Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning

RL4VLM@NeurIPS2024 / https://rl4vlm.github.io/

Yuexiang Zhai, Hao Bai*, Zipeng Lin*, Jiayi Pan*, Shengbang Tong*, Yifei Zhou*
Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, Sergey Levine

Berkeley UNIVERSITY OF CALIFORNIA

BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

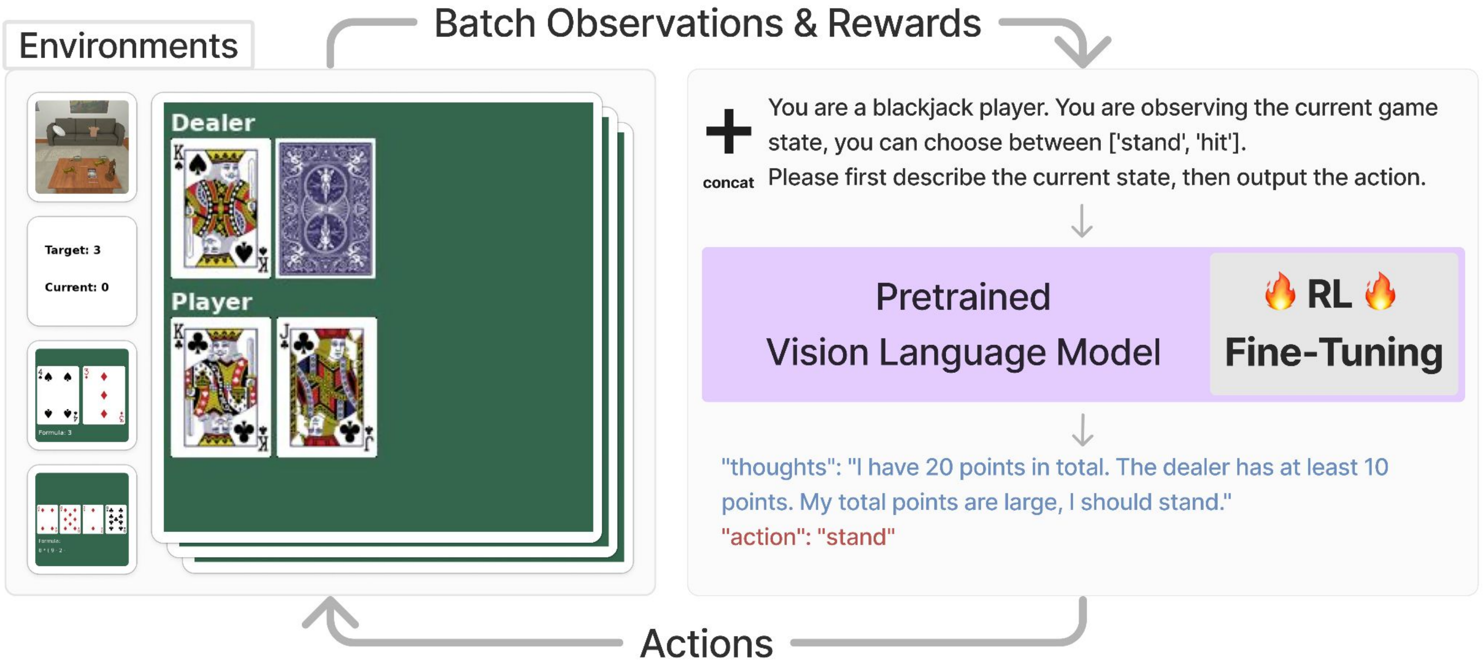
RL4VLM: training generative models as decision-making agents

- Overview
- Vision-language evaluation tasks
- Leveraging domain knowledge
- Results

RL4VLM: training generative models as decision-making agents

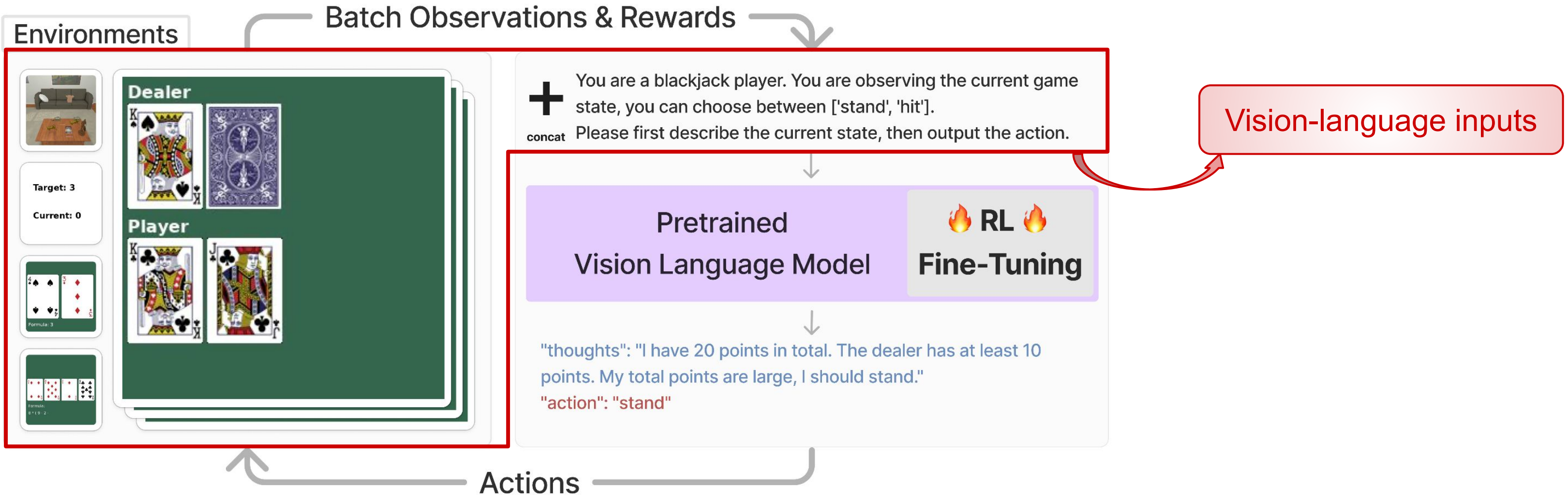
- Overview
- Vision-language evaluation tasks
- Leveraging domain knowledge
- Results

RL4VLM: training generative models as decision-making agents



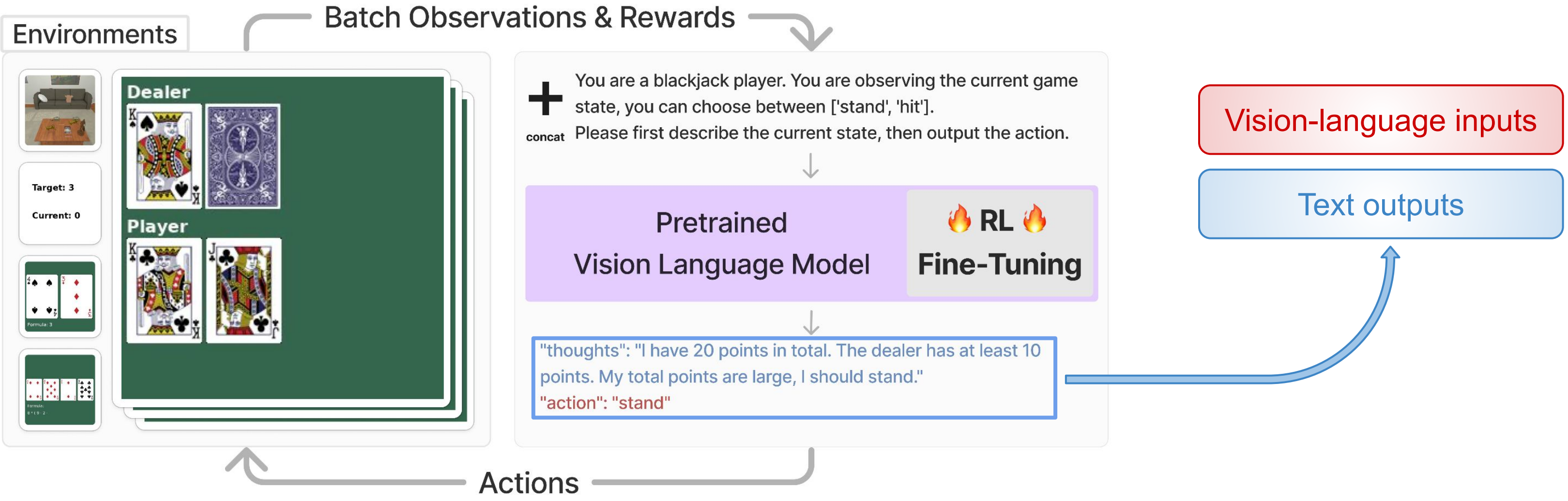
Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents



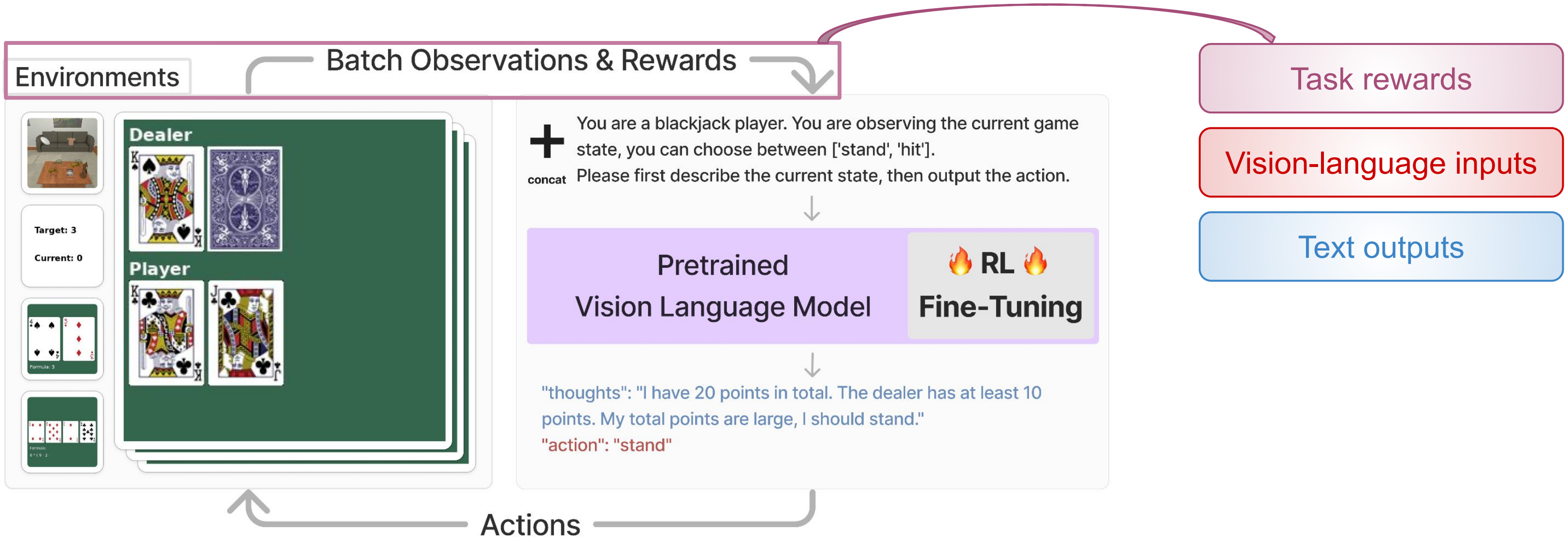
Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents



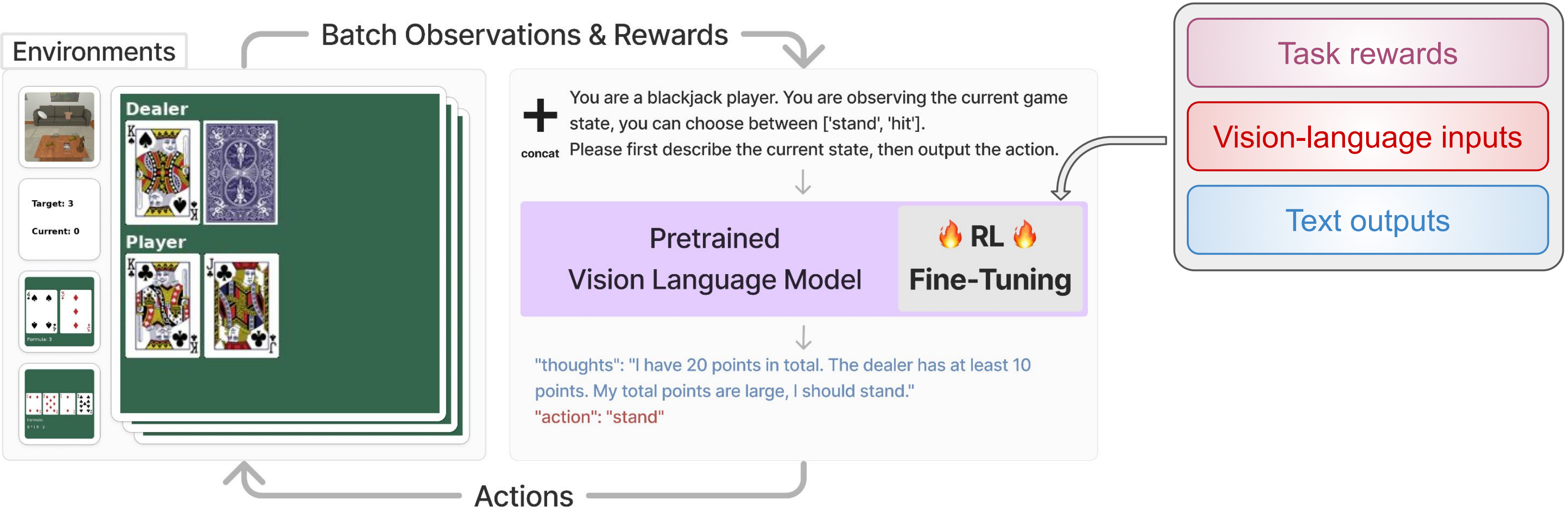
Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents



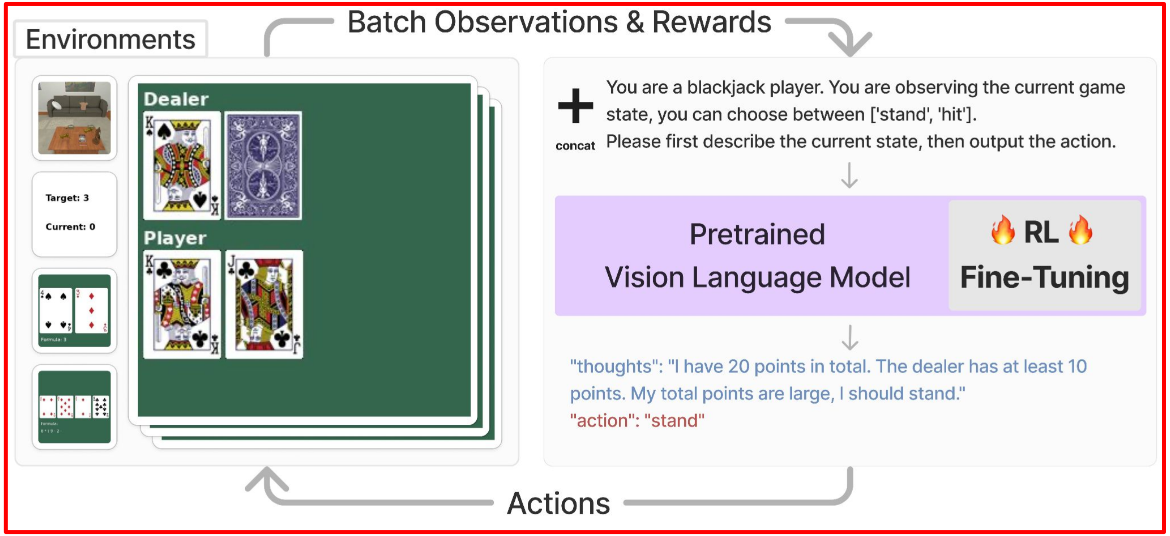
Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents



Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents



Task rewards

Vision-language inputs

Text outputs

End-to-end VLM training with RL

Zhai, Y., Bai, H., Lin, Z., Pan, J., Tong, S., Zhou, Y., Suhr, A., Xie, S., LeCun, Y., Ma, Y., Levine, S.. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. ArXiv 2024.

RL4VLM: training generative models as decision-making agents

- Overview
- Vision-language evaluation tasks
- Leveraging domain knowledge
- Results

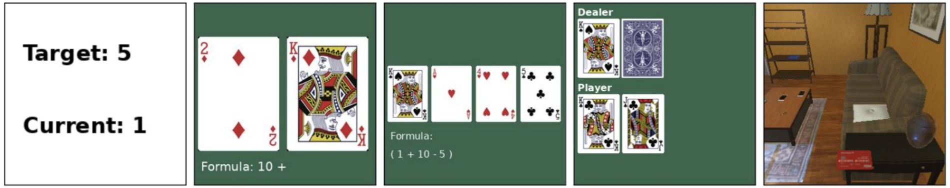
RL4VLM: vision-language based evaluation tasks

- Tasks requiring **fine-grained** visual recognition (a) - (d)
- Tasks requiring **visual semantic** reasoning (e)

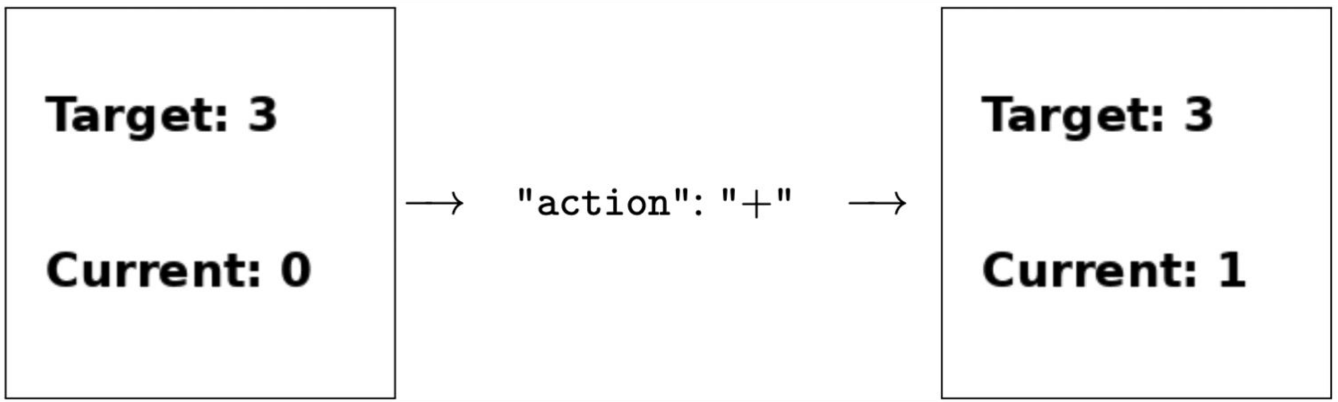
(a) NumberLine (b) EZPoints (c) Points24 (d) Blackjack (e) alfworld

RL4VLM: vision-language based evaluation tasks

- Examples of transitions with **text actions**



(a) NumberLine (b) EZPoints (c) Points24 (d) Blackjack (e) alfworld

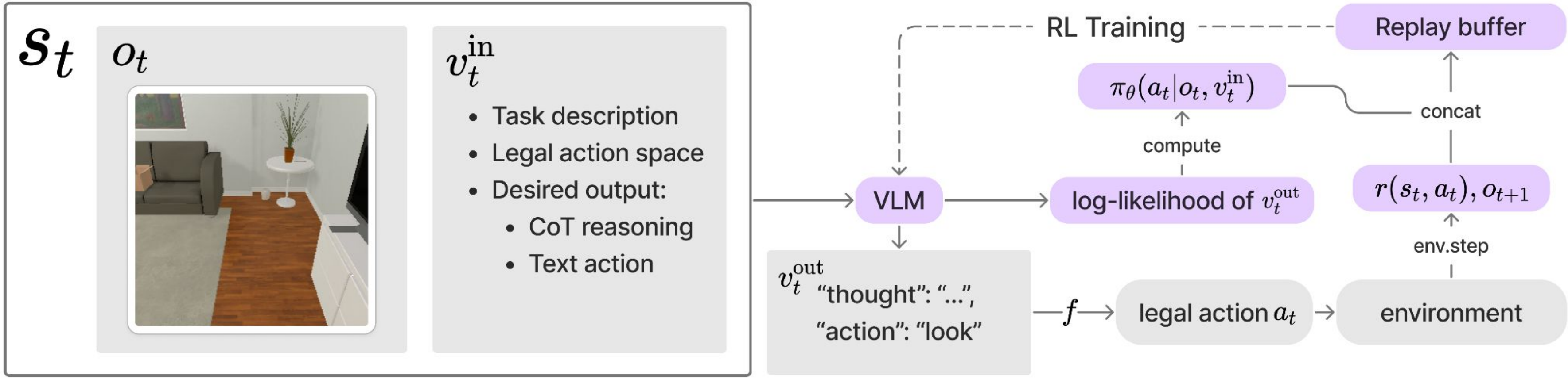


RL4VLM: training generative models as decision-making agents

- Overview
- Vision-language evaluation tasks
- Leveraging domain knowledge
- Results

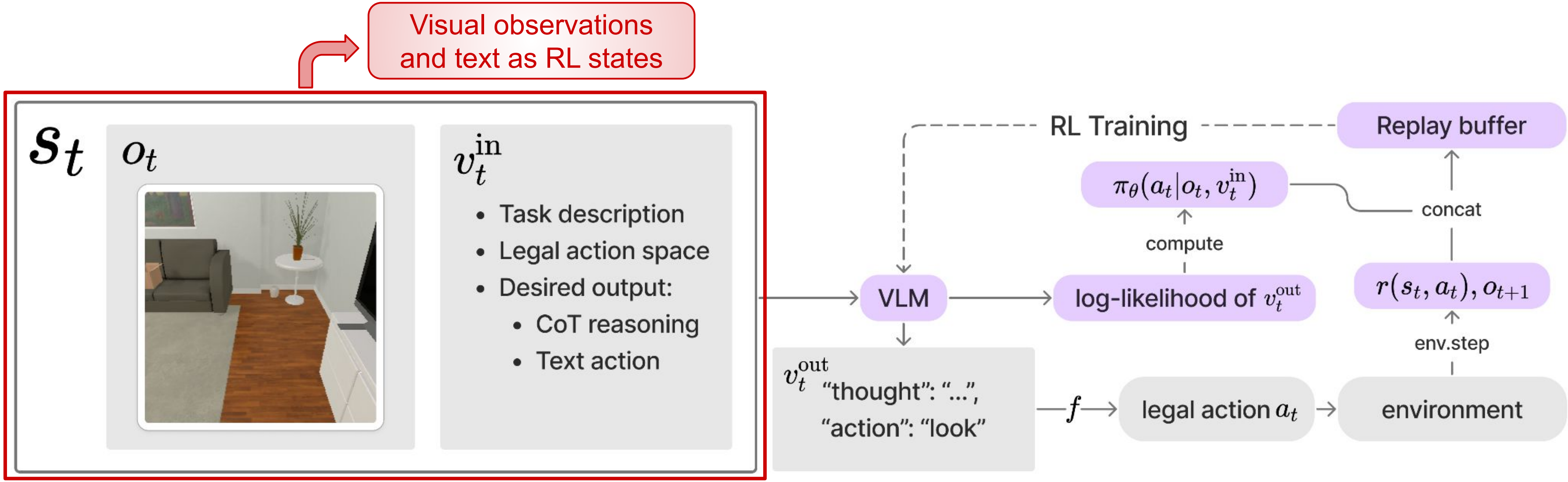
RL4VLM: formulating multimodal generative agent for RL training

- Each state contains a **visual** and **textual** input
- Parse **text outputs** into executable actions



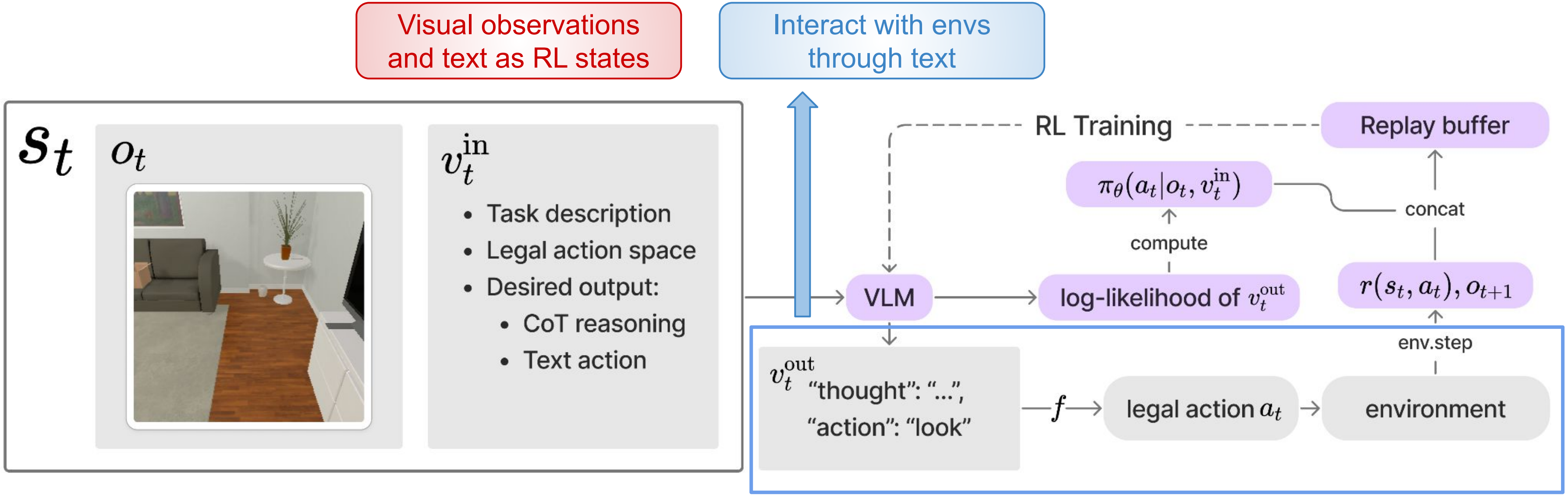
RL4VLM: formulating multimodal generative agent for RL training

- Each state contains a **visual** and **textual** input
- Parse **text outputs** into executable actions



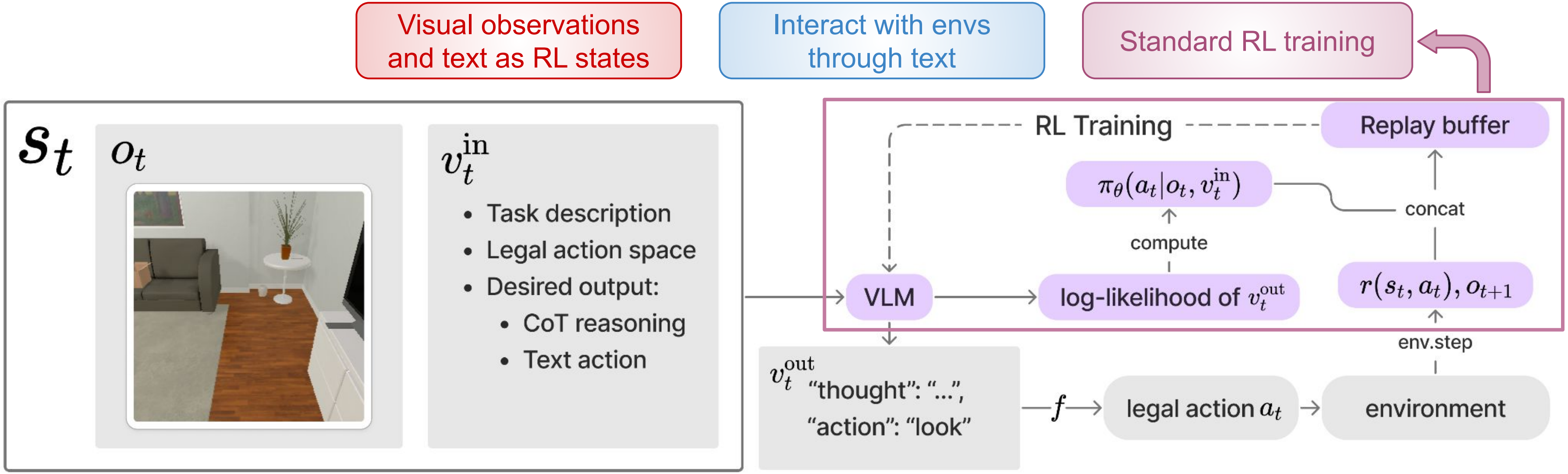
RL4VLM: formulating multimodal generative agent for RL training

- Each state contains a **visual** and **textual** input
- Parse **text outputs** into executable actions



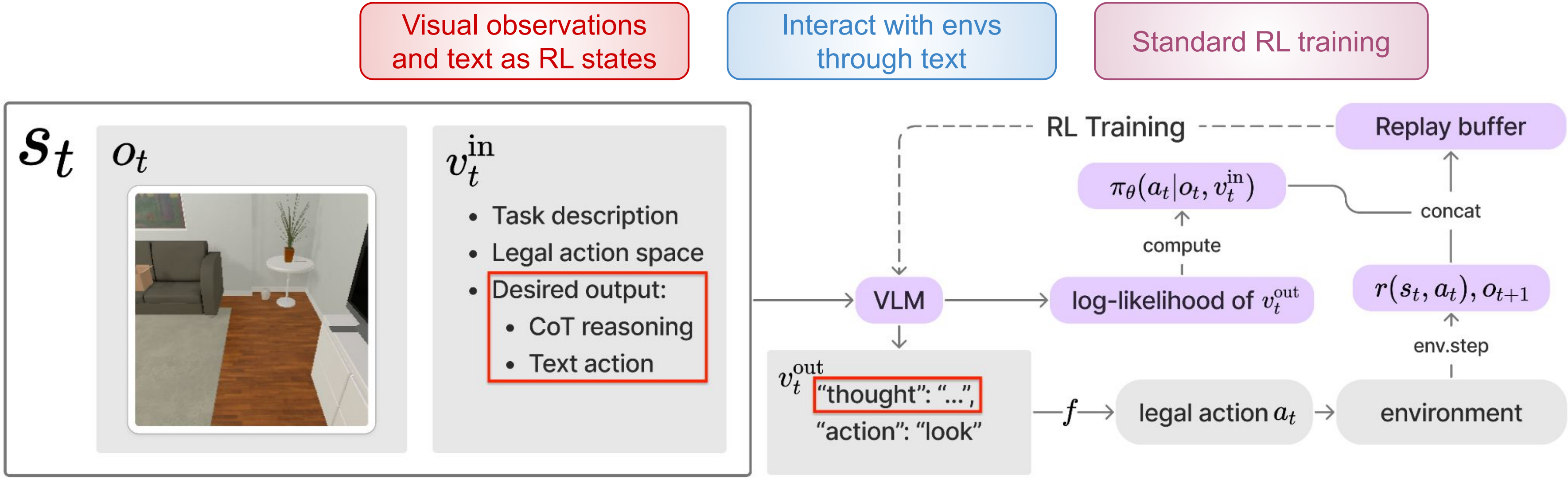
RL4VLM: formulating multimodal generative agent for RL training

- Each state contains a **visual** and **textual** input
- Parse **text outputs** into executable actions



RL4VLM: leveraging the domain information via CoT reasoning

- Each state contains a **visual** and **textual** input ➔ **In context learning** for domain knowledge
- Parse **text outputs** into executable actions ➔ **CoT reasoning** in output text



RL4VLM: an example of in context prompt and CoT output

CoT prompt v_t^{in} for task $\mathcal{M}$

You are trying to solve a task $\mathcal{M}$. You are observing the current status of the task. The action space of $\mathcal{M}$ is {text version of all legal actions $a \in \mathcal{A}$}. {Description of the task}. Your response should be a valid json file in the following format:

{
"thoughts": "{first describe the current status of the task, then think carefully about which action to choose}",
"action": {Choose an action "$a \in \mathcal{A}$"}
}

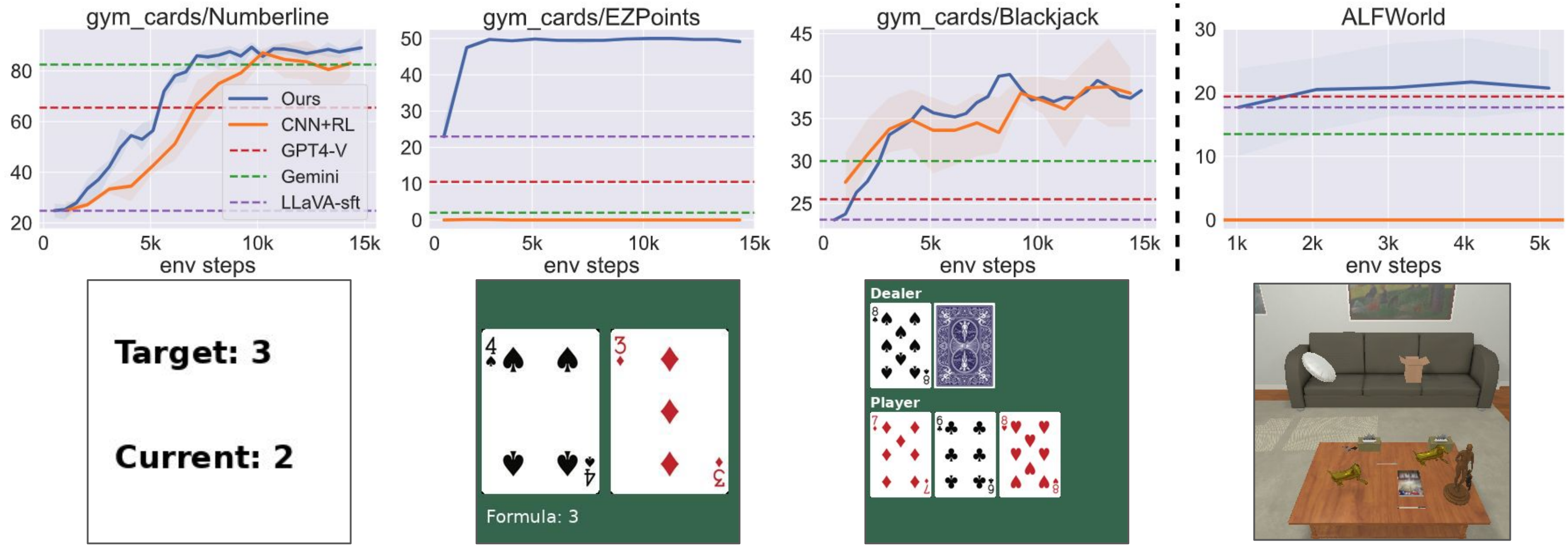
Formatted text output v_t^{out}

{
"thoughts": "I am solving task $\mathcal{T}$, given the current status of the task, I should choose a_t",
"action": "a_t"
}

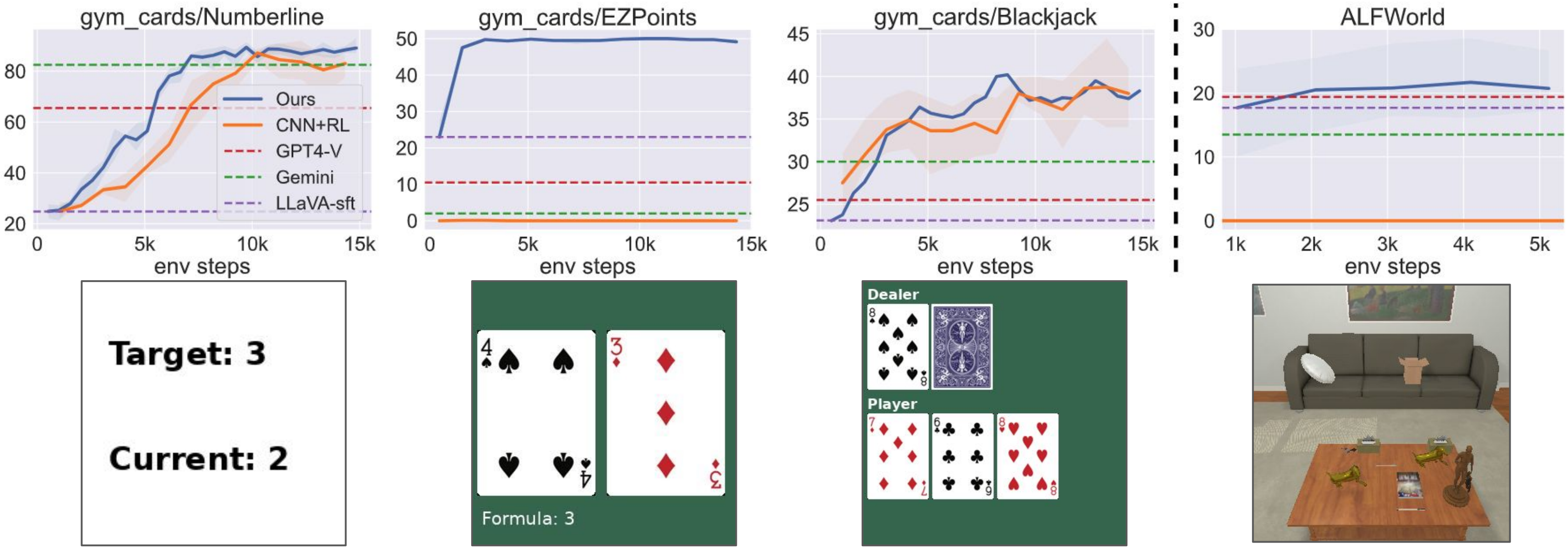
RL4VLM: training generative models as decision-making agents

- Overview
- Vision-language evaluation tasks
- Leveraging domain knowledge
- Results

RL4VLM: improving decision making capabilities of generative agents

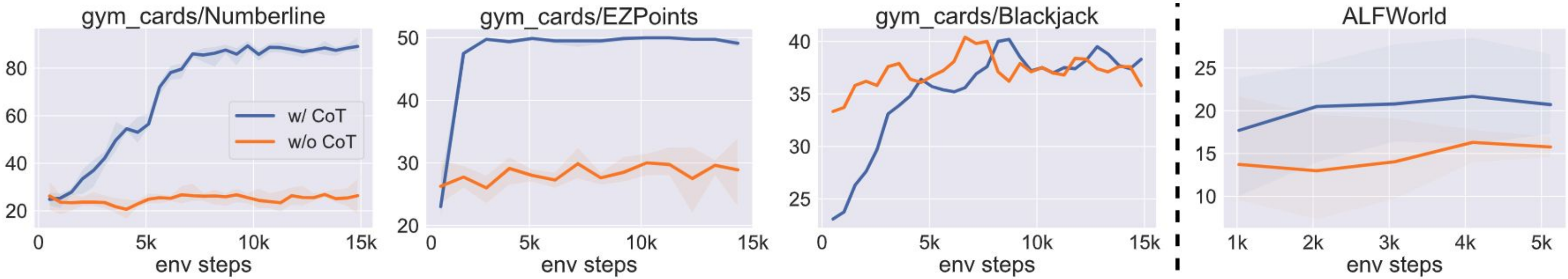


RL4VLM: improving decision making capabilities of generative agents



- Our method enables 7b models to **surpass** the performance of
  - Commercial models: **GPT4-V, Gemini**
  - **Supervised learning** based method (llava-7b-1.6)

RL4VLM: the importance of CoT reasoning



CoT prompt v_t^{in} for task $\mathcal{M}$

You are trying to solve a task $\mathcal{M}$. You are observing the current status of the task. The action space of $\mathcal{M}$ is {text version of all legal actions $a \in \mathcal{A}$}. {Description of the task}. Your response should be a valid json file in the following format:

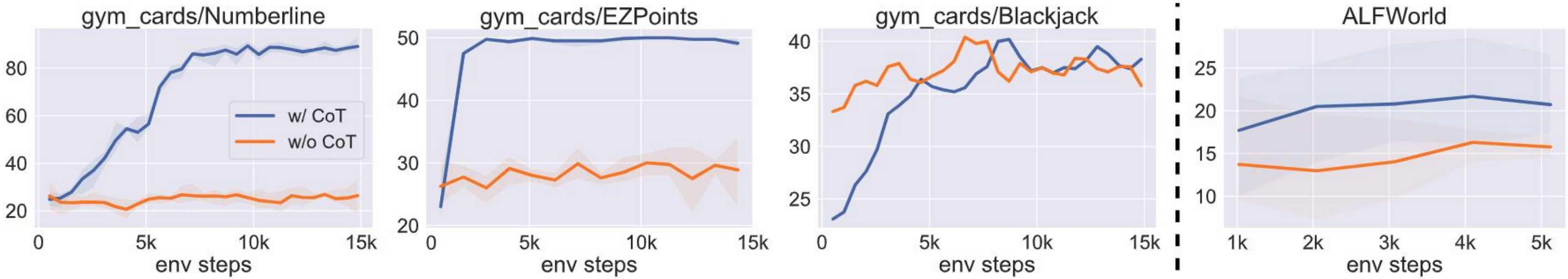
{
"thoughts": "{first describe the current status of the task, then think carefully about which action to choose}",
"action": {Choose an action "$a \in \mathcal{A}$"}
}

Formatted text output v_t^{out}

{
"thoughts": "I am solving task $\mathcal{T}$, given the current status of the task, I should choose a_t",
"action": "a_t"
}

CoT reasoning

RL4VLM: the importance of CoT reasoning



CoT prompt v_t^{in} for task $\mathcal{M}$
You are trying to solve a task $\mathcal{M}$. You are observing the current status of the task. The action space of $\mathcal{M}$ is {text version of all legal actions $a \in \mathcal{A}$}. {Description of the task}. Your response should be a valid json file in the following format:
{
"thoughts": "{first describe the current status of the task, then think carefully about which action to choose}",
"action": {Choose an action "$a \in \mathcal{A}$"}
}

Formatted text output v_t^{out}
{
"thoughts": "I am solving task $\mathcal{T}$, given the current status of the task, I should choose a_t",
"action": "a_t"
}

CoT reasoning

- CoT reasoning **enables efficient RL training** by exploiting domain knowledge

Outline

- Background and Motivation
- Training Large Generative Models as Decision-Making Agents
- Conclusions and Directions for Future Research

Conclusions and Limitations

- **First** end-to-end RL training framework for vision-language generative agent
  - Performance improvement
  - Leverage domain knowledge for efficient training via CoT
  - Without human feedback

Conclusions and Limitations

- First end-to-end RL training framework for vision-language generative agent
  - Performance improvement
  - Leverage domain knowledge for efficient training via CoT
  - Without human feedback
- Fail to improve performance when
  - Backbone model is not strong enough
  - Task is too hard