On VLM Reasoners

Abstract

In this article I investigate vision-language models (VLM) as reasoners. Going deeper with transformers-like layers as well as proper hyperparameter and other training choices, as well as stronger visual grounding from fused frozen pretrained backbones, led to strong improvements (up to 48% *gain in accuracy) over baselines on the SMART task, further underscoring the power of deep multimodal learning. Eight different algorithmic reasoning skills are evaluated -math, counting, path, measure, logic, spatial and pattern- and the smarterVLM which includes a novel QF layer improves upon best previous baselines in every skill.*

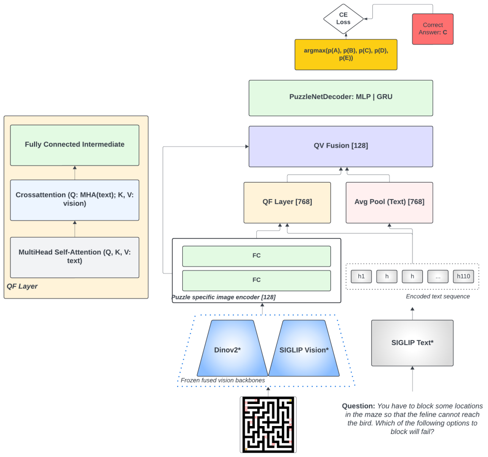
CE Loss
Correct Answer: C
argmax(p(A), p(B), p(C), p(D), p(E))
PuzzleNetDecoder: MLP | GRU
Fully Connected Intermediate
Crossattention (Q: MHA(text); K, V: vision)
MultiHead Self-Attention (Q, K, V: text)
QF Layer
QV Fusion [128]
QF Layer [768]
Avg Pool (Text) [768]
FC
FC
Puzzle specific image encoder [128]
h1
h
h
...
h110
Encoded text sequence
Dinov2*
SIGLIP Vision*
Frozen fused vision backbones
SIGLIP Text*
Question: You have to block some locations in the maze so that the feline cannot reach the bird. Which of the following options to block will fail?

Figure 1. The smarterVLM reasoner architecture (right) and the novel QF layer (left). Vision (DinoV2+SigLIP) and language (SigLIP) backbones are frozen. All other layers are trained from scratch.

1. Introduction/Background/Motivation

Opportunities for improvement of VLM were mentioned across several recent articles: problem-solving and algorithmic reasoning ability of transformers including VLMs is limited; there are still challenges on the vision modality; architectures for encoding, decoding and aligning are still to be explored. In the [17] several VLM design choices are explored. We take inspiration from their comprehensive approach to conduct two parallel and complementary investigations of our own in the VLM design space: reasoning ability and text-conditioned visual features. How we define intelligence (artificial or not) is still an open question. In "On the Measure of Intelligence" [10] the author conducts an in depth discussion and formulates a formal definition of intelligence based on algorithmic information theory, introducing the concept of algorithmic reasoning. Intelligence is also related to multimodal reasoning, as we humans use all our senses to get input for our higher abstractions to be built from. Better abstractions are akin to better mental representations. Deep neural networks excel at learning (artificial) representations [6]. **Why is this an important problem?** If we make neural networks better at algorithmic reasoning (I include here a few types of reasoning such as creative problem solving in math, physics, logic and coding algorithms, puzzles and IQ tests, learning, planning and decision making), this skill may be transferable to science, strategy, medical, law and commonsense, with far reaching real world impact. In the reasoning realm, many recent works focused on evaluating vision-language models on more general multimodal tasks [25, 35] or applying a chain of thought approach for in context learning [37,38]. In [4] large language models are pretrained to solve text only questions bringing on the full power of heavy-weight LLM, but without taking the visual signal into account. Then a separate line of work builds very large VLM akin to LLM. In [19] a vision-language architecture, the Query Transformer, adds transformer layers to frozen image and text encoders and learns in a contrastive pretraining paradigm on massive datasets. Llava [20, 21] versions [20] emerge as a multimodal instruction tuned large model finetuned on a science dataset. In [13] authors use an LLM and parameter-efficient visual instruction tuning, focusing on learning efficiently only the adapters, with early fusion of visual tokens in LLM layers. In [31] the deficiencies in visual grounding of large multimodal models is investigated and mixture of visual features are proposed to improve the vision modality. In [18], another pretrained and visual instruction tuned framework is proposed, employing Clip [29], Llama [32] and Perceiver adapters (from Flamingo) [1] as well as a dataset, MIMIC-IT. The pretrained vision encoder in DinoV2 [28] aims to leverage different techniques and diversity of images to pretrain backbones exactly for the purpose of using them as general-purpose features, as aimed in this investigation. However, it is not necessarily true that even this general purpose encoder will encode all the visual signal, so fusing with representations from a backbone pretrained in a different fashion, for instance the SigLIP [36] representation, may provide additional visual signal boost. Specifically, SigLIP improves on CLIP [29] for language-image pretraining by employing a sigmoid loss instead of the constrastive learning with softmax normalization and performs better across tasks.

2. The VLM Reasoning Problem

So how can we help (deep) artificial neural networks reason better? We can get inspired from the current works on the very large VLM and build upon them. In [9] experiments show that the visual signal is very important in solving complex multi-reasoning skill puzzles and that language only really large models lag behind. In the article "Are Deep Neural Networks SMARTer than Second Graders?" [9] another task, a Simple Multimodal Algorithmic Reasoning Task(SMART), is introduced with visuo-linguistic puzzles designed for children in the 6-8 age group (from the US Kangaroo Olympiad), which align with intelligence desiderata in [10]. The puzzles measure intelligence across 8 different skill classes: counting, math, logic, path, measure, logic, pattern. Problems include an image and a text question and are formulated as multiple choice. We can see a few examples in Figure 2 and more problems with full text in Appendix A, Figure 5. Vision-language models trained in [9] struggle to solve this task, especially transformers. In this Section I investigate how we can craft and train deep neural networks which employ transformer-like blocks and multimodal inputs from deep frozen transformers to reason better.

2.1. VLM Reasoners: Approach

In [2,3,27] the authors demonstrated how a deep learning module including transformers, which encode a sequence of image-and-text items using diverse representations composed on several axes-across modalities, across time steps, and across pooling methods, obtained massive results in sponsored search and recommendations. Building upon the ADPM in [2, 3, 27] and using tricks for properly training vision transformers in [12], a smarterVLM is built. In this article I make the following **contributions** on the vlm reasoning axis: 1. Architectural and deep learning training innovation: go deeper and make architectures work better with the multimodal input. 2. Introduce a novel QF-layer to learn a composite representation. 3. Improve the MLP decoders through gelu activations, residual connections and layer normalization. 4. Improve the sequence decoder by replacing the LSTM with a GRU. 5. Strengthen vision encoder by fusing two vision backbone: SigLIP [36] and DinoV2 [28] similarly to [17]. 6. Strengthen the text-vision alignment by using a frozen SigLIP language encoder together with the fused vision backbone, which does not overpower the visual signal. The improvements lead to up to 48% accuracy improvement in some skills on test set over best performing baseline from the SMART task.

2.1.1 Benchmark, Dataset and Challenges

A set of vision-language models are trained as benchmarks in [9] and **SMART-101 with 202K text-image pairs for train, validation and test dataset** is released. There are 101 origin puzzles and additional problems are generated programatically in each puzzle group for a total of 202,000 question-image pairs. Figure 5 clearly describes a training example problem. All the trained VLMs struggle on the SMART task, with transformers underperforming ResNet50 [15] based models. The learning tasks depend on the type of puzzle and are in the classification, regression and sequence generation category. Several image and text encoder backbones are considered. A puzzle specific set of image features are learned via an MLP and the text embeddings are aggregated using an LSTM layer. The decoder for the sequence generation is another LSTM layer. All image encoders are finetuned. Based on these characteristics, there are a few research opportunities worth exploring, especially since transformer based VLM reasoners are doing so poorly on the challenging SMART task in [9].

2.1.2 Methodology

I scope the problem to focus on the supervised learning formulation with a **classification loss.** For each image-question instance, **we predict the probability of one of five answer** options. When the options are a sequence, the GRU decoder decodes the sequence but the answer is still translated to one of $\{A, B, C, D, E\}$ options. Furthermore, I focus on training deep learning architectures from scratch for the SMART task with inputs from diverse pretrained **frozen** backbones. I focus the investigation on the eight skill classes counting, **(math, logic, algebra, path, pattern, measure, spatial**) rather than individual puzzle groups since these are of more general interest across domains and trademarks of intelligence. In [8] authors demonstrate that strong pretrained backbones can perform without meta-learning so I do not employ it, as they do in [9]. Since it is a classification loss, the **accuracy metric** calculated on validation set is used to tune the models and evaluate **method success**, and the accuracy for the five class classification is evaluated on the test set. The accuracy is calculated overall, and more interestingly, on skill class (counting, math etc.). Detailed information on development, training, evaluation and artifacts is in Appendix A. **Architec-**

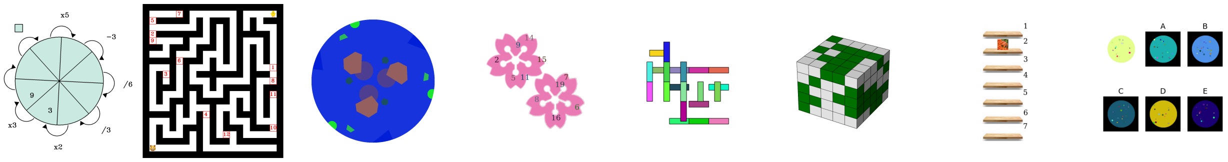

Figure 2. **Math Question**: *What do we need to put in the square to get a correct diagram?* **Answer Options:** A: -3; **B**: /9; C: x6; D: x2; E: 2; **Path Question with Sequence Answer**: *You have to block some locations in the maze so that the feline cannot reach the bird. Which of the following options to block will fail?* **Answer Options:** A: 1, 2, and 3; B: 4; **C**: 5, 6, and 7; D: 8 and 9 ; E: 10, 11, and 12. **Counting**; **Algebra**; **Measure**; **Spatial**; **Logic**; **Pattern.**

tural Innovation. I derive a composite representation via a novel layer, the QF layer, inspired from the ADPM in adsFormers [2, 3, 27] and the QFormer in [19, 40]. More recently, [17] and [31] combine multiple image representations to leverage diversity of signal in vision-language models, in a similar vein to the ADPM. The choice of the SigLIP text encoder is two-fold: first, due to the alignment with the SigLIP vision encoder; second, to tame the language power by not employing a large language model. In [31] we see that visual grounding is lacking in large multimodal models. Furthermore, in [9] the visual signal is quite important, the accuracy loss is more when removing the image rather than the text question, so the visual signal needs to be protected as it is critical. Details for the VLM reasoner architecture in Figure [**?**] are in Appendix A, including equations. The QF layer representation for the image-question input is concatenated to the average pooled text representation and the puzzle-specific image representation from the image encoder. The composite representation $compositeR$ takes as input three component representations, r_1, r_2, and r_3, defined as follows, with equations in the Appendix A.2. Text representation r_3 is an average pooled encoding of the question sequence of max length 110 tokens. Each token is first encoded using the frozen SigLIP text model into a representation of size 768. An image representation r_1 from the puzzle-specific image encoder block of dimension 128. The dimension is a hyperparameter selected via optimization. The image encoder consists of two feed forward layers with a gelu unit, with separate weights for each puzzle head, for the 101 separate puzzle groups. Each encoder takes as input the image representation from the two fused pretrained vision backbone, Dinov2 [28] and SigLIP [36], each of dimension 768. A QF representation, r_2, is produced by the optional QF layer which takes as input the encoded image representation r_1 and the SigLIP-encoded sequence of text tokens. The QF layer, passes the text sequence through a multi-head self-attention block, as in [33]. The resulting hidden representation is then fed to a cross attention layer as query, with key and values being the image encoder representation inspired from the q-former and VilBERT [24]. Finally, an intermediate stack of fully connected with residual connections [15] and layer normalisation [5] produces the QF text-and-vision representation. The the QVfusion layer as input the composite representation $CompositeR \in \mathbb{R}^{2*768+128}$ and passes through a three layer feed forward module with Gaussian error linear units in between [16], before being read by the puzzle specific decoder as seen in Figure 1. Finally the decoder, which is either a stack of three fully connected layers separated by gelu activations, or a gated recurrent neural network for sequence type answer puzzles, produces predictions fed to a cross-entropy loss. Each puzzle group calculates its own loss. The introduction of gelu units, layer normalisation, boost performance, as they do in most recent attention-based neural networks and not only allowing for a smoother loss landscape than relus and batchnorms.

2.2. VLM Reasoners: Experiments and Results

Several baselines from [9] are trained have results in 2 in Appendix A. I chose to move forward with the frozen BERT+Resnet50 as **baseline** for two reasons: 1. Note that the numbers are extremely close between the frozen and unfrozen variants but the frozen variant does better on Math, a skill of interest for this investigation; 2. This investigation focuses on a frozen backbones scope for efficiency reasons. Furthermore, keeping the backbone frozen afford a better comparison between models and the ability to reuse some hyperparameter settings such as batch size, number of epochs, optimizer, because if we choose to finetune the transformer based backbones as well, it does not make sense to employ the same techniques as for convnets. **Results on a subset of the SMART101** training set are discussed next. The biggest **challenges and limitations** for getting better models are as always in terms of **compute requirements**: to tune deep neural networks you need to run many experiments; transformers are data hungry as well so the GPUs for training large models with large memories and disk are needed for a long time. So I sampled and raised the bar on the model architectures rather than data and compute. A different split of $train : val : test = 60 : 20 : 20$ **and only 1000 out of the total 2000 question-image pairs per puzzle** group are utilized for fast training and insight, with a total budget of three epochs of training. This results in **474 training batches, 158 validation and 158 test**

batches of size 128. Since both the vision and text pretrained backbones are frozen, the additional model layers result in **29,623,375 trainable parameters** (unless dismissing the qf layer or experimenting with additional residual blocks, which some of the runs contained). Experiments are tracked in CometML [11], the multimodalai public project. The Table 3 in Appendix A shows results from intermediary experiments ran in the process of model development to make architecture, optimization, and hyperparameter decisions toward to final model; watching curves is what deep learning is all about.

As we can see in Table 1 of **best results**, some of the new models d**isplay massive gains in accuracy (eg. +48% gain over the baseline** in the counting skill). Recall that the baseline was chosen amid the original paper's choices as the strongest in math; and truly so, it was hardest to beat in math vs other skills. **The smartest VLM reasoner** was trained employing the following deep learning implements, building upon results obtained in [3, 27], [12], [30] and [26]:1. Adam optimizer with decoupled weight decay from [23] with weight decay of 0.2, $eps = 1e-8$, and $beta2 = 0.98$. 2. Cosine learning rate scheduler [22] with ten warmup steps, with the implementation in the HuggingFace repository [34]. 3. Clipping the gradient norm to no more than one. 4. Layer normalization throughout the architecture modules with $eps = 1e-6$. 5. A SigLIP frozen language backbone and fused DinoV2 and SigLIP vision backbone. 6. A composite hidden representation with a QF layer with two attention heads.7. Dropout probability of 0.2 anywhere it is used. 8. Gaussian error linear units; the MLP decoder is akin to SigLIP's MLP block. 9. Representation sizes of 128 and 256 within the various layers.10. GRU decoder for problems with sequence answer, as they are easier to train than LSTMs. **A vital insight** arose through the **training process**. In Figure 3 and 4, notice how the eight skill sets have different training dynamics and respond differently to learning rate choices, as well as to the cosine scheduler's learning rate decision throughout the training steps. This is something commonly seen in multitask learning [7, 12]. All experiments are run with seed 0 for the sake of **reproducibility** but I also evaluated the test accuracy standard deviation across a few seeds (0, 42, 7): mean overall test accuracy 20.8 ($\sigma = 0.16$), math skill mean accuracy of 9.73 ($\sigma = 0.38$), pattern mean accuracy 25.2 ($\sigma = 0.53$). As expected, we see more variability on smaller individual skill class sets. **Future direction**. Considering the different learning dynamics for the eight skill classes, a multitask learning approach with eight tasks may afford modulating the impact of eight weighted losses to account for the different dynamics. A mixture-of-experts approach [39] within the multitask learning framework could further help. Futhermore, we see that the qf learning helps, for some skills especially more so, and further deepening transformer-like specialized reasoning architectures, which read from frozen per-modality or multimodal encoders, can further help. Experimenting with further general purpose backbones, deeper or wider, is another potential avenue, since we see improvements from the fused Dinov2+siglip.

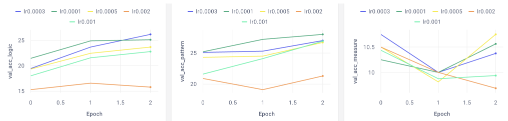

Figure 3. Validation accuracy curves per skill class -counting, math, spatial- for five different learning rates.

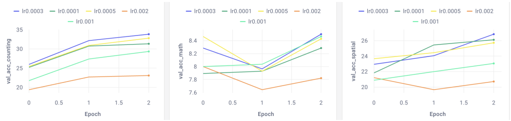

Figure 4. Validation accuracy curves per skill class -logic, pattern, measure- for five different learning rates.

3. Conclusion

Going deeper with transformer-like architectures, deep learning representations, and visual grounding lead to improvements across reasoning and general ability of VLMs and the path is open for further improvements.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Alaa Awad and Denisa Roberts. adSformers: Etsy Engineering Blog, 2023. 2, 3

[3] Alaa Awad, Denisa Roberts, Eden Dolev, Andrea Heyman, Zahra Ebrahimzadeh, Zoe Weil, Marcin Mejran, Vaibhav Malpani, and Mahir Yavuz. adSformers: Personalization from Short-Term Sequences and Diversity of Representations in Etsy Ads. *arXiv preprint arXiv:2302.01255*, 2023. 2, 3, 4

[4] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023. 1

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3, 6

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 1

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%)
SmarterVLM 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%)

Table 1. Test set skill class accuracy for a few top models and comparison to the baseline in first row (percentage change). From CometML multimodalai.

[7] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 4

[8] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019. 2

[9] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, K Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders?. arxiv. *Retrieved July*, 9:2023, 2022. 2, 3

[10] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 1, 2

[11] Comet.com. Comet.com home page, 2021. 4, 7

[12] Eden Dolev, Alaa Awad, Denisa Roberts, Zahra Ebrahimzadeh, Marcin Mejran, Vaibhav Malpani, and Mahir Yavuz. Efficient large-scale vision representation learning. *arXiv preprint arXiv:2305.13399*, 2023. 2, 4

[13] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 6

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3, 7

[17] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024. 1, 2, 3

[18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 1

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[22] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 4

[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 4

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3, 6

[25] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 1

[26] Denisa A Olteanu Roberts. Multilingual evidence retrieval and fact verification to combat global disinformation: The power of polyglotism. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 359–367. Springer, 2021. 4

[27] Denisa Anca Olteanu Roberts, Alaa Mohamad Awad, Eden Dolev, Andrea Laura Heyman, Marcin Mejran, Mahir Yafuz, and Vaibhav Malpani. ADPM US Patent App, 2024. 2, 3, 4

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3, 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[30] Denisa Roberts. Neural networks for lorenz map prediction: A trip through time. *arXiv preprint arXiv:1903.07768*, 2019. 4

[31] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 1, 3

[32] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 6

[34] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 4

[35] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 1

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2, 3, 6

[37] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*, 2024. 1

[38] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1

[39] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 43–51, 2019. 4

[40] Dongsheng Zhu, Xunzhu Tang, Weidong Han, Jinghui Lu, Yukun Zhao, Guoliang Xing, Junfeng Wang, and Dawei Yin. Vislinginstruct: Elevating zero-shot learning in multi-modal language models with autonomous instruction optimization. *arXiv preprint arXiv:2402.07398*, 2024. 3

A. Additional Information on VLM Reasoners

A.1. Example of Puzzles Problems and Image-Question Pairs For Each of the Eight Skills

In Figure 2 I showed a few examples of problems accross skill sets but ommited text question for most skill classes. In Figure 5 I include all the questions for a full understanding of train/val/test examples in the SMART 101 dataset.

A.2. Architectural Innovation Equations

The QF layer representation for the image-question input is concatenated to the average pooled text representation and the puzzle-specific image representation from the image encoder. The composite representation $compositeR$ takes as input three component representations, r_1, r_2, and r_3, defined as follows.

- Text representation r_3 is an average pooled encoding of the question sequence of max length 110 tokens. Each token is first encoded using the frozen SigLIP text model into a representation of size 768. Then

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

- An image representation r_1 from the puzzle-specific image encoder block of dimension 128 seen in 1. The dimension is a hyperparameter selected via optimization. The image encoder consists of two feed forward layers with a gelu unit, with separate weights for each puzzle head, for the 101 separate puzzle groups. Each encoder takes as input the image representation from the two fused pretrained vision backbone, Dinov2 [28] and SigLIP [36], each of dimension 768. Specifically, for an image X, r_1 is

$$\begin{aligned} r_1 &= FC_{1i}(Gelu(FC_{2i}(y))), \\ y &= Concat([Dino(x), SigLIP(x)]) \end{aligned}$$

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

- A QF representation, r_2, is produced by the optional QF layer which takes as input the encoded image representation r_1 and the SigLIP-encoded sequence of text tokens. The QF layer, passes the text sequence through a multi-head self-attention block, as in [33]. The resulting hidden representation is then fed to a cross attention layer as query, with key and values being the image encoder representation inspired from the q-former and VilBERT [24]. Finally an intermediate stack of fully connected with residual connections [15] and layer normalisation [5] produces the QF text-and-vision representation. Specifically,

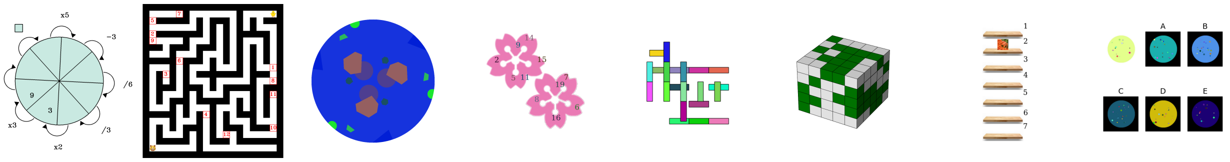

Figure 5. **Math Question**: *What do we need to put in the square to get a correct diagram?* **Answer Options:** A: -3; **B**: /9; C: x6; D: x2; E: 2; **Path Question with Sequence Answer**: *You have to block some locations in the maze so that the feline cannot reach the bird. Which of the following options to block will fail?* **Answer Options:** A: 1, 2, and 3; B: 4; **C**: 5, 6, and 7; D: 8 and 9 ; E: 10, 11, and 12. textbfCounting Question: *The entire pie is divided among several children. Each child receives a piece of pie, and each piece of pie looks identical. The maximum possible number of children there is:* **Answer Options:** A: 7; B: 2; C: 1; D: 4; **E**: 3. **Algebra Question**: *The entire pie is divided among several children. Each child receives a piece of pie, and each piece of pie looks identical. The maximum possible number of children there is:* **Answer Options:** A: 5; B: 4; C: 2; **D**: 0; E: 6. **Measure Question**:*A student had a few canes with a height of 1 cm and a length of 5cms. Using the canes, she built the arrangement illustrated. What is the width of the arrangement?* **Answer Options:** **A**: 20; B: 30; C: 15; D: 5; E: 35. **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. **Logic Question**: *Emily has 7 toy items: a remote, a hair brush, a truck, an eraser, a rubber duck, carrots, and a toe ring. She keeps each toy at a different row of the shelf. The carrots lower to toe ring. Remote lower to truck and toe ring higher to truck. Toe ring higher to rubber duck. She keeps carrots as shown. On which row can the rubber duck not be placed?* **Answer Options:** A: 4; B: 3; **C**: 7; D: 5; E: 6.**Pattern Question**: *Which picture on the right matches with the left, if we invert the colors?* **Answer Options:** A; B; C; D; **E**

$$r_2 = LayerNorm(x + Drop(FC(Gelu(FC(x)))))$$
$$X = MHCrossA(MHA([h_1, h_2, ..., h_{110}]), r_1)$$

Finally, the composite representation is

$$\begin{aligned} CompositeR &= CLayer([r_1, r_2, r_3]) \\ &= LayerNorm(Concat([r_1, r_2, r_3]). \end{aligned}$$

The the QVfusion layer as input the composite representation $CompositeR \in \mathbb{R}^{2*768+128}$ and passes through a three layer feed forward module with Gaussian error linear units in between [16], before being read by the puzzle specific decoder. Specifically, the QVFusion layer in 1 is

$$\begin{aligned} QVFusion(y) &= LayerNorm(GELU(y)) \\ y &= FC(GELU(FC(compositeR))). \end{aligned}$$

A.3. Additional Experiments for VLM Reasoners

In Figure 6 we can see training loss nicely descending over the training steps in the three epochs modulated by five different learning rates with a cosine scheduler which adapts the learning rate based on the step number. Note how the large 0.002 learning rate (orange) impacts learning negatively in the strongest way. Noticeable bumps in curves depend on the scheduler's change points.

For thoroughness of comparison in the developmental phase I proceeded with training and evaluation on a very small subset of data for quick insight and iteration: only 20 questions per puzzle, with a batch size of 16, a split ration of $train : val : test = 40 : 20 : 40$. Complete experiment results are available were tracked with CometML [11] and publicly accessible at vlm-reasoners. Results are included in Table 4 in the Appendix A. In

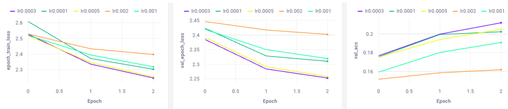

Figure 6. Epoch Train Loss, Validation Loss, and Validation Accuracy for five different learning rates. From CometML multimodalai.

SMART Baseline	Counting	Math	Logic	Path	Algebra	Measure	Spatial	Pattern	Overall
BERT+Resnet50	35.6	**26.4**	36.8	21.5	18.1	26.0	32.2	27.0	28.0
BERT+Resnet50(unfrozen)	35.7	20.8	39.6	22.2	18.4	28.2	33.7	30.6	28.2
BERT+MAE [14]	29.8	19.7	29.4	20.5	16.1	18.9	26.6	27.8	23.1
CLIP VL	35.5	8.6	27.1	17.9	11.8	16.0	26.8	26.3	22

Table 2. Skill class accuracy for original baselines with a 10hr budget training. All backbones are frozen unless noted otherwise.

	counting	math	logic	path	algebra	measure	spatial	pattern	vision	language
final_lr0.0006	23.6	8.4	19	18.6	9.5	10.6	23.9	22.4	dinov2siglip	siglip
baseline_mbert	23.4	8.1	18.9	17.9	10.3	10.2	23.8	20.8	resnet50	mbert
baseline_bertresnet50	23.4	8.1	19.2	17.8	10.7	10.3	24.8	20.5	resnet50	bert
lstm_decoder_siglipvision	24.6	7.9	17.9	17.9	10	9.1	22.6	21.4	siglip	siglip
lstm_decoder	27.7	8.4	21.3	18.6	11.2	9.2	23.3	25.1	dinov2siglip	siglip
qf_fusion_extra_residualinmlpdec	21.5	7.4	17.5	17.8	9.3	10.4	21.6	20.3	dinov2siglip	siglip
single_image_head	27.2	8.4	20.2	18.5	10.3	10.2	23.4	22.1	dinov2siglip	siglip
warmup0	29.8	7.4	21.3	20.3	10.1	10.8	22.8	26.6	dinov2siglip	siglip
warmup0.06	30	7.8	22.3	19	9.6	9.8	22.9	25.1	dinov2siglip	siglip
warmup0.01_no_extra_residuals	29.6	8.5	22.9	19.3	9.8	9.8	21.1	26.5	dinov2siglip	siglip
hardcode$_{n}um_{s}teps_{1}0$	29.9	8.1	17.8	18.8	9.5	9.7	23.2	19.8	dinov2siglip	siglip
final_lr0.001	29.3	8.5	22.8	19.1	10.3	9.9	23.1	26.9	dinov2siglip	siglip
noqf	32.8	8.5	23.8	19.7	11.5	10.3	25.6	25.4	dinov2siglip	siglip
batch64	26.1	8.2	21.4	18.8	10.1	10.8	22.3	26	dinov2siglip	siglip
reprsize256	25.3	8	21.5	18.5	10.3	10.6	23.8	26.3	dinov2siglip	siglip
hiddensize128	28.4	8.4	22.4	18.8	10.6	10.2	25.9	23.3	dinov2siglip	siglip
lneps1e-5	30	8.3	22.3	19.6	9.9	9.8	22.8	26.1	dinov2siglip	siglip
pdrop0.1	29.7	8.2	20.9	19.8	10	9.6	23.1	26.2	dinov2siglip	siglip
adamweps_beta2_defaults	29.5	8	21.1	19.1	10.7	9.7	23.2	26.8	dinov2siglip	siglip
final_lr0.002	23.1	7.8	15.8	18.8	10.2	9.7	20.7	21.3	dinov2siglip	siglip
1_mha_head_qf	29.7	8.2	23.1	19.4	10.5	10.8	23.2	22.7	dinov2siglip	siglip
final_lr0.0005_seed0	32.8	8.4	23.7	20.1	10.4	10.8	25.7	26.7	dinov2siglip	siglip
final_lr0.0001	31.3	8.3	25.1	19.4	10.4	10.6	26.1	28	dinov2siglip	siglip
final_lr0.0003	33.8	8.5	26.2	20.1	11.2	10.4	26.8	27	dinov2siglip	siglip

Table 3. Validation Accuracy per Skill Class for Architectural, Optimization and Hyperparameter Choices. From CometML multimodalai.

Description		**counting**	**math**	**logic**	**path**	**algebra**	**measure**	**spatial**	**pattern**	**num_qf_heads**	**pdrop**	**wd**	**word_embed**	**model_name**
run21_lr0.005		13.5	8.9	5.6	16.7	8.3	3.1	27.8	30	2	0.2	0.2	siglip	fused_dinov2_siglip
run20_lr0.0001		18.3	10.7	5.6	14.6	8.3	15.6	27.8	10	2	0.2	0.2	siglip	fused_dinov2_siglip
run19_pdrop0.1		13.5	10.7	2.8	18.8	15	3.1	30.6	30	2	0.1	0.2	siglip	fused_dinov2_siglip
extra_resid_run18		14.4	7.1	5.6	16.7	8.3	9.4	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
siglip_vision_run17		12.5	7.1	5.6	16.7	10	15.6	33.3	15	2	0.2	0.2	siglip	siglip
run16_feat64		13.5	7.1	5.6	16.7	11.7	6.3	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
repr_256_run15		18.3	8.9	5.6	16.7	10	9.4	30.6	20	2	0.2	0.2	siglip	fused_dinov2_siglip
run14_noqf		12.5	5.4	5.6	18.8	6.7	12.5	27.8	30	2	0.2	0.2	siglip	fused_dinov2_siglip
composite_back_run13		15.4	10.7	5.6	18.8	11.7	3.1	30.6	30	2	0.2	0.2	siglip	fused_dinov2_siglip
run12_rm_text_incomposite		16.3	7.1	5.6	14.6	6.7	12.5	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
run11_noim_repr_in_composite		10.6	5.4	5.6	16.7	3.3	9.4	30.6	20	2	0.2	0.2	siglip	fused_dinov2_siglip
run_full_single_image_w_qf and extra_resid		17.3	3.6	5.6	14.6	6.7	9.4	30.6	20	2	0.2	0.2	siglip	fused_dinov2_siglip
rm_a_gelu_in_mlp_decoder		13.5	10.7	5.6	16.7	6.7	3.1	30.6	30	2	0.2	0.2	siglip	fused_dinov2_siglip
residual_back_in		16.3	8.9	11.1	14.6	10	12.5	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
no_residual_in_qf		12.5	7.1	5.6	18.8	11.7	12.5	30.6	25	2	0.2	0.2	siglip	fused_dinov2_siglip
ivory_cliff_394										2	0.2	0.2	siglip	fused_dinov2_siglip
qf_128		17.3	5.4	5.6	16.7	6.7	12.5	25	25	2	0.2	0.2	siglip	fused_dinov2_siglip
768_dim_in_qf		15.4	5.4	8.3	16.7	6.7	9.4	30.6	20	2	0.2	0.2	siglip	fused_dinov2_siglip
scheduler_back_in		16.3	8.9	11.1	14.6	10	12.5	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
no_scheduler		8.7	3.6	8.3	6.3	1.7	3.1	16.7	25	2	0.2	0.2	siglip	fused_dinov2_siglip
relu_not_gelu		14.4	8.9	11.1	14.6	8.3	12.5	30.6	15	2	0.2	0.2	siglip	fused_dinov2_siglip
qf3heads		15.4	5.4	5.6	16.7	6.7	9.4	27.8	20	3	0.2	0.2	siglip	fused_dinov2_siglip
1qf_head		17.3	5.4	8.3	16.7	6.7	12.5	30.6	25	1	0.2	0.2	siglip	fused_dinov2_siglip
lstm_decoder		14.4	8.9	5.6	16.7	6.7	12.5	27.8	25	2	0.2	0.2	siglip	fused_dinov2_siglip
wd0		14.4	7.1	11.1	14.6	8.3	9.4	27.8	20	2	0.2	0	siglip	fused_dinov2_siglip
wd0.05		14.4	7.1	11.1	14.6	8.3	9.4	27.8	20	2	0.2	0.05	siglip	fused_dinov2_siglip
wd0.1		15.4	7.1	11.1	14.6	8.3	9.4	27.8	20	2	0.2	0.1	siglip	fused_dinov2_siglip
wd0.2		15.4	7.1	11.1	14.6	8.3	9.4	25	25	2	0.2	0.2	siglip	fused_dinov2_siglip

Table 4. Small Experimental Runs: Architecture and Hyperparameter Choices Impact on Skill Class Accuracy. From vlm-reasoners.