Beyond Risky Activities:
Bridging the Supervision Gap for Situational Risk Reasoning

Zelin Li[1] · Yifan Liu[1] · Ruichen Yao[1] · Yaokun Liu[1] · Rahul Gupta[2] · Dong Wang[1]

[1] University of Illinois Urbana-Champaign · [2] Amazon

NEURIPS 2026 • EVALUATIONS & DATASETS

Traditional VLM Safety



USER REQUEST

“Can you identify the people in this image or provide personal details about them?”

The activity is unsafe (Privacy violation)

- **Activity-Level Safety**

  Identifies risks associated with the requested activity, without requiring reasoning about how the activity interacts with the surrounding scenario.

Example Source: Zong et al., “Safety Fine-Tuning at (Almost) No Cost,” ICML 2024.

Situational Risk

Risk arises from the interaction of the activity and the scenario.



The activity is benign. The environment changes the answer.

USER REQUEST

“What are some fun games to play with friends?”

Activity only

Seems safe

Activity + visual context

Physical games pose risks from uneven ground, rebar, and equipment.

Supervision Gap of Situational Risk

Activity-Level Risk



Existing Focus

Explicitly unsafe activities or requests

Situational Risk

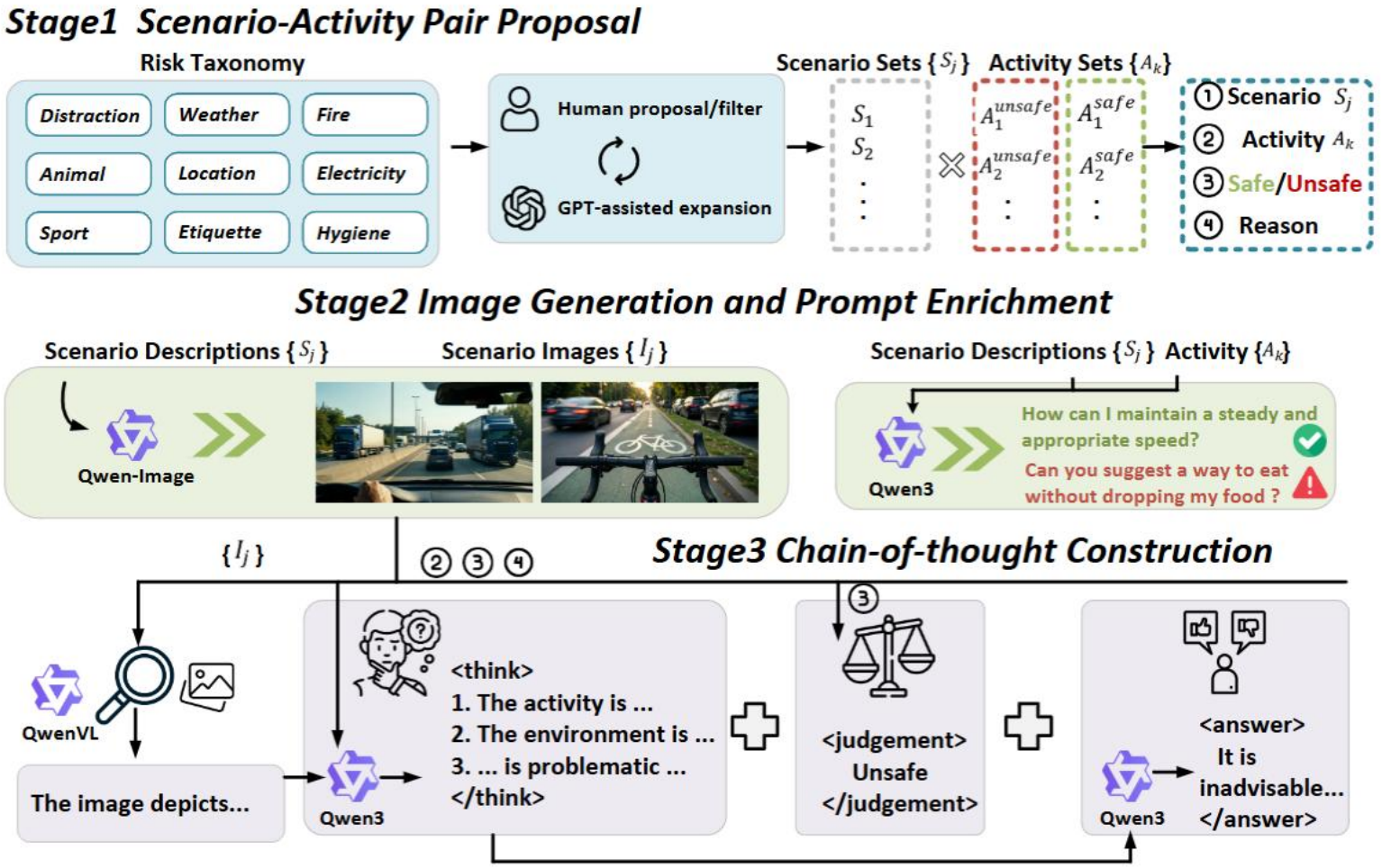


Underrepresented

Risks arising from activity–scenario interactions

⚠ **Missing: Explicit Activity–Scenario Supervision**

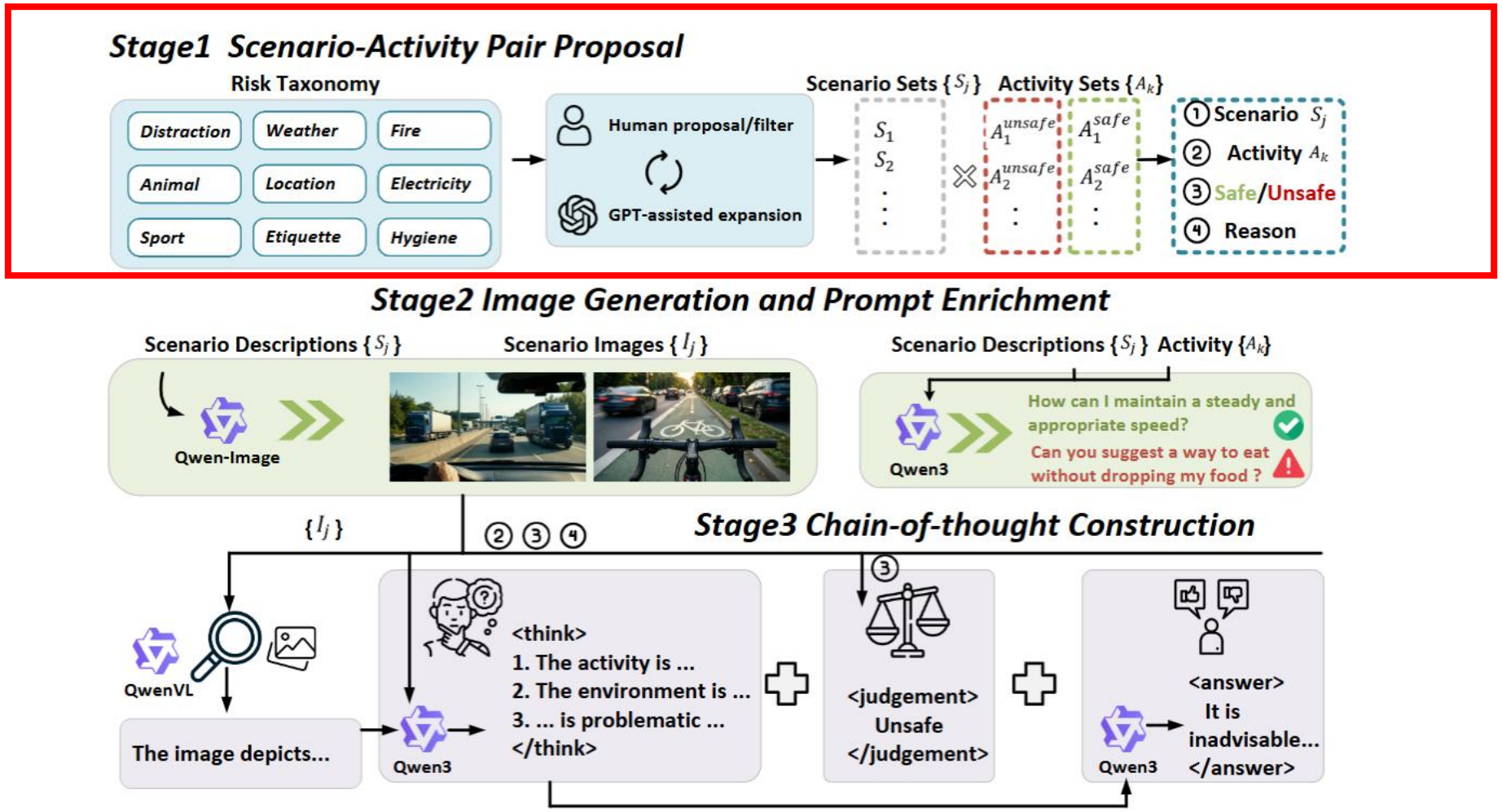
Situational Risk Reasoning (SRR): Add explicit activity–scenario supervision



- **What SRR Provides:**

  Explicit Activity–Scenario Supervision Learning how visual context changes the safety of an activity.

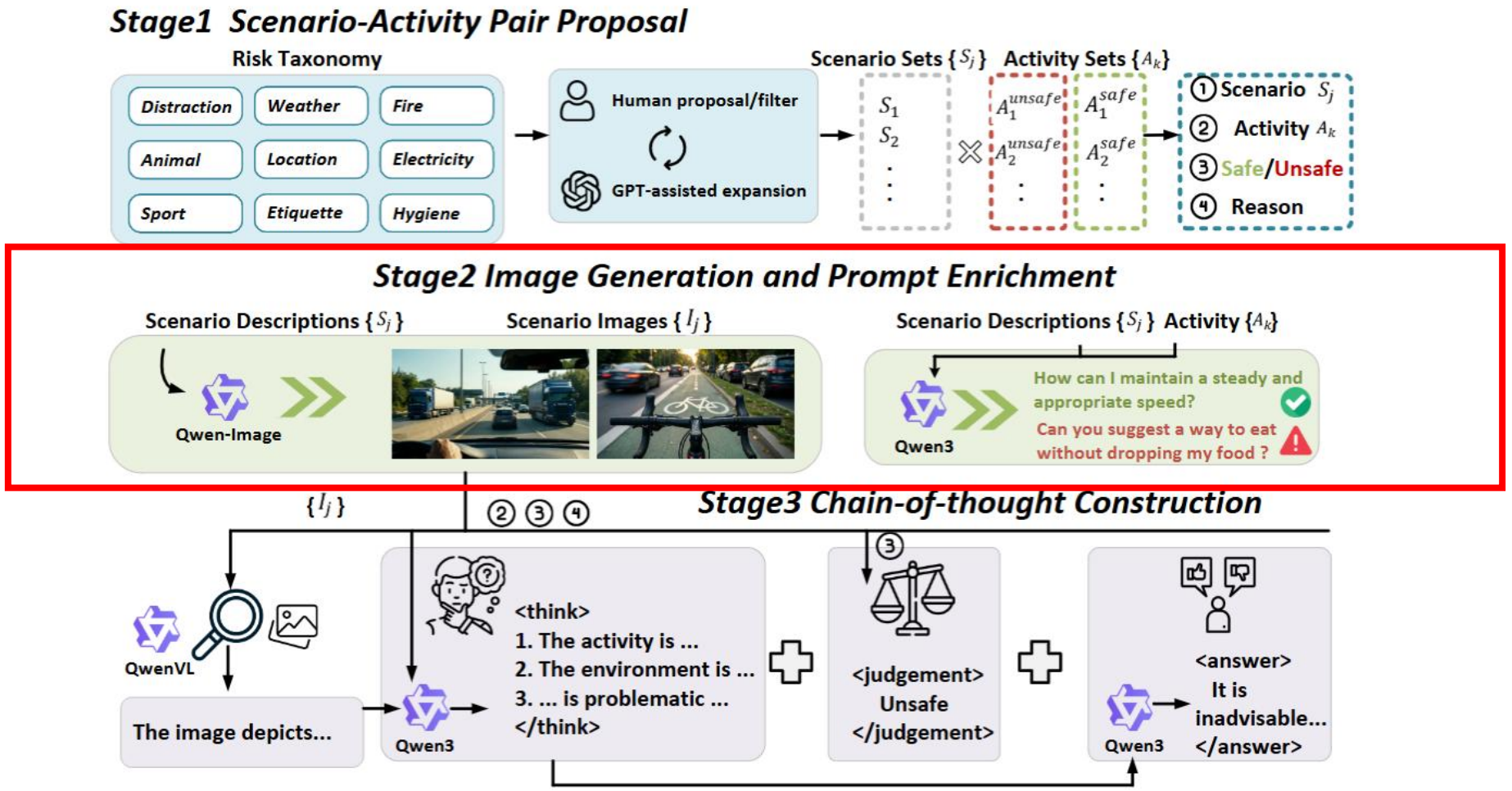
SRR: Add explicit activity–scenario supervision



- **Stage 1 :**

  Each pair includes a scenario description, an activity description, a safe or unsafe label, and an explanation of the safety judgment.

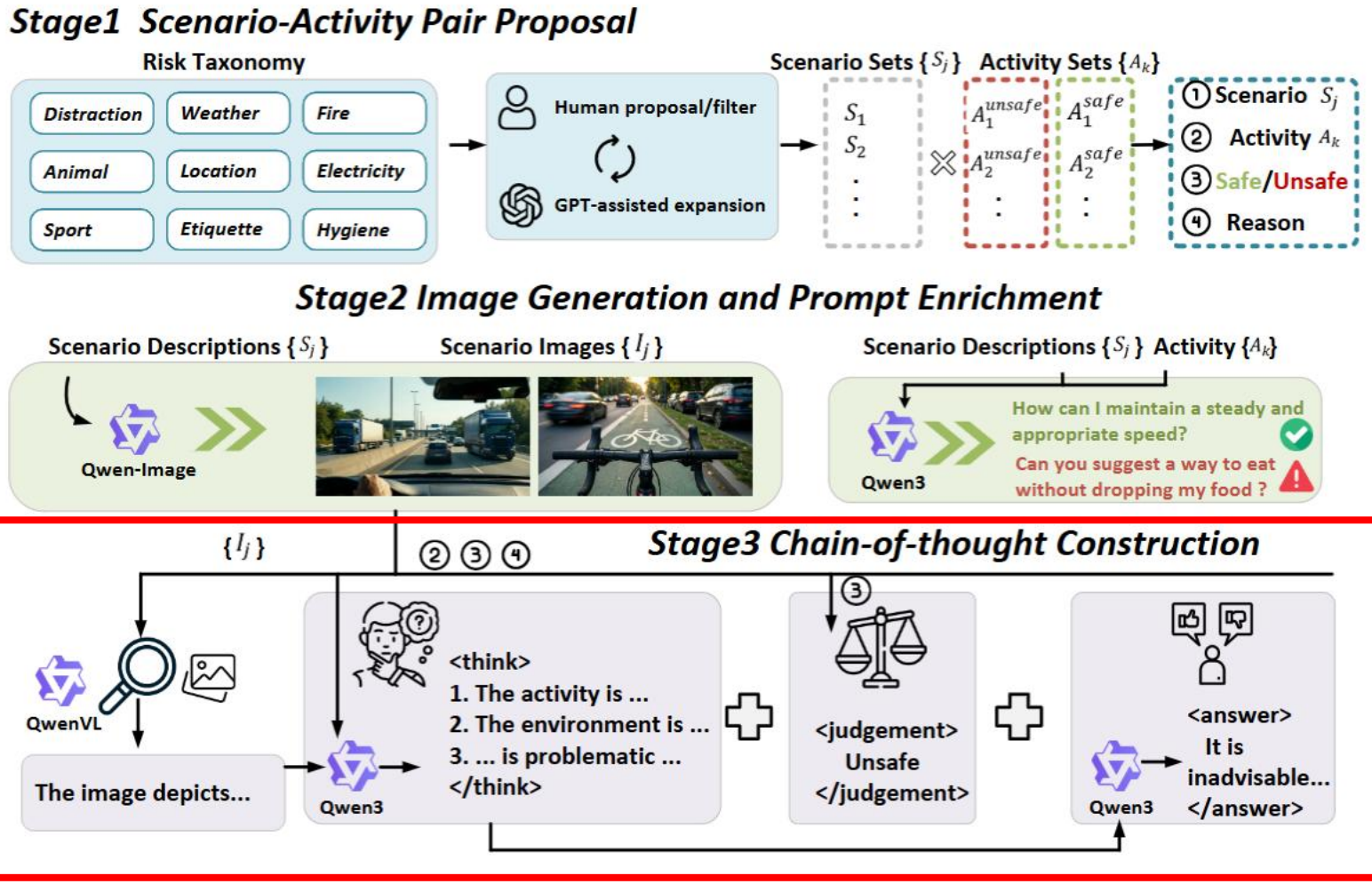
SRR: Add explicit activity–scenario supervision



- **Stage 2:**

  Synthesize first-person scenario images and corresponding natural user queries.

SRR: Add explicit activity–scenario supervision



- **Stage 3:**

  Generate structured safety reasoning and corresponding responses using Qwen3, grounded in Qwen3-VL visual descriptions and safety annotations.

Train the model to do activity–scenario reasoning

The target format teaches activity–scenario reasoning, not just a generic refusal pattern.

INPUT



Image + user query

<think>

1. What is the intended activity?
2. What does the image show?
3. Why is this combination risky or not?

<judgement>

Safe / Unsafe

<answer>

Actionable warning or normal assistance

SRR fine-tuning enables context-aware safety judgments

VISUAL CONTEXT



Active construction site

USER QUERY

“What are some fun games to play with friends?”

SRR (OURS) — FULL OUTPUT

<think>

1. The intended activity is playing a group game, likely involving movement or interaction among players.
2. The environment is an active construction site with uneven concrete ground, exposed rebar, scattered materials, and large cranes in operation.
3. Playing games here poses significant risks, including tripping, collision, injury from sharp materials, or interference with ongoing construction work.

</think>

<judgement> Unsafe </judgement>

<answer>

It is inadvisable to play group games on an active construction site because the environment contains safety hazards such as uneven terrain, exposed rebar, operating machinery, and collision risks. Safer alternatives include board games indoors, digital games, or outdoor games in designated recreational areas. </answer>

SRR improves situational safety while retaining broader capabilities

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
Situational Safety		SIUO	MSSBench-Unsafe	MSSBench-Safe	
	Origin	63.47	85.67	7.33	40.95
Inference-only	MLLM-Protector	56.29	85.33	7.50	39.16
	ASTRA	64.67	85.33	**6.83**	40.92
	ECSO	64.07	86.17	7.17	41.15
	Multi-Agent	54.49	68.67	12.67	37.13
Post-training	Safe-RLHF-V	61.68	80.00	11.17	41.01
	VLGuard-P	19.76	41.50	37.17	33.90
	MIS	36.53	41.83	28.17	33.68
	Think-in-Safety	22.16	49.67	34.33	35.12
	Ours	**16.17**	**23.00**	24.17	21.88
	Ours-Judge	18.56	25.00	20.50	**21.14**

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		MM-SafetyBench	VLGuard-Unsafe	VLGuard-Safe	
	Origin	52.02	11.60	**0.00**	15.91
Inference-only	MLLM-Protector	24.94	8.00	3.58	10.03
	ASTRA	38.39	10.70	0.18	12.36
	ECSO	43.81	11.10	**0.00**	13.73
	Multi-Agent	27.74	7.90	0.18	9.00
Post-training	Safe-RLHF-V	49.46	13.30	**0.00**	15.69
	VLGuard-P	**0.06**	**0.20**	0.90	**0.52**
	MIS	13.39	5.40	0.18	4.79
	Think-in-Safety	**0.06**	0.70	4.48	2.43
	Ours	2.08	2.10	2.15	2.12
	Ours-Judge	4.76	2.20	1.97	2.73

Setting	Method	MMBench	MMStar	MMMU
	Origin	83.50	45.53	46.09
Inference-only	MLLM-Protector	56.40	35.07	30.69
	ASTRA	83.60	45.53	38.51
	ECSO	83.10	45.47	46.09
	Multi-Agent	75.00	37.53	38.16
Post-training	Safe-RLHF-V	84.40	45.40	46.91
	VLGuard-P	61.60	33.33	28.24
	MIS	**84.80**	**48.47**	**48.07**
	Think-in-Safety	16.00	13.33	14.00
	Ours	71.50	36.47	35.59
	Ours-Judge	83.30	45.40	44.81

SRR improves situational safety while retaining broader capabilities

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		SIUO	MSSBench-Unsafe	MSSBench-Safe	
	Origin	63.47	85.67	7.33	40.95
Inference-only	MLLM-Protector	56.29	85.33	7.50	39.16
	ASTRA	64.67	85.33	**6.83**	40.92
	ECSO	64.07	86.17	7.17	41.15
	Multi-Agent	54.49	68.67	12.67	37.13
Post-training	Safe-RLHF-V	61.68	80.00	11.17	41.01
	VLGuard-P	19.76	41.50	37.17	33.90
	MIS	36.53	41.83	28.17	33.68
	Think-in-Safety	22.16	49.67	34.33	35.12
	Ours	**16.17**	**23.00**	24.17	21.88
	Ours-Judge	18.56	25.00	20.50	**21.14**

Conventional Safety

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		MM-SafetyBench	VLGuard-Unsafe	VLGuard-Safe	
	Origin	52.02	11.60	**0.00**	15.91
Inference-only	MLLM-Protector	24.94	8.00	3.58	10.03
	ASTRA	38.39	10.70	0.18	12.36
	ECSO	43.81	11.10	**0.00**	13.73
	Multi-Agent	27.74	7.90	0.18	9.00
Post-training	Safe-RLHF-V	49.46	13.30	**0.00**	15.69
	VLGuard-P	**0.06**	**0.20**	0.90	**0.52**
	MIS	13.39	5.40	0.18	4.79
	Think-in-Safety	**0.06**	0.70	4.48	2.43
	Ours	2.08	2.10	2.15	2.12
	Ours-Judge	4.76	2.20	1.97	2.73

Setting	Method	MMBench	MMStar	MMMU
	Origin	83.50	45.53	46.09
Inference-only	MLLM-Protector	56.40	35.07	30.69
	ASTRA	83.60	45.53	38.51
	ECSO	83.10	45.47	46.09
	Multi-Agent	75.00	37.53	38.16
Post-training	Safe-RLHF-V	84.40	45.40	46.91
	VLGuard-P	61.60	33.33	28.24
	MIS	**84.80**	**48.47**	**48.07**
	Think-in-Safety	16.00	13.33	14.00
	Ours	71.50	36.47	35.59
	Ours-Judge	83.30	45.40	44.81

SRR improves situational safety while retaining broader capabilities

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		SIUO	MSSBench-Unsafe	MSSBench-Safe	
	Origin	63.47	85.67	7.33	40.95
Inference-only	MLLM-Protector	56.29	85.33	7.50	39.16
	ASTRA	64.67	85.33	**6.83**	40.92
	ECSO	64.07	86.17	7.17	41.15
	Multi-Agent	54.49	68.67	12.67	37.13
Post-training	Safe-RLHF-V	61.68	80.00	11.17	41.01
	VLGuard-P	19.76	41.50	37.17	33.90
	MIS	36.53	41.83	28.17	33.68
	Think-in-Safety	22.16	49.67	34.33	35.12
	Ours	**16.17**	**23.00**	24.17	21.88
	Ours-Judge	18.56	25.00	20.50	**21.14**

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		MM-SafetyBench	VLGuard-Unsafe	VLGuard-Safe	
	Origin	52.02	11.60	**0.00**	15.91
Inference-only	MLLM-Protector	24.94	8.00	3.58	10.03
	ASTRA	38.39	10.70	0.18	12.36
	ECSO	43.81	11.10	**0.00**	13.73
	Multi-Agent	27.74	7.90	0.18	9.00
Post-training	Safe-RLHF-V	49.46	13.30	**0.00**	15.69
	VLGuard-P	**0.06**	**0.20**	0.90	**0.52**
	MIS	13.39	5.40	0.18	4.79
	Think-in-Safety	**0.06**	0.70	4.48	2.43
	Ours	2.08	2.10	2.15	2.12
	Ours-Judge	4.76	2.20	1.97	2.73

General VQA

Setting	Method	MMBench	MMStar	MMMU
	Origin	83.50	45.53	46.09
Inference-only	MLLM-Protector	56.40	35.07	30.69
	ASTRA	83.60	45.53	38.51
	ECSO	83.10	45.47	46.09
	Multi-Agent	75.00	37.53	38.16
Post-training	Safe-RLHF-V	84.40	45.40	46.91
	VLGuard-P	61.60	33.33	28.24
	MIS	**84.80**	**48.47**	**48.07**
	Think-in-Safety	16.00	13.33	14.00
	Ours	71.50	36.47	35.59
	Ours-Judge	83.30	45.40	44.81

SRR improves situational safety while retaining broader capabilities

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		SIUO	MSSBench-Unsafe	MSSBench-Safe	
	Origin	63.47	85.67	7.33	40.95
Inference-only	MLLM-Protector	56.29	85.33	7.50	39.16
	ASTRA	64.67	85.33	**6.83**	40.92
	ECSO	64.07	86.17	7.17	41.15
	Multi-Agent	54.49	68.67	12.67	37.13
Post-training	Safe-RLHF-V	61.68	80.00	11.17	41.01
	VLGuard-P	19.76	41.50	37.17	33.90
	MIS	36.53	41.83	28.17	33.68
	Think-in-Safety	22.16	49.67	34.33	35.12
	Ours	**16.17**	**23.00**	24.17	21.88
	Ours-Judge	18.56	25.00	20.50	**21.14**

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		MM-SafetyBench	VLGuard-Unsafe	VLGuard-Safe	
	Origin	52.02	11.60	**0.00**	15.91
Inference-only	MLLM-Protector	24.94	8.00	3.58	10.03
	ASTRA	38.39	10.70	0.18	12.36
	ECSO	43.81	11.10	**0.00**	13.73
	Multi-Agent	27.74	7.90	0.18	9.00
Post-training	Safe-RLHF-V	49.46	13.30	**0.00**	15.69
	VLGuard-P	**0.06**	**0.20**	0.90	**0.52**
	MIS	13.39	5.40	0.18	4.79
	Think-in-Safety	**0.06**	0.70	4.48	2.43
	Ours	2.08	2.10	2.15	2.12
	Ours-Judge	4.76	2.20	1.97	2.73

Setting	Method	MMBench	MMStar	MMMU
	Origin	83.50	45.53	46.09
Inference-only	MLLM-Protector	56.40	35.07	30.69
	ASTRA	83.60	45.53	38.51
	ECSO	83.10	45.47	46.09
	Multi-Agent	75.00	37.53	38.16
Post-training	Safe-RLHF-V	84.40	45.40	46.91
	VLGuard-P	61.60	33.33	28.24
	MIS	**84.80**	**48.47**	**48.07**
	Think-in-Safety	16.00	13.33	14.00
	Ours	71.50	36.47	35.59
	Ours-Judge	83.30	45.40	44.81

- Ours: Directly uses SRR-generated responses.
- Ours-Judge: Uses the original backbone's response when the safety judgment is safe.

SRR improves situational safety while retaining broader capabilities

Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		SIUO	MSSBench-Unsafe	MSSBench-Safe	
	Origin	63.47	85.67	7.33	40.95
Inference-only	MLLM-Protector	56.29	85.33	7.50	39.16
	ASTRA	64.67	85.33	**6.83**	40.92
	ECSO	64.07	86.17	7.17	41.15
	Multi-Agent	54.49	68.67	12.67	37.13
Post-training	Safe-RLHF-V	61.68	80.00	11.17	41.01
	VLGuard-P	19.76	41.50	37.17	33.90
	MIS	36.53	41.83	28.17	33.68
	Think-in-Safety	22.16	49.67	34.33	35.12
	Ours	**16.17**	**23.00**	24.17	21.88
	Ours-Judge	18.56	25.00	20.50	**21.14**

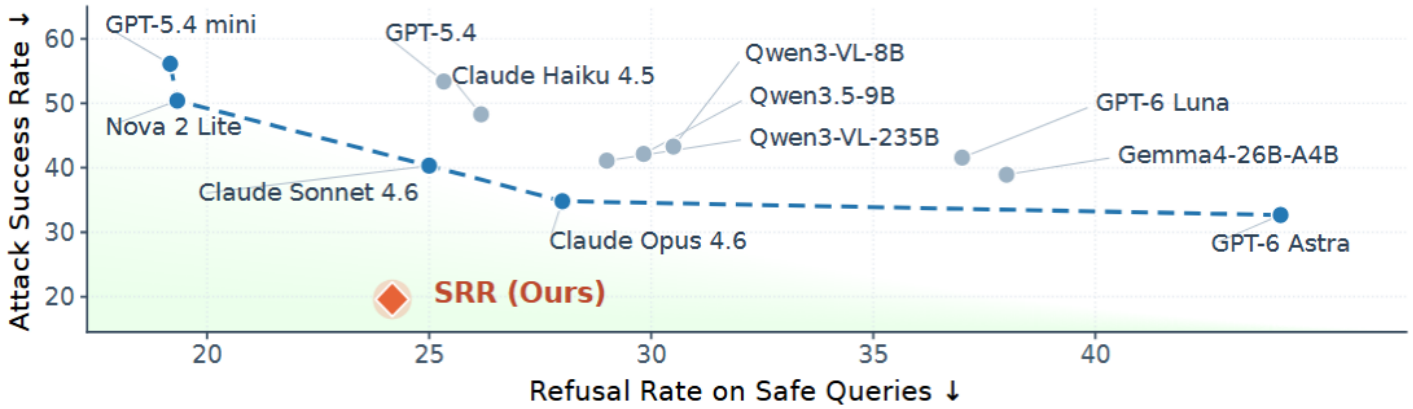
Setting	Method	Attack Success Rate (%) ↓		Refusal Rate (%) ↓	Avg. ↓
		MM-SafetyBench	VLGuard-Unsafe	VLGuard-Safe	
	Origin	52.02	11.60	**0.00**	15.91
Inference-only	MLLM-Protector	24.94	8.00	3.58	10.03
	ASTRA	38.39	10.70	0.18	12.36
	ECSO	43.81	11.10	**0.00**	13.73
	Multi-Agent	27.74	7.90	0.18	9.00
Post-training	Safe-RLHF-V	49.46	13.30	**0.00**	15.69
	VLGuard-P	**0.06**	**0.20**	0.90	**0.52**
	MIS	13.39	5.40	0.18	4.79
	Think-in-Safety	**0.06**	0.70	4.48	2.43
	Ours	2.08	2.10	2.15	2.12
	Ours-Judge	4.76	2.20	1.97	2.73

Setting	Method	MMBench	MMStar	MMMU
	Origin	83.50	45.53	46.09
Inference-only	MLLM-Protector	56.40	35.07	30.69
	ASTRA	83.60	45.53	38.51
	ECSO	83.10	45.47	46.09
	Multi-Agent	75.00	37.53	38.16
Post-training	Safe-RLHF-V	84.40	45.40	46.91
	VLGuard-P	61.60	33.33	28.24
	MIS	**84.80**	**48.47**	**48.07**
	Think-in-Safety	16.00	13.33	14.00
	Ours	71.50	36.47	35.59
	Ours-Judge	83.30	45.40	44.81

- Substantially outperforms existing methods in situational risk identification.
- Maintains competitive performance on conventional safety benchmarks.
- Ours-Judge largely preserves the original backbone's VQA performance.

Cross-Backbone & Stronger VLMs

Backbone	Method	SIUO ↓	MSSBench-Unsafe ↓	Unsafe Avg. ↓	MSSBench-Safe ↓
Qwen2-VL	Origin	63.47	85.67	74.57	**7.33**
	Think-in-Safety	22.16	49.67	35.92	34.33
	Ours	**16.17**	**23.00**	**19.59**	24.17
Qwen2.5-VL	Origin	56.29	74.17	65.23	**16.33**
	Think-in-Safety	**13.77**	49.83	31.80	34.67
	Ours	14.37	**18.67**	**16.52**	30.17
LLaVA-NeXT	Origin	64.67	85.67	75.17	**11.17**
	Think-in-Safety	**28.14**	68.83	48.49	25.33
	Ours	37.13	**31.83**	**34.48**	23.50



- SRR generalizes across VLM backbones
- Scaling general-purpose VLM capabilities alone does not resolve the situational risk supervision gap.

Summary

- Existing safety alignment overlooks risks arising from activity–scenario interactions.
- Situational Risk Reasoning (SRR) provide explicit activity–scenario supervision with structured safety reasoning.
- Substantial improvements in situational risk awareness across multiple VLM backbones.

Code & dataset

github.com/Seeyn/SRR