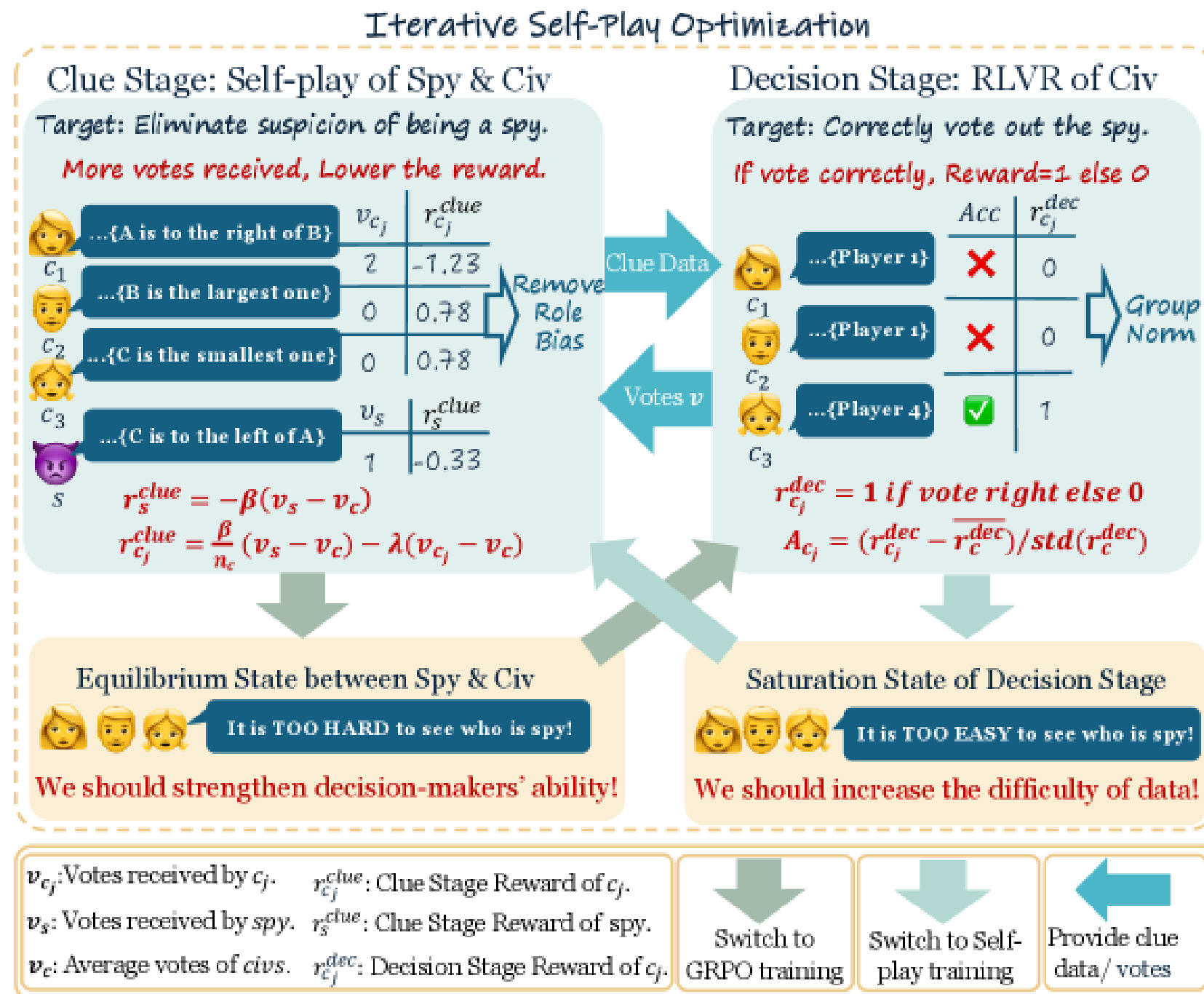


2.1 환경 및 데이터

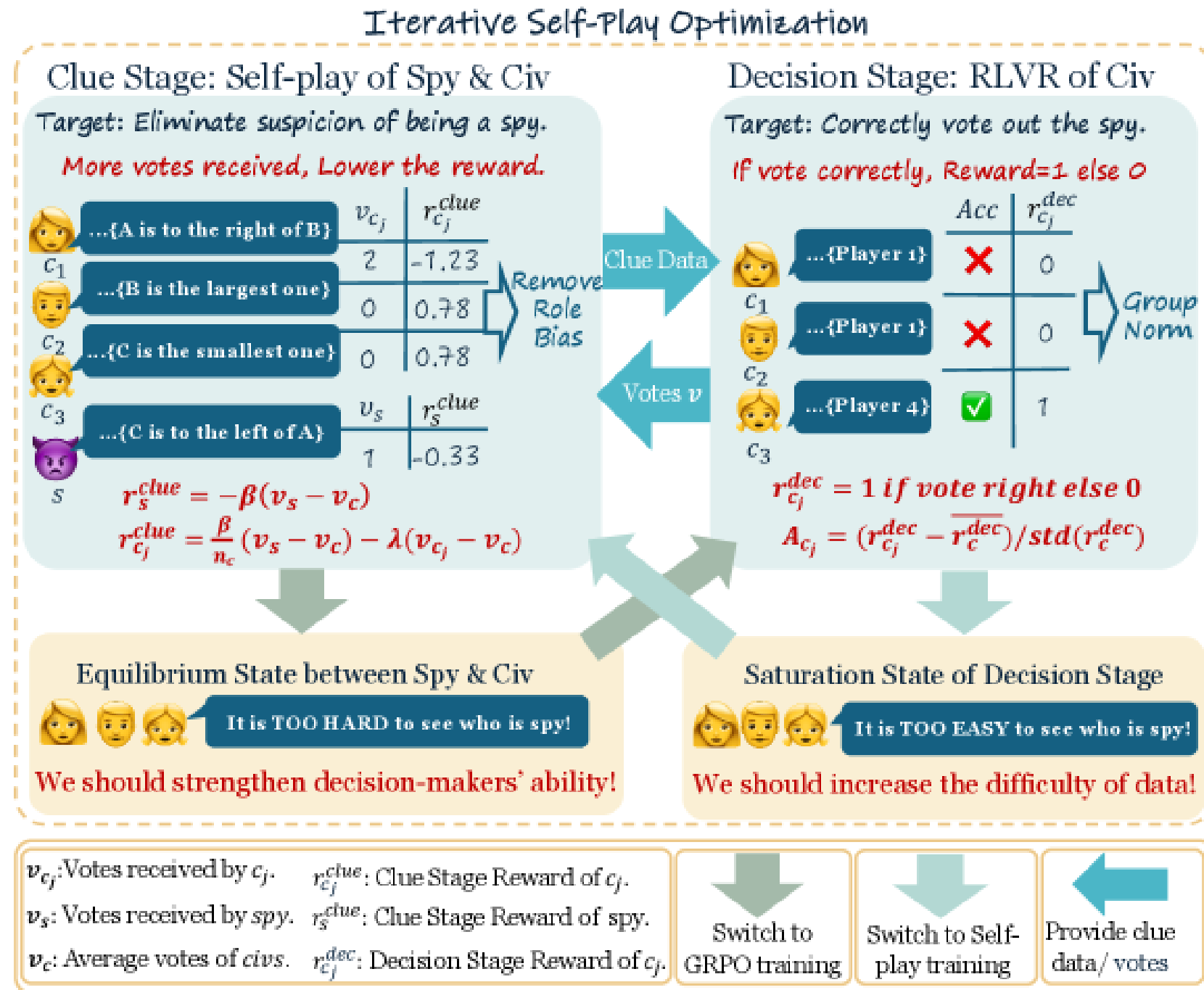
- 핵심 메시지: Vision-Zero는 라벨이 없는 이미지 쌍만으로도 다양한 도메인에서 학습 가능
 - N_c : 민간인에게 원본 이미지
 - N_s : 스파이에게 수정된 이미지
- 임의의 입력을 수용 → 다양한 도메인에 확장 가능
-  3가지 대표 데이터셋:
 - CLEVR: 자동 렌더링 (객체 색상·형태 변형)
 - 차트 데이터: ChartQA 기반 수치 변환
 - 실제 데이터: ImgEdit 편집 쌍 기반
 - 최근 이미지 편집 모델 발전 덕분에 데이터 생성 비용도 매우 저렴

2.2 Iterative Self-Play Policy Optimization

- 지속적인 성능 향상을 위해 Self-Play와 RLVR을 번갈아 수행하는 새로운 최적화 알고리즘을 제안



2.2 Iterative Self-Play Policy Optimization



- Self-Play는 전략적 사고를, RLVR은 검증 가능한 감독 신호를 제공
- 두 단계를 교차적으로 반복하여 지속적 학습.안정성.탐색성을 확보

2.2 Iterative Self-Play Policy Optimization

- 1단계 – 단서(Clue) 단계의 Self-Play 최적화

- 목표: 의심을 피하면서 추론 능력 향상
- Zero-sum 보상 설계: 스파이 \leftrightarrow 민간인의 경쟁 구조

$$r_s^{clue} = -\beta (v_s - v_c), r_{c_j}^{clue} = \frac{\beta}{n_c} (v_s - v_c) - \lambda (v_{c_j} - v_c), \quad j = 1, \dots, n_c.$$

- Role-Advantage Estimation (RAE) 적용으로 역할 정보 비대칭 완화

$$b_s = \alpha b_s + (1 - \alpha) r_s^{clue}, \quad b_c = \alpha b_c + (1 - \alpha) \frac{1}{n_c} \sum_{j=1}^{n_c} r_{c_j}^{clue}, \quad A_k^{clue} = r_k^{clue} - b_k, k \in \mathcal{K}$$

- KL regularization을 통해 정책 안정성 확보

$$\mathcal{L}^{clue}(\theta) = -\mathbb{E} \left[\frac{1}{n} \sum_{k \in \mathcal{K}} A_k^{clue} \log \pi_{\theta}^k(u_k | I_k, h) \right] + \tau_{clue} \mathbb{E} \left[\frac{1}{n} \sum_{k \in \mathcal{K}} D_{\text{KL}}(\pi_{\theta}^k \| \pi_{ref}^k) \right].$$

2.2 Iterative Self-Play Policy Optimization

- 2단계 – 결정(Decision) 단계의 RLVR 학습
 - 목표: 스파이 정확히 식별하기
 - 이산 보상 설계: 맞추면 +1, 모르겠으면 -0.5, 틀리면 -1

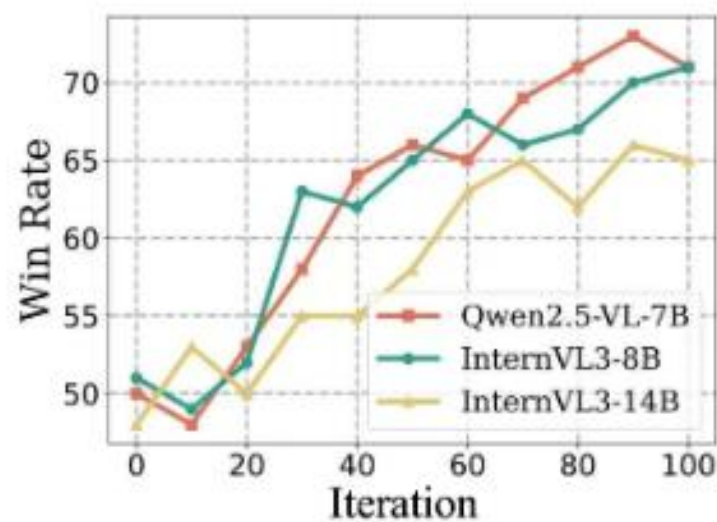
$$r_{c_t}^{dec} = +1 \text{ if } \hat{s}_{c_t} = s^*, -0.5 \text{ elif } \hat{s}_{c_t} = \emptyset, -1 \text{ else.}$$

- “n/a” 선택을 허용해 불확실성 인식 능력까지 학습
- Group Normalization을 통해 라운드 난이도 편차 제거
- KL 제약을 포함한 GRPO 목적 함수로 안정적인 학습

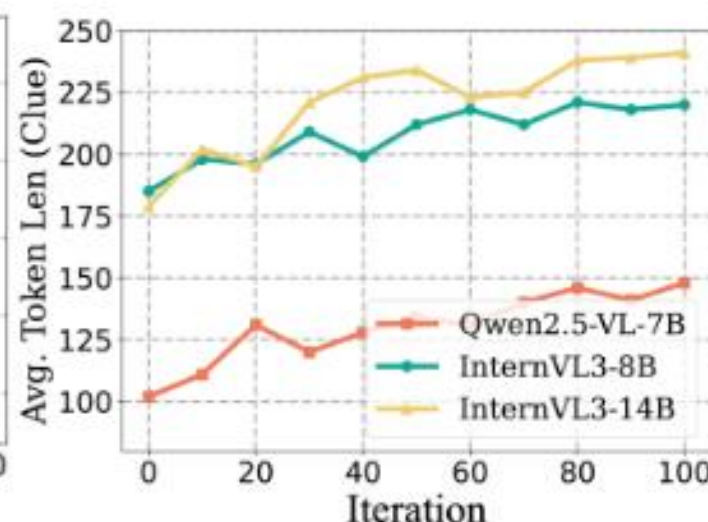
$$\mathcal{L}^{dec}(\theta) = -\mathbb{E}\left[\frac{1}{n_c} \sum_{i=1}^{n_c} A_{c_i}^{dec} \log q_{\theta}(\hat{s}_{c_i} | H)\right] + \tau_{dec} \mathbb{E}\left[\frac{1}{n_c} \sum_{i=1}^{n_c} D_{KL}(q_{\theta}(\cdot | H) \parallel q_{ref}(\cdot | H))\right].$$

2.2 Iterative Self-Play Policy Optimization

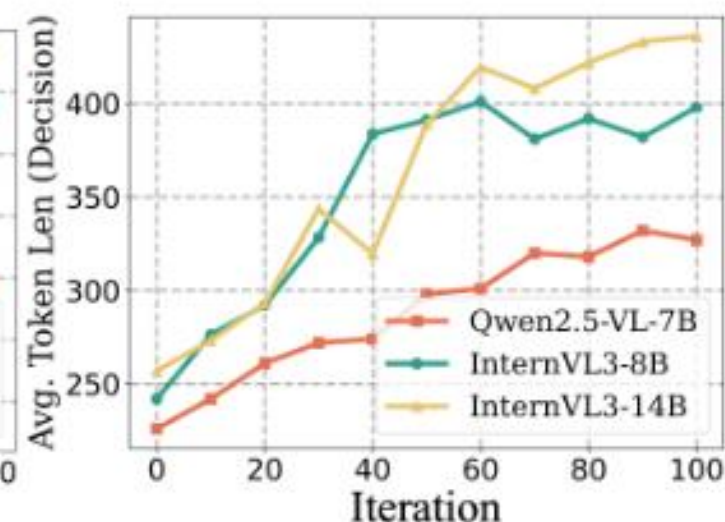
- Iterative-SPO의 효과: 안정적 학습과 성능 향상
 - 교대 학습을 통해 전략적 균형 정책과 지식 포화 모두 해결
 - RLVR의 감독 신호 덕분에 role collapse나 발산(divergence) 방지
 - Reasoning·Math 과제에서 기존 SOTA를 능가하는 성능 달성
 - Win Rate 및 Token 길이 변화(Fig. 6) → 학습 과정에서의 전략 복잡도 및 표현력 증가 시각화



(a) Winning Rate



(b) Avg. Token Length (Clue)



(c) Avg. Token Length (Decision)

2.2 Iterative Self-Play Policy Optimization

- VLMEvalKit에서 평가된 추론 및 수학에 대한 Vision-Zero 모델과 SOTA 모델의 성능 비교

Method	MathVista	MathVision	WeMath	MathVerse	LogicVista	DynaMath	Avg.
<i>Proprietary Model</i>							
[2pt/2pt] GPT4o	61.4	30.4	40.0	50.2	45.9	32.3	43.4
Gemini2.0-Flash	73.4	41.3	57.1	54.4	56.2	43.7	54.4
<i>Performance on Qwen2.5-VL-7B</i>							
[2pt/2pt] Qwen2.5-VL-7B	68.2	25.4	36.1	49.0	47.2	20.9	41.1
[2pt/2pt] R1-OneVision-7B	64.1	24.1	35.8	47.1	44.5	21.4	39.5
MM-Eureka-Qwen-7B	73.0	26.9	36.2	50.3	42.9	24.2	42.9
VLAA-Thinker-7B	68.0	26.4	36.0	51.7	47.2	21.9	41.9
OpenVLThinker-7B	70.2	25.3	36.5	47.9	44.3	21.2	40.9
ViGaL-Snake	70.7	26.5	–	51.1	–	–	–
ViGaL-Rotation	71.2	26.3	–	50.4	–	–	–
ViGaL-Snake+Rotation	71.9	27.5	36.9	52.4	46.5	22.9	43.0
VisionZero-Qwen-7B (CLEVR)	72.6	28.1	39.8	51.9	50.1	22.3	44.1
VisionZero-Qwen-7B (Chart)	72.2	27.6	39.2	52.1	50.6	21.9	43.9
VisionZero-Qwen-7B (Real-World)	72.4	28.0	39.5	52.2	50.3	22.1	44.1

2.2 Iterative Self-Play Policy Optimization

Algorithm 1 Iterative Self-Play Policy Optimization(Iterative-SPO)

```

1: Role set  $\mathcal{K} = \{\text{spy}\} \cup \{c_1, \dots, c_{n_c}\}$ ; reference policies  $\pi_{\text{ref}}^k$ ; hyperparams  $\beta, \lambda, \alpha, \tau_{\text{clue}}, \rho, \tau_{\text{acc}}^{\dagger}, \tau_{\text{err}}^{\dagger}, \tau_{\text{na}}^{\dagger}, \tau_{\text{na}}^{\dagger}, K_{\text{min}}, P$ ; learning rates  $\eta_{\theta}, \eta_{\theta'}$ .
2: Init RAE  $b_s \leftarrow 0, b_{\text{clv}} \leftarrow 0$ ; Stage switch metrics  $\text{acc} \leftarrow 0, \text{na} \leftarrow 0$ ; Stage  $m \leftarrow 0$  (Decision).
3: for  $t = 1, \dots, T$  do
4:   if  $m = 1$  then ▷ CLUE Stage
5:     Each player gives clue  $u_k \sim \pi_{\theta}^k(\cdot | I_k, h)$  based on the historical dialogue  $h$  and input picture  $I_k$ .
6:     Obtain votes from the decision stage  $v = (v_s, v_{c_1}, \dots, v_{c_{n_c}})$  and  $v_c \leftarrow \frac{1}{n_c} \sum_{j=1}^{n_c} v_{c_j}$ .
7:     Zero-Sum Rewards:  $r_s^{\text{clue}} \leftarrow -\beta(v_s - v_c); r_{c_j}^{\text{clue}} \leftarrow \frac{\beta}{n_c}(v_s - v_c) - \lambda(v_{c_j} - v_c)$  for  $j = 1, \dots, n_c$ .
8:     Role Advantage Estimation:  $b_s \leftarrow \alpha b_s + (1 - \alpha)r_s^{\text{clue}}, b_{\text{clv}} \leftarrow \alpha b_{\text{clv}} + (1 - \alpha)\frac{1}{n_c} \sum_j r_{c_j}^{\text{clue}}$ .
9:     RAE-based Advantages:  $A_s^{\text{clue}} \leftarrow r_s^{\text{clue}} - b_s; A_{c_j}^{\text{clue}} \leftarrow r_{c_j}^{\text{clue}} - b_{\text{clv}}$  for  $j = 1, \dots, n_c$ .
10:   else DECISION Stage ▷
11:     Each citizen casts vote  $\hat{s}_{c_i} \sim q_{\theta}(\cdot | H)$  based on the clue information  $H$  and the input image  $I_k$ .
12:     Reward:  $r_{c_i}^{\text{dec}} \leftarrow 1$  if  $\hat{s}_{c_i} = s^*$  (correct);  $r_{c_i}^{\text{dec}} \leftarrow -0.5$  if  $\hat{s}_{c_i} = \emptyset$  (unsure);  $r_{c_i}^{\text{dec}} \leftarrow -1$  else (wrong).
13:     Group-norm Advantage:  $A_{c_i}^{\text{dec}} = (r_{c_i}^{\text{dec}} - \mu_r) / (\sigma_r + \epsilon)$ 
14:     Policy update: Apply KL-regularized policy gradient as Eq. 3 or Eq. 6 to update  $\pi_{\theta}$  or  $q_{\theta}$ .
15:
16:   Stage Switch: Calculate average prediction accuracy  $\text{acc}_t$  and "n/a" rate  $\text{na}_t$  of players in the decision stage within a batch round:

$$\text{acc}_t = \frac{1}{B} \sum_i \mathbf{1}[\arg \max_y q_{\theta}(y | H_i) = s_i^*], \text{na}_t = \frac{1}{B} \sum_i q_{\theta}(\emptyset | H_i).$$

17:   Update EMAs  $\text{acc} \leftarrow \rho \text{acc} + (1 - \rho) \text{acc}_t; \text{na} \leftarrow \rho \text{na} + (1 - \rho) \text{na}_t, d \leftarrow d + 1$ .
18:   if  $m = 0$  and  $\text{acc} \geq \tau_{\text{acc}}^{\dagger}$  and  $\text{na} \leq \tau_{\text{na}}^{\dagger}$  and  $d \geq K_{\text{min}}$  then  $m \leftarrow 1, d \leftarrow 0$ ;
19:   if  $m = 1$  and  $(1 - \text{acc} \geq \tau_{\text{err}}^{\dagger} \text{ or } \text{na} \geq \tau_{\text{na}}^{\dagger})$  and  $d \geq K_{\text{min}}$  then  $m \leftarrow 0, d \leftarrow 0$ ;
20: return  $\theta, \theta'$ 

```


2.3 Advantage Analysis

- Vision-Zero의 주요 장점 (Advantage Analysis)
 - ① 도메인 비의존적 데이터 수용성
 - 이미지 차이를 활용하여 기존 고품질 데이터셋을 그대로 사용 가능
 - 다양한 입력에도 일반화 성능 확보
 - ② 시각.언어 통합 분석
 - 공간 관계 및 객체 특성을 동시에 고려 → 추론, 시각 이해, OCR 동시 향상
 - 텍스트 슛컷 편향 및 음의 전이 문제 완화
 - ③ 고효율 데이터 생성 전략
 - ChatGPT, NanoBanana 등의 편집 도구로 데이터셋을 빠르고 저렴하게 생성
 - 전통적 수작업 라벨링 대비 비용 절감 및 실용화 가속

3. Experiments

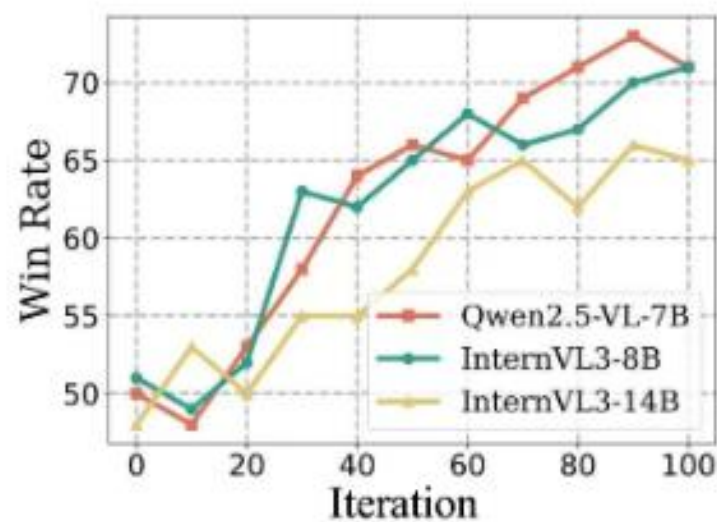
- Vision-Zero의 성능과 일반화를 검증하기 위해 다양한 모델, 과제, 그리고 강력한 SOTA 모델들과의 공정한 비교를 수행
 - **평가 모델:** Qwen2.5-VL-7B, InternVL3-8B, InternVL3-14B
 - **평가 과제:** 14개 과제 (Reasoning, Chart Analysis, Vision-Centric)
 - **비교 대상:** R1-OneVision, MM-Eureka, VLAA-Thinker, OpenVLThinker, ViGaL 등 최신 SOTA
 - **평가 환경:** 모든 비교는 **VLMEvalKit**에서 동일 조건으로 수행

3. Experiments

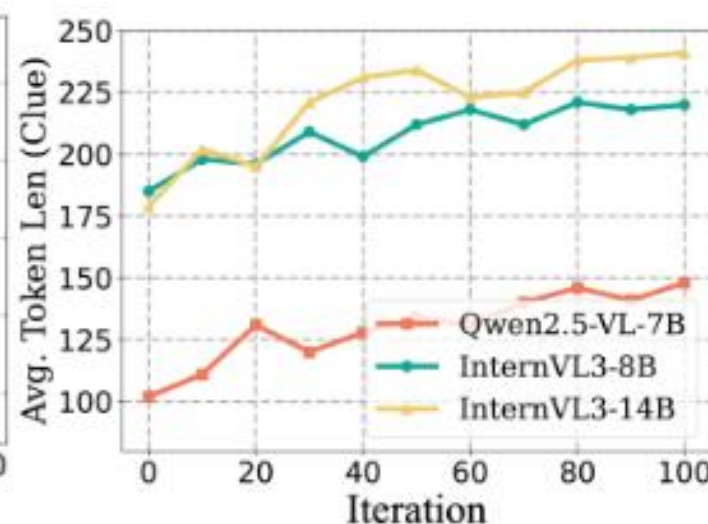
- Vision-Zero는 작은 데이터, 간단한 설정만으로도 안정적인 학습과 강력한 일반화 성능을 달성
 - **학습 구성:** 각 라운드에 민간인 $n_c = 4$, 단서 단계 발화 2회
 - **보상 파라미터:** $\beta = \lambda = 0.1$, RAE 감쇠 계수 $\alpha = 0.95$
 - **전환 조건:** $\tau_{acc}^{\uparrow} = 0.9$, $\tau_{err}^{\uparrow} = 0.4$, $\tau_{na}^{\uparrow} = 0.5$, $\tau_{na}^{\downarrow} = 0.1$, 최소 단계 유지 $K_{min} = 5$
 - **학습 환경:** 총 100 iteration, batch size = 1024, VLM-R1 프레임워크 기반
 - **일반화 평가:** InternVL3 모델은 **CLEVR** 데이터만 학습했음에도 다양한 과제에서 높은 성능 유지

3.1 Main Results

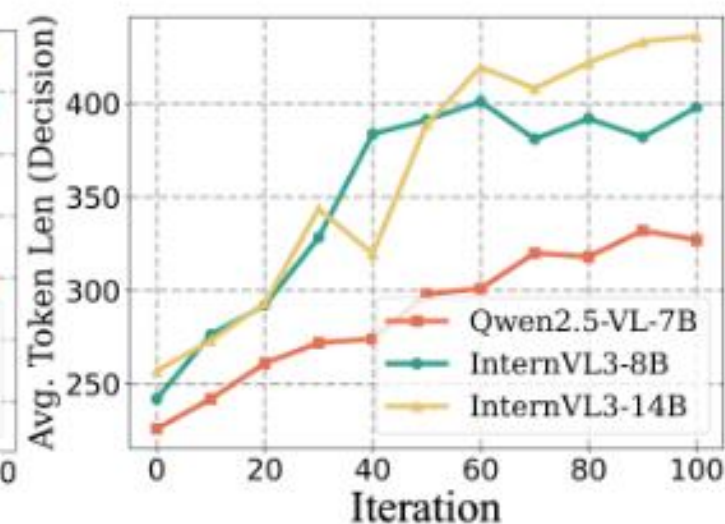
- 지속적 성능 향상: Iterative-SPO의 효과
 - Vision-Zero는 훈련이 진행될수록 승률(Win Rate)이 꾸준히 증가
 - Decision 단계에서 평균 토큰 길이 250 → 약 400으로 증가 → 더 복잡한 추론 전략 학습 의미**
- Iterative-SPO가 단순 성능 향상을 넘어 **추론 능력의 질적 성장**을 이끈다는 증거



(a) Winning Rate



(b) Avg. Token Length (Clue)



(c) Avg. Token Length (Decision)

3.1 Main Results

- 수학·추론 과제에서의 일반화 성능

Method	MathVista	MathVision	WeMath	MathVerse	LogicVista	DynaMath	Avg.
<i>Proprietary Model</i>							
[2pt/2pt] GPT4o	61.4	30.4	40.0	50.2	45.9	32.3	43.4
Gemini2.0-Flash	73.4	41.3	57.1	54.4	56.2	43.7	54.4
<i>Performance on Qwen2.5-VL-7B</i>							
[2pt/2pt] Qwen2.5-VL-7B	68.2	25.4	36.1	49.0	47.2	20.9	41.1
[2pt/2pt] R1-OneVision-7B	64.1	24.1	35.8	47.1	44.5	21.4	39.5
MM-Eureka-Qwen-7B	73.0	26.9	36.2	50.3	42.9	24.2	42.9
VLAA-Thinker-7B	68.0	26.4	36.0	51.7	47.2	21.9	41.9
OpenVLThinker-7B	70.2	25.3	36.5	47.9	44.3	21.2	40.9
ViGaL-Snake	70.7	26.5	–	51.1	–	–	–
ViGaL-Rotation	71.2	26.3	–	50.4	–	–	–
ViGaL-Snake+Rotation	71.9	27.5	36.9	52.4	46.5	22.9	43.0
VisionZero-Qwen-7B (CLEVR)	72.6	28.1	39.8	51.9	50.1	22.3	44.1
VisionZero-Qwen-7B (Chart)	72.2	27.6	39.2	52.1	50.6	21.9	43.9
VisionZero-Qwen-7B (Real-World)	72.4	28.0	39.5	52.2	50.3	22.1	44.1

3.1 Main Results

- 음의 전이(Negative Transfer) 문제 해결

Table 2: **Performance comparison between Vision-Zero and other state-of-the-art models on Chart/OCR and Vision-Centric benchmarks.** All models are evaluated using the open-source platform VLMEvalKit. Additional results on related datasets are provided in the Appendix A.4.

[2pt/2pt] Model	Chart / OCR				Vision-Centric			
	AI2D	ChartQA	OCR Bench	SEED-2	RealWorldQA	MMVP	BLINK	MuirBench
<i>Proprietary Model</i>								
[2pt/2pt] GPT-4o	84.4	85.7	73.9	72.0	75.4	86.3	68.0	68.0
Gemini2.0-Flash	87.2	79.3	85.5	71.2	73.2	83.0	63.5	64.6
<i>Performance on Qwen2.5-VL-7B</i>								
[2pt/2pt] Qwen2.5-VL-7B	84.7	86.1	88.3	70.4	68.1	76.8	55.2	58.2
[2pt/2pt] R1-OneVision-7B	82.2	–	81.0	66.4	58.0	61.3	48.7	46.3
MM-Eureka-Qwen-7B	84.1	77.3	86.7	68.2	66.1	74.3	54.0	61.1
VLAA-Thinker-7B	84.0	84.3	86.9	67.4	65.4	71.6	53.0	57.1
OpenVLThinker-7B	81.8	–	83.3	68.0	60.2	32.3	49.9	52.8
ViGaL-Snake+Rotation	84.5	79.9	86.8	69.1	66.5	74.6	55.6	57.8
VisionZero-Qwen-7B (CLEVR)	84.5	86.3	88.1	69.5	68.5	79.2	56.1	58.6
VisionZero-Qwen-7B (Chart)	85.8	87.2	89.0	70.9	68.2	77.9	57.2	59.4
VisionZero-Qwen-7B (Real-World)	84.8	86.3	88.5	69.8	68.5	79.5	57.5	59.8

3.1 Main Results

- 데이터셋 구축 비용 및 효율성

Table 3: **Comparison of dataset construction costs across methods.** Methods like R1-OneVision-7B use programmatic question-answer generation over real images with manual verification, taking months to a year. ViGaL collects gameplay data from two environments (Snake and CLEVR-based orientation game) over several weeks. In contrast, Vision-Zero employs simple image editing, using each image pair throughout entire game rounds, significantly reducing required samples. Moreover, since most baselines are trained primarily on pure reasoning and mathematical tasks, some models actually exhibit performance degradation on the comprehensive benchmark MMMU.

	Data Cost				Training		Performance	
Method	Data Type	Num	Prepare Method	Cost	Method	Interact	MMMu	MMMu _{pro}
Qwen2.5-VL-7B	–	–	–	–	–	–	54.3	37.0
[2pt/2pt] R1-OneVision-7B	Real-World Data	155k	Programmatic construction with human checks.	A few months	SFT+GRPO	X	51.9	32.6
VLAA-Thinker-7B		25k			SFT+GRPO	X	48.2	31.9
OpenVLThinker-7B		12k			SFT+GRPO	X	54.8	22.1
MM-Eureka-Qwen-7B		15k			GRPO	X	55.8	36.9
[2pt/2pt] ViGaL-Snake	Synthetic Data	72k	Collected in game environment via PPO policy	A few weeks	RLOO	X	55.8	36.6
ViGaL-Rotation							54.1	37.7
ViGaL-Snake+Rotation							58.0	37.4
VisionZero-Qwen-7B (CLEVR)	Synthetic	2k	Batch render scenes	≈ 6 GPUh	Alternating Self-play+ GRPO	✓	58.8	37.7

3.2 Ablation Studies

- 수학·추론 과제에서의 일반화 성능

Table 4: **Model generalizability of Vision-Zero.** We train InternVL3-8B and InternVL3-14B within the Vision-Zero using the CLEVR-based dataset, and evaluate on eight reasoning benchmarks.

Model	MathVista	MathVision	WeMath	MathVerse	LogicVista	DynaMath	Avg.
<i>Performance on InternVL3-8B</i>							
[2pt/2pt] InternVL3-8B	60.4	21.3	26.8	32.2	40.5	26.8	34.7
VisionZero-InternVL3-8B	62.2	24.2	28.7	32.9	41.8	29.2	36.5
<i>Performance on InternVL3-14B</i>							
[2pt/2pt] InternVL3-14B	74.1	33.8	42.3	43.3	51.6	30.1	45.8
VisionZero-InternVL3-14B	75.4	34.8	44.9	45.1	53.1	31.3	47.4

- Table 4: InternVL3-8B, 14B에서도 범용적 성능 향상 확인 → 일반화 능력 검증

3.2 Ablation Studies

- Iterative-SPO의 우수성

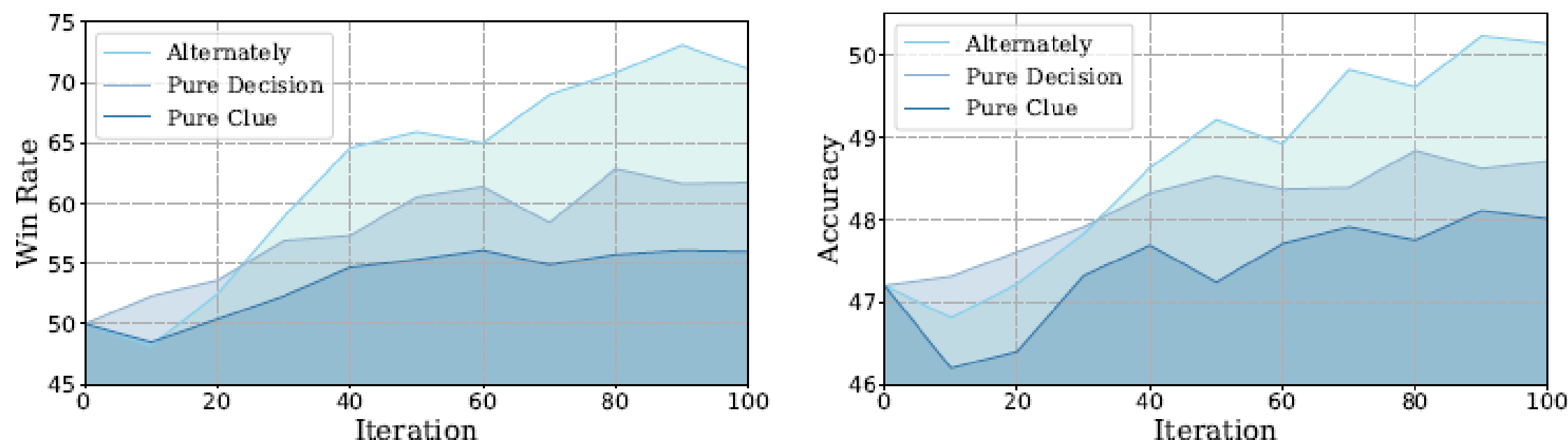


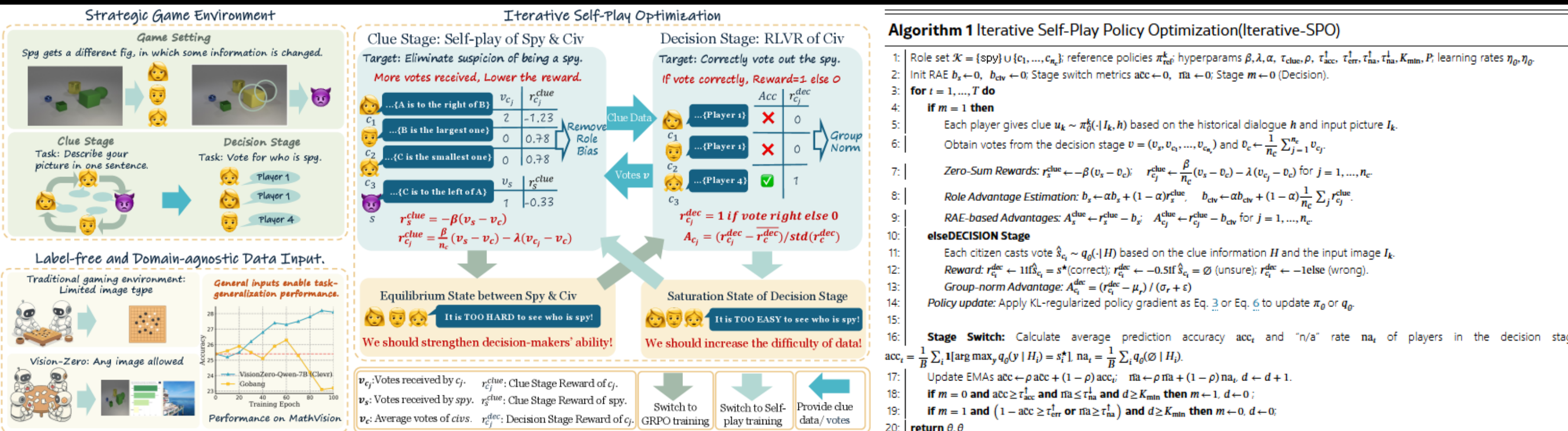
Figure 7: **Performance Comparison between Iterative-SPO and pure Self-play / pure RLVR training.** (left) Winning Rate (right) Performance on LogicVista. We evaluate under three settings: (1) Iterative-SPO; (2) Pure Decision: Clue stage frozen, training only Decision stage via RLVR; (3) Pure Clue: Decision stage frozen, training only Clue stage via Self-Play.

- 단일 모드(Self-Play 또는 RLVR)보다 지속적 성능 향상 달성
- Pure Self-Play: 검증 가능한 보상 부족 → 조기 수렴
- Iterative-SPO: 이를 극복하여 LogicVista에서 +2% / +1% 추가 향상

4. Conclusion

- Vision-Zero: **최초의 Zero-Human-in-the-Loop Self-Improvement 프레임워크 제안**
- **전략적 Self-Play 환경 + 도메인 비의존적 입력으로 학습 확장성 확보**
- **Iterative-SPO: Self-Play와 RLVR을 번갈아 수행하여 안정성과 성능 향상 동시 달성**
- Reasoning, Chart/OCR, Vision-centric 과제에서 **SOTA 성능 및 범용성 확보**
- 데이터 구축 비용을 획기적으로 절감하며 **경제적이고 실용적인 학습 패러다임 제시**

Vision-Zero : Scalable VLM Self-Improvement via Strategic Gamified Self-Play



- Vision-Zero는 인간 주석 없이도 VLM을 자가 개선하는 전략 게임 기반 Self-Play 학습 프레임워크
- Iterative-SPO를 통해 Self-Play와 RLVR을 번갈아 학습하여 지속적인 성능 향상과 데이터 비용 절감을 달성
- 기존 방법 대비 데이터 비용은 줄이면서도 reasoning·OCR·vision tasks 전반에서 SOTA 달성
- 다양한 이미지 쌍을 활용한 게임 환경은 도메인 특화 없이 일반화 능력을 강화하며, 새로운 추론 전략의 자발적 학습을 가능하게 함
- 또한 복수 능력을 통합적으로 향상시켜 단일 태스크 학습에서 발생하는 음의 전이 문제를 효과적으로 완화

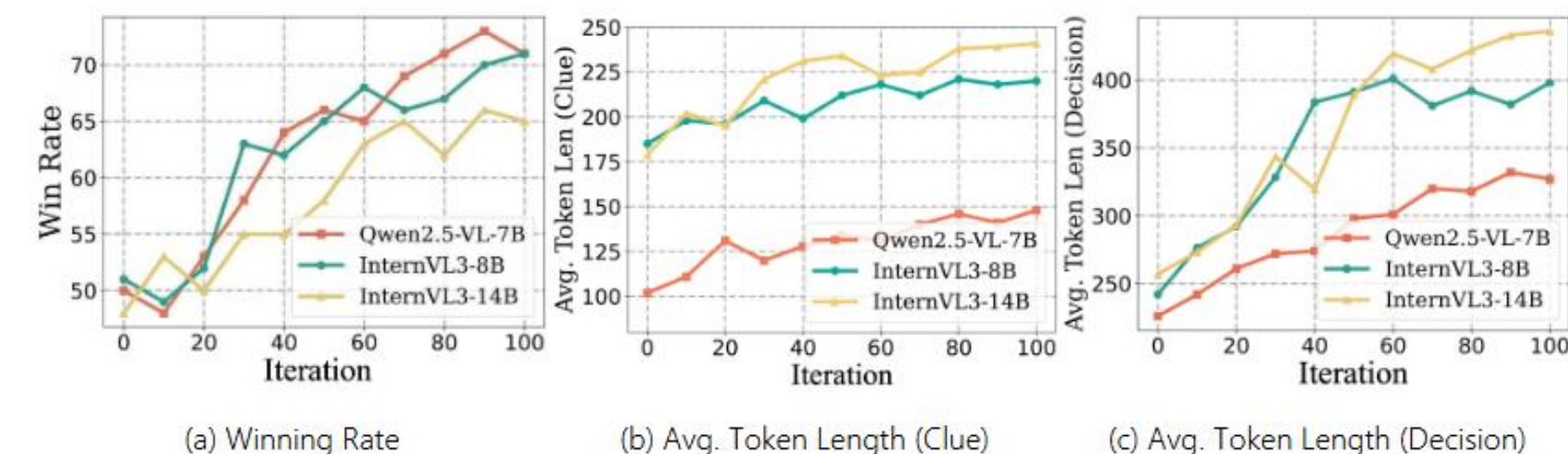


Table 3: Comparison of dataset construction costs across methods. Methods like R1-OneVision-7B use programmatic question-answer generation over real images with manual verification, taking months to a year. ViGaL collects gameplay data from two environments (Snake and CLEVR-based orientation game) over several weeks. In contrast, Vision-Zero employs simple image editing, using each image pair throughout entire game rounds, significantly reducing required samples. Moreover, since most baselines are trained primarily on pure reasoning and mathematical tasks, some models actually exhibit performance degradation on the comprehensive benchmark MMMU.

Method	Data Cost			Training		Performance	
	Data Type	Num	Prepare Method	Method	Interact	MMMu	MMMu _{pro}
Qwen2.5-VL-7B	—	—	—	—	—	54.3	37.0
[2pt/2pt] R1-OneVision-7B	Real-World Data	155k	Programmatic construction with human checks.	SFT+GRPO	X	51.9	32.6
VLAA-Thinker-7B		25k		SFT+GRPO	X	48.2	31.9
OpenVLThinker-7B		12k		SFT+GRPO	X	54.8	22.1
MM-Eureka-Qwen-7B		15k		GRPO	X	55.8	36.9
[2pt/2pt] ViGaL-Snake	Synthetic Data	72k	Collected in game environment via PPO policy	RLOO	X	55.8	36.6
ViGaL-Rotation						54.1	37.7
ViGaL-Snake+Rotation						58.0	37.4
VisionZero-Qwen-7B (CLEVR)	Synthetic	2k	Batch render scenes	Alternating Self-play+GRPO	✓	58.8	37.7