

Multimodal Reasoning

for Visual-Centric Long-Horizon Tasks

Zhengyuan Yang



Textual Reasoning with LLMs

- Last year: merely a next-token predictor
- Gold medal-level in IOI and IOM
- 99.95% Codeforces
- Reliably performing hour-long tasks



Scaling Post-Training LLMs for Reasoning

- Consistent, on-policy reward for large-scale online RL training with a good prior
- Reward Model (RM)
 - RLHF (human feedback): human expectations
 - RLAIF (AI feedback): LLM/LMM scoring
 - RLVR (verifiable reward): rule-based



Zooming Into the Reasoning Success



• Train LLMs for reasoning: exploration and consolidation

Thinking by Generating More Token: Chain-of-Thought (CoT)



Image Credit: "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models"

Can We Generate Even More Tokens? Test-time Compute



"Token info density": even more tokens and even better? Yes!

- Not only longer, but more: self-consistency
- Longer beyond pretrain distribution: s1, LIMO
- Longer and more: MCTS, Beam search
- Revision, and other multiple role collaboration

Internalize the Explored Strategies: RL - Why?



Consolidation to think more effectively: RL training

- Expensive: Memorize the correct way of thinking, success with fewer tokens
- Self-improving: Explore further to solve impossible problems before

Internalize the Explored Strategies: RL - How?



- Correctness into a scalar reward, e.g., string matching and relative score
- Pre-trained/mid-trained LLMs as prior
- Problems as environments

Summary on LLM Side





From Text-only to Multimodal: Unique Challenges



- 1. Foundation. Where to start: multimodal model and problems?
- 2. Exploration. What makes effective exploration for multimodal?
- 3. Internalization. What to learn from RL training that generalizes?

In This Talk

1. Foundation. Where to start for multimodal<u>models</u> and <u>problems</u>?

2. Exploration. What makes

effective exploration for

multimodal?

textualized visual CoT

to native multimodal

Exploration Internalization

Foundation

n 3. Internalization. What to learn from RL training that generalizes? <u>unique problems for</u> visual-centric reasoning

Where to Start for Multimodal Model and Problem?



Model: from <u>Texturized Visual Output</u> to Native Multimodal



- Texturized visual output
- Still text-heavy pretrain

- Multimodal-in, text-out
- LLM pretrain, add vision encoder

Model: from Texturized Visual Output to Native Multimodal



- Multimodal-in, text-out
 Multimodal-in, multimodal-out
- LLM pretrain, add vision encoder Multimodal in pretraining

What makes "native multimodal models"

Model: from Texturized Visual Output to Native Multimodal



- In this talk:
 - Texturized visual output
 - Native multimodal models

Problem: Can We Find Good Samples for Exploration?



What Makes a Good Sample to Explore on?



Explored trajectory trains the model better

- Unrelated topic
- Too hard 🗹
- Too easy 🗹
- More criteria? Models own weaknesses

How to Find Good Samples?



• Explicit force model to think longer

RL Training with the Good Samples



How "Good" are these Samples?

Models	Data size	Math/Vista testmini	MathVision mini	MathVerse mini	MMMU	MMStar	MMBench	MM-Vet	AI2D	Avg.	
	Proprietary Models										
OpenAI-GPT-40 OpenAI-01	_	63.8 73.9	36.8 58.2	50.2 57.0	69.1 77.6	64.7 -	83.4	69.1 _	84.6 _	65.21 _	
7B-le	vel General	and Rea	soning	Vision-L	anguag	e Mode	ls				
LLaVA-Onevision-7B InternVL2.5-8B Qwen2.5-VL-7B-Instruct LLaVA-Cot-11B Mulberry-7B Vision-R1-7B OpenVLThinker-7B MM-EUREKA-Qwen-7B	- - 100k 260k 210k 59.2k 15k	63.2 64.4 67.8 54.8 63.1 <u>73.5</u> 70.2 73.0	17.4 22.0 23.6 16.3 - 30.7 29.6 31.9	26.2 39.5 44.5 33.9 39.6 <u>51.9</u> 47.9 50.3	48.8 54.9 50.6 46.2 <u>55.0</u> 50.5 51.9 52.3	61.7 62.8 61.7 57.6 61.3 60.2 63.2 64.1	80.8 82.7 80.7 75.0 79.2 78.9 81.3 79.3	57.5 68.8 66.0 60.3 63.7 65.6 66.9 64.9	81.4 83.3 82.6 78.7 80.1 80.4 82.7 81.4	54.63 59.80 59.69 52.85 - 61.46 61.71 62.15	
	Our 7B-level Reasoning Model										
ThinkLite-VL-7B-Random11k ThinkLite-VL-7B Δ (Ours - Random selection) Δ (Ours - Open 7B SoTA)	11k 11k - -	71.9 75.1 +3.2 +1.6	26.1 32.9 +6.8 +1.0	47.3 52.1 +4.8 +0.2	51.7 55.5 +3.8 +0.5	62.7 65.0 +2.3 +0.9	81.1 81.4 +0.3 -1.3	65.5 67.8 +2.3 -1.0	80.9 83.6 +2.7 +0.3	60.89 64.18 +3.29 +2.03	
72B-le	evel General	and Rea	asoning	Vision-I	anguag	ge Mode	els				
LLaVA-Onevision-72B InterVL2.5-78B Qwen2.5-VL-72B-Instruct QvQ-72B	- - -	67.5 72.3 <u>74.8</u> 71.4	29.3 34.9 <u>35.2</u> 32.7	39.1 51.7 <u>53.3</u> 48.6	56.8 <u>68.7</u> 63.4 70.3	66.1 <u>68.9</u> 68.4 67.2	85.9 87.2 <u>87.4</u> 86.3	63.7 72.3 <u>76.3</u> 75.9	85.6 87.9 87.2 86.6	61.75 67.99 <u>68.25</u> 67.37	
Our 72B-level Reasoning Model											
ThinkLite-VL-72B-Random7.5k	7.5k	76.4	37.1	57.5	65.8	71.3	87.6	76.7	86.9	69.91	
ThinkLite-VL-72B Δ (Ours - Random selection) Δ (Ours - Open 72B SoTA)	7.5k - -	79. 7 +3.3 +4.9	43.8 +6.7 +8.6	64.3 +6.8 +11.0	68.3 +2.5 -2.0	72.0 +0.7 +3.1	88.2 +0.6 +0.8	+0.6 +1.0	87.7 +0.8 -0.2	72.67 +3.06 +4.42	

Qwen-2.5-VL-7BQwen-2.5-VL-72B11k samples7.5k samples

Out of the initial 70k samples

- SOTA benchmark performance at both 7Blevel and 72B-level
- Clear Improvements over base model and baselines (random-selection, full set, majority vote, online filtering, etc.)

When Selected Samples Remain Effective?

- Good sample is the one that fits the model the best, not even strong-to-weak
- Onlineness is not as strict as in rollouts;
 OK to select before each epoch, instead of strictly online



What are "Good" Samples: An Entropy View

- Higher entropy for the model
- Future:
 - Data value attribution
 - From selection to generating questions



What Makes Effective Exploration for Multimodal?



Are There Visual Cues that Can be Texturized?

(a) Question: What is the difference in percentage between the years 1979 and 1986?



Qwen-VL

In 1979, the gross enrollment ratio in tertiary education is approximately 5%. In 1986, the gross enrollment ratio is approximately 1%. The absolute difference is: 5% - 1% = 4%. Therefore, the answer is 4%.

Point-RFT

Be line representing Guinea's enrollment ratio starts at 5% in 1979.

Othe line representing Guinea's enrollment ratio reaches 2% in 1986.

To find the difference, subtract the value in 1986 from the value in 1979: 5% - 2% = 3%.

The difference in percentage between the years 1979 and 1986 is 3%.

Therefore, the answer is 3.

• Visual grounding: Thinking While Pointing

Credit: "Point-RFT: Improving Multimodal Reasoning with Visually Grounded Reinforcement Finetuning"

How to Invoke Pointing?



- Lightweight SFT for textualized pointing format
- RL to explore the pointing usage

Credit: "Point-RFT: Improving Multimodal Reasoning with Visually Grounded Reinforcement Finetuning"

• From Document Domain

• To General Graphics

Question: I am at Airport S. How can I get to Dongshankou?



Qwen-VL

Preight transport No. Mile New Monthly, 191

From Airport S: Take the Orange Line (Line 3) towards Tianhe.

Question: How many bar has a same value?

At Jiahewanggang Station: Change to the Blue Line (Line 2).

At Dongshankou Station: You will arrive at your destination.

Point-RFT

Start at Airport S: This station is located at the top of the map, near the center.

Take Line 3: Line 3 is represented by the orange color. Follow the orange line from Airport S towards the center of the map.

Transfer at Guangzhou East Railway Station: At Guangzhou East Railway Station, transfer to Line 1, which is the yellow line.

Continue to Dongshankou: After transferring to Line 1, continue along the yellow line towards the bottom of the map.

Question: On April 21, weather will be?

Qwen-VL

Point-RFT

The bar representing Finland has a value of

10 The bar representing Hungary also has a

Since both Finland and Hungary have the same

value of 0.01 million tonne-kilometres, there are

value of 0.01 million tonne-kilometres.

3 bar has a same value.

Therefore, the answer is 3.

0.01 million tonne-kilometres.

two bars with the same value.

Therefore, the answer is 2.



Owen-VL

Owen-VI

Point-RFT

18, 75, 16, 76, 24, and 71, The sure of the

17. 14. M. 51. 42. and 14. The sam of the

repairing the name, 275 (Rep/Lean Repl)

9 - 34 - 71 - 275.

18 + 40 + 54 + 381

barafare. the answer is his

Question: is the sum of all values in the

lep/Lean Rep more than Dem/Lean Dem?

Younger Americans more skeptical that

U.S. impacts world problems positively

Products is world world by annual to a second without U.S.

27 000 000 000

42

24

40

Anne Dart-more responses tel attenti. Roverse furnes el 2.5. altate constantes ser le 24. Mich.

21

Warferman

Total

35.45

50-64

March E. K.

Dent-Lake Dark

THE REAL AND CONTUR

The image shows the weather forecast for the week starting from April 16. However, the question asks about April 21.

Since the image only provides information up to April 21, we cannot determine the weather for April 21 from this image alone

Point-RFT

(1) The image shows the weather forecast for the week starting from April 16. However, the question asks about April 21.

233 (3) (5) (6) To find the weather for April 21, we need to count 5 days from Monday, April 16. Counting forward, we have:

- April 17: Tuesday
- April 18: Wednesday
- April 19: Thursday
- April 20: Friday
- April 21: Saturday

According to the image, on April 21, the weather is forecasted to be 12°C with a partly sunny icon.

Credit: "Point-RFT: Improving Multimodal Reasoning with Visually Grounded Reinforcement Finetuning"

New Visual Content during Exploration? Vision Tools



What color is the dog?



from PIL import Image
img = Image.open("input.jpg")
crop = img.crop(100,100,400,400)
crop.save("output.jpg")

Sample credit: "V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs" Method: "O3", "Thinking with images"

New Visual Content during Exploration? Vision Tools

- Great community works
- Action space
 - o Vision tools
 - o Different domains
 - Data and training stages
- Also natural in video and robotics

- V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs
- OpenThinkIMG: Learning to Think with Images via Visual Tool Reinforcement Learning
- ReFocus: Visual Editing as a Chain of Thought for Structured Image Understanding
- Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning
- Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL
- GRIT: Teaching MLLMs to Think with Images
- UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning
- PixelThink: Towards Efficient Chain-of-Pixel

Can Model Imagine Visual Content during Exploration?



 Instead of vision tools for new visual input, model visual generation

Image credit: "Mind's Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models" "CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models"

Thinking in Latent Space

- Mixture of multi modality CoT tokens
- Thinking in latent space





- Promises in LLMs
- But we still need native multimodal models: effective exploration is rooted in pre-training

Image credit: "Training Large Language Models to Reason in a Continuous Latent Space"

What to Learn from RL Training that Generalizes?



Thinking Strategies Learned in LLMs

Textual reasoning

- Language, math, coding
- Human designed for abstraction: communication and reasoning
 O Widely recorded. Mimic human in pretraining
- More tokens for better calculation, planning, programming, etc.

Extending Textual Reasoning with Visual Input

- Great community works
 - o Visual math
 - o Visual documents
 - o Visual coding
 - o Grounding, localization
- Textual CoT for reasoning with visual inputs

- Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models
- SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement
- LLaVA-CoT: Let Vision Language Models Reason Step-by-Step
- VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model
- MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning
- Visual-RFT: Visual Reinforcement Fine-Tuning
- Mulberry: Empowering MLLM with o1-like Reasoning and Reflection via Collective Monte Carlo Tree Search
- VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning
- Too many to enumerate, sorry for missing any :)

Textual reasoning

- Language, math, coding
- Human designed for abstract reasoning: mimic human in pretraining
- More tokens for better calculation, planning, programming, etc.

Multimodal (e.g., vision)

• Visual perception strategies text-only does not work

Image credit: "More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models"

Textual reasoning

- Language, math, coding
- Human designed for abstract reasoning: <u>mimic humans in pretraining</u>
- More tokens for better calculation, planning, programming, etc. /

Multimodal (e.g., vision)

• Visual perception strategies: mimic - humans not always record how

Textual reasoning

- Language, math, coding
- Human designed for abstract reasoning: <u>mimic humans in pretraining</u>
- More tokens for better calculation, planning, programming, etc. /

Multimodal (e.g., vision)

- Visual perception strategies: <u>mimic humans not always record how</u>
 - Perception proxy task: requires good perception to solve
 - Leveraging existing vision models to guide the steps
 - O Invoke pretrained capabilities in native multimodal models

Textual reasoning

- Language, math, coding
- Human designed for abstract reasoning: mimic humans in pretraining
- More tokens for better calculation, planning, programming, etc.

Multimodal (e.g., vision)

- Visual perception strategies: mimic humans not always record how
 - <u>Perception proxy task: requires good perception to solve</u>
 - Leveraging existing vision models to guide the steps
 - Invoke pretrained capabilities in native multimodal models

What Makes a Good RL Task? Challenging yet Easily Verifiable

Challenging yet Easily Verifiable Tasks for Visual Perception?

Image credit: "Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models" "Towards vqa models that can read."

Challenging yet Easily Verifiable Tasks for Visual Perception?

The image showcases a social gathering of Caucasian individuals, both male and female, ranging from middle age to about 60, seated at multiple tables inside a room that appears to be a café or restaurant. The café's walls are a light brown to mustard yellow, adorned with an eclectic mix of picture frames and flags, including one particularly striking black flag with curved white stitching that reads both "true" and "false." There is a tall vertical window on the left side, offering a view of trees and parked cars outside. Hanging from the ceiling are two distinct light fixtures: a black wrought iron chandelier with six gold-colored bulbs, and a single glass pendant light with a black wire. Additionally, a lamp occupies the corner on the left side. Near this window, a woman dressed in black and wearing glasses is seated alone with an iPad on the table, a coffee cup beside her, and she is gazing out the window. Nearby, a group of four individuals, predominantly young men, are engaged in conversation and one is looking at his phone. To the right, there are smaller tables, where pairs of people, including some young women, are conversing. At one table in the lower right corner, a man with headphones and a blue jacket looks down, perhaps immersed in his own world. The atmosphere is lively, with a mix of discussions and some quiet moments of individual focus.

Descriptive Captioning:

• Challenging in visual

perception, but how to grade?

Question: what **number** is on the bike on the right? ---- A: the number is **317**

Visual Question Answering:

• Easily verifiable, but omitting most visual elements

Image credit: "Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models" "Towards vqa models that can read."

Proxy RL Task: Visual Caption Hallucination Critic (ViCrit)

ViCrit Task: Which noun phrase is replaced by the hallucinated noun phrase in this image description?

Perturbed caption: (209 words)

The image captures a detailed overhead view of a grey and white granite or formica tabletop. Central to the composition, towards the bottom, is a rectangular yellow styrofoam tray commonly found in grocery store meat departments. The tray is neatly lined with two rows of raw chicken drumsticks.

Just to the right edge of the tray, a small black jar labeled "Fly and Saucy Medium" in red and white lettering is positioned. Adjacent to this, a red-capped container with a label reading "Zatarain's Creole Seasoning" is placed prominently. Near the top right corner of the tray, a large bottle of Frank's RedHot Wing Sauce, filled to the brim with its fiery red contents, stands upright.

In the background, slightly to the right, a charming small statue of a brown corgi dog adds a whimsical touch. The dog wears a sign around its neck, albeit unreadable, and is flanked by salt and pepper shakers nestled in tiny ornate dog bowls.

In the upper right corner of the image, a brown bottle with a white plastic wrap around its top can be seen. Its white label, featuring "Mark's Lone Star Certified Good Stuff Honey Barbecue" in white lettering against a tan background, completes the scene with a hint of delectable flavor.

Visual hallucination (GT): (2 words)

corgi dog

- Same visual perception difficulty; but easily verifiable
- Incentivize with ViCrit proxy task outcome reward

Image credit: "ViCrit: A Verifiable Reinforcement Learning Proxy Task for Visual Perception in VLMs" "Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models"

Does the Learned Perception Strategies Generalizes?

	Hallucination Benchmark					General benchamrk								
Model Size	Model	CHAIRs	CHAIRi	MMHal	MathVista testmini →	MathVision mini →	MathVerse mini →	MMMU	MMStar →	MM-Vet →	Blind →	Charxiv reasoning →	Avg.	
-	GPT-40 o1	-	-	-	63.8 73.9	36.8 58.2	50.2 57.0	69.1 78.2	64.7 -	69.1 -	50.4 57.0	52.7 55.1	57.10 _	
7B	Molmo-7B-D-0924 LLaVA-OneVision-7B InterVL2.5-8B Qwen2.5-VL-7B-Instruct	36. 35. 29. 28.	7 6.0 0 5.5 2 5.4 0 5.1	3.03 3.12 3.65 3.74	54.1 63.2 64.4 67.8	19.5 17.4 22.0 23.6	23.2 26.2 39.5 44.5	40.2 48.8 54.9 50.6	52.6 61.7 62.8 61.7	59.2 57.5 68.8 66.0	43.3 40.1 47.6 49.3	30.8 31.3 32.9 41.4	40.38 43.28 49.11 50.61	
	ViCrit-RL-7B Δ (Ours - Qwen2.5-7B)	25.	2 4.5 8 -0.0	3.77 5 +0.03	70.7 +2.9	25.7 +2.1	46.3 +1.8	52.0 +1.4	61.9 +0.2	67.1 +1.1	52.6 +3.3	47.8 +6.4	53.01 +2.40	
72B	Molmo-72B-0924 LLaVA-OneVision-72B InterVL2.5-78B Qwen2.5-VL-72B-Instruc	28 27 25 1 26	8 5.7 4 4.9 9 5.2 4 4.8	3.54 3.71 3.89 3.82	61.1 67.5 72.3 74.8	24.7 29.3 34.9 35.2	30.9 39.1 51.7 53.3	48.3 56.8 68.7 63.4	58.4 66.1 68.9 68.4	65.5 63.7 72.3 76.3	46.9 49.6 59.8 61.3	35.2 38.2 42.4 45.5	46.38 51.29 58.75 59.78	
	ViCrit-RL-72B Δ (Ours - Qwen2.5-72B)	21	0 3.9 4 -0.9	3.91 +0.09	77.3 +2.5	40.1 +4.9	59.8 +6.5	66.0 +2.6	69.8 +1.4	77.1 +0.8	65.8 +4.5	49.4 +3.9	63.16 +3.38	

Trained mostly on **natural images** =>

"in-domain": task, image domain

Does the Learned Perception Strategies Generalizes?

		Halluc	cination 1	Benchmark				Gene	ral benc	hamrk			
Model Size	Model	CHAIRs ←	CHAIRi ←	MMHal →	MathVista testmini →	MathVision mini →	MathVerse mini →	MMMU	MMStar →	MM-Vet →	Blind →	Charxiv reasoning →	Avg.
-	GPT-40 01	-	-	-	63.8 73.9	36.8 58.2	50.2 57.0	69.1 78.2	64.7 -	69.1 -	50.4 57.0	52.7 55.1	57.10 -
7B	Molmo-7B-D-0924 LLaVA-OneVision-7B InterVL2.5-8B Qwen2.5-VL-7B-Instruct	36.7 35.0 29.2 28.0	6.0 5.5 5.4 5.1	3.03 3.12 3.65 3.74	54.1 63.2 64.4 67.8	19.5 17.4 22.0 23.6	23.2 26.2 39.5 44.5	40.2 48.8 54.9 50.6	52.6 61.7 62.8 61.7	59.2 57.5 68.8 66.0	43.3 40.1 47.6 49.3	30.8 31.3 32.9 41.4	40.38 43.28 49.11 50.61
	Δ (Ours - Qwen2.5-7B)	-2.8	-0.6	+0.03	+2.9	+2.1	+1.8	+1.4	+0.2	+1.1	+3.3	+6.4	+2.40
72B	Molmo-72B-0924 LLaVA-OneVision-72B InterVL2.5-78B Qwen2.5-VL-72B-Instruct ViCrit-RL-72B	28.8 27.4 25.9 26.4 21.0	5.7 4.9 5.2 4.8 3.9	3.54 3.71 3.89 3.82 3.91	61.1 67.5 72.3 74.8 77.3	24.7 29.3 34.9 35.2 40.1	30.9 39.1 51.7 53.3 59.8	48.3 56.8 68.7 63.4 66.0	58.4 66.1 68.9 68.4 69.8	65.5 63.7 72.3 76.3 77.1	46.9 49.6 59.8 61.3 65.8	35.2 38.2 42.4 45.5 49.4	46.38 51.29 58.75 59.78 63.16
	Δ (Ours - Qwen2.5-72B)	-5.4	-0.9	+0.09	+2.5	+4.9	+6.5	+2.6	+1.4	+0.8	+4.5	+3.9	+3.38

Trained mostly on **natural images** => (w/o VQA-ORM RL)

Math, STEM

Abstract image, Chart

How does the Learned Perception Strategies Generalizes?

Question: Subtract all small blue rubber objects. Subtract all brown shiny balls. How many objects are left?

Question: Can you find the missing shape in this picture puzzle?

t Qwen-2.5-VL-72B-Instruct

To solve the problem, we need to follow these steps:\n\n1. Identify all objects in the image.\n2. Identify the small blue rubber objects and subtract them.\n3. Identify the brown shiny balls and subtract them.\n4. Count the remaining objects.\n\nStep 1: Identify all objects in the image.\n- **There are 9 objects in total: 2 cubes, 1 red cylinder, 1 yellow sphere, 1 green sphere, 1 blue sphere, 1 brown shiny sphere, and 1 cyan cube**.\n\nStep 2: Identify the small blue rubber objects and subtract them.\n- There is 1 small blue rubber sphere.\n- Subtracting this leaves us with 8 objects.\n\nStep 3: Identify the brown shiny balls and subtract them.\n- There is 1 brown shiny sphere.\n- Subtracting this leaves us with 7 objects.\n\nStep 4: Count the remaining objects.\n- After subtracting the small blue rubber sphere and the brown shiny sphere, we are left with 7 objects.

The final answer is: 7

ViCrit-RL-72B

<think>\nTo solve the problem, I need to follow these steps:\n\n1. Identify all objects in the image.\n2. Subtract all small blue rubber objects.\n3. Subtract all brown shiny balls.\n4. Count the remaining objects.\n\nStep 1: Identify all objects:\n- Small yellow metallic cube\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyae metallic cube\n- Small brown shiny sphere\n- Large yellow rubber sphere\n- Small green rubber sphere\n- Small blue rubber sphere\n\nStep 2: Subtract all small blue rubber cylinder\n- Large cyae objects:\n- There is 1 small blue rubber sphere\n- Removing this leaves 7 objects.\n\nStep 3: Subtract all brown shiny balls:\n- There is 1 small brown shiny sphere.\n- Removing this leaves 6 objects.\n\nStep 4: Count the remaining objects:\n- Small yellow metallic cube\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere.\n- Removing this leaves 6 objects.\n\nStep 4: Count the remaining objects:\n- Small yellow metallic cube\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere.\n- Removing this leaves 6 objects.\n\nStep 4: Count the remaining objects:\n- Small yellow metallic cube\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere\n- Small green rubber sphere\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere\n- Small green rubber sphere\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere\n- Small green rubber sphere\n- Small green rubber sphere\n- Small green rubber sphere\n- Small purple rubber sphere\n- Small red rubber cylinder\n- Large cyan metallic cube\n- Large yellow rubber sphere\n- Small green rubber s

Qwen-2.5-VL-72B-Instruct

To find the missing shape in the picture puzzle, let's analyze the pattern:\n\n1. The first shape is a **black hexagon**.\n2. The second shape is a **white hexagon**.\n3. The fourth shape is a **white trapezoid**.\n\nThe pattern seems to alternate between black and white shapes, and the shapes themselves are changing. The sequence of shapes is hexagon, hexagon, and then trapezoid.\n\nGiven this pattern, the **missing shape should be a black trapezoid to maintain the alternating color pattern and the sequence of shapes**.\n\nLooking at the options:\n- A is a black pentagon.\n- B is a white pentagon.\n- C is a black hexagon.\n- D is a black trapezoid.\n\nThe correct option is D, as it is a black trapezoid. The final answer is: D

ViCrit-RL-72B

<think>\nTo find the missing shape in the picture puzzle, let's analyze the pattern:\n\n1. The first shape is a black heptagon (7-sided polygon).\n2. The second shape is a white hexagon (6-sided polygon).\n3. The third shape is missing.\n4. The fourth shape is a white trapezoid.\n\nWe need to determine the pattern in terms of the number of sides and the color of the shapes.\n\n- The heptagon has 7 sides and is black.\n- The hexagon has 6 sides and is white.\n- The trapezoid has 4 sides and is white.\n\nThe pattern seems to be alternating between black and white shapes, and the number of sides is decreasing. Following this pattern:\n\n- The next shape should have 5 sides (since 7, 6, 5, 4 is a decreasing sequence).\n- The color should be black (since the sequence is black, white, black, white).\n\nLooking at the options:\n- A is a black pentagon (5 sides).\n- B is a white pentagon (5 sides).\n- C is a black hexagon (6 sides).\n- D is a black trapezoid (4 sides).\n\nThe correct shape is a black pentagon, which is option A.\n

• Explore the way to perceive an image vs. Mimic text space pattern (SFT)

Image credit: "ViCrit: A Verifiable Reinforcement Learning Proxy Task for Visual Perception in VLMs"

Perceive All Visual Details: ViCrit-Bench

- 4 image categories, 8 hallucinate types
- Benchmark number well foreshadow the VLM performance

"ViCrit: A Verifiable Reinforcement Learning Proxy Task for Visual Perception in VLMs"

Models	General Task Avg.	Overall
OpenAI-GPT-40	_	23.3
OpenAI-o1	-	45.8
OpenAI-o3	-	47.7
Gemini-2.0-Flash	-	19.3
Gemini-2.5-Flash	-	44.4
Gemini-2.5-Pro	-	45.2
Molmo-7B-D-0924	40.48	9.6
LLaVA-OneVision-7B	43.28	12.4
InternVL-2.5-8B	49.11	20.0
Qwen-2.5-VL-7B	50.61	21.9
ViCrit-RL-7B	53.01	35.6
Molmo-72B	46.38	18.2
LLaVA-OneVision-72B	51.29	24.5
InternVL-2.5-78B	58.75	32.7
Qwen-2.5-VL-72B	59.78	42.4
ViCrit-RL-72B	63.16	43.0

Textual reasoning

- Language, math, coding
- Human designed for abstract reasoning: mimic human in pretraining
- More tokens for better calculation, planning, programming, etc.

Multimodal (e.g., vision)

- Visual perception strategies: humans not always document how
 - Perception proxy task: requires good perception to solve
 - Leveraging existing vision models to guide the steps
 - O Invoke pretrained capabilities in native multimodal models

Can We Leverage the Powerful Vision Models?

- LMMs are not good enough in perception
- There are so many powerful vision models

How to Effectively Select? Vision Value Model based on CLIP

- Vision expert to guide decoding steps during process
- Vision value model (VisVM): Learn how good the image description is in the future

Image credit: "Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension"

Discussion: from H(text|image) to H(image|text)

- ViCritic proxy task outcome reward
- Vision value model based on CLIP Pushing the limit of multimodal reasoning with text output

Credit: Aditya Ramesh

Discussion: from H(text|image) to H(image|text)

- ViCritic proxy task outcome reward
- Vision value model based on CLIP Pushing the limit of multimodal reasoning with text output

Great, but an upper bound: non-verbal

Credit: Aditya Ramesh "SITE: towards Spatial Intelligence Thorough Evaluation"

Discussion: from H(text|image) to H(image|text)

- ViCritic proxy task outcome reward
- Vision value model based on CLIP Pushing the limit of multimodal reasoning with text output
- Great, but an upper bound: non-verbal
- Non-verbal problems: Native multimodal models with text and visual output

From H(text|image) to H(image|text)

Better Image Generation with Enriched Image Prompt

Image credit: "T2I-R1: Reinforcing Image Generation with Collaborative Semantic-level and Token-level CoT"

Scaling Test-time Compute for Better Image Prompt

Image credit: "ImageGen-CoT: Enhancing Text-to-Image In-context Learning with Chain-of-Thought Reasoning"

From Text to Multimodal Tokens

Semantic-level CoT Prompt **Token-level CoT T2I-R1 Janus-Pro** The image depicts a beautiful and intricate igloo ... A structure made of The igloo has an aerodynamic dome shape, showcasing ice, traditionally built Its surface appears smooth and translucent by the Inuit people The entrance to the igloo consists of a gently sloping tunnel... A specific type of The image portrays tulips, the famous flower that is grown in the flower cultivated in country where Amsterdam is located. the country where The vibrant colors of these flowers, ranging from yellow and red to Amsterdam is located purple... An image of a circular A circular emblem depicting the balance and harmony of Taoist emblem representing philosophy features the Yin and Yang symbol prominently in its balance and harmony in center. The emblem showcases a harmonious arrangement... Taoist philosophy Underneath the train, on the bottommost part of the platform, lies a a pig on the bottom of tiny pig with a brown-yellowish coat. The pig is curled up, resting a train peacefully, surrounded by the iron structure of the train tracks and the dimly lit platform ...

• Trained by the final reward on visual output (preference, count, VQA, etc.)

Image credit: "T2I-R1: Reinforcing Image Generation with Collaborative Semantic-level and Token-level CoT"

From Visual Generation to Non-Verbal Reasoning

- Imagining visual states (latent, symbolic, image) for non-verbal reasoning
- Gap: <u>Visual Thinking!</u> (modality interleave) learned in pretraining

Image credit: "Unfolding Spatial Cognition: Evaluating Multimodal Models on Visual Simulations"

Summary

1. Foundation.

Model+Problems. Scaling compute for problem selection is also beneficial!

3. Internalization.

Unique tasks and assets to internalize perception strategies and visioncentric reasoning (ViCrit task, VisVM)

2. Exploration. Visual-centric!

From textualized CoT to

native multimodal

Exploration Internalization

Foundation

"Man is a tool-using animal."

Reasoning to act; Acting to reason better.

Credit: Thomas Carlyle

Thank you!

Zhengyuan Yang

zhengyuan.yang13@gmail.com zhengyuan.info

Openings for full-time and intern :) Looking forward to chatting with you!