On Vision-Language Models

Denisa Roberts
Georgia Tech
droberts308@gatech.edu

Ahmad Shapiro
Georgia Tech
ahmad.shapiro@gatech.edu

Abstract

In this article [1] *we investigate two research conjecture sets concerning vision-language multimodal models design: 1. VLM architectures as reasoners; 2. Text-aware image encoding in VLMs. Code repository for the project can be found at https://github.com/droberts308/DL7643-project-vlms*

0.1. Introduction

Opportunities of improvement related to multimodal / VLM models are mentioned across several 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 still to be explored; efficient training with compression (LoRa, quantization), etc. In the [9] recent article on prismatic models, several distinct 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 on second grade math puzzles and text-conditioned visual features.

1. Literature Review

It does not seem that deep neural networks are just yet smarter than second graders but they are still transforming industries. In [18] Olteanu Roberts develops dilated convolutions and LSTM modules to solve trajectories commonly seen in ODE problem solving. In [19] Olteanu Roberts develops deep learning algorithms commonly used in quantum machine learning and compression. In [15] Olteanu Roberts trains multilingual transformers that do better than English only ones at English-only natural language reasoning tasks. In [2], Awad, Olteanu Roberts et al. develop deep learning modules from scratch and learn representations leading to improved sponsored search. In [7], Dolev, Awad, Olteanu Roberts et al. learn image representations efficiently with residual nets and vision transformers, reusable across retrieval and ranking tasks in search and recommendations.

in "BLIP-2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models" [11] an interesting vision-language architecture, the Query Transformer, or Q-former, adds transformer layers to frozen image and text encoders. THis is great candidate architecture to utilize (Q-former). The article does contrastive pretraining, trains in 6+3 days and has code and models.

In Llava: Visual Instruction Tuning [13] and v1.5 in "Improved baselines for visual instruction tuning" [12] the topic is interesting, since they fine-tune end to end and utilize a Multimodal Science dataset. They mention that data, code and model are publicly available. They combine GPT-4 and visual tuning to get "Large Language Visual Assistant" (LLava). Model tuned on Science QA, which has train of 12726, valid of 4241 and test 4241, but no math.

"Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models" [9] has models, training and eval code and utilizes flash attention which requires Ampere and later NVIDIA GPUs. The article sports an interesting vision encoder fusing idea (SigCLIP and Dino); 4 axes of model improvement (VLM). Model outperforms Llava15, InstructBLIP. VLM: get visual patches and projects into Llama space (LM). Predicts next token. They don't do two-stage training (alignment and tuning). They study: optimization procedure (single stage); image (fused frozen); language models; scaling for data and training (add diverse and longer). Prismatic model checkpoint has 7B and 13B. Suggest auxiliary objectives in training for better vision. They find finetuning the visual backbone degrades but mention that Fuyu achieves good perf with finetuning. Also GCD-CLIP [17] finetunes CLIP for better perf. Some other architectures to consider to improve on this: qformer [11], perceiver resampler in flamingo [1]).

On efficiency side, in "DoRA: Weight-Decomposed Low-Rank Adaptation" [14] methods claim better than LoRa on accuracy with the lack of inference overhead that LoRa displays (as compared to adapters or prompting in

[1] Overleaf project draft

terms of PEFT methods). Tested with vision-language as well; visual instruction tuning and Llava. Decomposes pretrained weights on direction and magnitude. Gets inspiration from weight norm. Does an analysis of gradients akin to grad cam to understand fundamental differences between FT and PEFT. Eval on commonsense reasoning. Gets a little close to my work on qr decomposition, and fourier transforms. Key idea to use: weight decomposition in trainining.

In LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model [8] they use Llama as LLM and parameter-efficient visual instruction tuning. Does early fusion of visual tokens in early LLM layers. Unlocks more parameters (as compared to Llama-adapter v1) - norm, bias, scale. Optimization strategy allows small-scale image text and instruction dataset. Exibits stronger language-only instruction following abilities. Eval on ScienceQA.

In ”VisLingInstruct: Elevating Zero-Shot Learning in Multi-Modal Language Models with Autonomous Instruction Optimization” - no models, no dataset - [25] they make good use of Q-Formers. Can use as proxy to Instruct-Blip, since newer. Presents, in addition to incontext learning with instructions and FlanT5 and Vicuna instruction tuned language models, an improved vision feature extraction method.

In “Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs” [20] investigates how vision part of LMM is falling behind, opportunity for LMM to improve. Has a dataset MMVP. No model, but gives a method to improve vision modality. Uses CLIP and DinoVision interleaved and additive tokens; similar to Prismatic. Has Llava backbone.

In LLEMMA: an open language model for mathematics [3] from MathAI workshop at NeurIPS’23, a pretrain dataset is derived. They continued pretraining of Llama 7B on, Proof-pile-2, which is available. They released the model; integrated in Lean math proof net. They start with Llama2 (inlcuding 7b); language only. Good details in paper; this is a competitor of Wizardmath from Neurips23.

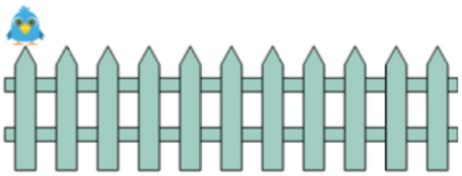
In “Multimodal Chain-of-Thought Reasoning in Language Models” [24] CoT method helps deal with hallucinations of LMM. Input is : Question and Context (language), images (vision), OPtions for Answers (A, B). Output is : rationale, answer. Has model and code. Finetunes T5 (a small ¡ 1bn) in 2 stages. Fuses text and image features, done recently as well, with a nice algo description. Vision CLIP, DETR, ResNet; text T5 is pretrained on Unified QA. Not clear what is their finetuning dataset, it said crawled. In “MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI” [22] - an eval dataset which includes math too similar to MathVista. Several fields, including Science and Engineering, Medical etc. Hard problems with 11.5K questions. They eval 14 open source LMM; including instruction tuned LLava; LLama Adapter; Instruct BLIP; in-context learning: including Otter, OpenFlamingo. They mention MathVista and GAIA as other eval datasets.

In “Otter: A Multi-Modal Model with In-Context Instruction Tuning” [10] - references to Blip-2, Llama-adapter, Lllava; has models; integrated with Huggingface. Has a dataset MIMIC-IT. Perceiver based; CLIP + LLama. Finetunes only the perceiver resampler and cross-attn layers.

2. On the Algorithmic Reasoning Abilities of Vision-Language Multimodal Transformers

2.1. Motivations and Challenges

How we define intelligence (artificial or not) is still an open question. In a related work in the multimodal domain, the article “Are Deep Neural Networks SMARTer than Second Graders?” [4] another task, a Simple Multimodal Algorithmic Reasoning Task(SMART), is introduced with visuo-linguistic puzzles designed for children in the 6-8 age group (the US Kangaroo Olympiad style), which comes close to testing for of the listed intelligence desiderata. The starting challenge consists of 101 unique puzzles with a picture and a question, as seen in Figure A.1.1



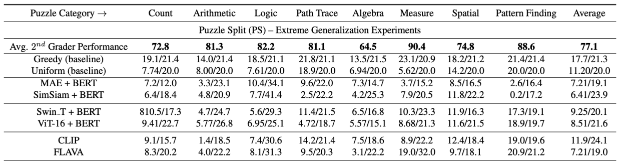
Question: *Bird Bobbie jumps on a fence from the post on the left end to the other end. Each jump takes him 4 seconds. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 jump back, and so on. In how many seconds can Bobbie get from one end to the other end?*
Answer Options: A: 64 B: 48 C: 56 D: 68 E: 72

To solve such a puzzle one needs a mix of several skills including arithmetic, algebra and spatial reasoning among others.

2.2. Benchmarks and Challenges

A set of vision-language models are trained as benchmarks and a dataset created pragmatically is released (SMART-101 with 200K puzzles for train and test). All the trained VLMs struggle on the SMART task, with transformers underperforming ResNet50-based models. In Figure **??** we can see how the VLM does as compared to a second grade human kangaroo olympic (not well):

Interestingly in Figure A.1.2 is an ablation of a ResNet50, which points to the fact that the vision modal-



Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
Puzzle Split (PS) – Extreme Generalization Experiments									
Avg. 2^{nd} Grader Performance	**72.8**	**81.3**	**82.2**	**81.1**	**64.5**	**90.4**	**74.8**	**88.6**	**77.1**
Greedy (baseline)	19.1/21.4	14.0/21.4	18.5/21.1	21.8/21.1	13.5/21.5	23.1/20.9	18.2/21.2	21.4/21.4	17.7/21.3
Uniform (baseline)	7.74/20.0	8.00/20.0	7.61/20.0	18.9/20.0	6.94/20.0	5.62/20.0	14.2/20.0	20.0/20.0	11.20/20.0
MAE + BERT	7.2/12.0	3.3/23.1	10.4/34.1	9.6/22.0	7.3/14.7	3.7/15.2	8.5/16.5	2.6/16.4	7.21/19.1
SimSiam + BERT	6.4/18.4	4.8/20.9	7.7/41.4	2.5/22.2	4.2/25.3	7.9/20.5	11.8/22.2	0.2/17.2	6.41/23.9
Swin_T + BERT	810.5/17.3	4.7/24.7	5.6/29.3	11.4/21.5	6.5/16.8	10.3/23.3	11.9/16.3	17.3/19.1	9.25/20.1
ViT-16 + BERT	9.41/22.7	5.77/26.8	6.95/25.1	4.72/18.7	5.57/15.1	8.68/21.3	11.6/21.5	18.9/19.7	8.51/21.6
CLIP	9.1/15.7	1.4/18.5	7.4/30.6	14.2/21.4	7.5/18.6	8.9/22.2	12.4/18.4	19.0/19.6	11.9/24.1
FLAVA	8.3/20.2	4.0/22.2	8.1/31.3	9.5/20.3	3.1/22.2	19.0/32.0	9.7/18.1	20.9/21.2	7.21/19.0

ity is quite important in this task (take it out, the drop in performance is larger than when removing text).

Method	$S_{\text{acc}} \uparrow$	$O_{\text{acc}} \uparrow$
Instance split		
R50 + BERT	**42.8**	**50.2**
No meta learning/MTL	29.7	37.3
Image only (no question)	28.3	36.3
Question only (no image)	15.1	23.2
Single image head	25.0	34.3

Table 5. Ablation studies using the R50 + BERT model.

The reasoning VLM architecture proposed in the article is displayed in Figure A.1.2.

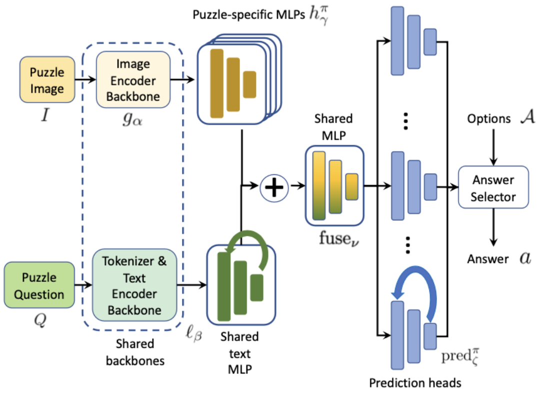


Figure 3. An illustration of our learning and reasoning model.

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 a GRU layer. The decoder for the sequence generation is another GRU 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.

2.3. Planned Investigations for the VLM reasoner

Specifically, proposed research investigations on the VLM reasoning path are as follows:

- Train the IS baseline with ResNet50+BERT, which is the winning model. However, to keep the problem clear and within resource availability, only consider the IS supervised learning task with no meta learning and compare baseline with updates on only 1 epoch of training (instead of the model 100) with frozen visual backbone.

- Strengthen vision encoder. In [9] good results are obtained with frozen/non-finetuned vision encoder with two fused vision backbone: SigLip [23] and DinoV2 [16]. Experiment with this technique for the vlm reasoner.

- Stretch (possibly not do): In recent VLMs a series of alignment modules such as the Q-Former or Perceiver [25] led to improved results on question answering and other multimodal tasks. Experiment with enhancing/replacing the MLP and/or the GRU modules in the VLM reasoner.

- Train and evaluate on SMART101.

2.4. [Ahmad's Path] : Don't take vision features for granted : An Efficient Text Aware Image Encoding

Recent VLM Research can be divided into two different but yet very similar axes. Vision language alignment and Instruction following capabilities.

Alignment Approaches can be ordered as follows in terms of their complexity:

1. Simple linear projection or MLP projection such as LLava series which project the vision encoder features into the same space of language model embedding, so that the image features can be passed as a soft prompt input for the language model.

2. Extra complex modules such as Q-Former in BLIP series.

3. Integration of vision features through cross-attention with language model such as Flamingo and Otter Series.

4. A complete unified approach such as Fuyu-8B which a transformer decoder which accepts both text and images as inputs.

After alignment, the instruction tuning process doesn't differ among most of the previous works. The only difference might be in the dataset collection procedure, the prompting strategy and the sampling process.

All of the following approaches use a frozen vision encoder except for Fuyu which relies on a vision-encoder free pipeline. Some of the approaches fine-tune full parameters of the language model. While others uses a some parameter efficient methods such as LoRa and complicated variant of prefix tuning such as Llama Adapter v2.

After surveying the literature, it became obvious that relying on a frozen vision encoder is one of the weaknesses of current VLM design aspects. On the other hand, fine-tuning the visual encoder while training the whole system might not be the best case as shown in Prismatic. This can be attributed to the fact that the vision encoder doesn't interact directly with the textual input. Only two approaches got close to that conclusion, Fuyu8B which removed the whole visual encoder. And BLIP2 which condition the visual encoder output on the instruction text but inside the Q-former. I argue that the textual context should be fused in the visual encoder from the start, not after the feature extraction process Because the feature that LLM might be looking for, might not be present in the visual encoder features itself.

2.4.1 Planned Contributions

My proposed architecture is shown in Figure 4.

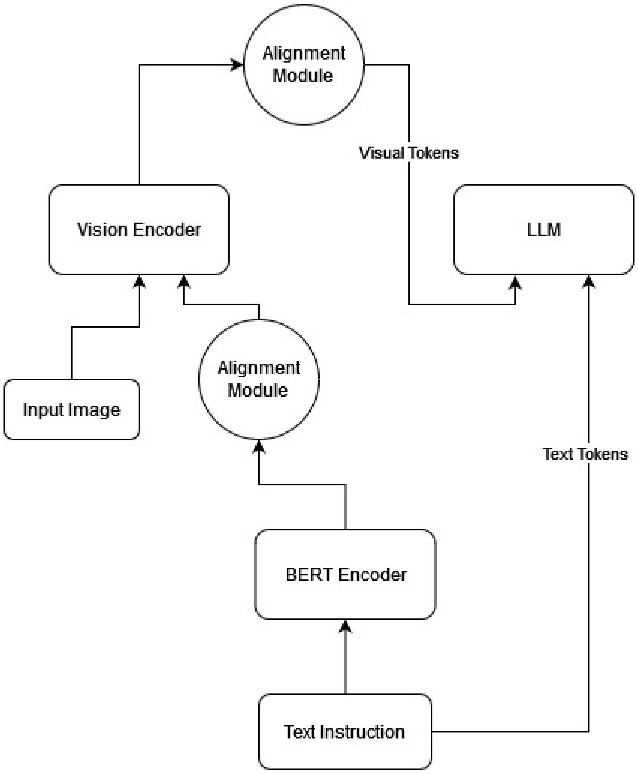


Figure 1. Proposed Architecture

I plan to follow the dataset recipe from LLava 1.5 combined with LLavaR for the tuning and alignment part. And transform the dataset from LLavaRLHF to DPO format and do DPO alignment after training. I don't plan to follow the two stage pretraining like it has been done in all previous work, instead, as the authors of Prismatic showed that combining the two stages might be more beneficial. For the alignment components in the figures i plan to start from the simple possible and increase complexity as needed :

1. MLP Projection
2. Zero-Gated Prefix Tuning (LLama Adapter V2)
3. DoRa or LoRa variants.

BERT and LLM can be replaced with Flan T5 encoder and decoder respectively to ensure that the text representation passed to both the vision encoder and language decoder come from the same distribution. I plan if i had time to try both approaches, using bert and off the shelf decoder based LLM and use T5 or any encoder decoder based LLM.

For the Visual Encoder part, adapting it effeciently to a task won't be that complex because I'm planning to use ViT based visual encoders such as CLIP or DINOv2 or a an ensemble of both.

In additions to the benchmarks in LLava Series, and Pristmatic Models. I plan to benchmark my results also on :

1. MMVP Benchmark "Eyes Wide Shut" which address multiple visual preception tasks
2. MagnifierBench From OtterHD which address smaller tress on minute details and spatial relationships of small objects.

3. Conclusion

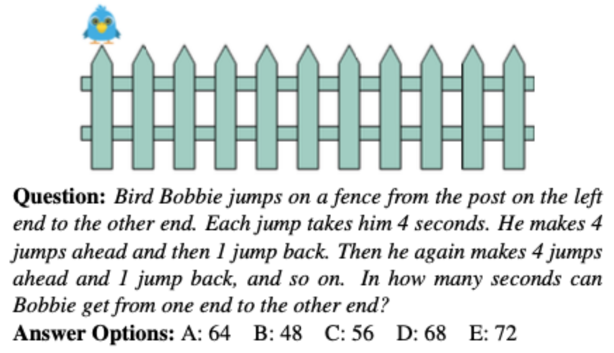
A. Project Proposal From Graded Submission

During the literature review, brainstorming and project timelines/planning meetings during the months of February 2024 and early March 2024, we came up with a longer pool of potential ideas related to vlms, from wich we selected based on personal preferences. This section of the appendix includes the initial list included in the Project Proposal. Based on feedback from the professor, the final implementations on each path are a further smaller subset of each path to keep it manageable within time and space toward submission deadline of April 29, 2024.

A.1. [Denisa's Path]: On the Algorithmic Reasoning Abilities of Vision-Language Multimodal Transformers

A.1.1 Problem

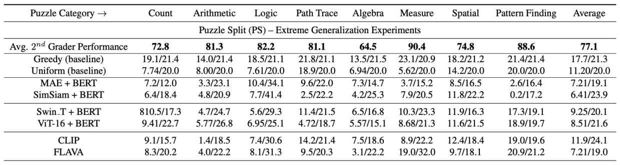
How we define intelligence (artificial or not) is still an open question. In "On the Measure of Intelligence" [5] the author conducts an in depth discussion on this open question. They formulate a new formal definition of intelligence (and dataset) based on algorithmic information theory and introduce the concept of algorithmic reasoning, a topic of current interest with a dedicated workshop at CVPR 2024. Specifically, scope, generalization difficulty, priors and experience are listed as desiderata for intelligent systems. In a related work in the multimodal domain, the article "Are Deep Neural Networks SMARTer than Second Graders?" [4] another task, a Simple Multimodal Algorithmic Reasoning Task(SMART), is introduced with visuo-linguistic puzzles designed for children in the 6-8 age group (the US Kangaroo Olympiad style), which comes close to testing for of the listed intelligence desiderata. The starting challenge consists of 101 unique puzzles with a picture and a question, as seen in Figure A.1.1


Question: *Bird Bobbie jumps on a fence from the post on the left end to the other end. Each jump takes him 4 seconds. He makes 4 jumps ahead and then 1 jump back. Then he again makes 4 jumps ahead and 1 jump back, and so on. In how many seconds can Bobbie get from one end to the other end?*
Answer Options: A: 64 B: 48 C: 56 D: 68 E: 72

To solve such a puzzle one needs a mix of several skills including arithmetic, algebra and spatial reasoning among others.

A.1.2 Benchmarks and Challenges

A set of vision-language models are trained as benchmarks and a dataset created pragmatically is released (SMART-101 with 200K puzzles for train and test). All the trained VLMs struggle on the SMART task, with transformers underperforming ResNet50-based models. In Figure **??** we can see how the VLM does as compared to a second grade human kangaroo olympic (not well):



Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure	Spatial	Pattern Finding	Average
	Puzzle Split (PS) – Extreme Generalization Experiments								
Avg. 2^{nd} Grader Performance	**72.8**	**81.3**	**82.2**	**81.1**	**64.5**	**90.4**	**74.8**	**88.6**	**77.1**
Greedy (baseline)	19.1/21.4	14.0/21.4	18.5/21.1	21.8/21.1	13.5/21.5	23.1/20.9	18.2/21.2	21.4/21.4	17.7/21.3
Uniform (baseline)	7.74/20.0	8.00/20.0	7.61/20.0	18.9/20.0	6.94/20.0	5.62/20.0	14.2/20.0	20.0/20.0	11.20/20.0
MAE + BERT	7.2/12.0	3.3/23.1	10.4/34.1	9.6/22.0	7.3/14.7	3.7/15.2	8.5/16.5	2.6/16.4	7.21/19.1
SimSiam + BERT	6.4/18.4	4.8/20.9	7.7/41.4	2.5/22.2	4.2/25.3	7.9/20.5	11.8/22.2	0.2/17.2	6.41/23.9
Swin_T + BERT	810.5/17.3	4.7/24.7	5.6/29.3	11.4/21.5	6.5/16.8	10.3/23.3	11.9/16.3	17.3/19.1	9.25/20.1
ViT-16 + BERT	9.41/22.7	5.77/26.8	6.95/25.1	4.72/18.7	5.57/15.1	8.68/21.3	11.6/21.5	18.9/19.7	8.51/21.6
CLIP	9.1/15.7	1.4/18.5	7.4/30.6	14.2/21.4	7.5/18.6	8.9/22.2	12.4/18.4	19.0/19.6	11.9/24.1
FLAVA	8.3/20.2	4.0/22.2	8.1/31.3	9.5/20.3	3.1/22.2	19.0/32.0	9.7/18.1	20.9/21.2	7.21/19.0

Interestingly in Figure A.1.2 is an ablation of a ResNet50, which points to the fact that the vision modality is quite important in this task (take it out, the drop in performance is larger than when removing text).

Method	$S_{\text{acc}} \uparrow$	$O_{\text{acc}} \uparrow$
Instance split		
R50 + BERT	**42.8**	**50.2**
No meta learning/MTL	29.7	37.3
Image only (no question)	28.3	36.3
Question only (no image)	15.1	23.2
Single image head	25.0	34.3

Table 5. Ablation studies using the R50 + BERT model.

The reasoning VLM architecture proposed in the article is displayed in Figure A.1.2.

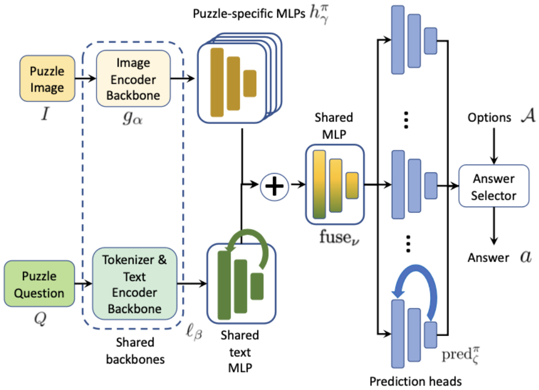


Figure 3. An illustration of our learning and reasoning model.

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 a GRU layer. The decoder for the sequence generation is another GRU 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.

A.1.3 Planned Investigations for the VLM reasoner

Specifically, proposed research investigations on the VLM reasoning path are as follows:

- Train at least one of the baseline VLM reasoners in the article and aim to replicate their results (ResNet50+BERT seems to do the best).

- Strengthen vision encoder. In [9] good results are obtained with frozen/non-finetuned vision encoder with two fused vision backbone: SigLip [23] and DinoV2 [16]. Experiment with this technique for the vlm reasoner.
- In recent VLMs a series of alignment modules such as the Q-Former [25] as seen in Figure A.1.3 and the Perceiver Resampler [1] in Figure A.1.3 led to improved results on question answering and other multimodal tasks. Experiment with enhancing/replacing the MLP and the GRU modules in the VLM reasoner.
- Consider a flamingo like decoder for the sequence generation learning task (certain puzzles require a sequence answer) [1].
- Train and evaluate on SMART101.

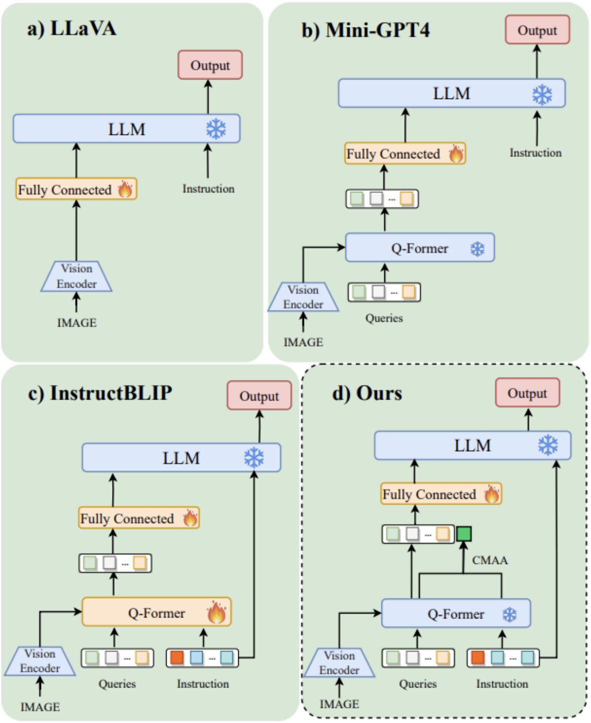


Figure 1: The structural comparison among the alignment modules of different MMLMs. The orange modules in the figure represent open weights, while the blue modules indicate frozen weights.

Figure 2. Example Q-former usage from VisLingInstruct article.

Other stretch goals that I may consider in next steps:

- Consider strengthening the vision encoder with a Llama2 [21] and instruction tuned VLM text encoder.

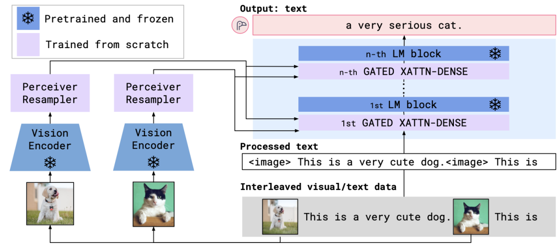


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

Figure 3. The Perceiver Resample based VLM architecture.

- Reformulate problem in Chollet's article on intelligence [5] to develop, train and evaluate a multimodal model on the Abstraction and Reasoning Corpus (ARC) task.

A number of other ideas already listed in the general idea section are also to be considered in future steps toward improving algorithmic reasoning of vlm/multimodal models. Other limitations beyond timeboxing are compute resources, which the new VLM in the Llava1.5 [12] and Prismatic [9] style and new architectures such as including FlashAttention2 [6] require (Ampere family NVIDIA GPUs with large memories). I intend to use Colab and school resources (PACE) and limit to what is possible in terms of architectures.

A.1.4 Motivations - Why VLM reasoners?

In my motivation for pursuing algorithmic reasoning line of multimodal deep learning architectures research I rely on a few conjectures:

- Intelligence is related to multimodal reasoning. If a person is deaf and cannot hear more than 50% of what is being said the speech modality input is supplemented with reading faces and other visual aids (vision modalities), captions(text) as well as other modalities (all the other senses as well as enhanced reasoning and computation abilities). So it is worth striving to improve multimodal deep learning architectures and their reasoning abilities. As a reviewer for MathAI Workshop at NeurIPS'23, I noticed that the work was almost exclusively with language models, with other modalities being almost nonexistent. I consider mathematical reasoning a subset of algorithmic reasoning.
- Training leads to learning and enhanced reasoning. As a former math and physics olympic I know that math and physics olympiad expert competitors practice more than non-experts in environments that value

the pursuit. Polgar sisters chess players grandmasters were brought up to perform at chess. Mozart and Beethoven were brought up to perform at music. More generally, a partially deaf person learns to make sense of the world especially if deafness occurred at the pre-verbal stage, through years of world training. In fact evolution is training and learning; we learned to grow better brains (because we needed to). So training transformers to enhance reasoning makes evolutionary sense.

- Intelligence is related to better abstractions and those are related to better representations. Expert chess players, thespians, martial artists, mathematicians, coders have fine grained relevant representations they can reason with to create/imagine new plays rapidly. Improving the representations derived with deep learning architectures is worth pursuing (better image, text etc. representations as well as their cross-play).

- 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) they will be better at science QA, medical and law QA and some other types of reasoning such as commonsense reasoning.

A.2. [Ahmad's Path] : Don't take vision features for granted : An Efficient Text Aware Image Encoding

Recent VLM Research can be divided into two different but yet very similar axes. Vision language alignment and Instruction following capabilities.

Alignment Approaches can be ordered as follows in terms of their complexity:

1. Simple linear projection or MLP projection such as LLava series which project the vision encoder features into the same space of language model embedding, so that the image features can be passed as a soft prompt input for the language model.

2. Extra complex modules such as Q-Former in BLIP series.

3. Integration of vision features through cross-attention with language model such as Flamingo and Otter Series.

4. A complete unified approach such as Fuyu-8B which a transformer decoder which accepts both text and images as inputs.

After alignment, the instruction tuning process doesn't differ among most of the previous works. The only difference might be in the dataset collection procedure, the prompting strategy and the sampling process.

All of the following approaches use a frozen vision encoder except for Fuyu which relies on a vision-encoder free pipeline. Some of the approaches fine-tune full parameters of the language model. While others uses a some parameter efficient methods such as LoRa and complicated variant of prefix tuning such as Llama Adapter v2.

After surveying the literature, it became obvious that relying on a frozen vision encoder is one of the weaknesses of current VLM design aspects. On the other hand, fine-tuning the visual encoder while training the whole system might not be the best case as shown in Prismatic. This can be attributed to the fact that the vision encoder doesn't interact directly with the textual input. Only two approaches got close to that conclusion, Fuyu8B which removed the whole visual encoder. And BLIP2 which condition the visual encoder output on the instruction text but inside the Q-former. I argue that the textual context should be fused in the visual encoder from the start, not after the feature extraction process Because the feature that LLM might be looking for, might not be present in the visual encoder features itself.

A.2.1 Planned Contributions

My proposed architecture is shown in Figure 4.

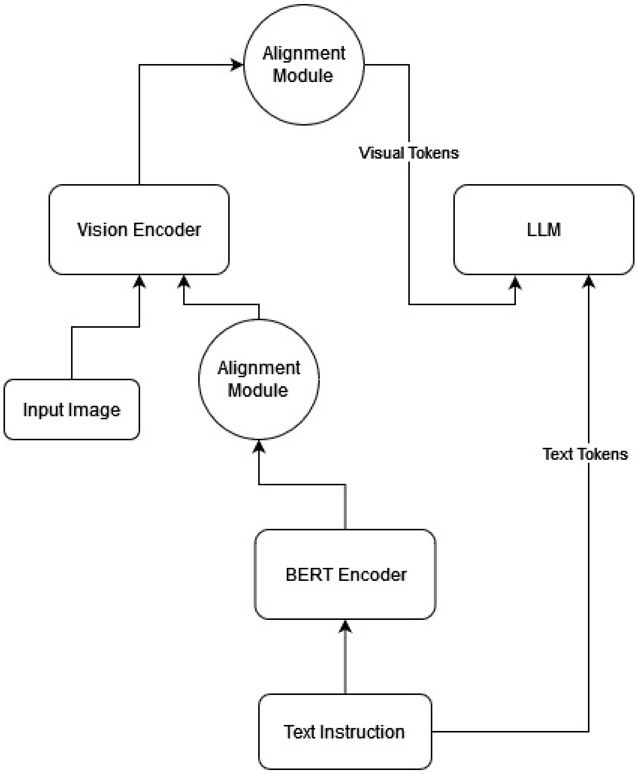


Figure 4. Proposed Architecture

I plan to follow the dataset recipe from LLava 1.5 combined with LLavaR for the tuning and alignment part. And transform the dataset from LLavaRLHF to DPO format and do DPO alignment after training. I don't plan to follow the two stage pretraining like it has been done in all previous work, instead, as the authors of Prismatic showed that combining the two stages might be more beneficial. For the alignment components in the figures i plan to start from the simple possible and increase complexity as needed :

1. MLP Projection
2. Zero-Gated Prefix Tuning (LLama Adapter V2)
3. DoRa or LoRa variants.

BERT and LLM can be replaced with Flan T5 encoder and decoder respectively to ensure that the text representation passed to both the vision encoder and language decoder come from the same distribution. I plan if i had time to try both approaches, using bert and off the shelf decoder based LLM and use T5 or any encoder decoder based LLM.

For the Visual Encoder part, adapting it effeciently to a task won't be that complex because I'm planning to use ViT based visual encoders such as CLIP or DINOv2 or a an ensemble of both.

In additions to the benchmarks in LLava Series, and Pristmatic Models. I plan to benchmark my results also on :

1. MMVP Benchmark ”Eyes Wide Shut” which address multiple visual preception tasks
2. MagnifierBench From OtterHD which address smaller tress on minute details and spatial relationships of small objects.

A.3. Rough Initial Pool of Considered Ideas on Improving Multimodal Models

We can start from Prism codebase [9] (and Llava 1.5 codebase [12]) or another codebase and baseline model, each of the two contributions can use something different. A list of research questions/hypotheses to possibly pick and choose from:

- Maybe adapt Dora for efficient finetuning
- Maybe improve the connnecting architecture in Prism and Llava1.5 (MLP) (consider a Q-former adaptation from Blip-2 or Perceiver variation).
- Maybe improve the vision feature module. Consider technique in Eyes Wide Shut. Consider encoder from SigLip and a vision model like DINOv2 for vision feature combination as in Prism (possibly other vision encoders). Possibly consider finetuning some of the vision encoder too or just some of its layers (Fuyu inspiration).
- Consider the bias tuning mechanism from Llama-Adapter v2.
- Consider continuing pretraining of one of the base models on Pile2dataset from Llema for improved mathematical reasoning ability which can possibly improve the overall reasoning ability.
- Can we use two images for input? I am not sure of a dataset. The issue would also be the context. We could use the patch mix technique from the World Model LMM paper (for videos and long context). The two images can be two-shot from the video.
- consider the technique in MM-CoT to reduce hallucinations.
- If necessary, we can augment at inference time with another module from a pretrained open source off the shelf model for the input augmentation, the way Llamma-adapter V2 does. They call them ”experts” to overcome the shortcoming of the visual abilities of their models.
- Mix the image and text patches - adapt early instead of late (to include the textual prompt early). Look at the big world model how they do with the video and at Fuyu. (dataset to train?)
- Mix in multilingual for improved ability in multilingual reasoning? Dataset to train on? TO mix with the image-caption pairs and visual instruction datasets from Llava1.5 and blip2?
- What does the Perceiver do that we could improve the architecture of token mixing with?
- Can evaluate on ARB dataset with images that were not used for problem solving, ScienceQA and the new MMMU dataset with diverse images and hard problems.
- Can evaluate the COCOt style with contrastive chain of thought with multiple images as in a visual story.

A few candidate datasets for training (pretraining and/or finetuning) are made available in Llava, Llema for math, world model (LWM), Blip, Aya from Cohere for multilingual.

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, 6

[2] 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. 1

[3] 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. 2

[4] 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, 5

[5] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 4, 6

[6] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 6

[7] Eden Dolev, Alaa Awad, Denisa Roberts, Zahra Ebrahimzadeh, Marcin Mejran, Vaibhav Malpani, and Mahir Yavuz. Efficient large-scale visual representation learning and evaluation. 1

[8] 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. 2

[9] 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, 3, 6, 8

[10] 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. 2

[11] 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

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 6, 8

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

[14] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024. 1

[15] 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. 1

[16] 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. 3, 6

[17] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023. 1

[18] Denisa Roberts. Neural networks for lorenz map prediction: A trip through time. *arXiv preprint arXiv:1903.07768*, 2019. 1

[19] Denisa AO Roberts and Lucas R Roberts. Qr and lq decomposition matrix backpropagation algorithms for square, wide, and deep–real or complex–matrices and their software implementation. *arXiv preprint arXiv:2009.10071*, 2020. 1

[20] 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. 2

[21] 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. 6

[22] 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. 2

[23] 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. 3, 6

[24] 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. 2

[25] 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. 2, 3, 6