Learning to Reason Iteratively and Parallelly for Complex Visual Reasoning Scenarios

Shantanu Jaiswal[1,2] Debaditya Roy[1] Basura Fernando[1] Cheston Tan[1]

NeurIPS 2024

1 Agency for Science, Technology and Research SINGAPORE

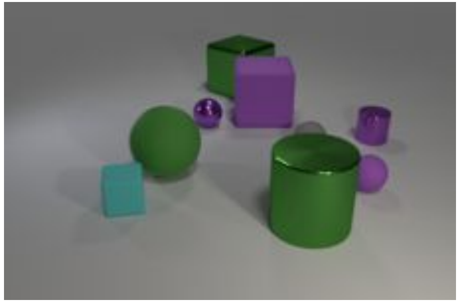
2 Carnegie Mellon University

Code to be released at: https://github.com/shantanuj/IPRM_Iterative_and_Parallel_Reasoning_Mechanism

Motivation

1. Complex visual reasoning scenarios require **compositional multi-step processing** and **higher-level reasoning capabilities** beyond immediate perception and knowledge of the world.

Are both the ball to the right of other balls and the black helmet made of plastic?



What is the color of small object in front of green object with the max occurring shape?

Examples of compositional multi-step reasoning tasks on images (GQA and CLEVR-Humans)

Motivation

1. Complex visual reasoning scenarios require **compositional multi-step processing** and **higher-level reasoning capabilities** beyond immediate perception and knowledge of the world.

Did they put down the camera before or after the longest occurring action?

Example of compositional spatiotemporal and situational reasoning (AGQA and STAR)

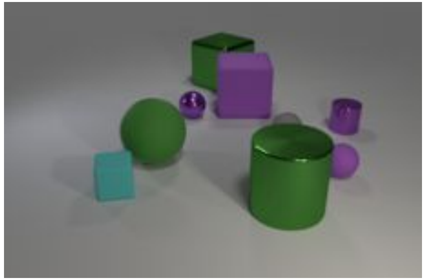
Motivation

- These tasks are less reliant on world knowledge, and may not be sufficiently addressed **through scaling pretraining of models alone.**
- **Architectural refinements** may also be needed.

Motivation

- These tasks are less reliant on world knowledge, and may not be sufficiently addressed **through scaling pretraining of models alone.**
- **Architectural refinements** may also be needed.
- Hence, we focus on designing a new neural reasoning architecture that **combines iterative and parallel computational priors** to support complex reasoning capabilities.

Iterative and Parallel Computation



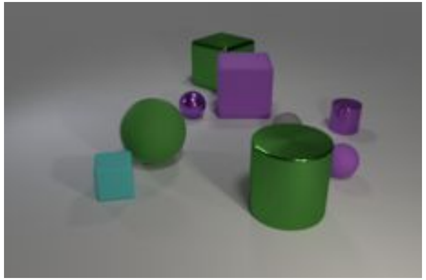
What is the color of small object in front of green object with the max occurring shape?

***Steps**: i) "count shapes" -> ii) "compute max shape"*
-> iii) find green object with target shape ..
-> vi) get color of small object

<u>Iterative computation:</u>

1. Enables **breaking down a problem** into appropriate sub tasks**.**
2. **Reason in a step-by-step manner** by utilizing memory (similar to in RNNs) to store and compose results.

Iterative and Parallel Computation



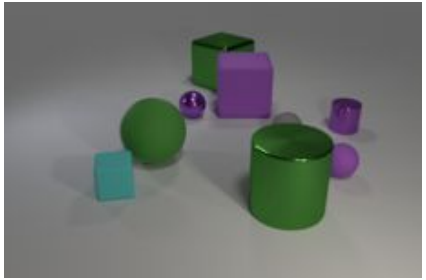
What is the color of small object in front of green object with the max occurring shape?

Iterative computation (similar to in RNNs):

Limitations:

- **Always performs operations sequentially** and can attend to a limited view at each time.
- Hence, independent operations that can be computed simultaneously are still computed sequentially (e.g. counting diff shapes in above example).

Iterative and Parallel Computation



What is the color of small object in front of green object with the max occurring shape?

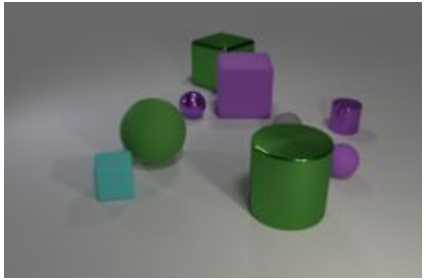
Sub-steps for counting 'shapes' to compute max shape:
i) Cubes -> ii) Cylinders -> iii) Spheres
(= 3 sequential time steps without forgetting prev. counts)

<u>Iterative computation (similar to in RNNs):</u>

Limitations:

- Independent operations that can be computed simultaneously are still computed sequentially.
- **Computation and memory retention demand grows** with number of operations (e.g. counting shapes scales with num of shapes in scene).

Iterative and Parallel Computation



What is the color of small object in front of green object with the max occurring shape?

Instead of iteratively, count shapes parallely:

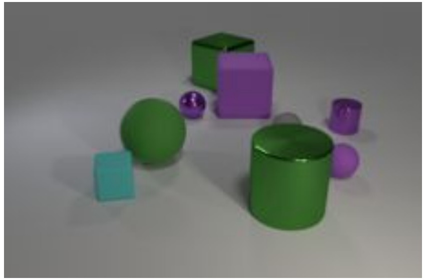
i) Cubes x 3
ii) Cylinders x 2
iii) Spheres x 4

(= 1 sequential time step and counts maintained separately)

Parallel computation (similar to in Transformers):

1. **Reason simultaneously** over independent operations and different reasoning paths.
2. Allows parallelly processing multiple operations or stimuli (e.g. co-occurring events in videos) in a more efficient and robust manner.

Iterative and Parallel Computation



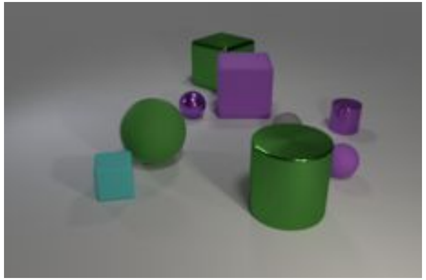
What is the color of small object in front of green object with the max occurring shape?

Parallel computation (similar to in Transformers):

Limitations:

- Does not explicitly model compositional computation to store and compose results in a step-by-step manner.
- (e.g. *max shape -> green obj -> in front of*)

Iterative and Parallel Computation



What is the color of small object in front of green object with the max occurring shape?

Iterative computation (similar to in RNNs):

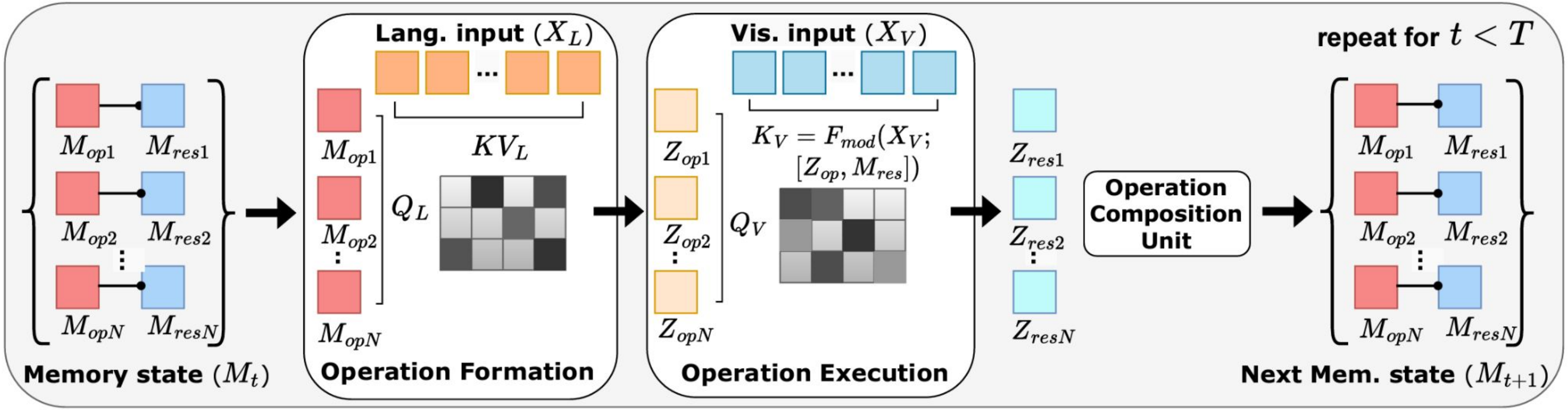
1. Enables **breaking down a problem** into appropriate sub tasks**.**
2. **Reason in a step-by-step manner** by utilizing memory to store and compose results.

Parallel computation (similar to in Transformers):

1. **Reason simultaneously** over independent operations and different reasoning paths.
2. Allows **parallelly processing multiple operations or stimuli** (e.g. co-occurring events in videos) in a more efficient and robust manner.

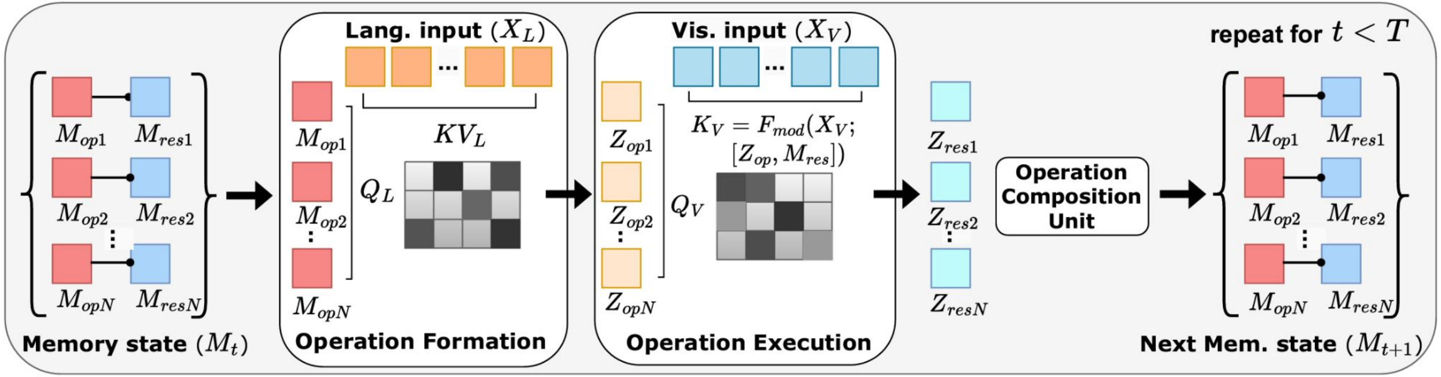
Iterative and Parallel Reasoning Mechanism (IPRM)

Given vision input (X_V) and language input (X_L), IPRM maintains an internal working memory (M_t) and performs the following computations for T iterations and N parallel operations.



IPRM computation flow (detailed in next slides)

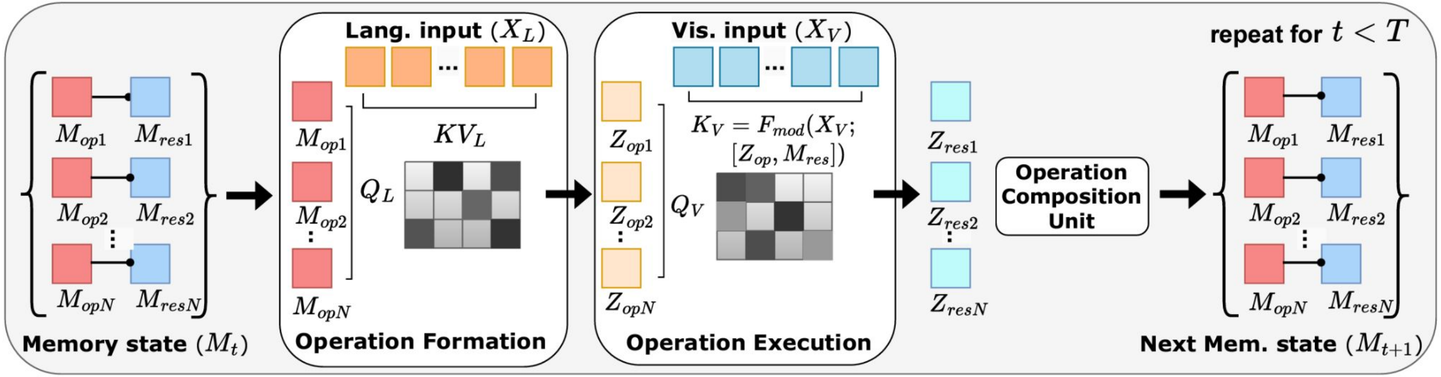
Iterative and Parallel Reasoning Mechanism (IPRM)



For T iterations, do:

1) **Operation Formation**: Form 'N' parallel operations by attending to language input (X_L) conditioned on previous operations $M_{op,t-1}$ in working memory.

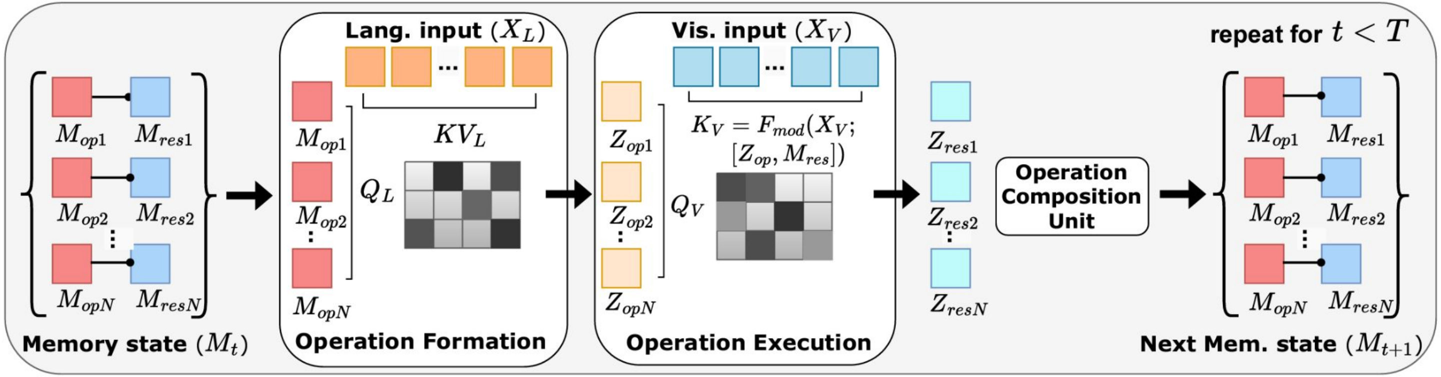
Iterative and Parallel Reasoning Mechanism (IPRM)



For T iterations, do:

2) **Operation Execution**: Execute the parallel operations by attending to visual input (X_v) conditioned on the formed operations (Z_{op}) and previous results $M_{res,t-1}$ in working memory.

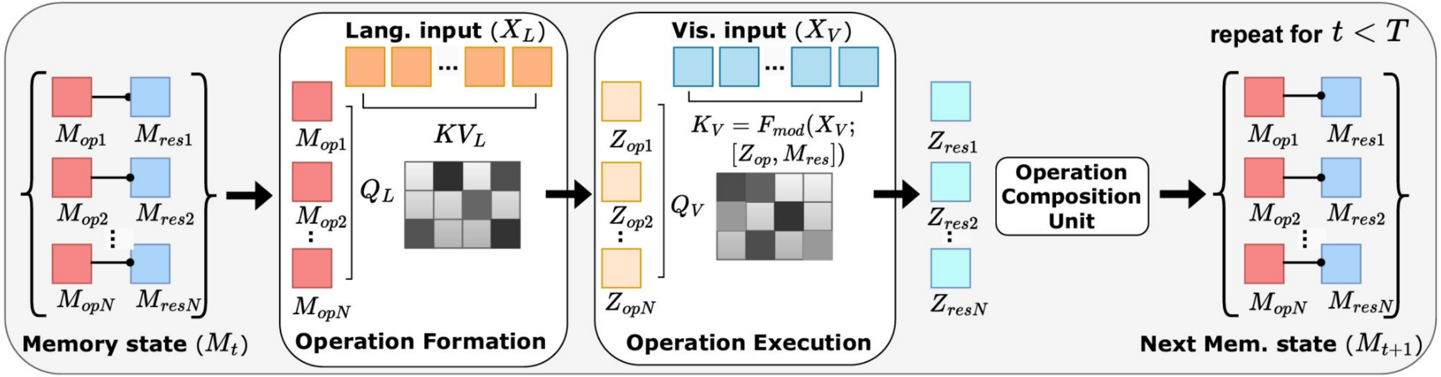
Iterative and Parallel Reasoning Mechanism (IPRM)



For T iterations, do:

3) **Operation Composition**: Update the working memory by composing the new parallel operations and results with one-another and integrating with the previous working memory state M_{t-1}.

Iterative and Parallel Reasoning Mechanism (IPRM)



For T iterations, do:

1) **Operation Formation**: Form 'N' parallel operations by attending to language input (X_L) conditioned on previous operations $M_{op,t-1}$ in working memory.
2) **Operation Execution**: Execute the parallel operations by attending to visual input (X_V) conditioned on the formed operations (Z_{op}) and previous results $M_{res,t-1}$ in working memory.
3) **Operation Composition**: Update the working memory by composing the new parallel operations and results with one-another and integrating with the previous working memory state M_{t-1}.

Experiments

Video reasoning benchmarks

- STAR (Situational Reasoning)
- AGQAv2 (Compositional spatio-temporal reasoning)
- CLEVRER-Humans (Causal reasoning)

Image reasoning benchmarks

- GQA (Compositional reasoning on real-world images)
- CLEVR-Humans (Compositional reasoning generalization to unseen language forms and novel reasoning skills)
- CLEVR-CoGenT (Compositional reasoning generalization on novel attribute compositions)

Experiments

Video reasoning benchmarks

<u>STAR</u>: *What did the person do with the bottle?*
<u>AGQA</u>: *Did they put down the camera before or after the longest occurring action?*

Model	Int.	Seq.	Pred.	Feas.	Avg.
LRR*[5]	73.7	71.0	71.3	65.1	70.3
LRR (w/o surrogate)	54.5	48.7	44.3	45.5	48.2
All-in-One [72]	47.5	50.8	47.7	44.0	47.5
Temp[ATP][7]	50.6	52.8	49.3	40.6	48.3
MIST [18]	55.5	54.2	54.2	44.4	51.1
InternVideo (8) [75]	62.7	65.6	54.9	51.9	58.7
SeViLA-BLIP2 [86]	63.7	70.4	63.1	**62.4**	64.9
Concat-Att-4L	68.1	71.4	66.6	55.2	65.3
Cross-Att-4L	67.5	72.1	64.4	58.5	65.6
IPRM	**71.8**	**77.7**	**71.0**	59.1	**69.9**

STAR

Metric	HCRN[39]	AIO[72]	Temp[7]	MIST[18]	GF [4]	IPRM
obj-rel	40.3	48.3	50.2	51.7	55.0	**57.8**
superlative	33.6	37.5	39.8	42.1	44.6	**48.0**
sequencing	49.7	49.6	48.3	67.2	53.2	**75.6**
exist	50.0	50.8	51.8	60.3	59.1	**62.4**
duration	43.8	45.4	49.6	**54.6**	52.8	50.7
act. recog.	5.5	19.0	19.0	19.7	14.2	**20.0**
open	36.3	-	-	50.6	56.1	**58.6**
binary	48.0	-	-	58.3	54.2	**62.3**
all	42.1	48.6	49.8	54.4	55.1	**60.4**

AGQA

- Improves state-of-art by close to 5% on both STAR and AGQA and **outperforms transformer-based vision-language attention modules and models such as BLIP2.**
- Particularly beneficial for "Prediction" (+7% on STAR) and "Sequencing" question types (+5% on STAR and +8% on AGQA).

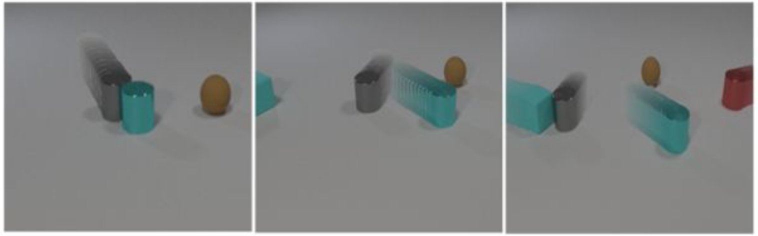
Iterative and parallel computation for visual reasoning

Table 2: Comparison of methods for CLEVRER-Humans [51] (Opt. is per option acc. and Qs. is per question acc.). IPRM achieves state-of-art across settings.

Model	Zero-shot		Finetune		Scratch	
	Opt.	Qs.	Opt.	Qs.	Opt.	Qs.
NS-DR[84]	51.0	32.0	-	-	-	-
VRDP[13]	50.9	31.6	-	-	-	-
CNNLSTM[51]	50.3	30.0	51.7	34.2	51.5	30.8
CNNBERT[51]	52.9	32.0	52.0	30.2	50.1	30.4
ALOE[12]	54.0	26.9	51.8	31.7	52.7	32.1
IPRM	**61.7**	**38.9**	**74.1**	**53.0**	**62.0**	**38.3**

CLEVRER-Humans:

- Increases state-of-art across zero-shot, fine-tuned and trained from scratch settings.

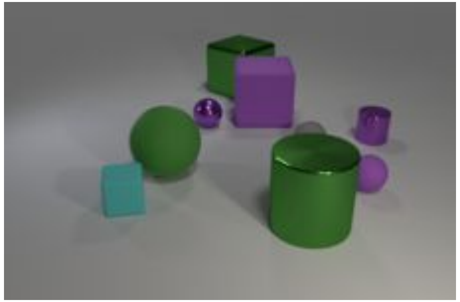


Is collision b/w cyan and gray cylinder responsible for collision b/w gray cylinder and cyan cube?

Experiments

Image reasoning benchmarks

Model	Extra supv.	CLV-Hum		CLV-CoGen		CLOSURE
		ZS	FT	ValA	ValB	ZS Avg.
PG+EE [35]	Programs	54.0	66.6	96.6	73.7	75.6
NS-VQA [85]	Programs	-	67.8	**99.8**	63.9	**77.2**
FiLM [60]	None	56.6	75.9	98.3	78.8	56.9
MAC [28]	None	57.4	81.5	99.0	78.3	73.8
MDETR [36]	Bound. Box	59.9	81.7	**99.8**	76.7	-
IPRM	None	**63.8**	**85.5**	99.1	**80.3**	75.6



What is the color of the small object in front of green object with the max occurring shape?

- Improves state-of-art on CLEVR-Humans and CLEVR-CoGenT without requiring extra supervision.
- Achieves strong zero-shot performances suggesting better generalizability of reasoning skills.

Experiments

Data efficient learning and better zero-shot generalization

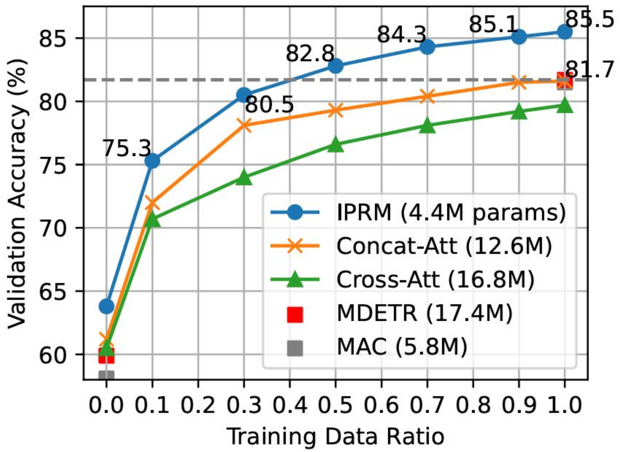


Figure 5: IPRM performance on CLEVR-Humans at different training data ratios of Cross- and Concat-Att.

- More **data-efficient learning** and **better zero-shot performances** than prior state-of-art (MDETR) and transformer-based modules.
- IPRM when trained with **only 50% data** exceeds prior state of art (MDETR) and also requires lesser parameters.

Experiments

Image reasoning benchmarks

Are both the ball to the right of other balls and the black helmet made of plastic?

Table 4: Performance comparison on GQA with imageQA methods and large-scale models that do not utilize ground-truth scene graphs. * indicates large-scale pretrained VL model. **Utilizes ground truth scene graphs, programs and bounding boxes for auxiliary training.

	LCGN [25]	MCAN[87]	LXMERT*[66]	12-in-1*[49]	OSCAR*[46]	CFR** [55]	IPRM
GQA	55.8	57.4	60.0	60.0	**61.6**	72.1	60.3

- Achieves highest performance on GQA amongst imageQA methods (that are trained only on GQA without additional supervision or pretraining).
- Performs competitively with larger-scale pretrained vision-language models.
- Achieves 87.2% when trained with ground truth bounding boxes and attributes, suggesting further benefits possible through stronger visual backbones.

Experiments

Performs strongly at longer program lengths (proxy for reasoning steps)

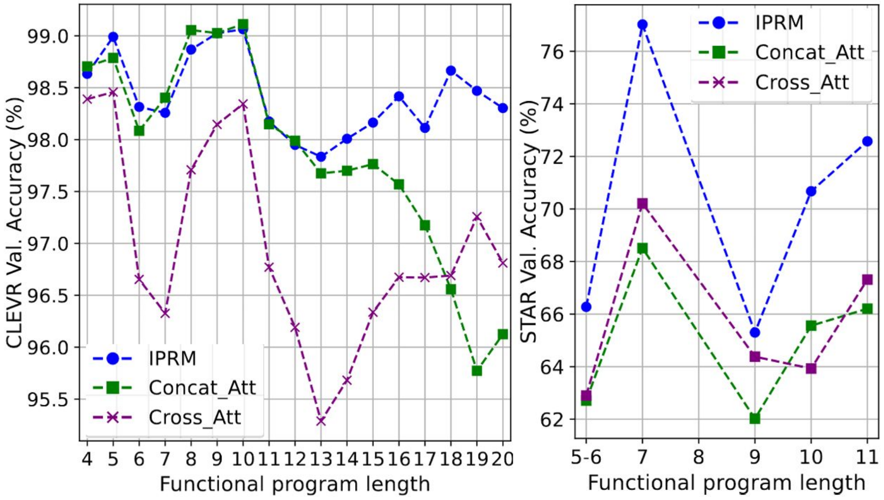


Figure 4: Acc. of IPRM (blue) across program lengths for CLEVR (left) and STAR (right). IPRM has significantly higher accs. at longer program lengths.

- Maintains **strong performances at longer program lengths** (indicative of more complex questions) compared to transformer-based vision-language attention modules.

Experiments

Further results on CLIP backbones

Table 8: **Left:** Comparison of IPRM with prominent vision-language attention mechanisms with CLIP VIT-L/14 backbones on CLEVR-Humans, GQA and NLVRv2 benchmarks (‘4L’ indicates 4 att layers; ‘x’ indicates model did not converge). **Right:** Results with other CLIP variants VIT-B and VIT-L@ 336 on GQA and NLVRv2.

Model (CLIP VIT-L/14 bbone)	+Param	+GFLOPs	GQA TestD	NLVR2 Test	CLV-H	
					ZS	FT
Wt-Proj-Fusion	0.6M	0.1	53.5	60.8	58.5	74.4
Cross-Att (2L)	9.2M	1.5	55.1	62.1	-	-
Concat-Att (2L)	7.2M	4.4	55.3	60.5	-	-
Cross-Att (4L)	17.6M	3.1	57.4	54.4	60.3	80.0
Concat-Att (4L)	13.6M	8.9	58.7	55.9	61.2	81.1
Cross-Att (6L)	26.0M	4.5	56.8	x	60.8	80.4
Concat-Att (6L)	19.7M	13.3	57.4	x	62.0	81.8
IPRM	5.2M	5.9	**59.2**	**65.1**	**64.3**	**84.6**

Model (CLIP VIT-B/16 bbone)	GQA TestD	NLVR2 Test
Wt-Proj-Fusion	51.4	59.9
Cross-Att	54.6	56.6
Concat-Att	56.0	57.4
IPRM	55.9	60.8

Model (CLIP VIT-L/14@336)	GQA TestD	NLVR2 Test
Wt-Proj-Fusion	54.0	61.1
Cross-Att	57.4	58.4
Concat-Att	57.3	59.1
IPRM	59.0	65.3

- Adding IPRM is more effective than adding further transformer-based attention blocks.
- Requires lesser parameters and comparable FLOPs.

Interpretability and Visualization of Intermediate Steps

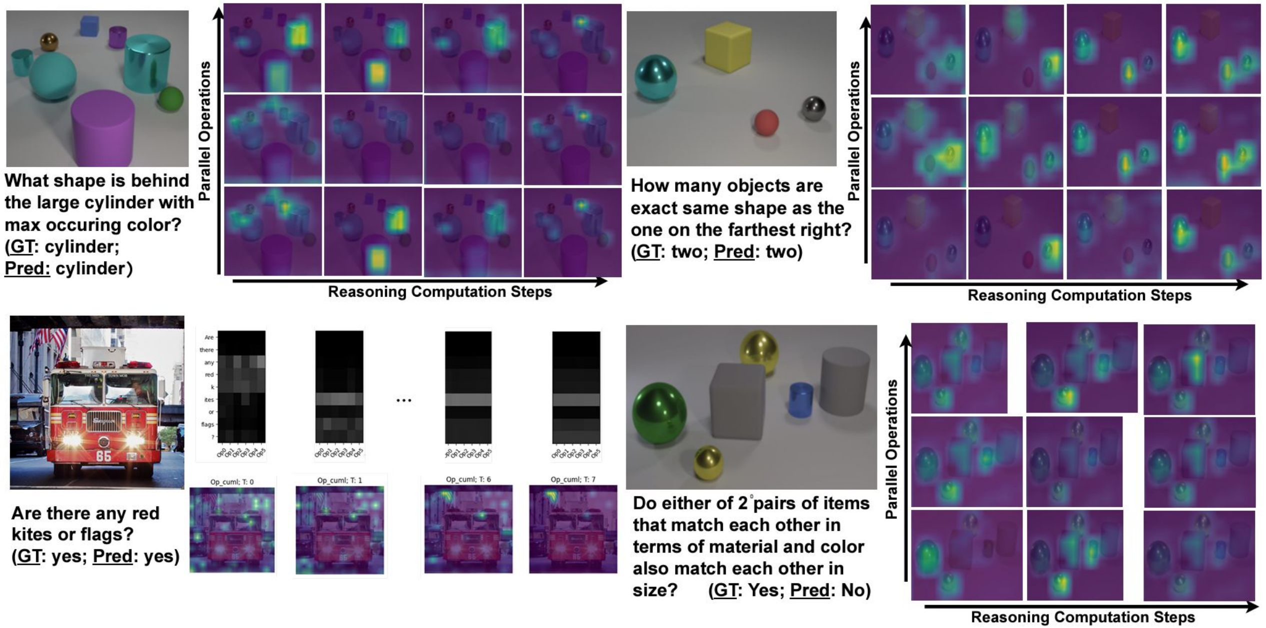


Figure 7: **Condensed reasoning visualization of IPRM**. In the top two examples, IPRM correctly utilizes both parallel and iterative computation to arrive at the correct answer. The bottom left example shows IPRM's cumulative lang. and visual attentions when solving a real-world GQA example. The bottom right example, shows an error case where IPRM seems to misunderstand question and outputs wrong ans. with less relevant attentions. See appendix for further reasoning visualizations and error

Model Ablations

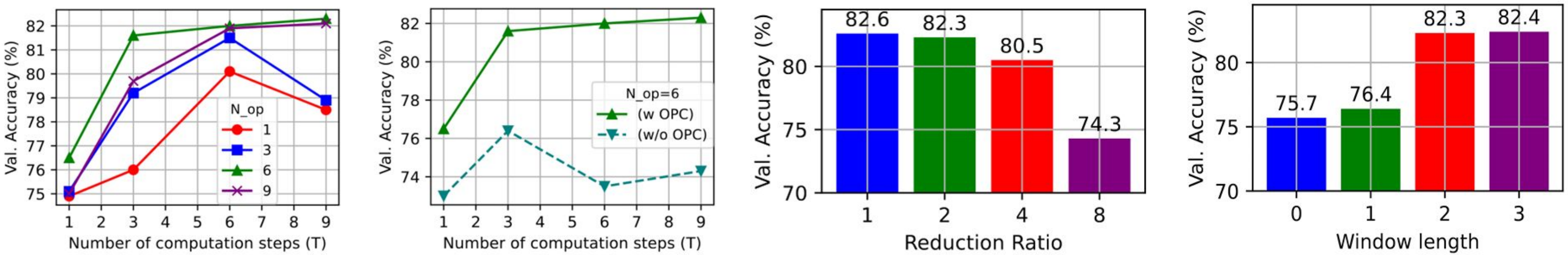


Figure 6: IPRM Model ablations in order: **(i)** Impact of number of parallel operations (N_{op}) vs computation steps (T). **(ii)** Impact of Operation Composition Block (OPC). **(iii)**: Impact of reduction ratio (r) and **(iv)** memory window length (W).

Conclusion

1. Introduced a new neural reasoning architecture (IPRM) to better support complex visual reasoning capabilities.
2. Can be conveniently integrated with conventional transformer and non-transformer based vision and language backbones
3. Outperforms transformer-based modules while being more parameter efficient, having comparable FLOPs and retaining parallelizability benefits.
4. While currently studied in context of visual reasoning, future work can look into application of IPRM for language and embodied reasoning tasks as well.
    a. X_L= reasoning task (e.g. a question, a task specification prompt, etc.)
    b. X_V = reasoning stimuli (e.g. embodied scene, language document, etc)