## - 2026-02-17

: [RLVR], [ ], [Agent], [Benchmark]

---

## 1. Experiential Reinforcement Learning

: Experiential Reinforcement Learning : Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, Jieyu Zhao : Microsoft **ArXiv ID**: 2602.13949 : [ ], [ ]

: $T$ RL $r_t$ $t$ RL $\mathbf{E}_\tau[\sum_{t=0}^{T}\gamma^t r_t]$
$r_t$ —— (credit assignment)

: RLHF → DPO → GRPO → RLVR self-reflection RL
" " ____ " "

: - reflection - initial attempt → refined attempt

: - - reflection

: RL Hindsight Experience Replay (HER) +81% +11%

1. **Experience-Reflection-Consolidation** : single-step RL
2. **Structured Behavioral Revision**: " " reflection
3. : consolidation policy reflection

- : PPO, GRPO, DPO RL
- : Sparse-reward control environments, agentic reasoning benchmarks
- :

reflection

: RL " " RL " " ERL " " reflection

: Self-reflection RL CoT

- : agentic benchmark LLM
- :
- : Implicit REINFORCE

---

## 2. Embed-RL: Reinforcement Learning for Reasoning-Driven Multimodal Embeddings

: Embed-RL: Reinforcement Learning for Reasoning-Driven Multimodal Embeddings : Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, Yansong Tang : Tsinghua University **ArXiv ID**: 2602.13823 : [ ], [ ]

: query target multimodal embedding $e$ $sim(e_q, e_t)$ CoT reasoning retrieval

: $\mathcal{L} = -\log \frac{\exp(sim(q,t^+)/\tau)}{\sum \exp(sim(q,t)/\tau)}$ reasoning quality retrieval quality

1. **Traceability CoT (T-CoT)**: reasoning trace multimodal cues
2. **Embedder-Guided RL**: Embedder reward signal Reasoner

: T-CoT retrieval " "

1. **EG-RL** : Embedder reward Reasoner
2. **T-CoT**: multimodal cues retrieval
3. **baseline**

- **Reward** : Embedder reward retrieval
- : multimodal cues
- : E5, BGE

[ ] RL + Multimodal

---

## 3. A Critical Look at Targeted Instruction Selection

: A Critical Look at Targeted Instruction Selection: Disentangling What Matters (and What Doesn't) : Nihal V. Nayak, Paula Rodriguez-Diaz, Neha Hulkund, Sara Beery, David Alvarez-Melis : Harvard DCML **ArXiv ID**: 2602.14696 : [ ], [ ], [ ]

: instruction $\mathcal{D}$ query $\mathcal{Q}$ $\mathcal{D}$ $\mathcal{S}$ LLM $\mathcal{Q}$

: subset $\mathcal{S}$ query $\mathcal{Q}$ " "

$$\min_{\mathcal{S}\subseteq\mathcal{D},|\mathcal{S}|=k} \text{dist}(\mathcal{S},\mathcal{Q})$$

: gradient-based 1. (data representation): semantic embedding vs gradient-based 2. (selection algorithm): greedy, importance sampling, etc.

1. : approximate distance minimization
2. : generalization bound
3. : gradient-based + greedy round-robin

- : instruction selection
- : LLM
- :

[ ] [ ] — " "

- entanglement
-
- zero-shot baseline

---

## 4. WebWorld: A Large-Scale World Model for Web Agent Training

: WebWorld: A Large-Scale World Model for Web Agent Training : Zikai Xiao, Jianhong Tu, Chuhang Zou, Yuxin Zuo, Zhi Li, Peng Wang, Bowen Yu,

Fei Huang, Junyang Lin, Zuozhu Liu : Qwen (Alibaba) **ArXiv ID**: 2602.14721
: [ ], [Agent], [Benchmark]

: Web agent trajectories

: $W$ agent $W$

1. (1M+ interactions)
2. (code, GUI, game)

1. : 1M+ open-web interactions
2. **WebWorld-Bench**: 9 dual metrics
3. : GUI

- : WebWorld comparable to Gemini-3-Pro
- : Qwen3-14B WebArena +9.2%
- : GPT-5

[ ] — + Agent 1M trajectories

| | |
|---|---|
| WebArena | ~10K |
| WebShop | ~2K |
| **WebWorld** | **1M+** |

| | | |
|---|---|---|
| Experiential RL | RLVR | [ ], [ ] |
| Targeted Instruction Selection | | [ ], [ ] |

| | | |
|---|---|---|
| Embed-RL | RL + Multimodal | [ ], [ ] |
| WebWorld | Agent/Benchmark | [ ] |

1. **RLVR** : ERL explicit self-reflection RL RLHF → DPO → GRPO → RLVR
2. : Harvard gradient-based representation instruction selection
3. **Agent** : WebWorld

- DeepSeek/OpenAI RL
- MoRL (Peking)