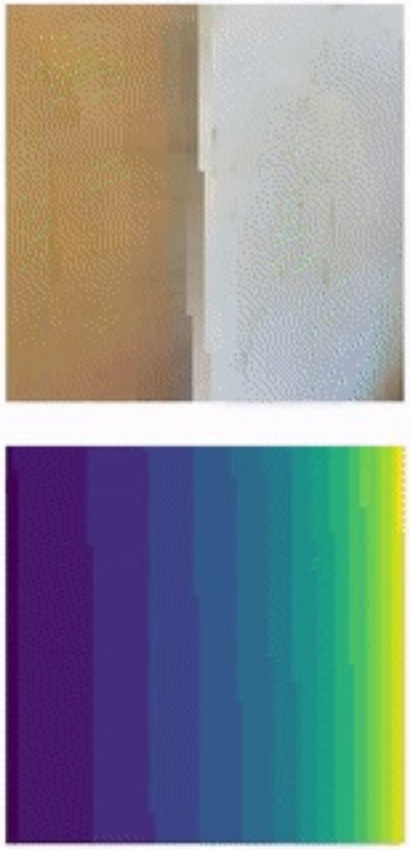
Visualizing, evaluating and transferring reasoning patterns in VQA,

Christian Wolf
September 29th, 2021

The group in Feb. 2020: Corentin Kervadec, Steeven Janny, Edward Beeching, Fabien Baradel, Théo Jaunet, Quentin Possamaï.

Topological neural maps



[Beeching, Dibangoye, Simonin, Wolf, ECCV 2020]

Self-supervision for Deep-RL

1st ranked at Multi-ON Challenge (CVPR 2021)

Rank	Team	Progress	PPL	Success	SPL
1	Lyon	67	44	55	35
2	SGoLAM	64	38	52	32
3	VIMP	57	36	41	26

1st place: Team Lyon, *Pierre Marza, Laetitia Matignon, Olivier Simonin, Christian Wolf*, Learning mapping and spatial reasoning with auxiliary tasks.

[Marza, Matignon, Simonin, Wolf, Submitted to ICRA 2021]

Navigation in real environments

NAVER LABS EUROPE

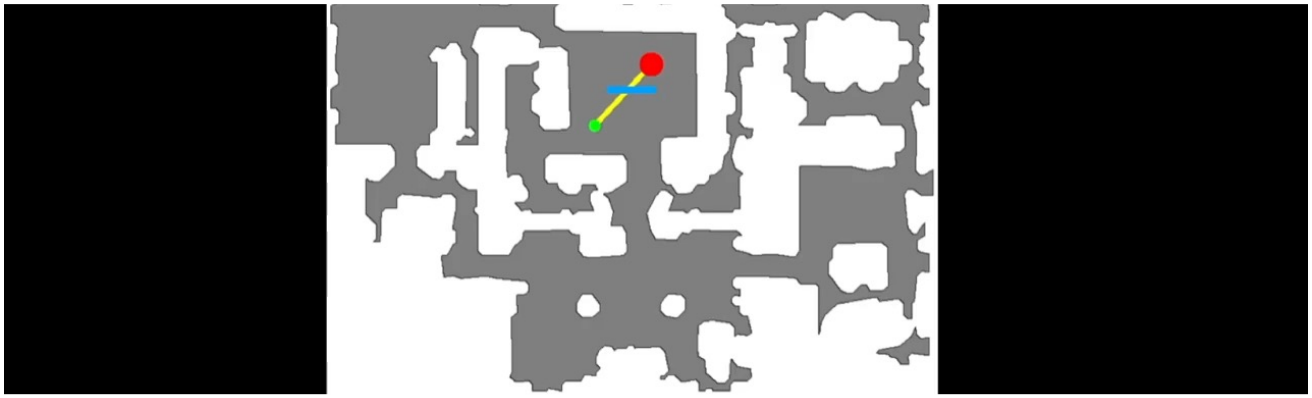

Assem Sadek


Guillaume Bono


Boris Chidlovskii


Christian Wolf

Navigation in real environments

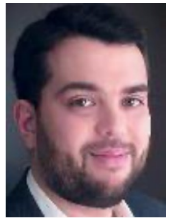


Vision and Language Reasoning

orange™


Corentin Kervadec

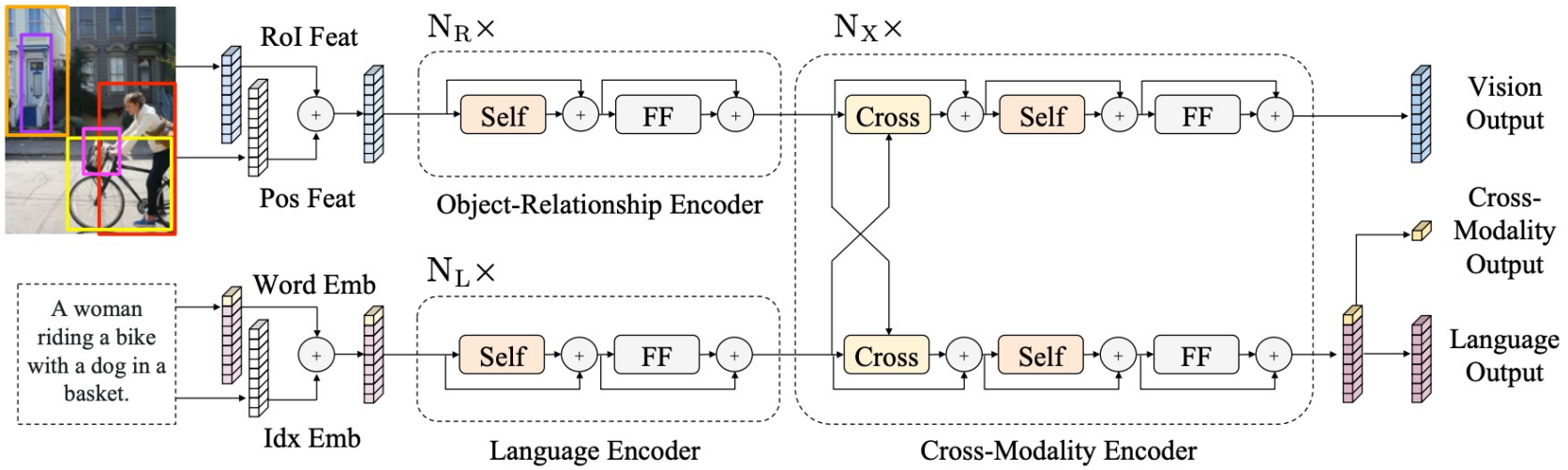

Grigory Antipov


Moez Baccouche


Christian Wolf

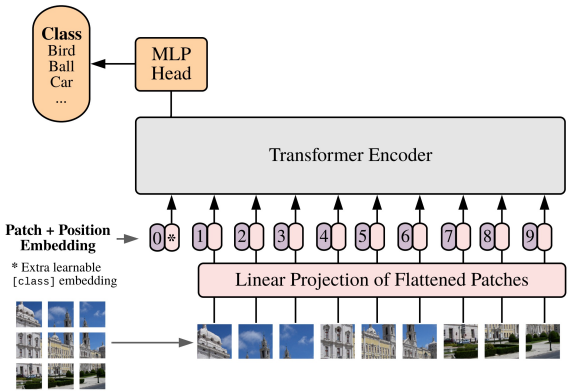
LXMERT

A vision and language encoder with self-attention and cross-attention.

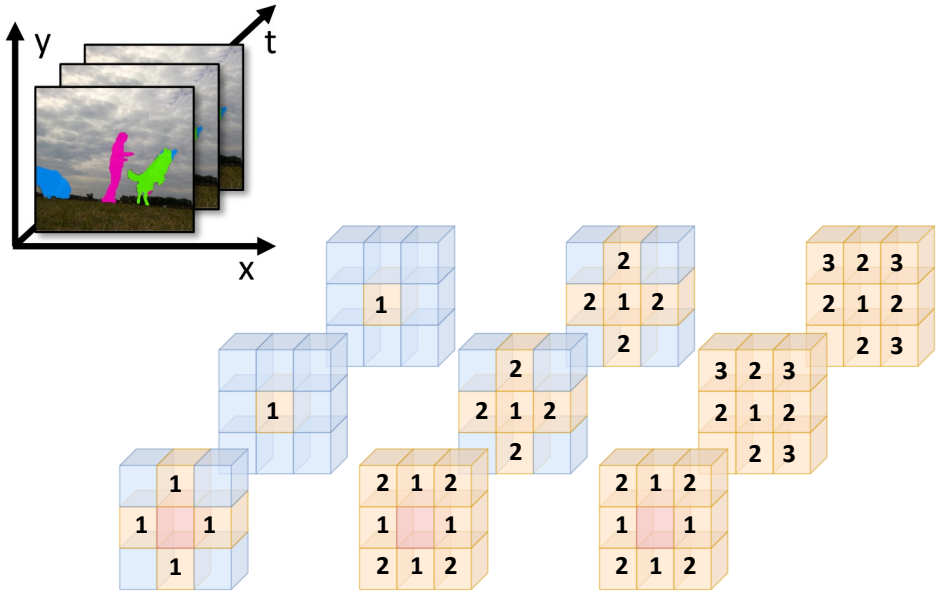


Tan, H. and Bansal, M. LXMERT: Learning cross-modality encoder representations fromtransformers. EMNLP-IJCNLP 2019.

Transformers for images and videos

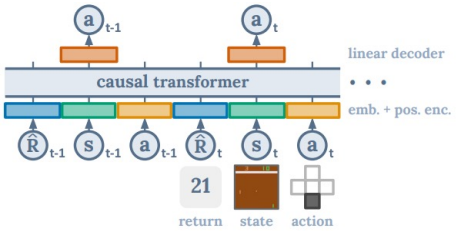


A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021.

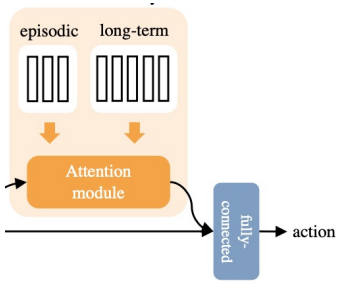


[Duke, Ahmed, Wolf, Arabi, Taylor, CVPR 2021]

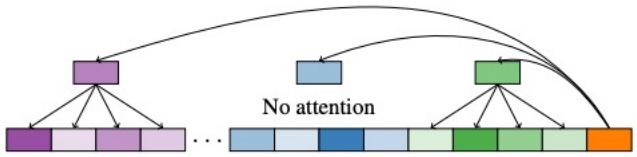
Transformers for robotics



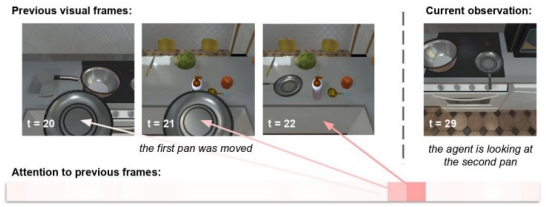
L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, I. Mordatch. Decision Transformer: Reinforcement Learning via Sequence Modeling arxiv 6/2021



L. Mezghani, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, Piotr B., K. Alahari, Memory-Augmented Reinforcement Learning for Image-Goal Navigation, arxiv 1/2021

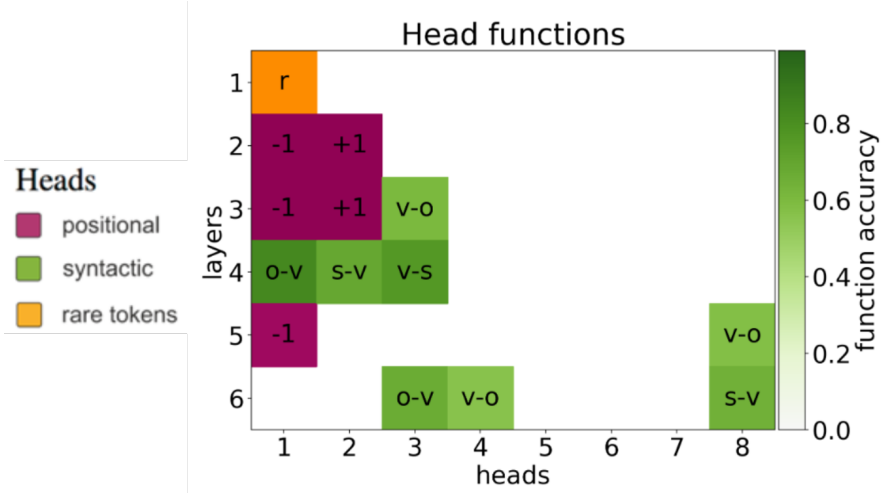


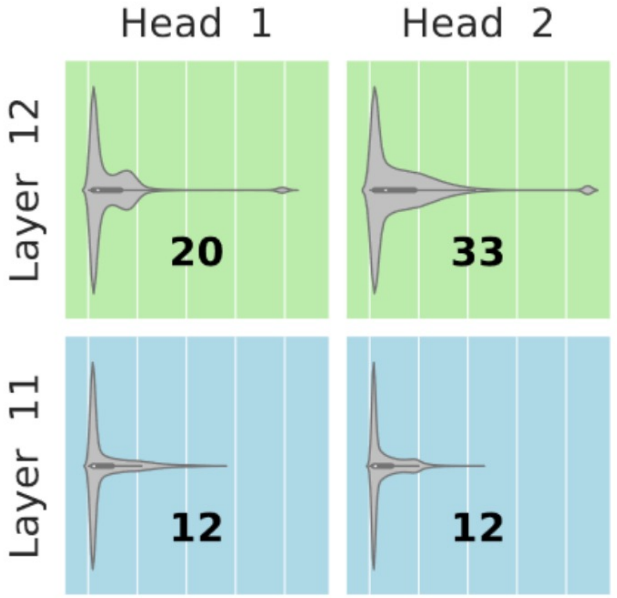
A.K. Lampinen, S.C.Y. Chan, A. Banino, F. Hill, Towards mental time travel: a hierarchical memory for reinforcement learning agents., arxiv 5/2021



A. Pashevich, C. Schmid, C. Sun, Episodic Transformer for Vision-and-Language Navigation, ICCV 2021.

Visualization of attention maps



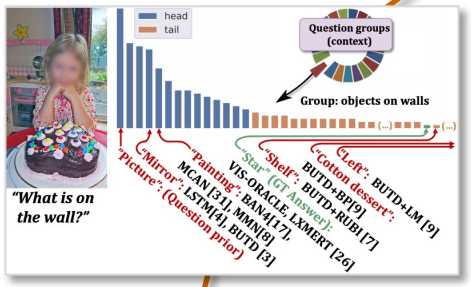


Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I., Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. ACL 2019.

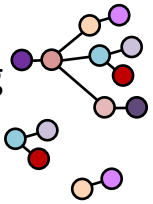
Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M.,Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter. Hopfield networks is all you need. Pre-print arXiv:2008.02217, 2020.

VQA

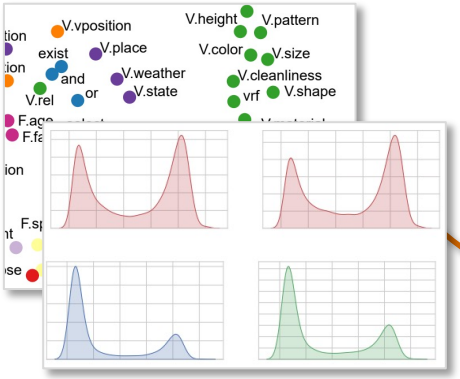
How can we evaluate biases in learning? (CVPR 2021a)



Can we supervise reasoning programs? (NeurIPS 2021)



How can we visualize and transfer reasoning? (CVPR 2021b, VIS 2021)



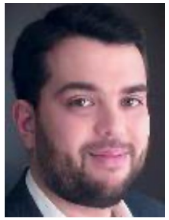
Can we weakly supervise word-object alignment? (ECAI 2020)

Can we ground object detection through language (under preparation)

orange™


Corentin Kervadec

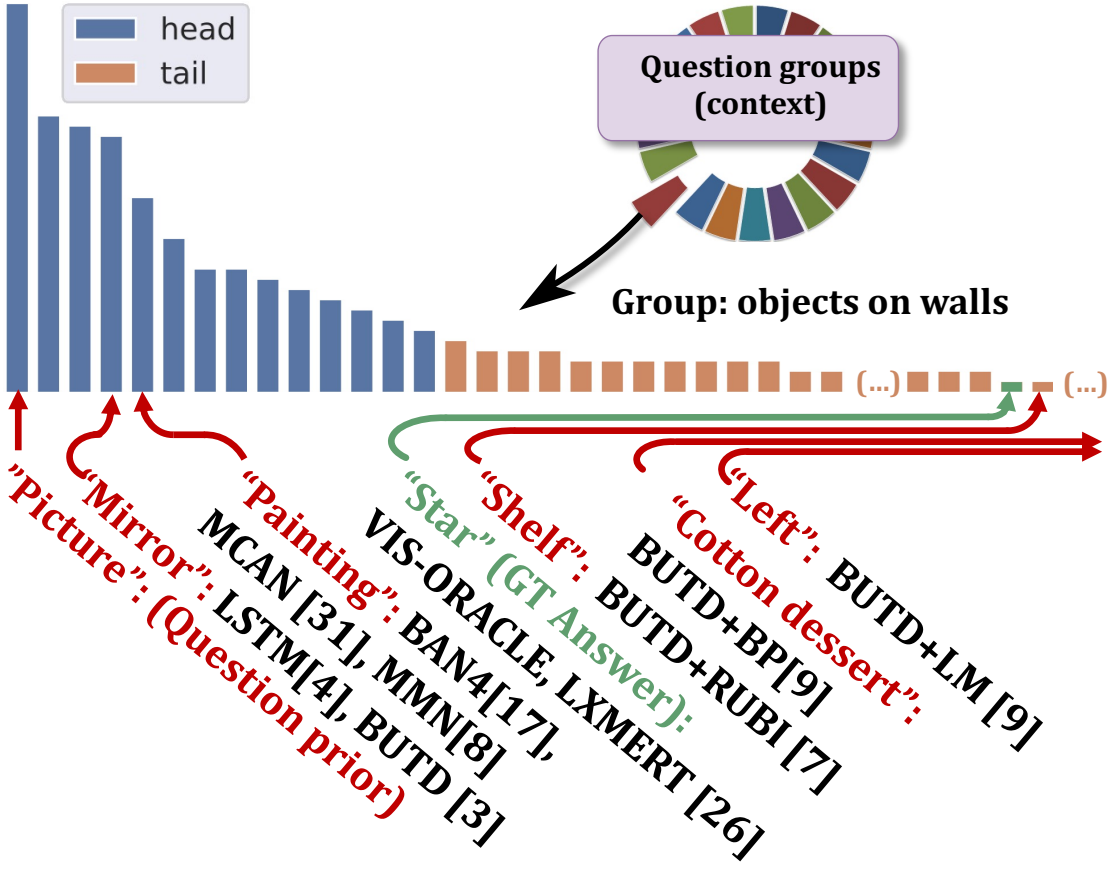

Grigory Antipov


Moez Baccouche


Christian Wolf

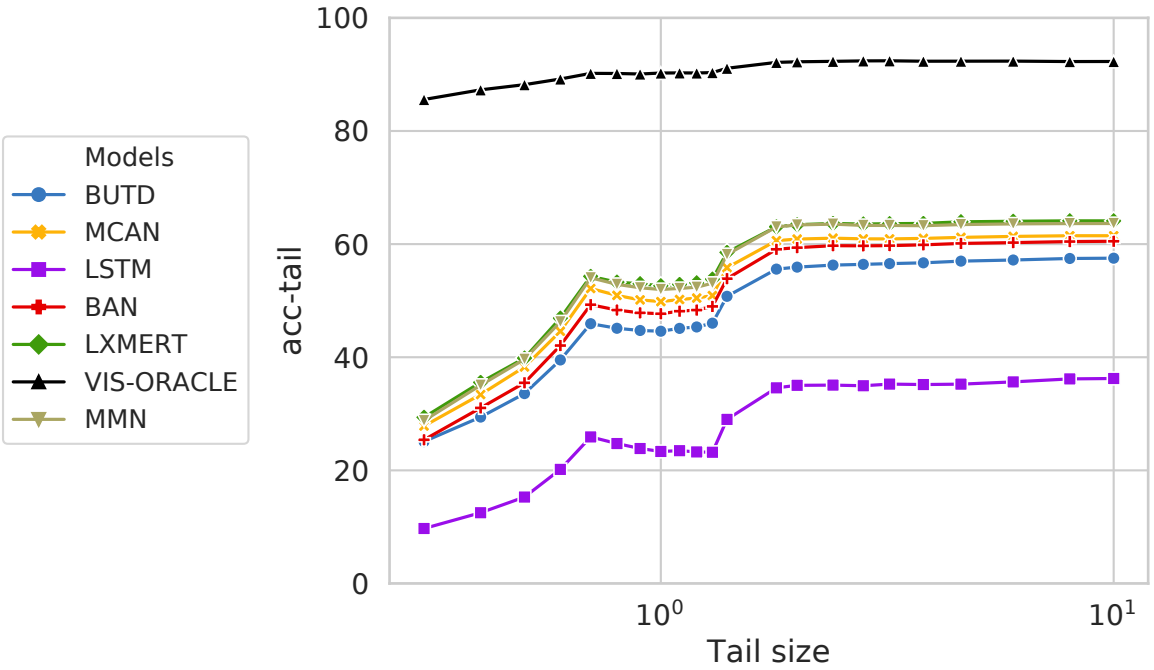
Roses Are Red, Violets Are Blue... but Should VQA Expect Them To?

"What is on the wall?"



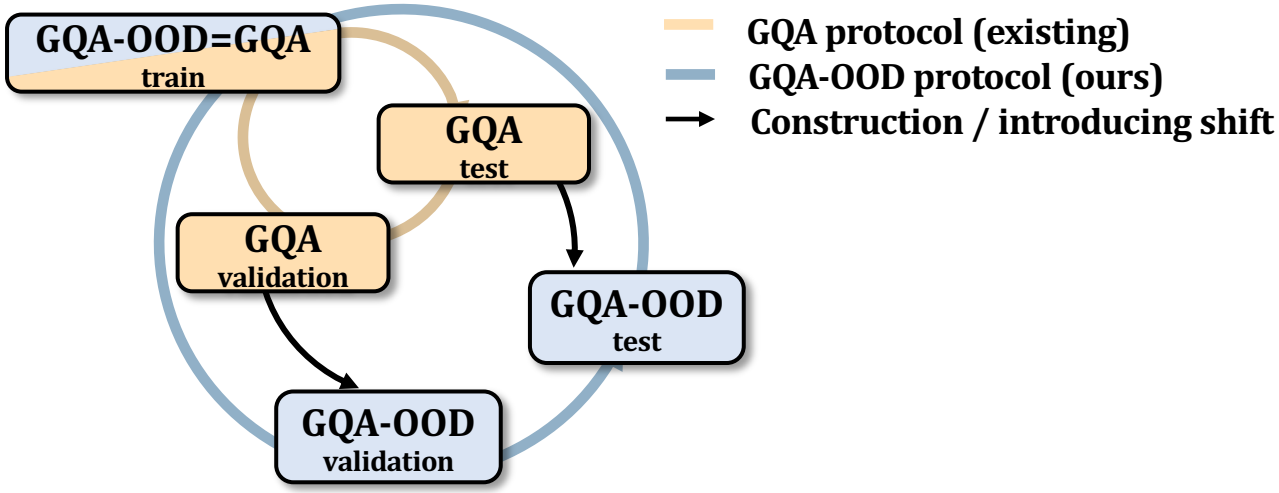
[Kervadec, Antipov, Baccouche, Wolf, CVPR 2021a]

Reasoning vs. bias exploitation



[Kervadec, Antipov, Baccouche, Wolf, CVPR 2021a]

OOD split


GQA-OOD=GQA
train
GQA
test
GQA
validation
GQA-OOD
test
GQA-OOD
validation
GQA protocol (existing)
GQA-OOD protocol (ours)
Construction / introducing shift

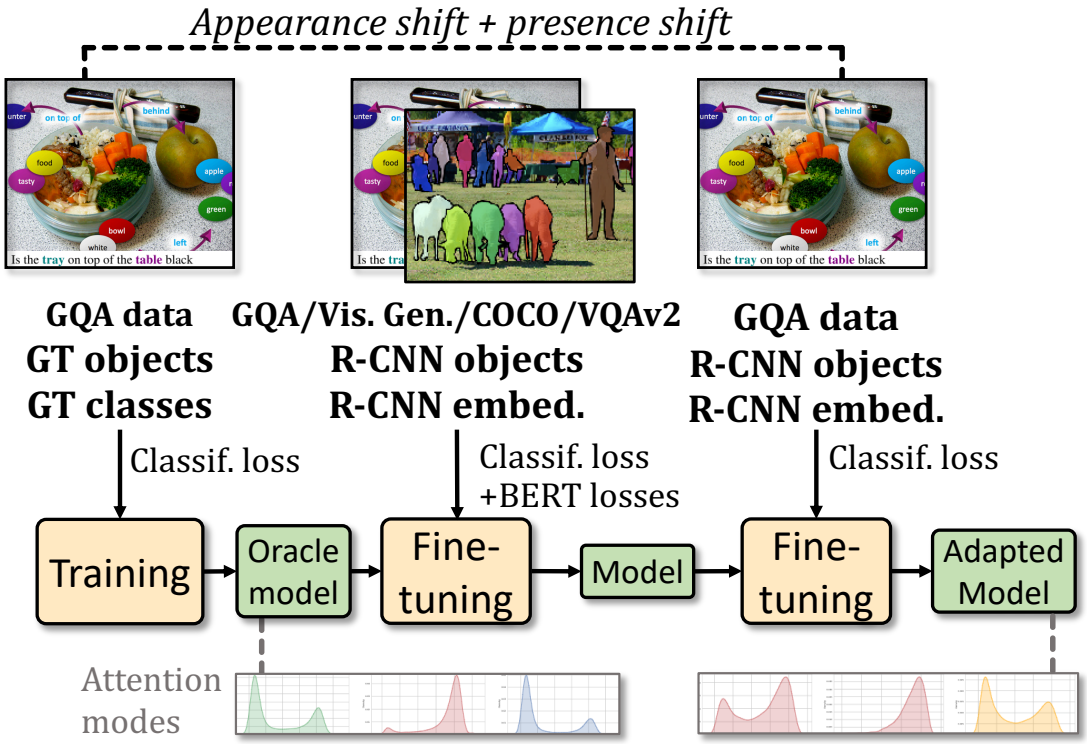
Calibrating the metric

Model	Baseline benchm. Tot. Acc.	Proposed benchmark (Acc-tail) α=1.2	α=0.5	α=0.3
BUTD [3] + *bal*	$60.7_{\pm 0.4}$	$45.4_{\pm 0.3}$	$33.8_{\pm 0.5}$	$24.6_{\pm 0.5}$
BUTD [3] + *all*	$59.8_{\pm 0.1}$	$41.9_{\pm 0.1}$	$29.5_{\pm 0.3}$	$18.3_{\pm 0.6}$
Δ (relative):	-1.4%	-7.7%	-12.9%	-25.7%

Results

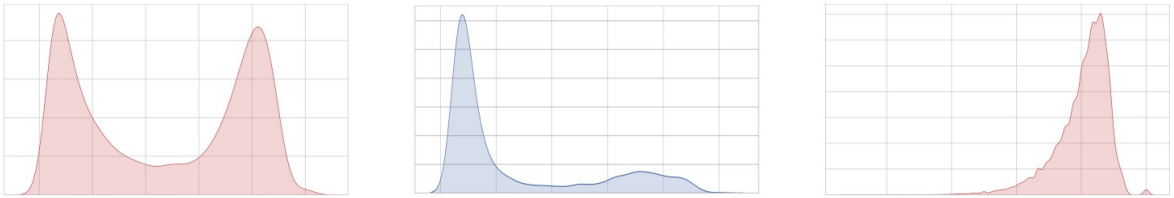
Model	VQA2	GQA		VQA-CP2	GQA-OOD
	overall	overall	dist.	overall	acc-tail
Q. Prior	32.1	*27.0*	*55.6*	8.8	*17.8*
LSTM [4]	43.0	*39.1*	*3.6*	*22.1*	*24.0*
BUTD [3]	63.5	*51.6*$\pm$0.3	*1.8*	40.1	*42.1*$\pm$0.9
MCAN [29]	***66.1***	*56.3*$\pm$0.2	*1.6*	***42.5***	*46.5*$\pm$0.5
BAN4 [18]	65.9	*54.7*$\pm$0.4	*1.6*	40.7	*47.2*$\pm$0.5
MMN [8]	-	**59.6**	1.8	-	*48.0*
LXMERT [24]	**69.9**	***59.6***	***1.5***	-	***49.8***
BUTD [3]	**63.5**	*51.6*$\pm$0.3	*1.8*	40.1	***42.1***$\pm$0.9
+RUBi+QB	-	***51.9***$\pm$1.1	***1.7***	***47.6***$\pm$3.7	***42.1***$\pm$1.0
+RUBi [7]	61.2	*43.6*$\pm$2.0	*1.9*	44.2	*35.7*$\pm$2.3
+LM [9]	56.4	*39.7*$\pm$0.7	*2.1*	52.0	*32.2*$\pm$1.2
+BP [9]	63.2	*39.6*$\pm$0.3	*2.2*	39.9	*30.8*$\pm$1.0

Oracle transfer



[Kervadec, Jaunet, Antipov, Baccouche, Vuillemot, Wolf, CVPR 2021b]

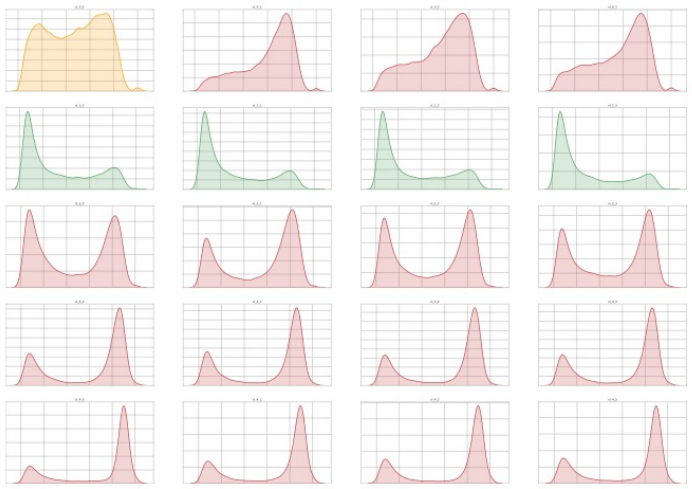
K-numbers and attention modes



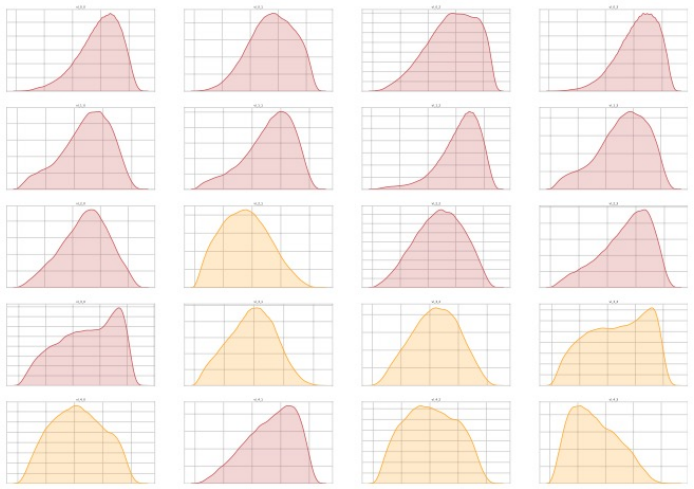
(a) Bimorph (b) Dirac (c) Uniform

Distribution of the number k of tokens required to reach 90% of the attention energy, on GQA-val.
X-axis: ratio of the tokens k w.r.t. the total number of tokens.

Analyzing reasoning patterns



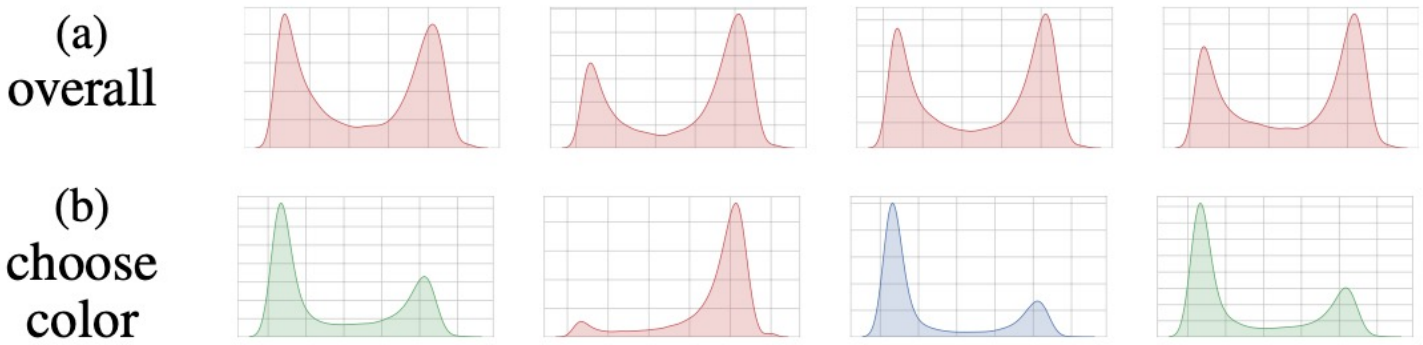
Oracle



Noisy/deployable

[Kervadec, Jaunet, Antipov, Baccouche, Vuillemot, Wolf, CVPR 2021b]

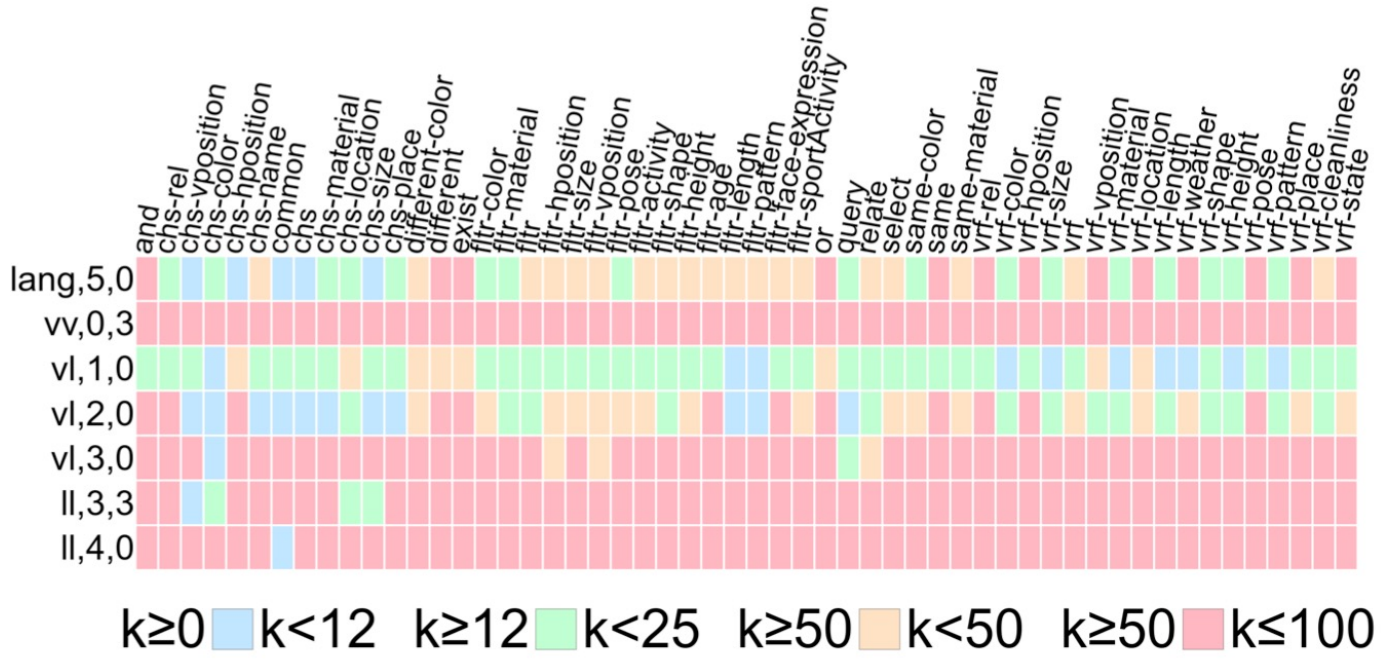
Attention and language tasks (1)



Samples of all functions vs. samples with questions involving “choose color”.
Oracle model, heads of third layer of $T_{\times}^{L \leftarrow V}$

Task dependence of attention head behavior:
Activation of the 1st, 2nd and 4th head, desactivation of the 3rd.

Attention and language tasks (2)

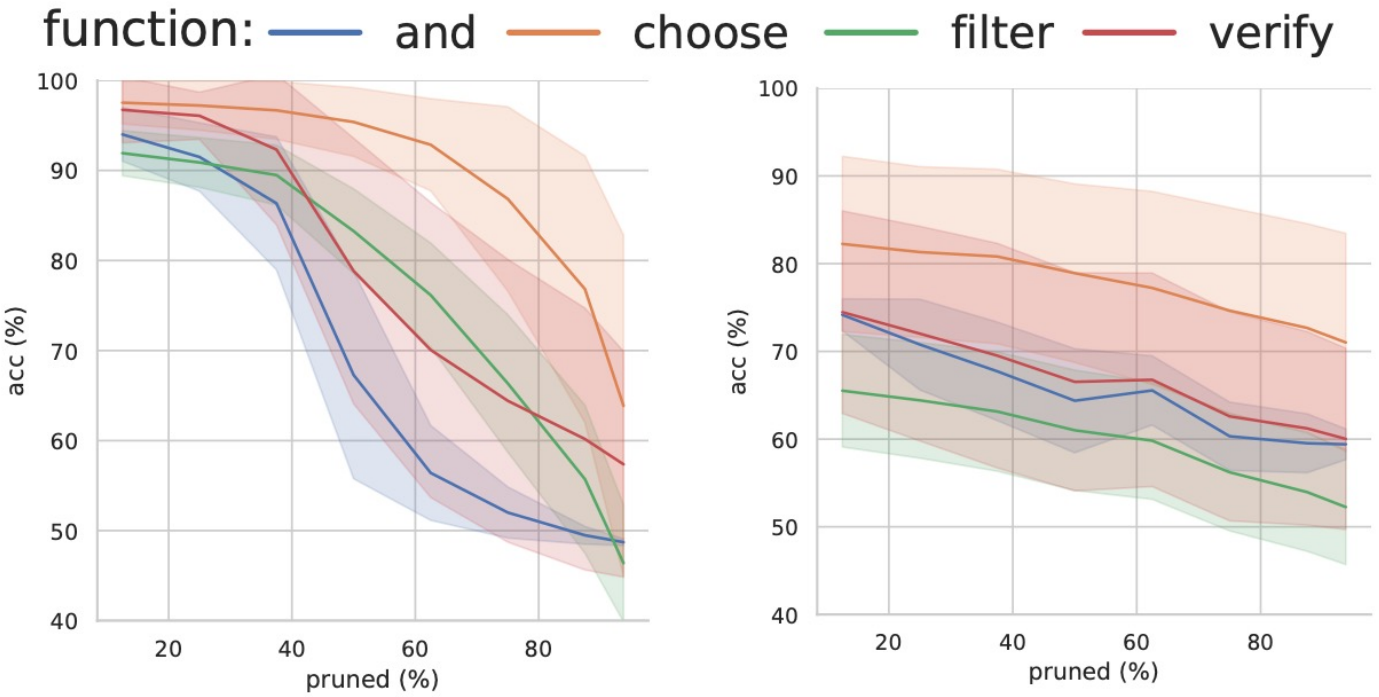


Attention heads behave differently depending on the function.
A given function causes different attention modes for different heads.

(Oracle model, median k-numbers)

Head pruning

Pruned attentions	n/a	L	V	L←V	V←L
Accuracy	91.5	37.9	91.4	52.8	68.1



Oracle

Noisy/deployable

How transferrable are Reasoning Patterns in VQA?

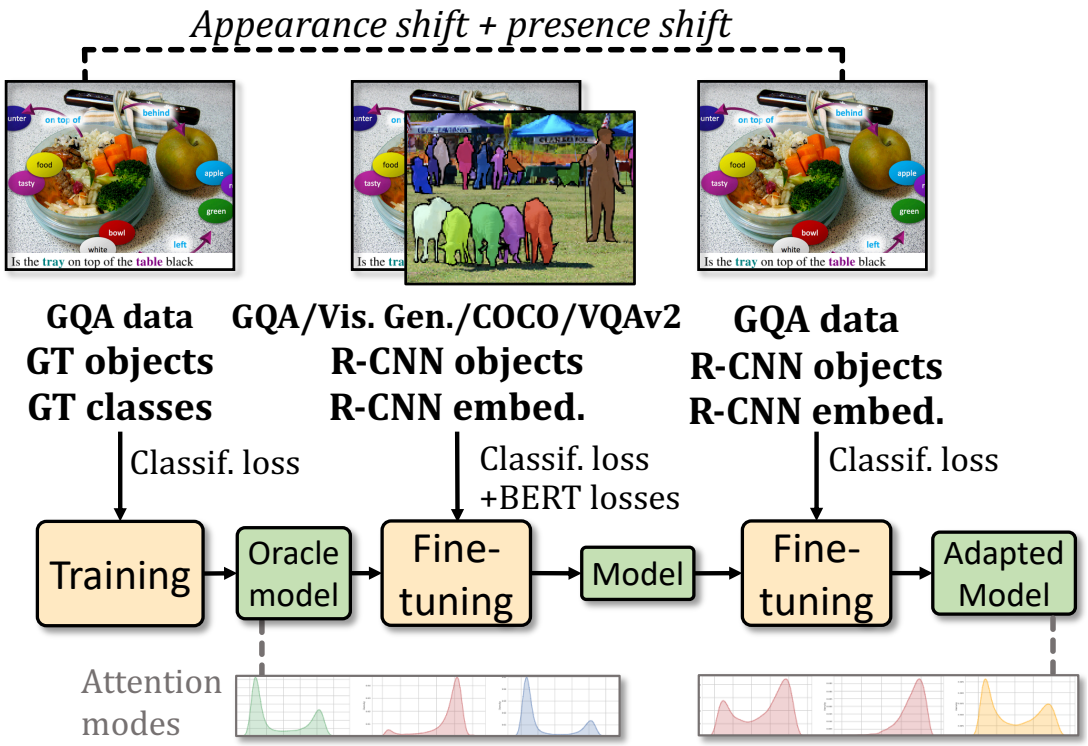
CVPR 2021

Corentin Kervadec Théo Jaunet Grigory Antipov
Moez Baccouche Romain Vuillemot Christian Wolf

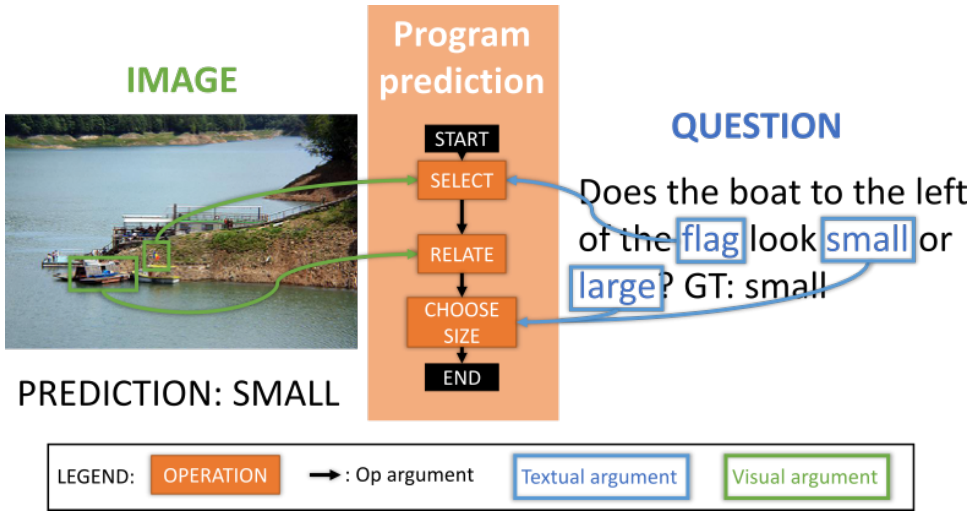
LIRIS, INSA-Lyon, Ecole Centrale de Lyon, Orange

https://visqa.liris.cnrs.fr

Are reasoning patterns transferrable?

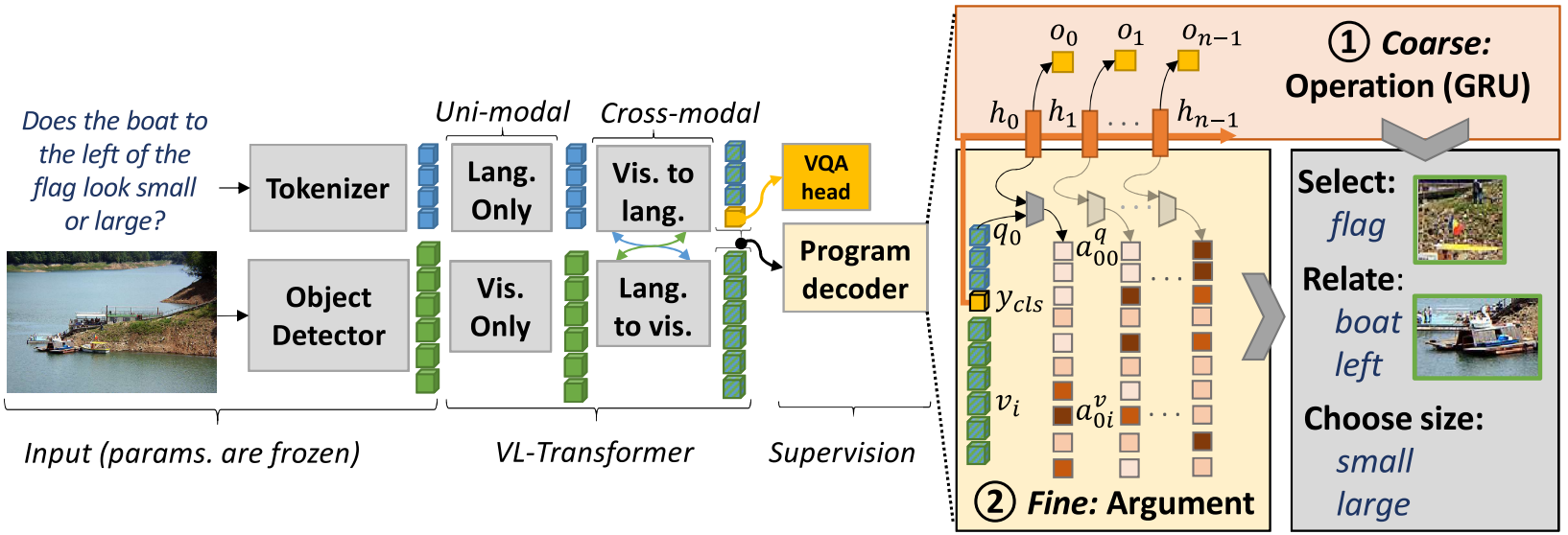


GT reasoning programs



[Kervadec*, Wolf*, Antipov, Baccouche, Nadri, NeurIPS 2021]

Program prediction



[Kervadec*, Wolf*, Antipov, Baccouche, Nadri, NeurIPS 2021]

Oracle + program supervision

	Model	Oracle transf.	Prog. sup.	GQA-OOD [20] acc-tail	GQA-OOD [20] acc-head	GQA [17] test-dev	GQA [17] binary*	GQA [17] open*	GQA [17] test-std	AUC† prog.
scratch	(a) Baseline			42.9	49.5	52.4	-	-	-	/
	(b) Oracle transfer	✓		48.2±0.3	54.6±1.1	57.0±0.3	74.5	42.1	57.3	/
	(c) Ours	✓	✓	**48.8**±0.1	**56.1**±0.3	**57.8**±0.2	**75.4**	**43.0**	**58.2**	97.1
+ lxmert	(d) Baseline			47.5	55.2	58.5	-	-	-	/
	(e) Oracle transfer	✓		47.1	54.8	58.4	77.1	42.6	58.8	/
	(f) **Ours**	✓	✓	**48.0**±0.6	**56.6**±0.6	**59.3**±0.3	**77.3**	**44.1**	**59.7**	96.4

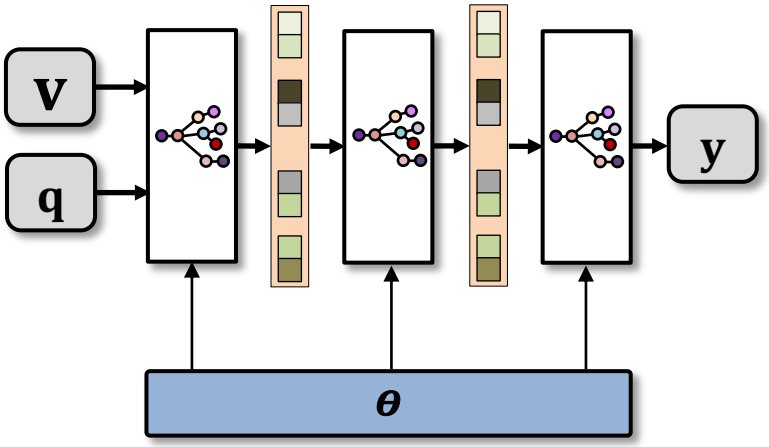
Table 1: Impact of program supervision on *Oracle transfer* [23] for vision-language transformers. LXMERT [36] pre-training is done on the GQA unbalanced training set. We report scores on GQA [17] (*test-dev* and *test-std*) and GQA-OOD (*test*). * binary and open scores are computed on the test-std; † we evaluate visual argument prediction by computing AUC@0.66 on GQA-val.

Comparison with SOTA

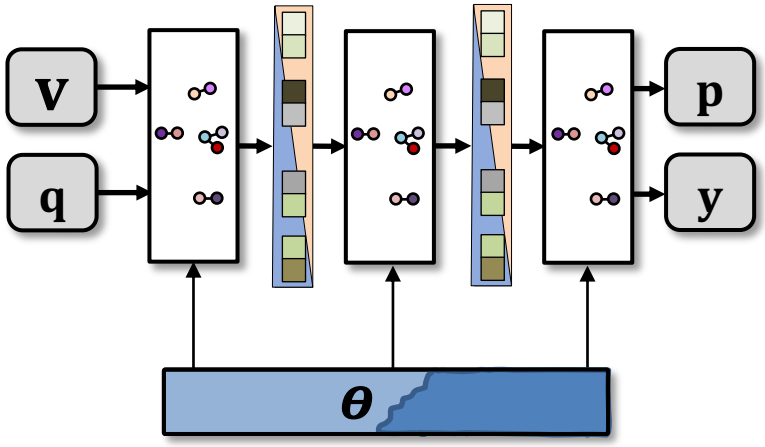
Method	Visual feats.	Additional supervision	Training data (M) Img	Sent	GQA-OOD [18] acc-tail	acc-head	GQA [15] bin.	open	all
BAN4 [22]	RCNN [2]	-	≈ 0.1	≈1	47.2	51.9	76.0	40.4	57.1
MCAN [37]	RCNN [2]	-	≈ 0.1	≈1	46.5	53.4	75.9	42.2	58.0
Oracle transfer [21]	RCNN [2]	-	≈0.18	≈1	48.3	55.5	75.2	44.1	58.7
MMN [6]	RCNN [2]	Program	≈0.1	≈15	48.0	55.5	78.9	44.9	60.8
LXMERT [30]	RCNN [2]	-	≈0.18	≈9	**49.8**	57.7	77.8	45.0	60.3
Ours	VinVL [39]	Program	≈0.1	≈15	49.1	**59.7**	80.1	48.0	63.0
NSM [14]	SG [14]	Scene graph	≈0.1	≈1	-	-	78.9	**49.3**	63.2
OSCAR+VinVL [39]	VinVL [39]	-	≈5.7	≈9	-	-	**82.3**	48.8	**64.7**

Compact version: embedding dim=128, 4 heads
26M params instead of 212M (LXMERT: dim=768, 8 heads).

Sample complexity



Classical training, CE loss on answers **y**



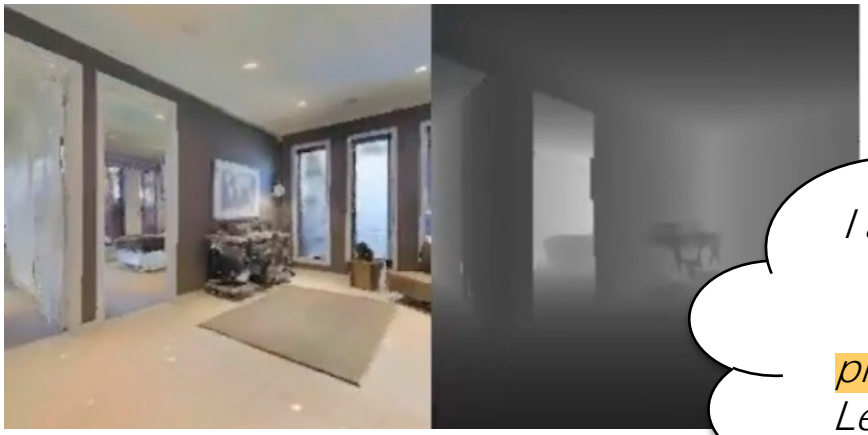
CE loss on **y** + program prediction **p**

Learned knowledge of the reasoning processes

Latent variables necessary for reasoning over multiple hops (used by reasoning processes)

Decomposition of the underlying (unknown) reasoning function

Outlook: language as Universal knowledge source



I am looking for a pillow and I a see a glimpse of a bedroom there, where pillows are frequently found. Let's check at this place first

Pierre Marza

Laetitia Matignon

Olivier Simonin

Christian Wolf

Conclusion

- Vision-Language reasoning as
  - A goal in it's own right
  - A mean to transfer knowledge to downstream tasks
- Short cuts and Clever-Hans effects are ubiquitous
- Out-of-distribution testing is important
- Cleaner data leads to easier learning of proper reasoning
- Challenges
  - Find correct learning signals
  - Create inductive biases
  - Identify and collect data sources (world models?)