

Toward Cognitive Supersensing in Multimodal Large Language Models

Paper accepted by NeurIPS 2026

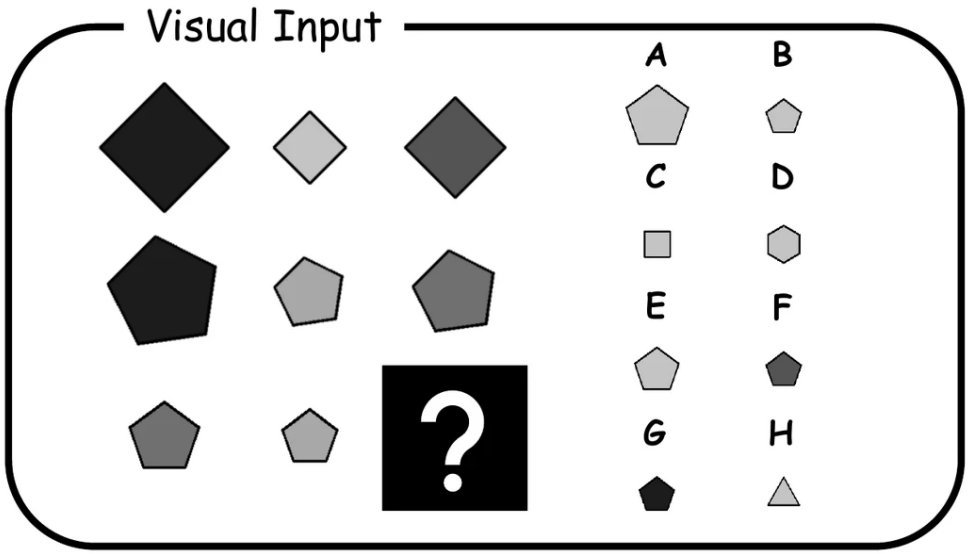
Presenter: Boyi Li

Perception and LANguage (PLAN) Lab & Rehg Lab
& PediaMed AI

A familiar test of visual intelligence

Think of the pattern questions in children's intelligence tests.

Today's multimodal models can describe what they see.
Can they also reason with what they imagine?



Correct: B · column subtraction

Show GPT-5's response **Reveal the answer**

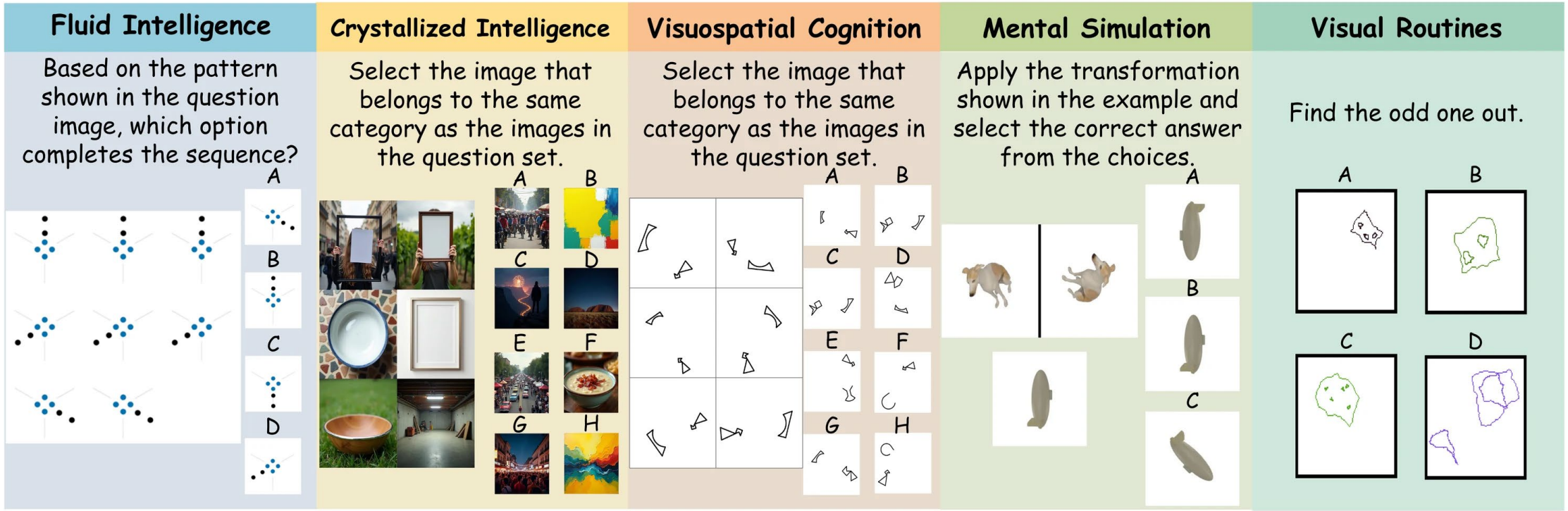
GPT-5 · full response

Step 1: View the question image as a 3×3 grid: each row keeps the same shape type and follows the same left→right progression in size and shading.
Step 2: From rows 1–2, the pattern is large & darkest → small & lightest → medium & dark (same structure for each row).
Step 3: In row 3, the missing cell must therefore be a medium-sized dark pentagon, which matches choice F.
So, the final answer is F.

CogSense-Bench

105.4K training pairs

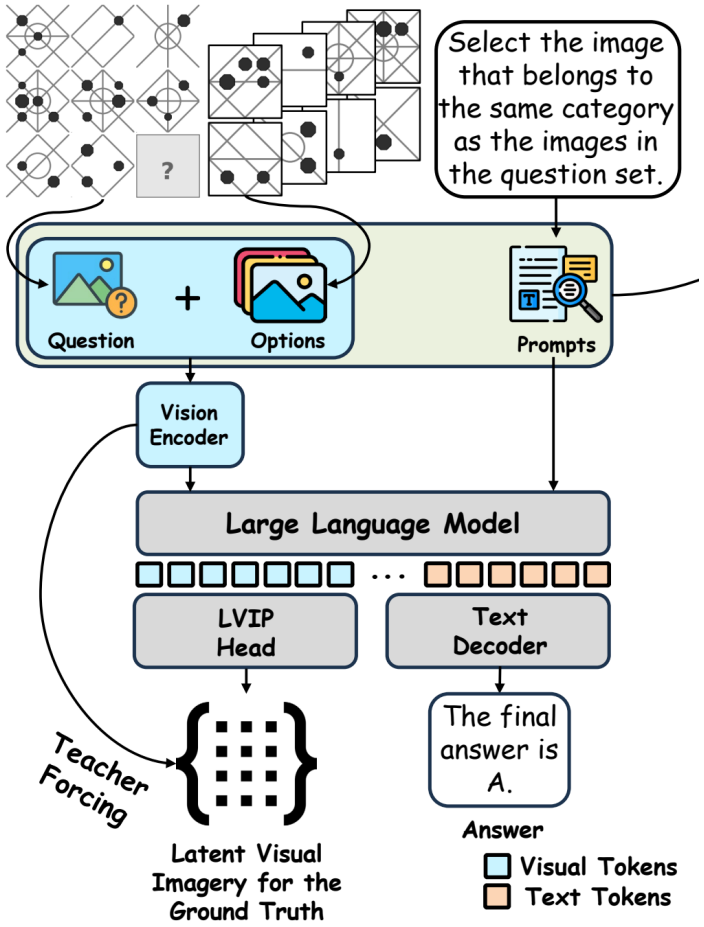
1,000 held-out test questions



Click a dimension to explore an example

Continue →

Model architecture · Two outputs from one backbone



Architecture from the paper

Question + option images
Vision encoder + projector

Prompt
Token embeddings

Shared backbone · Qwen3-VL-8B

Select option-image states and average pool

Text decoder

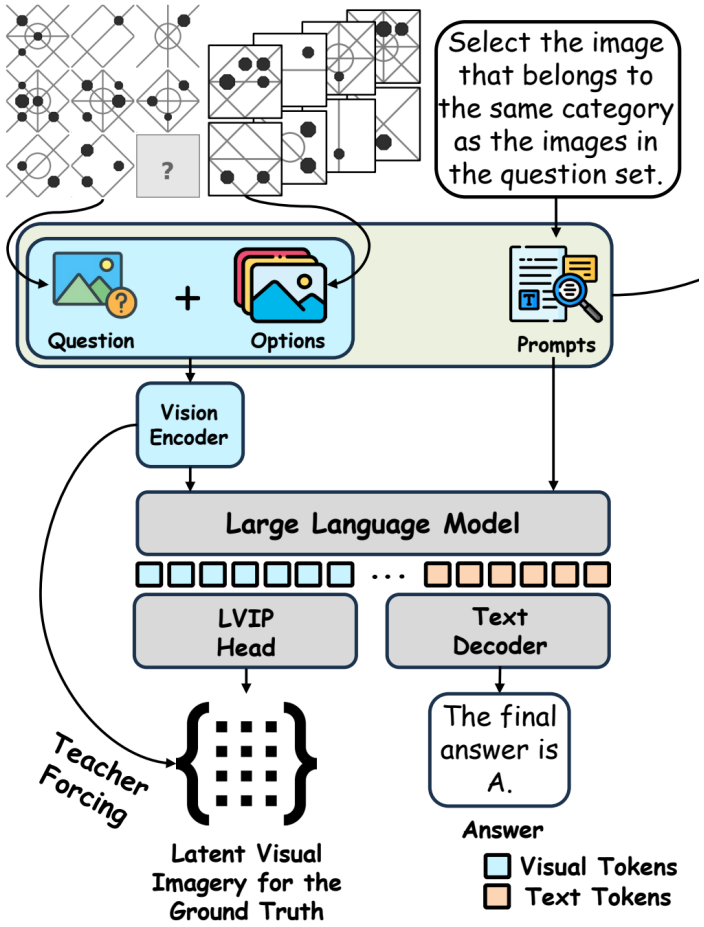
LVIP head · two-layer MLP

Rationale Z + final answer y

Answer-oriented visual latent

Language output

LVIP head · Prediction and visual target



Architecture from the paper

Prediction branch

Option-image hidden states

↓

Average pool over M tokens

↓

LVIP head · two-layer MLP

↓

$$\hat{\mathbf{h}}_y = g_\psi\left(\overline{\mathbf{H}}_{\text{opt}}\right) \in \mathbb{R}^{1\times d_{\text{vis}}}$$

Training target

Correct option image

↓

Frozen vision encoder

↓

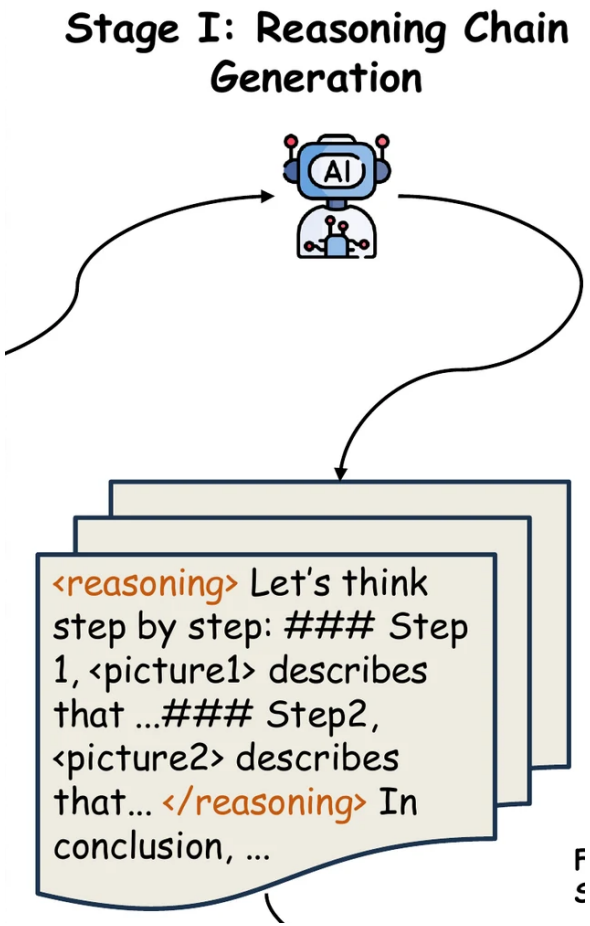
MLP on visual features

↓

$$\mathbf{h}_y = \text{MLP}\left(\mathbf{v}_y\right)$$

Prediction and target share the visual feature space.

Stage I · Build the reasoning corpus



Generate

GPT-5 writes a step-by-step rationale and predicts an answer for each question.

Filter

Remove chains with an incorrect answer or hallucinated content.

Output: (input, rationale, correct answer)

Stage II · Joint text and visual supervision

Shared backbone conditioned on input and teacher rationale

Text decoder → next-token probabilities

Option-token states → AvgPool → LVIP head

Text generation

$$\mathcal{L}_{\text{text}} = -\sum_{t=1}^{|x|} \log q_\theta(x_t | X, x_{<t}), x = (Z, y)$$

Visual latent alignment

$$\mathcal{L}_{\text{lvip}} = \text{MSE}(\hat{\mathbf{h}}_y, \mathbf{h}_y)$$

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|x|} \log q_\theta(x_t | X, x_{<t}) + \beta \text{MSE}(\hat{\mathbf{h}}_y, \mathbf{h}_y)$$

Stage III · Two rewards for rationale training

Initialize from Stage II and freeze the answer scorer and LVIP head.

Sample m rationale candidates from the trainable policy

Answer evidence

$$R_{\text{ans}}(Z; X, y) = \log q_{\theta_0}(y|X, Z)$$

Visual grounding

$$R_{\text{lvip}}(Z; X, y) = -\left\| g_\psi\left(\overline{\mathbf{H}}_{\text{opt}}(X, Z; \theta)\right) - \mathbf{h}_y \right\|_2^2$$

$$R(Z; X, y) = \alpha R_{\text{ans}}(Z; X, y) + \gamma R_{\text{lvip}}(Z; X, y)$$

The two rewards score answer evidence and visual consistency.

Stage III · Anchor rewards and policy training

Anchor rewards

Score sparse rationale prefixes at stride λ and interpolate between anchors.

$$\tilde{R}(\tau_{t+i}) = \tilde{R}(\tau_t) + \frac{i}{t^+ - t}\left[\tilde{R}(\tau_{t^+}) - \tilde{R}(\tau_t)\right]$$

$$\tilde{R}(\tau_t) = R(\tau_t; X, y)\ \text{(anchors)}, \qquad t^+ = \min(t + \lambda, L)$$

Policy training

Sample candidates, filter by the reference score, and optimize SubTB.

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} \mathbb{I}(Z_i)\mathcal{L}_{\text{SubTB}}(Z_i; \theta)$$

Update the rationale policy while keeping the scorer and LVIP head frozen.

Inference · Evidence scoring and answer selection

Sample multiple rationales and decode an answer for each.

Length-normalized evidence from the frozen scorer

$$S_i = \frac{1}{|Z_i| + |y_i|} \log q_{\theta_0}(y_i | X, Z_i)$$

Final prediction

$$i^* = \arg\max_i S_i; \quad \hat{y} = y_{i^*}$$

The scorer evaluates predicted answers without the correct test label.

Cognitive ability · Closed and open models

Model	Fluid	Crystal.	Spatial	Simul.	Routines	Avg.
Human reference	82.7	91.3	88.5	97.9	78.7	88.4
Gemini 2.5 Flash	23.2	40.2	31.0	40.2	45.3	36.3
GPT-o3	4.7	51.4	20.4	38.7	43.0	32.3
GPT-5.2	29.4	35.9	57.5	60.0	37.6	40.3
Claude Sonnet 4	22.5	31.3	26.6	58.0	34.4	32.6
Grok 4 Fast	13.0	45.4	41.6	21.3	37.6	31.7
Llama-4-Scout-17B	20.3	29.9	35.4	48.7	41.9	31.8
Gemma-3-27B	18.5	29.4	39.8	55.3	43.0	32.7
Ministral-3-14B	9.4	10.6	10.6	39.3	31.2	16.5
Intern-S1	6.6	21.0	12.4	17.5	42.6	17.4
Qwen3-VL-30B	30.8	34.0	37.2	56.0	40.9	37.4
CogSense-8B	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

Accuracy (%) · Human reference: 20 volunteers, 100 sampled questions

Cognitive ability · Fine-tuned comparisons

Model	Fluid	Crystal.	Spatial	Simul.	Routines	Avg.
Qwen3-VL-8B (FT)	51.1	76.6	63.7	59.3	41.9	62.3
InternVL3.5-8B (FT)	52.2	80.2	63.7	59.3	41.9	63.9
Qwen3.5-9B (FT)	56.5	89.1	61.9	62.7	46.2	69.1
CogSense-8B	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

+11.5 points

over the same backbone with standard SFT

+4.7 points

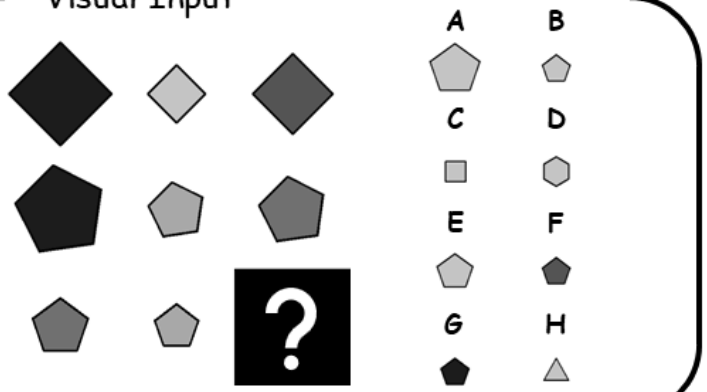
over Qwen3.5-9B (FT)

Qwen3-VL-8B backbone · 8 × H200 · AdamW · learning rate / weight decay: 10^{-5}

Qualitative analysis · Complete model responses

Q: Based on the pattern shown in the question image, which choice completes the pattern?

Visual Input



Response from Gemini 2.5 Flash

Step 1: By examining the properties across the rows, three rules emerge: All items in Row 3 are Pentagons. Each row contains Small, Medium, and Large. Row 3 is missing Large. Each row contains Light, Medium, and Dark. Row 3 is missing Dark/Black.
Step 2: Comparing Row 2 and Row 3 reveals a shift pattern: the elements of Row 2 are moved one position to the left to create Row 3, with the first element wrapping around to the end. This confirms the missing piece corresponds to the first element of Row 2: a Large, Black Pentagon.
Step 3: No option perfectly matches the derived "Large, Black Pentagon." G is the closest one.

Answer: G

Errors occur during reasoning

Response from GPT-5

Step 1: View the question image as a 3×3 grid: each row keeps the same shape type and follows the same left→right progression in size and shading.
Step 2: From rows 1–2, the pattern is large & darkest → small & lightest → medium & dark (same structure for each row).
Step 3: In row 3, the missing cell must therefore be a medium-sized dark pentagon, which matches choice F.

So, the final answer is F.

Errors occur during reasoning

Response from CogSense-8B

Step 1: The first row consists of diamonds, and the second row consists of pentagons. Since the third row starts with pentagons, the missing figure must also be a pentagon.
Step 2: There is a subtraction relationship between the columns for both size and color shade (Value 1 - Value 2 = Value 3). Size: Large (3) - Small (1) = Medium (2). Color: Black (3) - Light Grey (1) = Dark Grey (2).
Step 3: Apply Logic to the Third Row: Size: Medium (2) - Small (1) = Small (1).Color: Dark Grey (2) - Light Grey (1) = Light Grey (1). The missing figure must be a Small Light-Grey Pentagon.

Looking at the options, B is the correct answer.

General ability · General and spatial benchmarks

Benchmark	Qwen3-VL-8B	CogSense-8B	Δ
GQA	71.4	71.8	**+0.4**
ScienceQA	92.6	92.6	**0.0**
RealWorldQA	71.5	71.9	**+0.4**
BLINK	64.7	65.3	**+0.6**
HallusionBench	61.1	60.5	**−0.6**
MMSI-Bench	28.0	32.4	**+4.4**
ViewSpatial-Bench	38.6	40.8	**+2.2**
ERQA	46.8	48.8	**+2.0**

General scores remain close to the backbone; the largest gains are spatial.

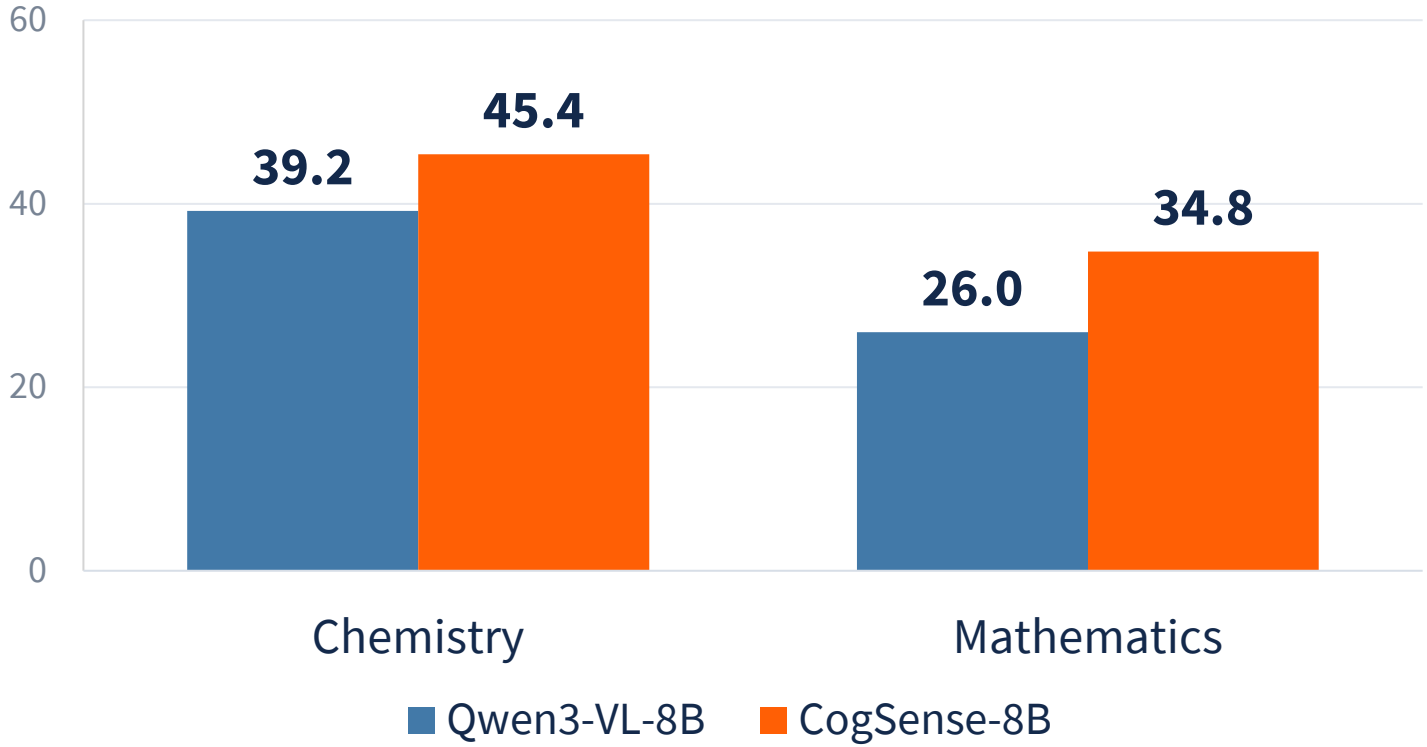
Ablation · Contributions of LVIP and rationale RL

Variant	Fluid	Crystal.	Spatial	Simulation	Routines	Avg.
Base	31.2	34.8	31.0	45.3	40.9	35.5
SFT	51.1	76.6	63.7	59.3	41.9	62.3
SFT + LVIP	55.4	88.6	61.1	61.3	44.1	68.0
SFT + GRPO	55.8	79.9	63.7	63.3	43.0	65.5
SFT + LVIP + GRPO	59.1	89.9	64.6	65.3	46.2	70.8
SFT + GRPO + LR	56.5	80.5	64.6	64.0	44.1	66.2
SFT + LVIP + GRPO + LR	59.8	90.2	65.5	66.7	47.3	71.5
SFT + LR	57.2	80.7	65.5	65.3	45.2	66.9
CogSense: SFT + LVIP + LR	**63.8**	**91.0**	**69.0**	**68.0**	**50.5**	**73.8**

LVIP: +5.7 over SFT **Full recipe: 73.8** **Full recipe + GRPO: 71.5**

All variants use Qwen3-VL-8B · LR = latent-rationale optimization

EMMA · Out-of-domain science and math



+6.2

points in chemistry

+8.8

points in mathematics

Accuracy (%) · selected EMMA subsets with images in questions and answer options

Thank you

Questions?

Boyi Li

pediamed.ai/Cognition-MLLM/cogsense