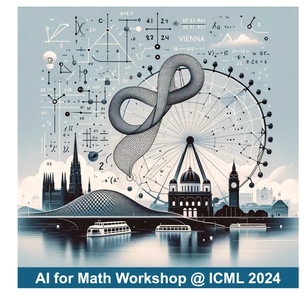
Smart Vision-Language Reasoners

Denisa Roberts[1], Lucas Roberts[2,3],
[1]New York University, [2]Yext Inc, [3]Scipy Core Maintainer.

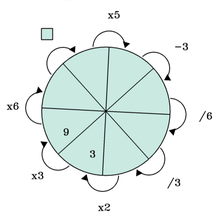


Introduction

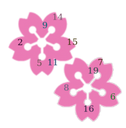
Problem Formulation:
Introduced in Cherian (2022) the Simple Multi-modal Algorithmic Reasoning Task (SMART) tests vision-language models' intelligence along eight meta-cognitive dimensions: math, algebra, counting, measure, path, spatial, logic, and pattern.

Motivations:

- Mathematical reasoning and general problem-solving use, indeed require, intelligence.
- Intelligence is related to multimodal reasoning.



Math Question: *What do we need to put in the square to get a correct diagram?*
Answer Options: A: -3; **B: /9**; C: x6; D: x2; E: 2;



Algebra Question: *Unique values are written on two flowers. One value on a flower is covered. The sums of the values on the two flowers are equal. What is the covered value?*
Answer Options: A: 5; B: 4; C: 2; **D: 0**; E: 6.

Spatial Question: *Cristina made a setup using some green blocks and 94 white blocks. How many of these white blocks are not visible in the figure?*
Answer Options: A: 28; B: 61; **C: 64**; D: 90; E: 79.

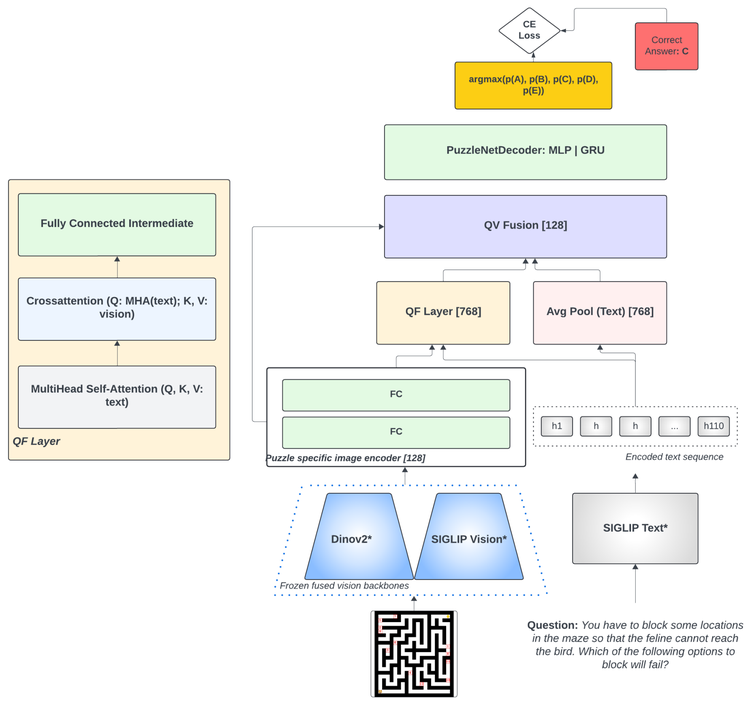
- Intelligence is related to better abstractions and those are related to learning better representations.
- Vision-Language Models (VLM) struggle in there ability to reason multimodaly, especially in their ability to use the diagram.

Key Contributions:

- Introduce a novel multimodal QF-layer to learn a hidden representation from the vision and language modalities.
- Strengthen the vision modality by learning an adaptive visual representation on top of two fused frozen vision backbones, SigLIP and DinoV2.
- Strengthen the text-vision alignment by using a frozen SigLIP language encoder which does not overpower the visual signal.
- Include a composite hidden representation through the concatenation of language-only representations, an adaptive image-only representation learned on top of the fused frozen foundation backbones, and the QF multimodal layer representation which includes a language-vision cross-attention sublayer.
- Improve the general architecture through GELU activations, residual connections, and layer normalization.
- The smarter VLM show up to 48% accuracy gain across several of the meta reasoning skills measured by the challenging SMART task.

Methodology

Smarter VLM Reasoner Architecture. The smarterVLM reasoner architecture is shown in the right panel and the novel QF layer in the left. Both the fused vision (DinoV2+SigLIP) and the language (SigLIP) backbones are frozen. All other layers (QF-layer, adaptive visual layer, and MLP and GRU decoders) are trained from scratch.



Composite Deep Representation Details. The smarter VLM reasoner architecture benefits from learning a composite representation,

$$CR = LN(Concat([r_1, r_2, r_3])),$$

as a concatenation of 3 components, r_1 (adaptive vision-only), r_2 (QF text-and-vision multimodal) and r_3 (text-only):

$$r_1 = FC_{1i}(GELU(FC_{2i}(y)))$$

for $i \in \{1, \ldots, 101\}$, a distinct puzzle group, where

$$y = Concat([Dino(x), SigLIP(x)]),$$

$$r_2 = LN(x + Drop(FC(GELU(FC(x))))),$$

where

$$x = MHCrossA(MHA([h_1, h_2, ..., h_{110}]), r_1),$$

and

$$r_3 = AveragePooling([h_1, h_2, ..., h_{110}]),$$

where h are SigLIP-encoded text tokens.
The QV-Fusion layer is:

$$LN(GELU(FC(GELU(FC(CR))))).$$

Experiments & Results

- For efficiency, we train the smarter vision-language reasoners on a subset of the Smart-101 dataset (scan QR code on poster for a link to the dataset).

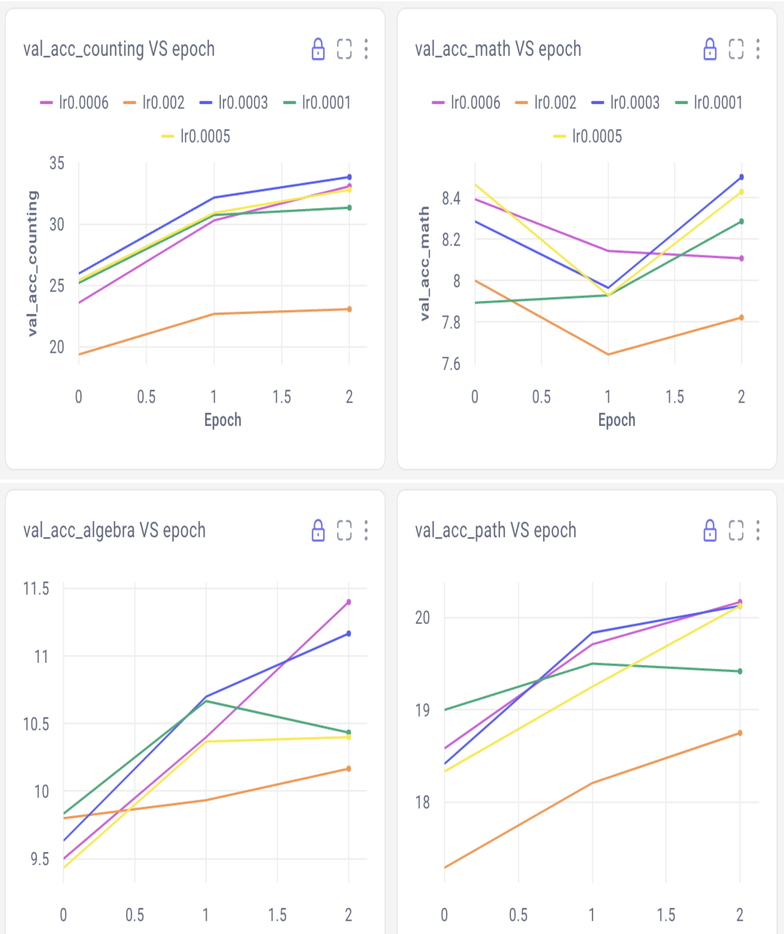
Code, Dataset, & Model

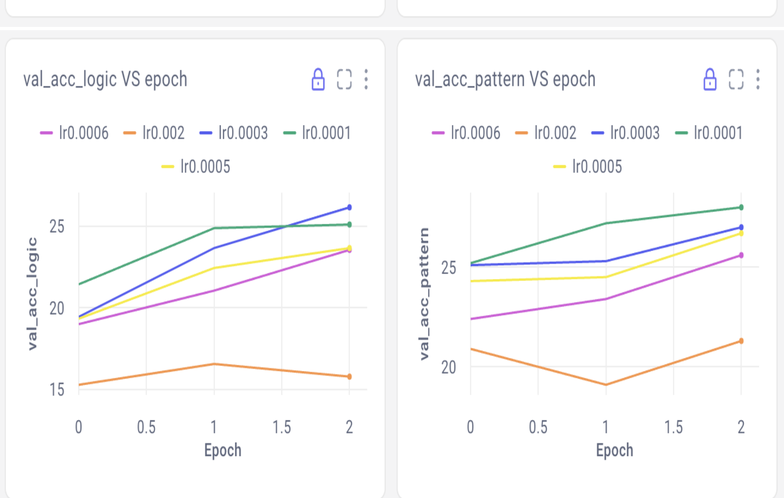
- Ablation studies indicate that the use of both the fused DinoV2+SigLIP vision backbone and the QF multimodal layer with cross-attention improve reasoning ability along the eight meta-cognitive skills. The use of cross-attention improves the ability of the reasoner to make use of the puzzle's visual cues.

Statistics of the SMART-101 dataset

Skill class	Algebra	Arithmetic	Spatial	Logic	Measure	Path	Pattern	Count
-	13.9%	12.9%	11.9%	11.9%	8.9%	7.9%	6.9%	25.7%

Validation accuracy curves for 5 learning rates





Quantitative results on test dataset by skill

The test data broke out by skill is shown in the table, the model with the best improvement in the skill (column) relative to the baselines is shown in bold.

Neural Net	Counting	Math	Logic	Path	Algebra	Measure	Spatial	Pattern	Overall
BERT+ResNet50	23.4(-)	9.6(-)	17.9(-)	17.5(-)	10.5(-)	9.9(-)	25.8(-)	20.3(-)	17.1(-)
SmarterVLM lr0.001	29.0(+24%)	9.9 (+3%)	21.2 (+18%)	17.9(+2%)	10.8 (+3%)	11.1 (+12%)	23.2 (-10%)	25.7 (+27%)	19.12 (+12%)
SmarterVLM lr0.0005	32.9(+41%)	10.0(+4%)	22.8(+27%)	**19.5(+11%)**	11.2(+7%)	**11.6(+17%)**	26.3(+2%)	25.8(+27%)	20.86(+22%)
SmartestVLM lr0.0003	**34.7(+48%)**	9.5(-1%)	**25.7(+44%)**	**19.5(+11%)**	**11.3(+8%)**	11.1(+12%)	**26.7(+3%)**	**27.4(+35%)**	**21.59(+26%)**
SmarterVLM (no QF)	32.3(+38%)	**10.3(+7%)**	23.3(+30%)	18.8(+7%)	10.0(-5%)	10.1(+2%)	25.8 (+0%)	23.6(+16%)	20.14(+18%)

[2] *This research effort conducted independently of Yext employment*