Can Large **Vision-Language Models** Correct Semantic Grounding Errors **By Themselves**?

Yes*, *come to our poster!

Yuan-Hong Liao[1], Rafid Mahmood[2,3], Sanja Fidler[1,2], David Acuna[2]

[1] University of Toronto & Vector Institute, [2] NVIDIA, [3] University of Ottawa

Can LLMs Self-Correct w/o fine-tuning?

Prior works say,

*"In the context of reasoning, our research indicates that LLMs **struggle to (instrinsically) self- correct** their responses **without external feedback**, and at times, their performance even degrades after self-correction."*, Huang et al., ICLR 2024

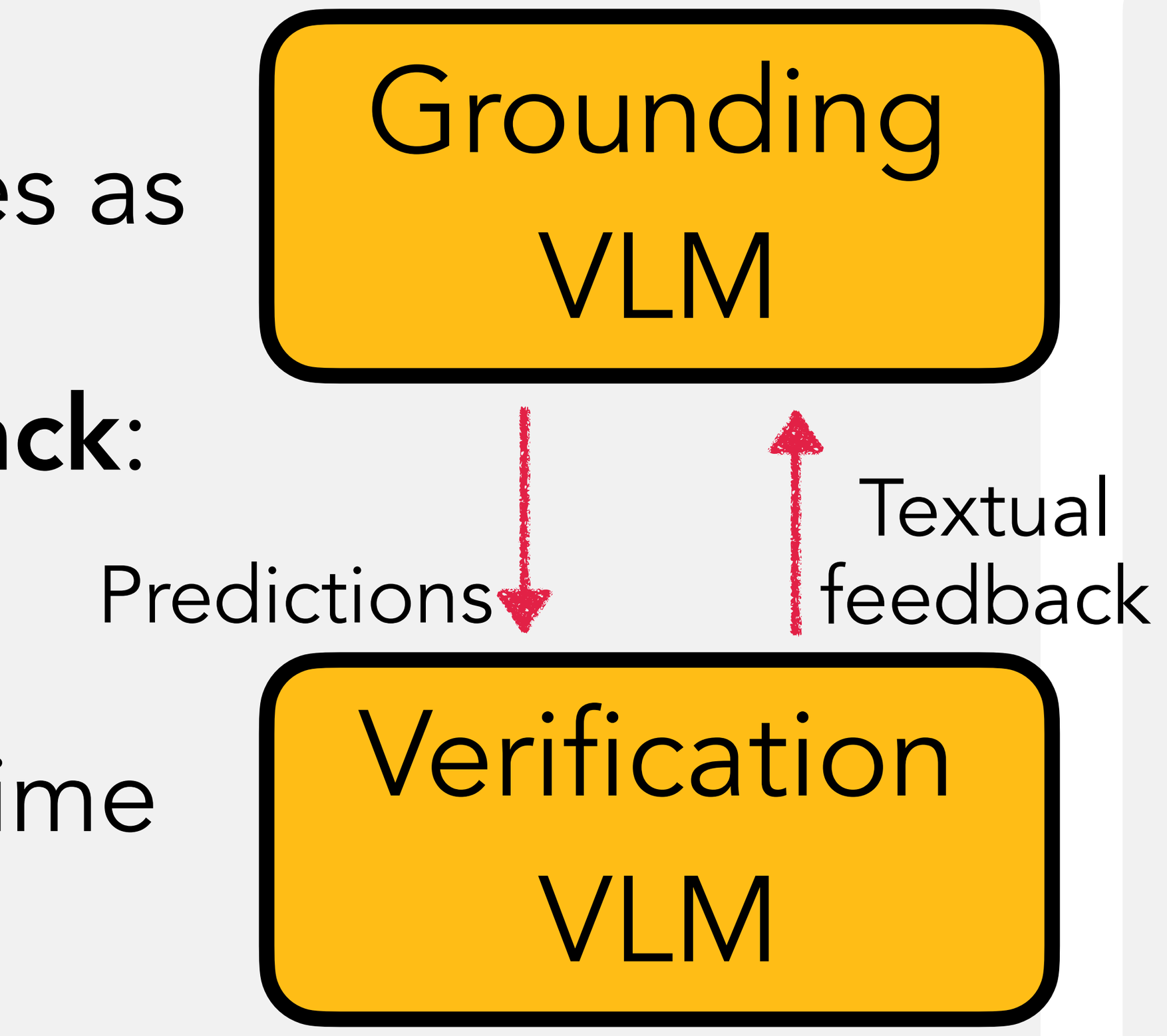
*"There is still **no consensus** on the question of when LLMs can correct their own mistakes, as recent studies also report negative results."*, Kamoi et al., EMNLP 2024

Our Research Questions: Can VLMs Self-Correct? If so, under what context?

Our Key Contributions

💡 Our Idea:

- **Agentic**: Use VLMs themselves as verifiers.
- **Simplicity -> Reliable feedback**: Verification is **easier** than generation
- **Performances**: Achieve test-time scaling! (see our results 🚀)



- Approach: Agentic, Training-free
- Task: Semantic grounding
- VLMs: Open-source VLMs, GPT-4V, and GPT-4o

Main Results

Experimental setup

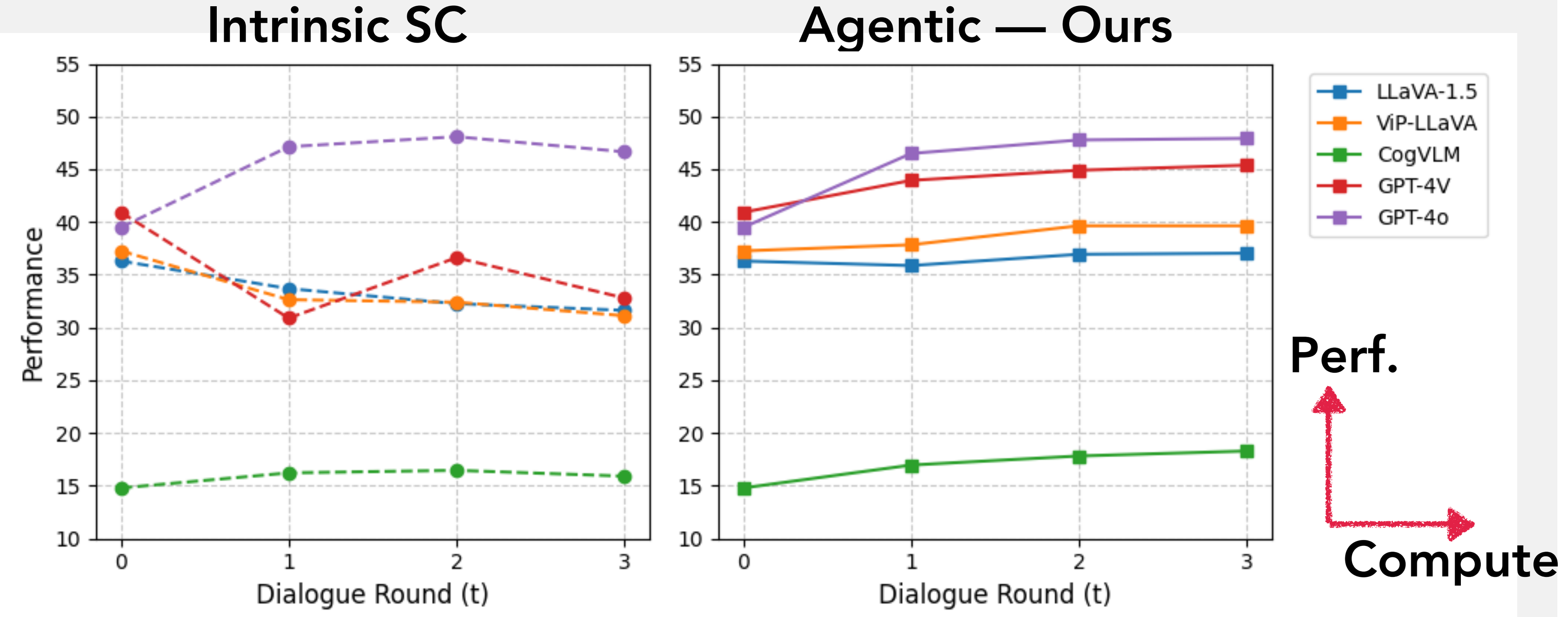
- COCO-inst-seg, ADE-inst-seg
- Baseline: intrinsic self-correction

Intrinsic SC prompt

User: {{ Question }}

VLM: {{ Prediction }}

User: Please carefully review your answer and refine your answer if necessary.



Key results

- Test-time scaling:
  Trade performances with #tokens
- Better VLMs, better gains:
  Gains of GPT-4o > Gains of GPT-4V

Self-correction := VLMs *receive* feedback + VLMs *provide* feedback

Task: Semantic grounding

User: What is inside the specified region?

Automated feedback-based verification

Agent: It is a modern style dinning table in a living room.

System: *Does the specified region contain a modern style dining table in a living room?*

Verifier: No the specified region does not.

System: *A verifier disagrees with your prediction. Review the image again.*

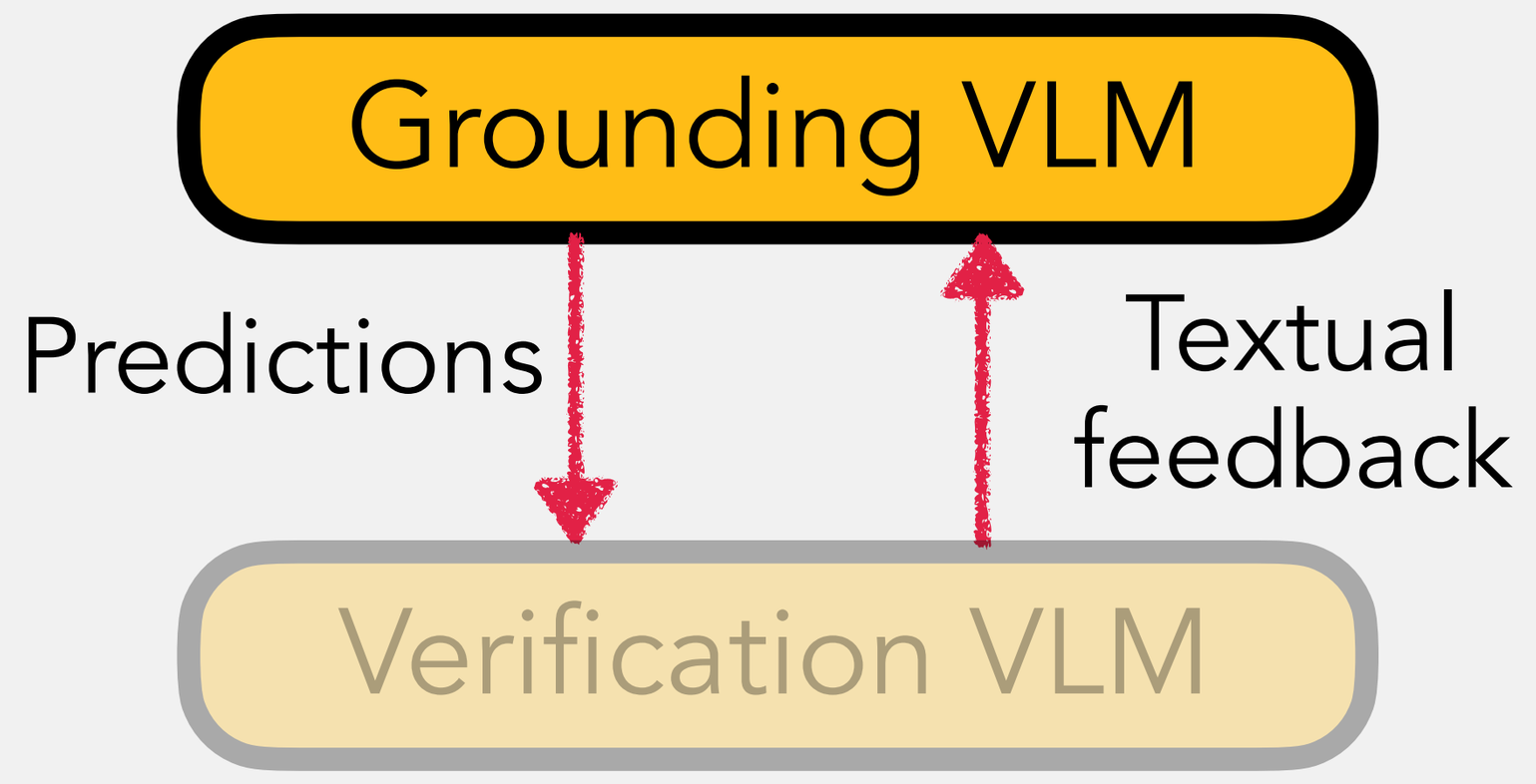
Agent: It is a modern style armchair.

System: *Does the specified region contain a modern style armchair?*

Verifier: Yes.

It is a modern style dinning table in a living room.

VLMs receive feedback



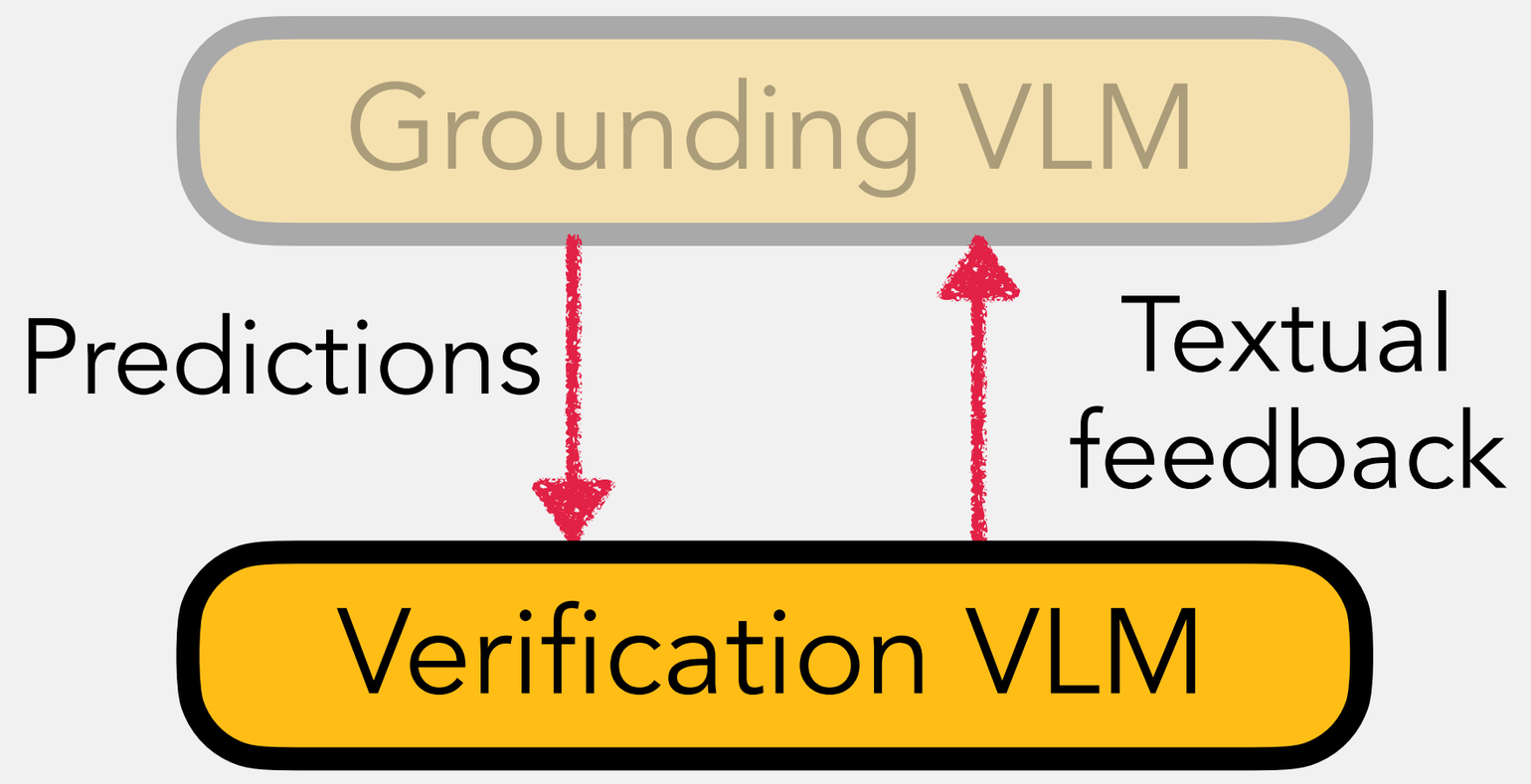
*What (noiseless) feedback? Even **binary** one improves +10%*

	LLaVA-1.5	ViP-LLaVA	CogVLM
Base prediction	35.86	35.86	15.98
+ Binary Feedback	41.04	40.36	16.25
+ Class Label Feedback	94.8	74.99	77.04

How to give feedback? Visual markers ⭕ works the best

	Zero-shot CoT	Visual Prompt	LLaVA-1.5	ViP-LLaVA	CogVLM
Base prediction	No	No	35.86	35.86	15.98
+ Binary Feedback	No	No	41.04	40.36	16.25
	Yes	No	43.3	42	18.25
	Yes	SoM	42.41	44.53	18.64
	Yes	Visual marks	**45.38**	**45.21**	**19.46**

VLMs provide (binary) feedback



How to provide binary feedback? All better than intrinsic SC. Best prompts depends on VLM.

	Visual prompt	LLaVA-1.5	ViP-LLaVA	CogVLM
Intrinsic Self-Correction	N/A	51.12	48.19	21.87
VLM Binary Verification	Visual marks	56.16	**60.47**	39.16
	RoI crop	**61.71**	58.18	**40.68**
	Visual marks + RoI crop	61.14	59.6	39.79

Interested in System-2 thinking in VLMs?

Check our related papers:

1. 📏 EMNLP'24: Enhancing quantitative spatial reasoning with no extra training.
   training-free, **spatial-reasoning**
2. 💭 arXiv'25: Synthesizing System-2 reasoning traces for System-1 Perception
   synthetic-data-generation, **cognitive-behaviors**