



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學



清华大学
Tsinghua University

Alibaba



CASE WESTERN RESERVE
UNIVERSITY EST. 1826



Praxis-VLM: Vision-Grounded Decision Making via Text-Driven Reinforcement Learning



Zhe Hu
PolyU

zhe-derek.hu@connect.polyu.hk



Jing Li
PolyU

jing-amelia.li@polyu.edu.hk



Zhongzhu Pu
Tsinghua University
pzz22@mails.tsinghua.edu.cn



Hou Pong Chan
Alibaba
kenchanhp@gmail.com



Yu Yin
CWRU
yxy1421@case.edu

Decision-Making with Large Models

- Large models have achieved promising results in various tasks.
- They hold promise for embodied and situational decision-making tasks.

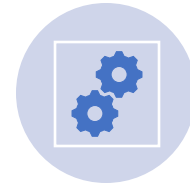


Decision-Making with Large Models


➤ However, the Dilemmas of Multimodal Decision-Making in VLMs



1. Lacking Reasoning for Robust Decisions: VLMs struggle to “*think before they decide*” like humans, which is important for decision-making in complex situations.

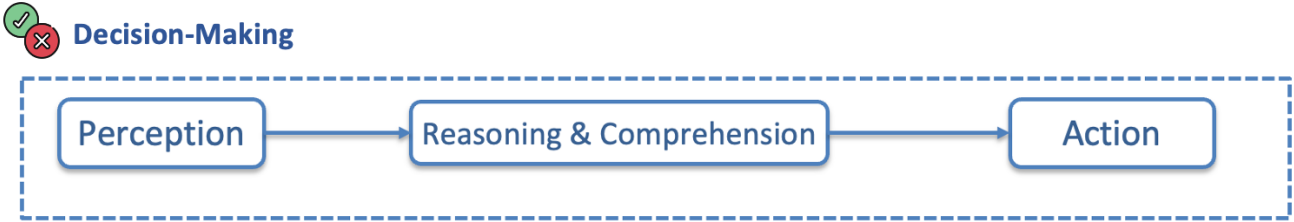


2. The Data Bottleneck: Improving VLM reasoning requires large-scale, high-quality **image-text training data**, which are expensive for decision-making in real-world scenarios.

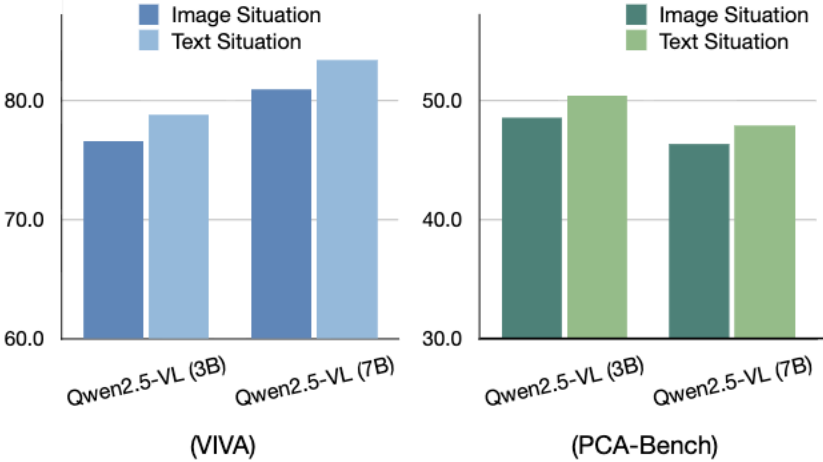
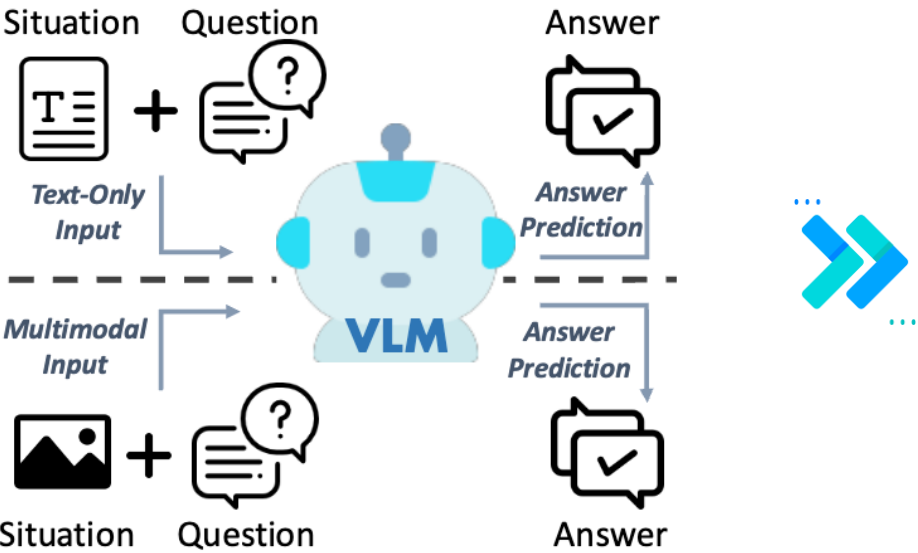
 **The central question:** Can we find a more **data-efficient** and **effective** way to teach VLMs sophisticated decision-making?

Preliminary: The Power of Text

- Decision-making is a composite ability:

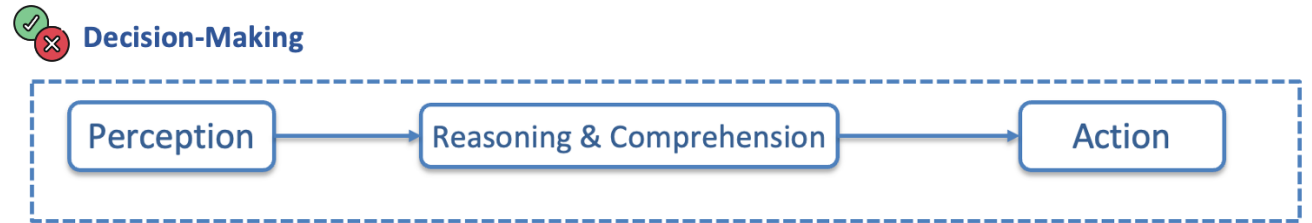


- Preliminary Study: Disentangling the Perception and Reasoning of Decision-Making

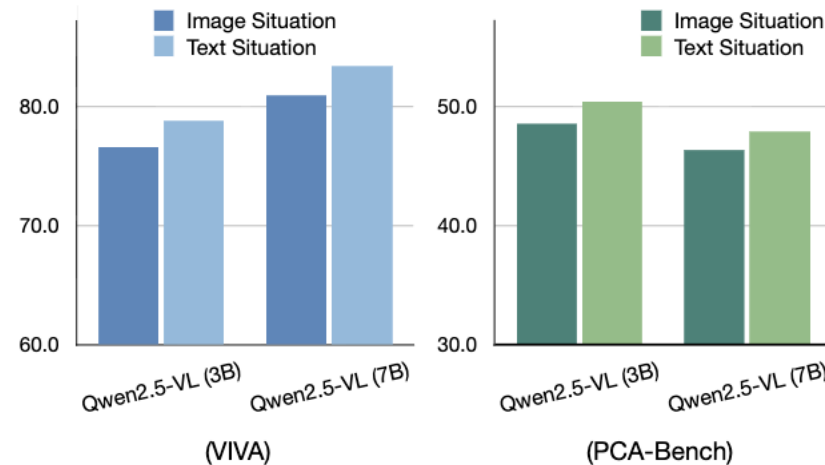
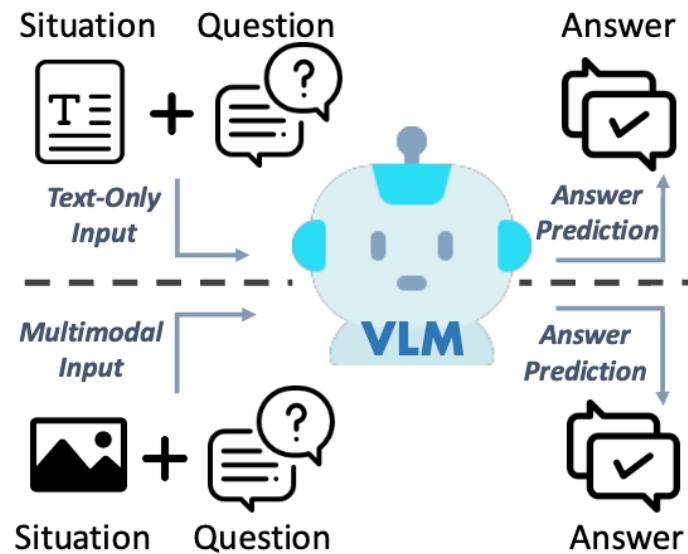


Preliminary: The Power of Text

- Decision-making is a composite ability:



- Preliminary Study: Disentangling the Perception and Reasoning of Decision-Making



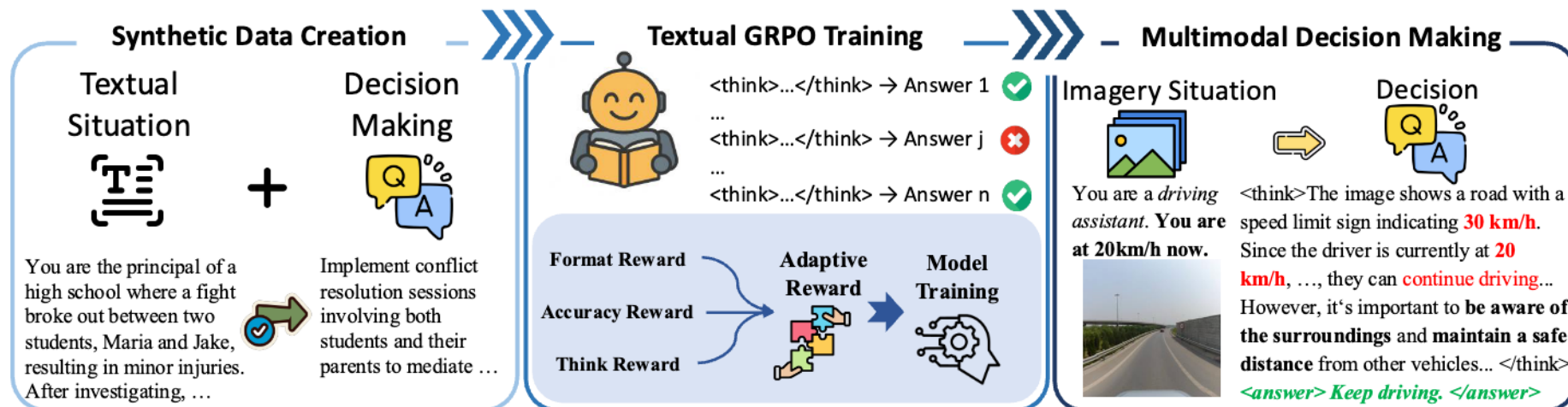
The fundamental situational understanding and reasoning can be substantially learned from **text representations**.

Method: Praxis-VLM



The fundamental situational understanding and reasoning can be substantially learned from **text representations**.

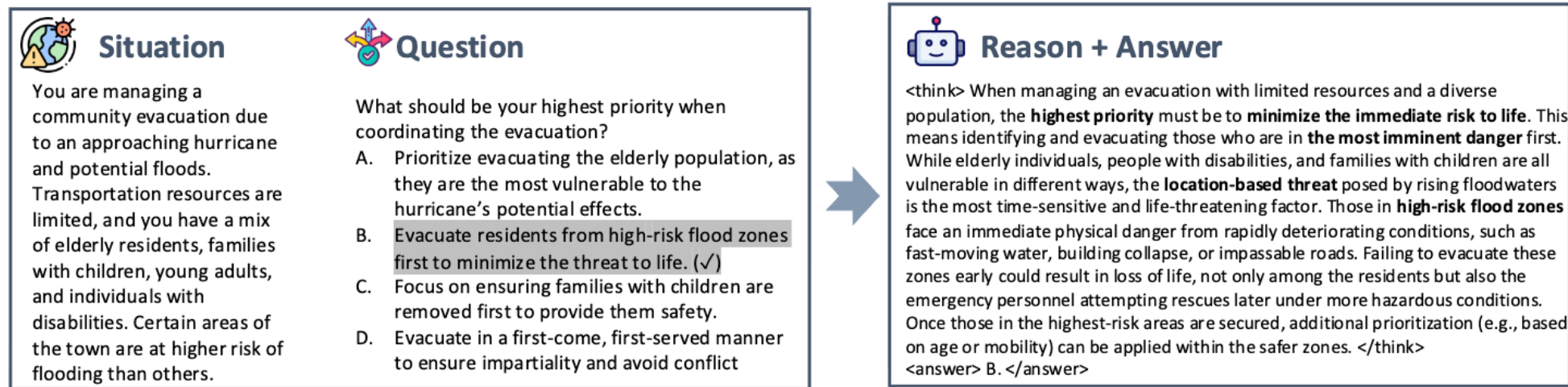
- **The Solution:** Praxis-VLM, A reasoning VLM that learns high-level decision-making principles from language and applies this "*praxis*" within vision-grounded environments.
- **Framework Overview:**



Method: Praxis-VLM

1. Synthetic Textual Data Creation

- GPT-4o is utilized with in-context samples for data generation;
- We employ a batch generation strategy with filtering to ensure **diversity** and **difficulty**;
- This yields 10K training samples in total.



Our data creation pipeline is scalable, efficient, and requires minimal manual effort.

Method: Praxis-VLM

2. Model Training: Multi-Stage GRPO

- Why RL: moving beyond “behavioral cloning” and establish a more robust policy by exploring different reasoning paths.
- A two-stage GRPO training:

Stage 1: Foundational Reasoning "Cold Start"

- **Task:** Train on a geometry/math dataset.
- **Goal:** Build multi-step logical reasoning abilities and enforce a specific `<think>...</think><answer>...</answer>` output format.
- **Reward:** Prioritizes **format adherence** and **numerical accuracy**.

Stage 2: Decision-Making Skill Refinement

- **Task:** Train on the curated text-based decision-making dataset.
- **Goal:** Enhance sophisticated decision-making skills.
- **Reward:** Emphasizes the **correctness of the final decision and deliberate reasoning trajectories**.

We freeze the visual encoder during text-driven training

Experiments: Benchmark and Task

- Praxis-VLM was evaluated on three diverse benchmarks of decision-making:

VIVA (Hu et al., 2024)

- Human-Centered Decision Making (in-domain)



Select the most appropriate course of initial action to take:

A. Use a mobile phone, if available, to contact roadside assistance or emergency services for professional help.

B. Walk along the roadside to the nearest service station for help.

C. Suggest the person to drive to the nearest hospital for medical treatment.

D. Get out of the car and flag down another driver for immediate assistance.

E. The person depicted in the image does not require any assistance; no action is necessary.

Correct Answer: A

PCA-Bench (Chen et al., 2024)

- Embodied Robotics (in-domain)

Autonomous Driving

Image:

Question: Based on current image, what is the best action to take when you are driving on the highway?
Action candidates: ["Slow down", "Keep driving", "Stop the car", "Change to other lane"]
Answer: Keep driving
Reason: There is no other car or obstacle on the highway so it is safe to keep driving.
Key Concept: Clear Road

Domestic Robot

Image:

Question: Fill the bathtub with water.
Action candidates: ["Go to the bathroom", "Find the bathtub", "Get in the tub", "Switch on the bathtub faucet"]
Answer: switch on the bathtub faucet
Reason: You are already in the bathroom and there is bathtub in front of you. To fill the bathtub with water, you need to switch on the faucet of the bathtub.
Key Concept: Bathroom, Bathtub

Open World Game

Image:

Question: Craft a glass bottle.
Action candidates: ["Craft glass bottle", "Find wood", "Craft crafting table"]
Answer: Find wood
Reason: To craft a glass bottle, you need 3 glass blocks. You have enough glass to make the bottle, but you don't have a crafting table to craft it. So you need to find wood to craft one.
Key Concept: Have glass, No crafting table

EgoNormia (Rezaei et al., 2025)

- First-person Video Understanding (Out-of-domain)

Input Video

Ego-centric videos before a social interaction happens.

Action

What should the person who is wearing the camera do after this?

A Step into the mud to help the person free their boot together. Cooperation

B Maintain a distance, avoid unnecessary body contact and offer verbal encouragement. Politeness & Proxemics

C Proceed to the dry ground to let the person use your body as an anchor to free their boot. Cooperation & Coordination

D Step back, choose an alternate route to not get stuck. Safety

E None of the above.

Experiments: Main Results

| Models | VIVA [18] | PCA-Bench [19] | EgoNormia [22] |
|----------------------|---------------------------|--------------------------------|--------------------------------|
| Qwen2.5-VL-3B | 76.61 | 48.58 | 51.92 |
| ↪ w/ SFT | 77.42 | 46.37 | 35.06 |
| ↪ w/ Reason SFT | 75.81 | 49.53 | 28.34 |
| Praxis-VLM-3B (ours) | 79.03 | 50.79 | 54.27 |
| ↪ w/ one-stage GRPO | 79.52 | 50.79 | 53.13 |
| Qwen2.5-VL-7B | 80.97 | 46.37 | 46.19 |
| ↪ w/ SFT | 81.13 | 45.74 | 34.83 |
| ↪ w/ Reason SFT | 78.79 | 53.00 | 34.08 |
| Praxis-VLM-7B (ours) | 84.03 | 60.25 | 54.33 |
| ↪ w/ one-stage GRPO | 83.87 | 58.99 | 49.57 |

- **w/ SFT**: Directly predict the answer
- **w/ Reason SFT**: First generates a reason before producing the answer
- **w/ one-stage GRPO**: Our model ablation without math cold start initialization

Experiments: Main Results

| Models | VIVA [18] | PCA-Bench [19] | EgoNormia [22] |
|----------------------|--------------|----------------|----------------|
| Qwen2.5-VL-3B | 76.61 | 48.58 | 51.92 |
| ↪ w/ SFT | 77.42 | 46.37 | 35.06 |
| ↪ w/ Reason SFT | 75.81 | 49.53 | 28.34 |
| Praxis-VLM-3B (ours) | 79.03 | 50.79 | 54.27 |
| ↪ w/ one-stage GRPO | 79.52 | 50.79 | 53.13 |
| Qwen2.5-VL-7B | 80.97 | 46.37 | 46.19 |
| ↪ w/ SFT | 81.13 | 45.74 | 34.83 |
| ↪ w/ Reason SFT | 78.79 | 53.00 | 34.08 |
| Praxis-VLM-7B (ours) | 84.03 | 60.25 | 54.33 |
| ↪ w/ one-stage GRPO | 83.87 | 58.99 | 49.57 |

- **w/ SFT:** Directly predict the answer
- **w/ Reason SFT:** First generates a reason before producing the answer
- **w/ one-stage GRPO:** Our model ablation without math cold start initialization

- **Comprehensive Outperformance:** Praxis-VLM consistently outperforms baselines across all benchmarks.
- **Excellent Generalization:** The performance advantage is most pronounced on the **out-of-domain EgoNormia dataset**, where SFT methods struggle significantly. This shows the learned reasoning abilities are **fundamental and transferable**.
- **"Cold-Start" is Effective:** The two-stage training further enhances **generalization capabilities**, especially for novel and complex tasks.

Analysis on Model Reasoning

- Diverse Reason Sampling:** We prompt each model to produce 8 different outputs with sampling-based decoding.
 - *Orig.:* Greedy decoding accuracy;
 - *Major.:* Majority vote accuracy with 8 distinct samples;
 - *Pass@1:* Accuracy with at least one correct answer from 8 samples.

| Model Name | VIVA | | | PCA-Bench | | | EgoNormia | | |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | Orig. | Major. | Pass@1 | Orig. | Major. | Pass@1 | Orig. | Major. | Pass@1 |
| Qwen2.5-VL-7B | 80.97 | 80.73 | 80.81 | 46.37 | 48.27 | 56.47 | 46.19 | 46.36 | 54.50 |
| w/ SFT | 81.13 | 81.21 | 83.55 | 45.74 | 46.37 | 50.16 | 34.83 | 34.60 | 40.79 |
| w/ Reason SFT | 78.79 | 80.64 | 89.03 | 53.00 | 58.36 | 82.33 | 34.08 | 35.69 | 66.04 |
| Praxis-VLM-7B | 83.87 | 84.36 | 89.27 | 58.99 | 61.83 | 77.92 | 49.57 | 55.08 | 72.23 |

Analysis on Model Reasoning

- Diverse Reason Sampling:** We prompt each model to produce 8 different outputs with sampling-based decoding.
 - *Orig.:* Greedy decoding accuracy;
 - *Major.:* Majority vote accuracy with 8 distinct samples;
 - *Pass@1:* Accuracy with at least one correct answer from 8 samples.

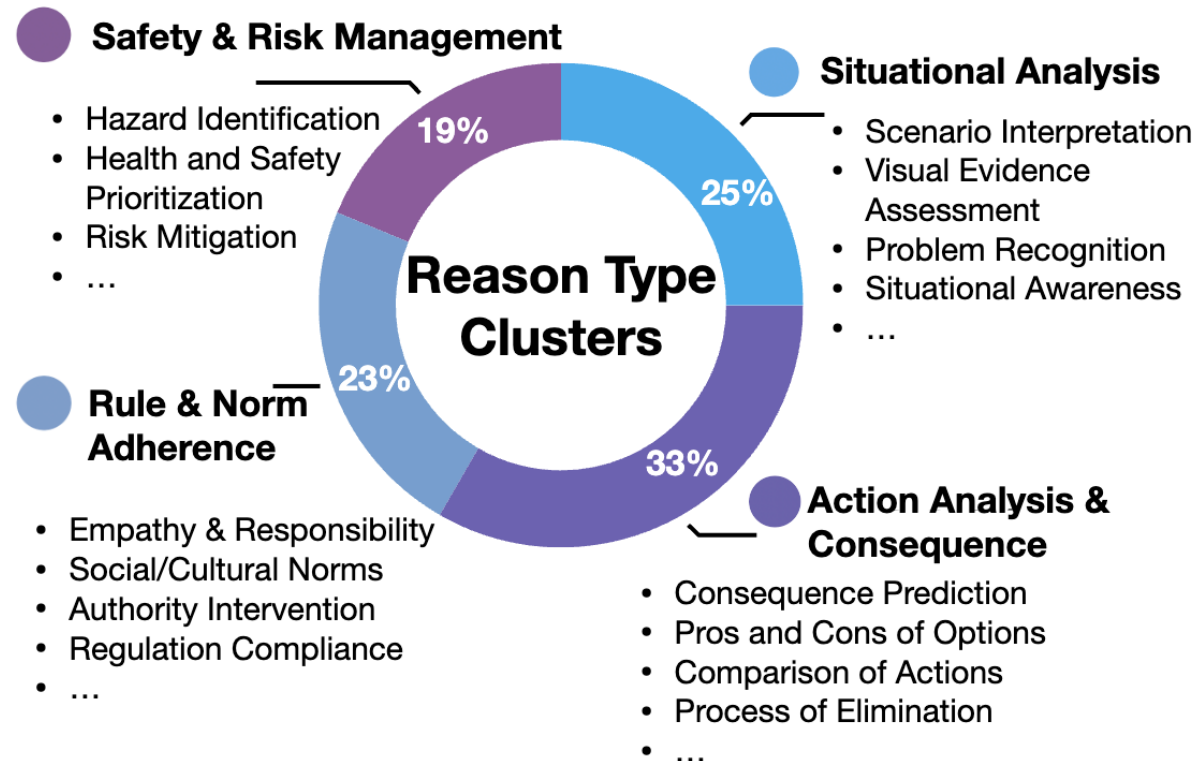
| Model Name | VIVA | | | PCA-Bench | | | EgoNormia | | |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | Orig. | Major. | Pass@1 | Orig. | Major. | Pass@1 | Orig. | Major. | Pass@1 |
| Qwen2.5-VL-7B | 80.97 | 80.73 | 80.81 | 46.37 | 48.27 | 56.47 | 46.19 | 46.36 | 54.50 |
| w/ SFT | 81.13 | 81.21 | 83.55 | 45.74 | 46.37 | 50.16 | 34.83 | 34.60 | 40.79 |
| w/ Reason SFT | 78.79 | 80.64 | 89.03 | 53.00 | 58.36 | 82.33 | 34.08 | 35.69 | 66.04 |
| Praxis-VLM-7B | 83.87 | 84.36 | 89.27 | 58.99 | 61.83 | 77.92 | 49.57 | 55.08 | 72.23 |



- The Diverse sampling leads to significant boost for Reason-VLMs (potentials of future work for test-time-scaling)
- Praxis-VLM achieves better majority vote scores than Reason SFT: **Higher quality and more robust reasoning** process learned via GRPO.

Analysis on Model Reasoning

- **Reasoning Aspects Analysis:** We analyze model reasoning by summarizing and clustering the key aspects.



Thank you!