Roses are Red, Violets are Blue... But Should VQA expect Them To?

Corentin Kervadec
T. Jaunet G. Antipov M. Baccouche R. Vuillemot C. Wolf

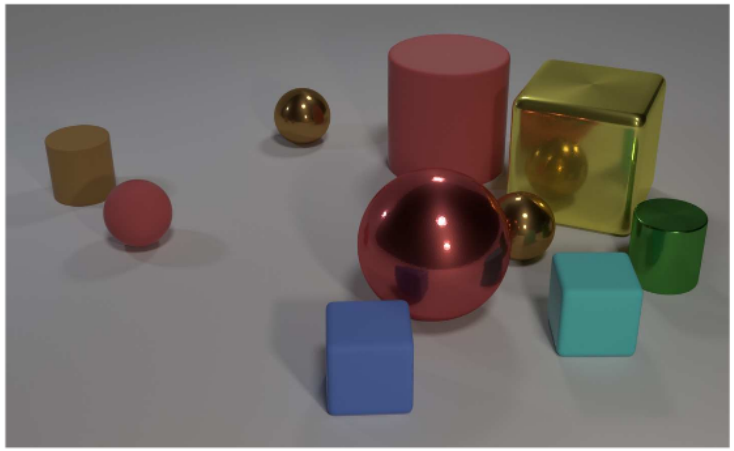
PhD Student @ Orange & LIRIS, INSA Lyon

Visual reasoning

Algebraically manipulating words and visual objects to answer a new question [Bot14]

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

(a) GQA [HM19]



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

(b) CLEVR [JHvdM+17]

Figure: Using Visual Question Answering (**VQA**) to evaluate reasoning skills.

Biases and Reasoning

Reasoning *vs.* **shortcut learning**

"decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions" [GJM+20]

Biases and Reasoning

Reasoning *vs.* **shortcut learning**

"decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions" [GJM+20]

What is the person holding?

Answer: Paper
Pred: Banana.

Also known as: biases, educated guesses, *etc*...

GQA-OOD: Measuring biases in VQA [KABW21]

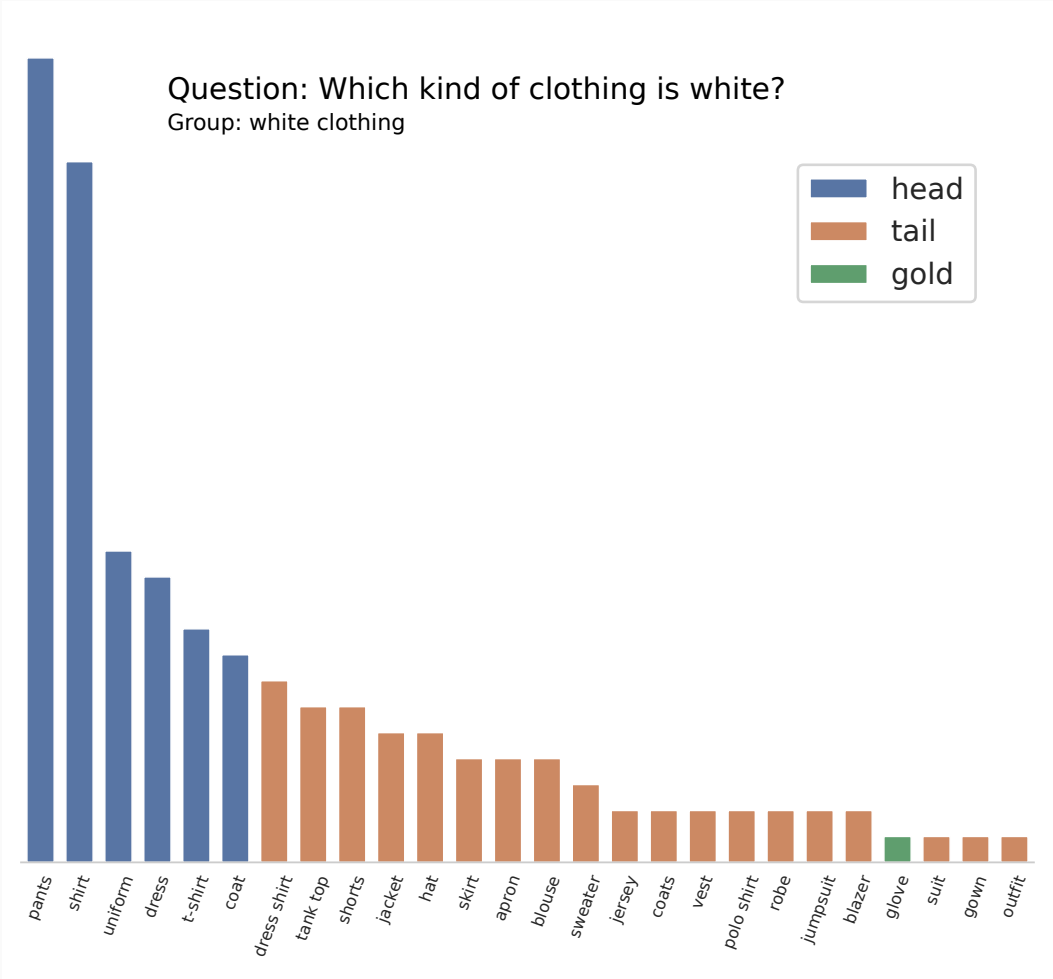
Roses are Red, Violets are Blue...
But Should VQA expect Them To?

Corentin Kervadec, Grigory Antipov, Moez Baccouche and Christian Wolf.
CVPR'21

GQA-OOD: Measuring biases in VQA [KABW21]

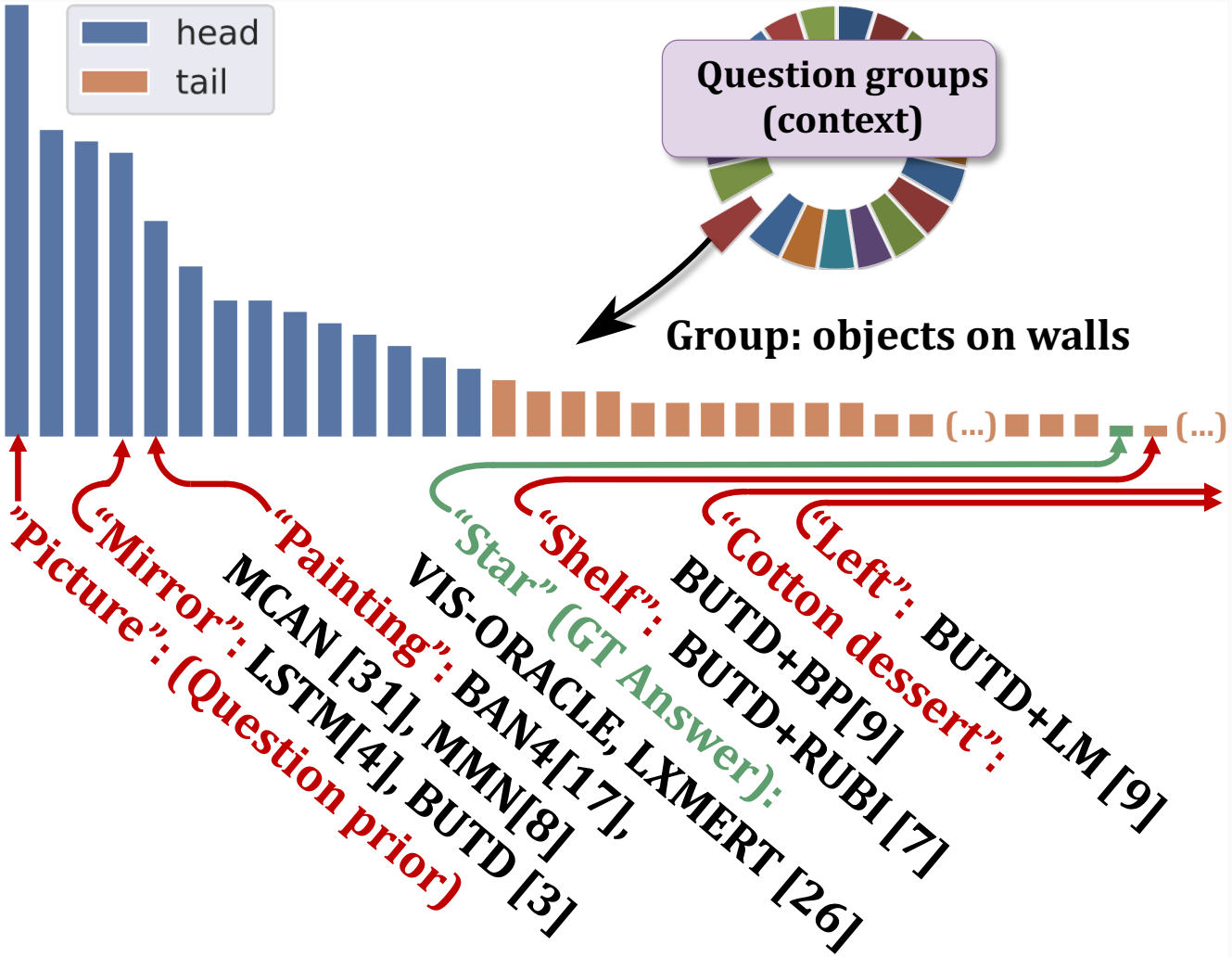
In VQA, questions and concepts are naturally unbalanced.

→ many biases



GQA-OOD: Measuring biases in VQA [KABW21]

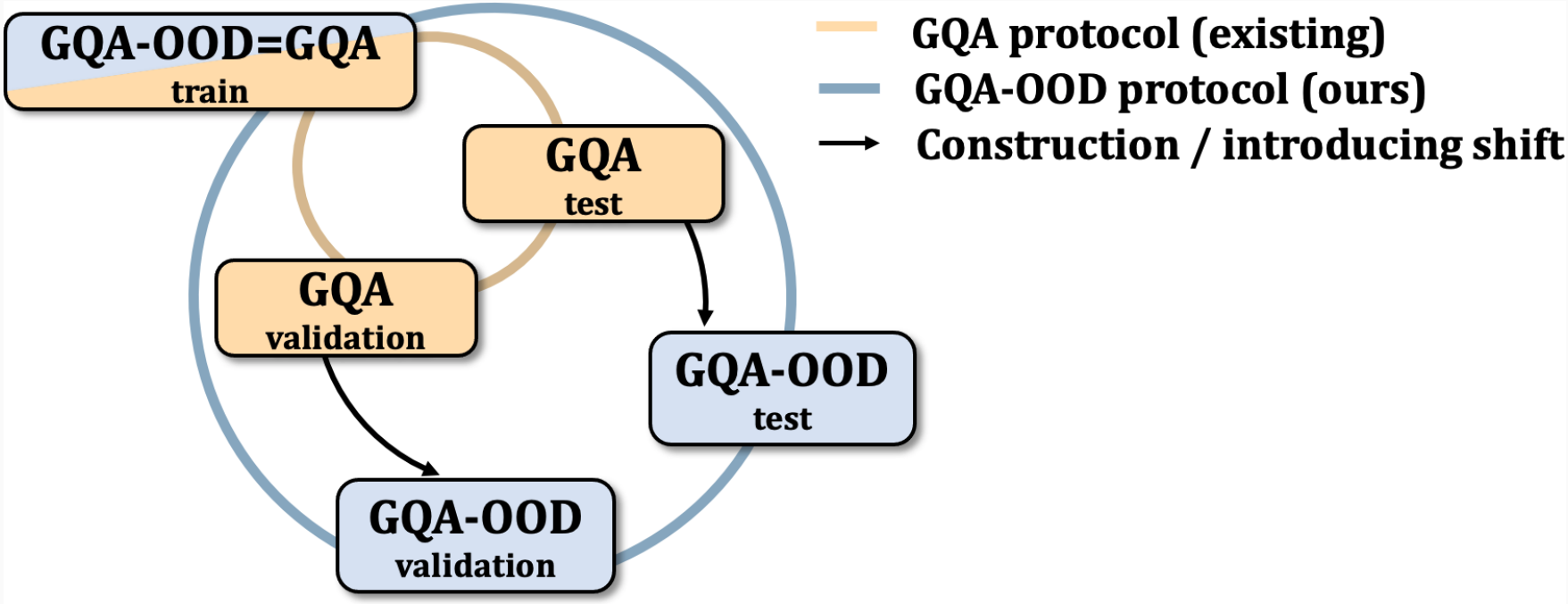
"What is on the wall?"



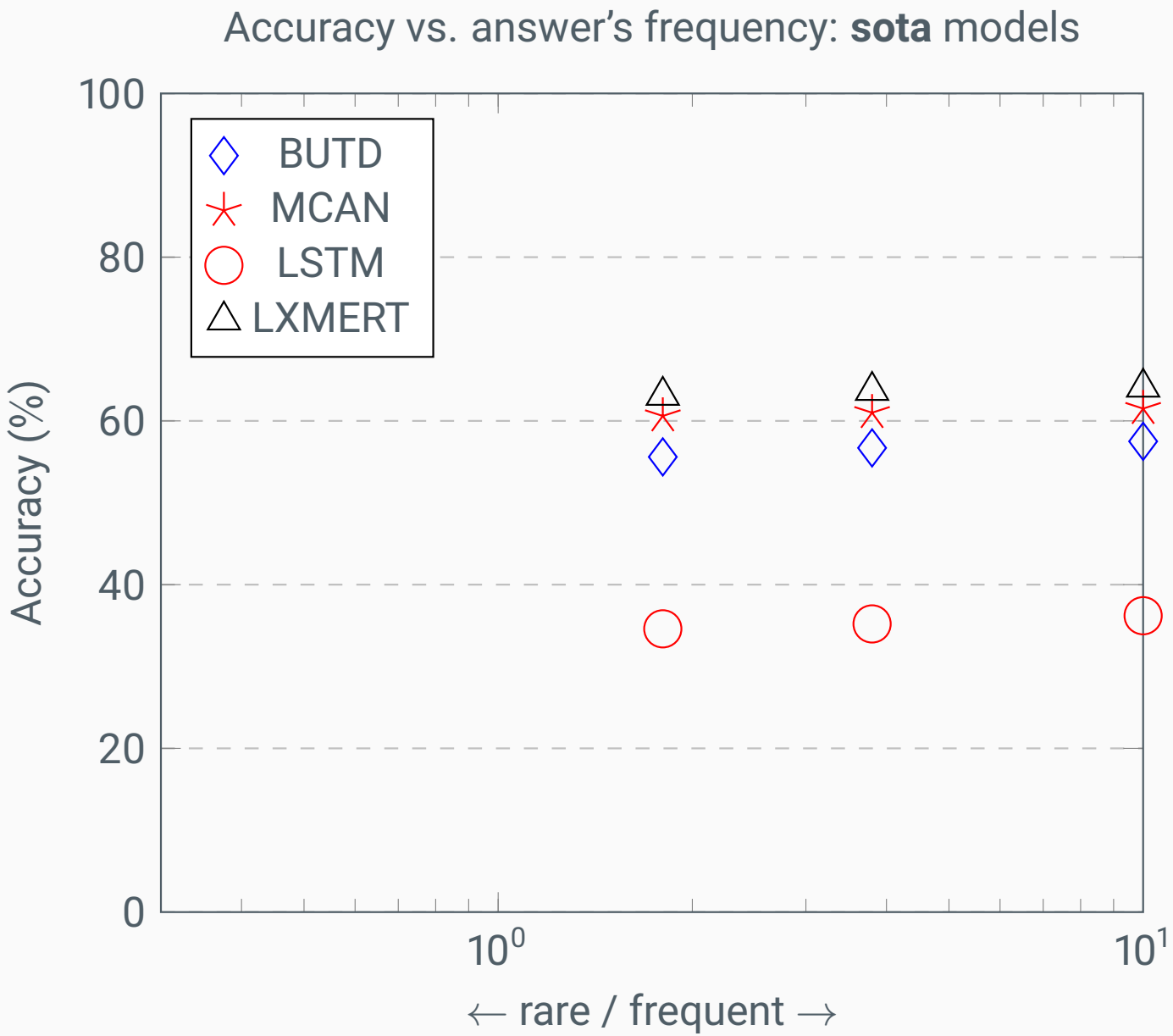
GQA-OOD: Measuring biases in VQA [KABW21]

GQA-OOD (Out-Of-Distribution)

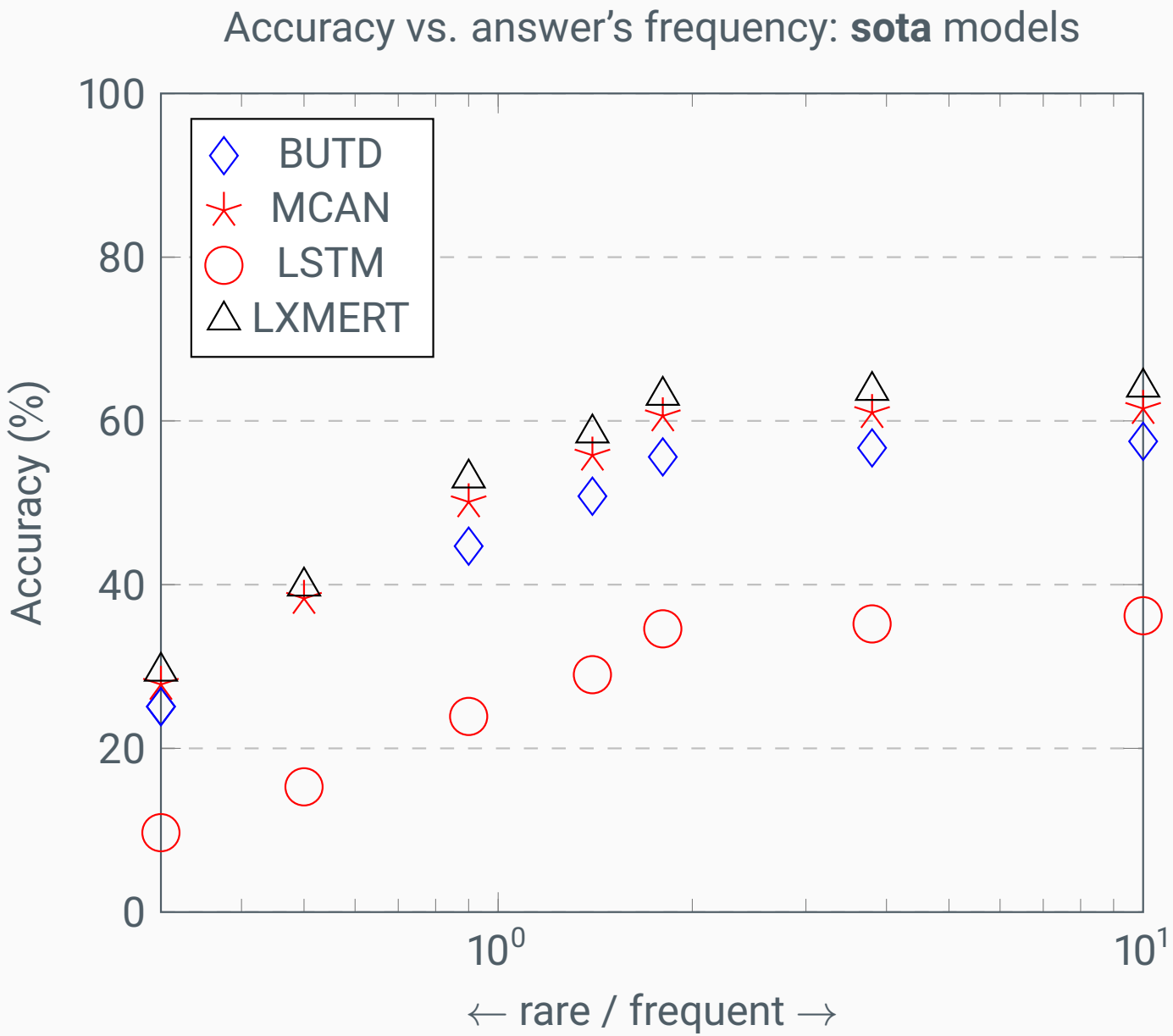
We measure and compare accuracy over both rare and frequent question-answer pairs



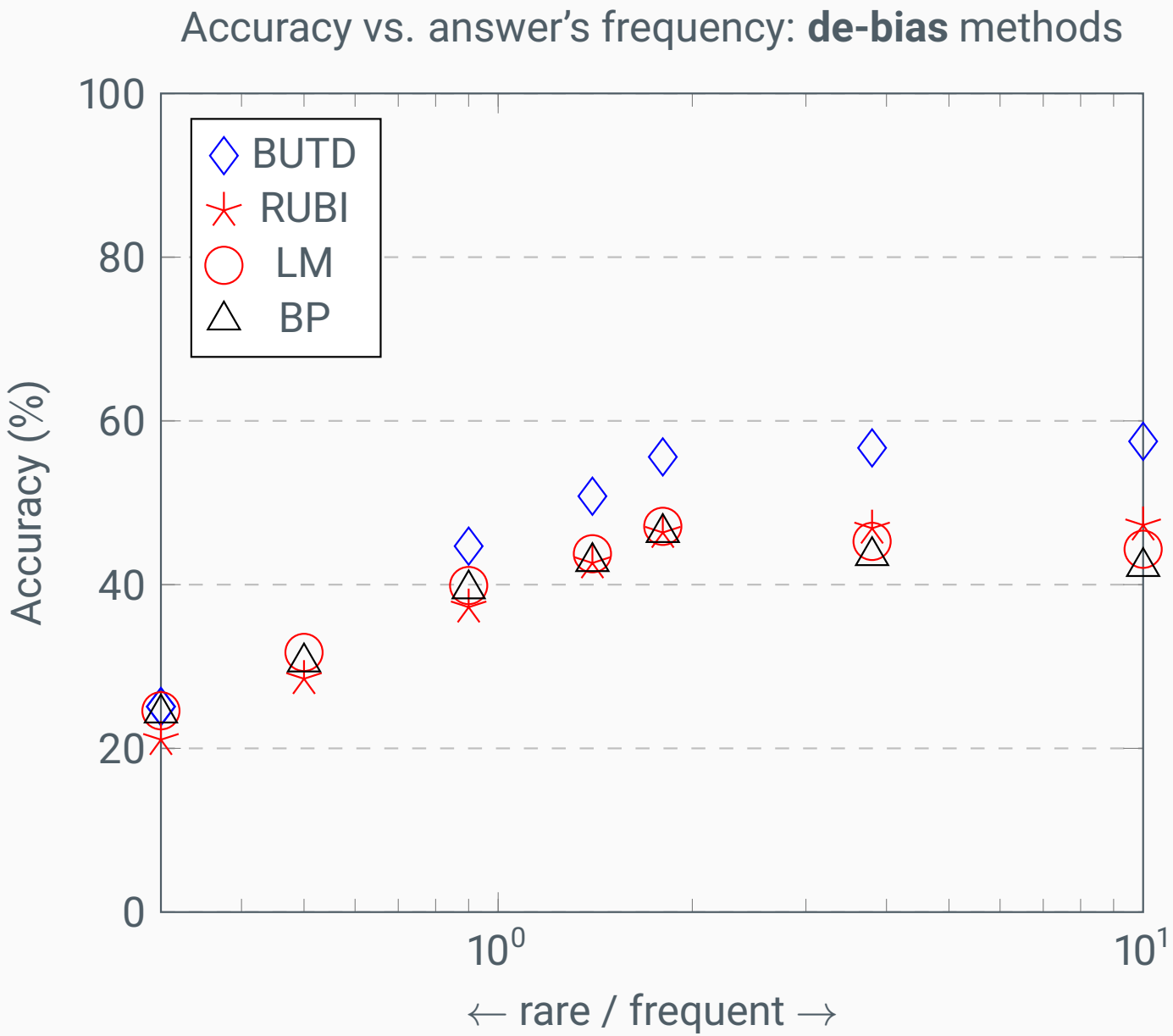
GQA-OOD: Measuring biases in VQA [KABW21]



GQA-OOD: Measuring biases in VQA [KABW21]



GQA-OOD: Measuring biases in VQA [KABW21]



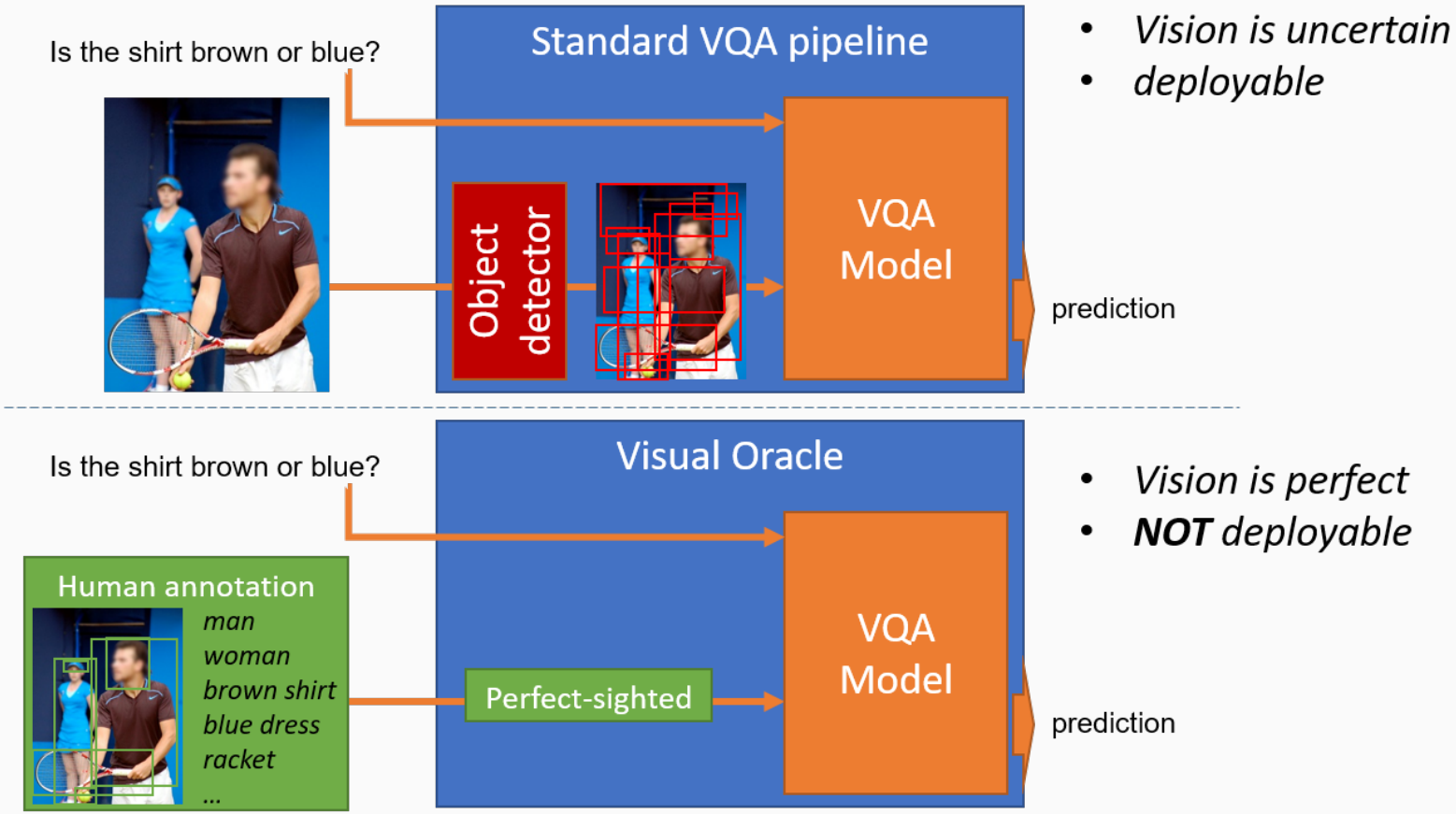
How Transferable are Reasoning Patterns in VQA?

Corentin Kervadec, Théo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot and Christian Wolf. **CVPR'21**

Reasoning Patterns [KJA+21]

Hypothesis

Shortcut learning is in part caused by the visual uncertainty



Reasoning Patterns [KJA+21]

Visual oracle is less prone to learn shortcuts:

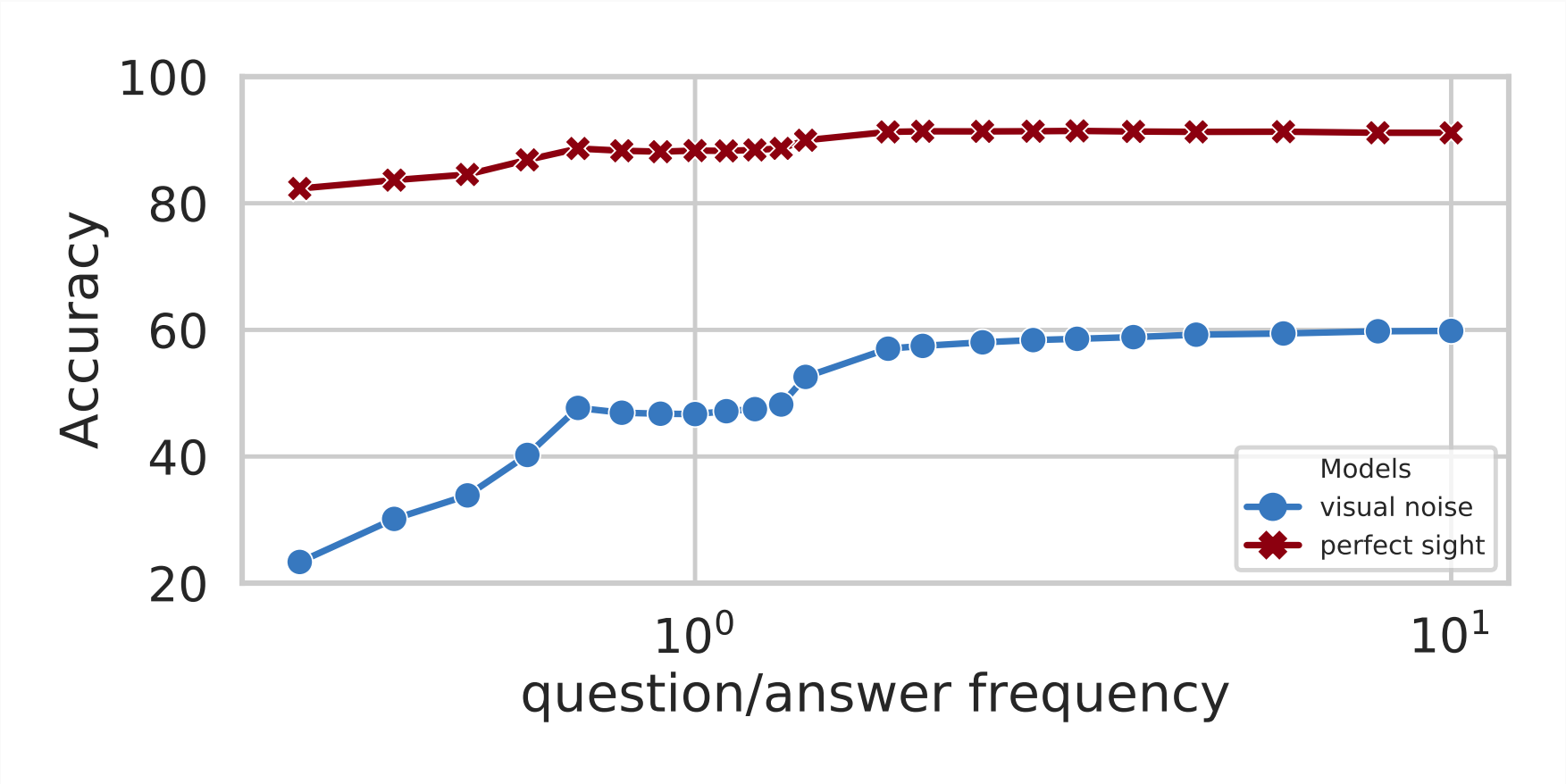
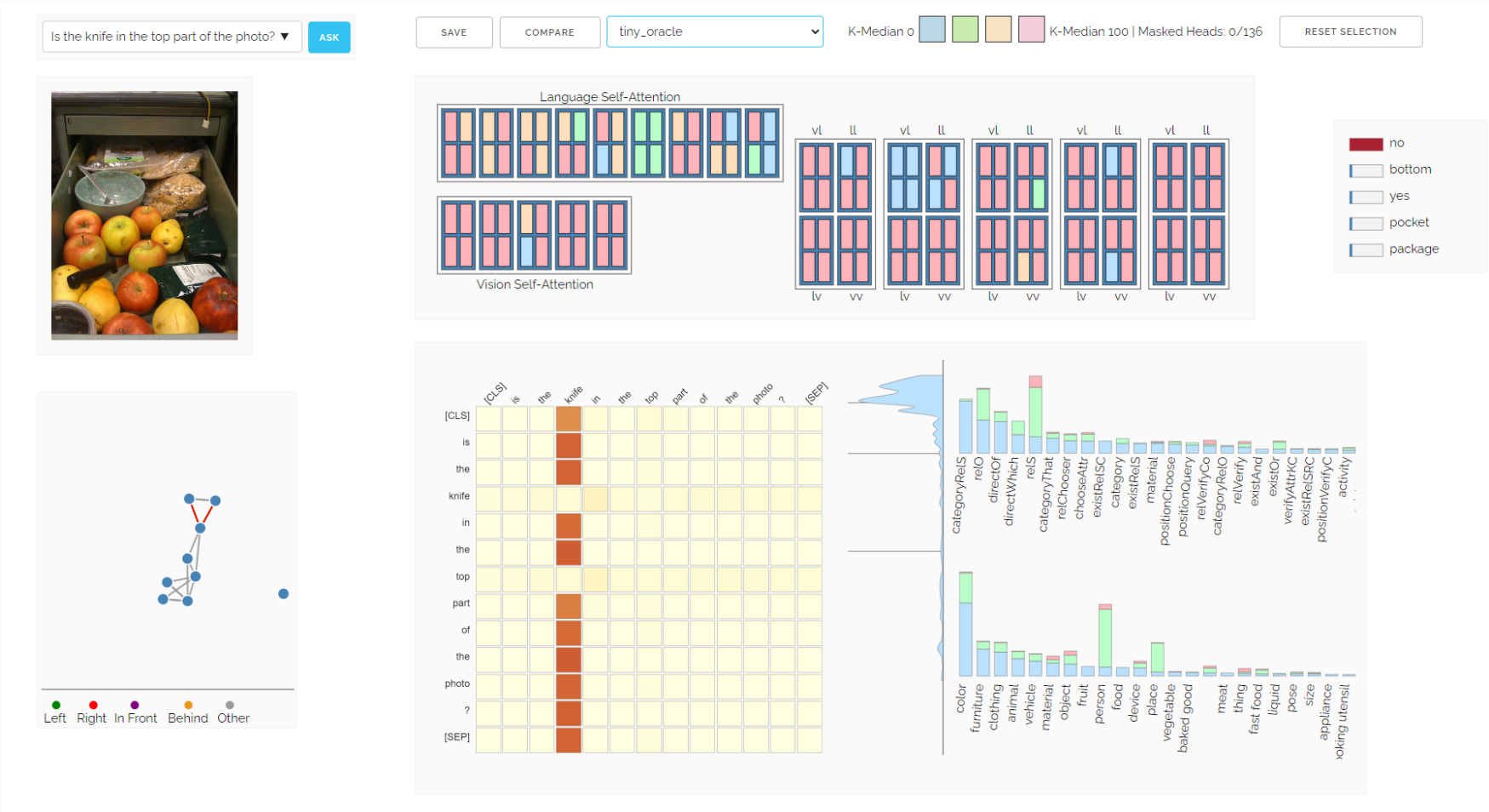


Figure: Comparison of the out-of-distribution generalization: a perfectly-sighted oracle model vs. a standard noisy vision based model (GQA-OOD benchmark [KABW21]).

Reasoning Patterns [KJA+21]

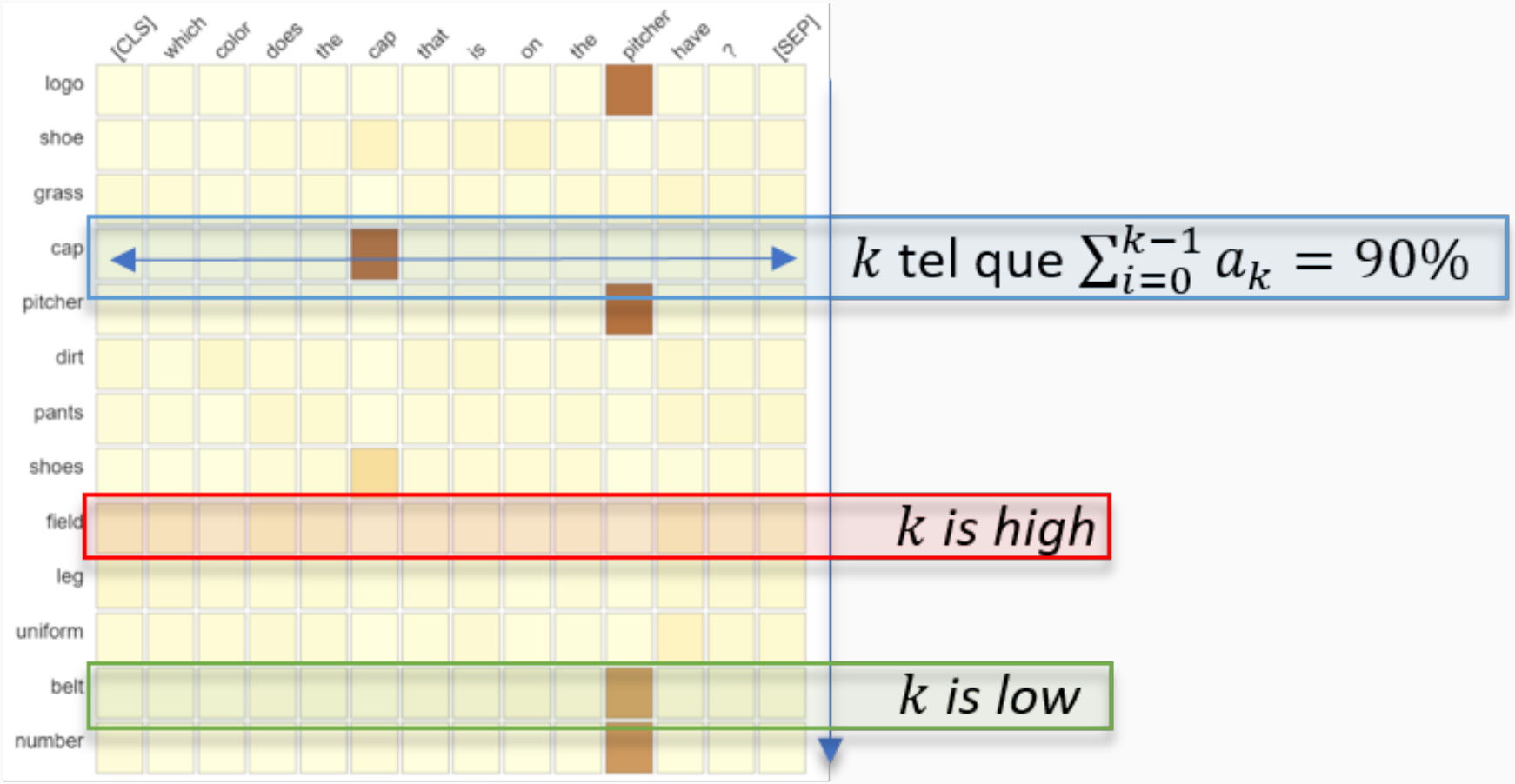
Interactive tool

`https://visqa.liris.cnrs.fr/`

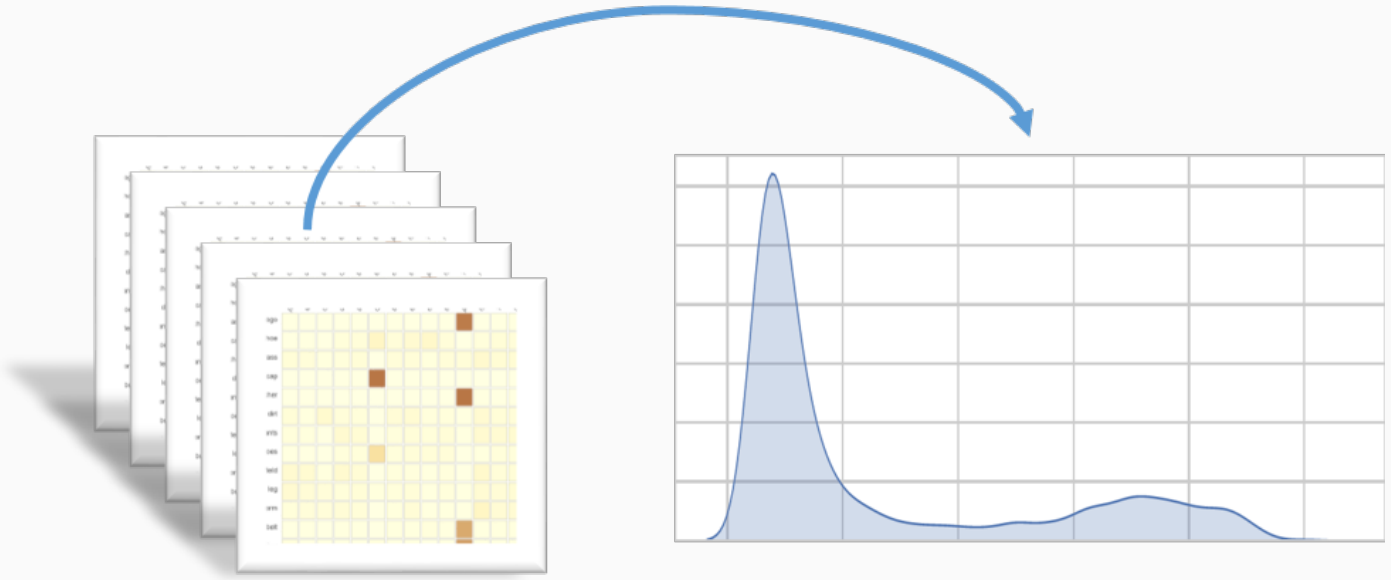


Reasoning Patterns [KJA+21]

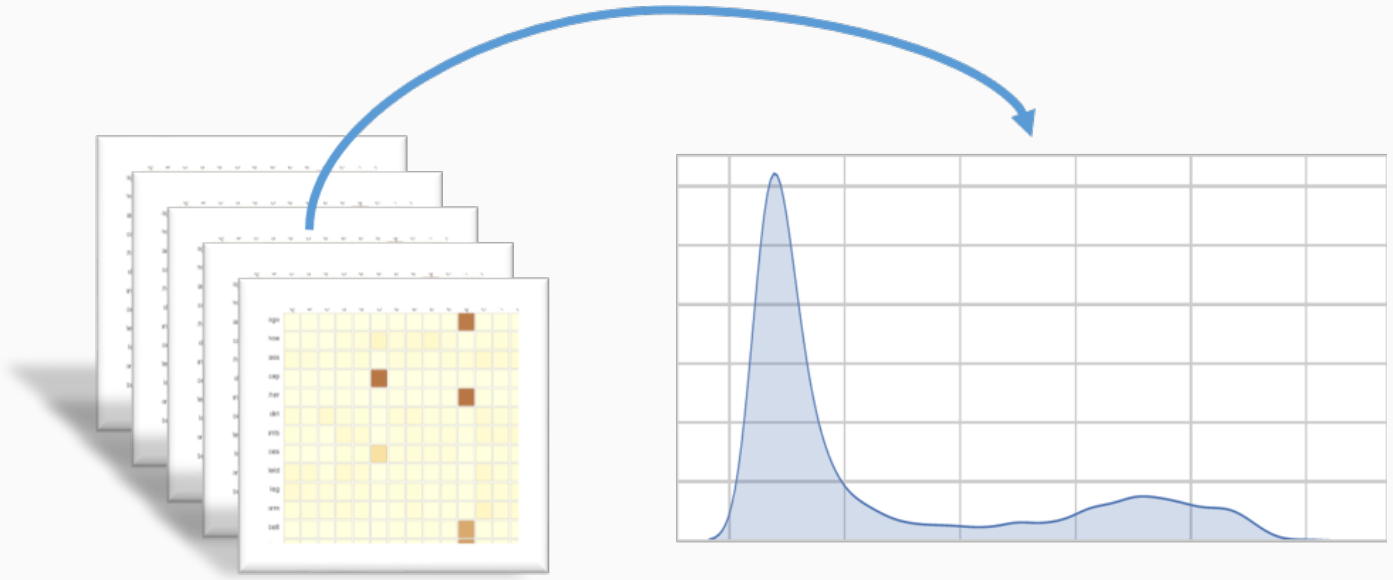
Measuring attention modes

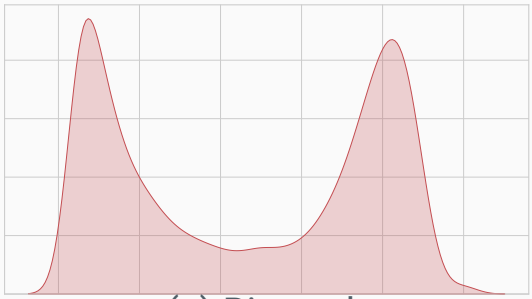


Reasoning Patterns [KJA+21]

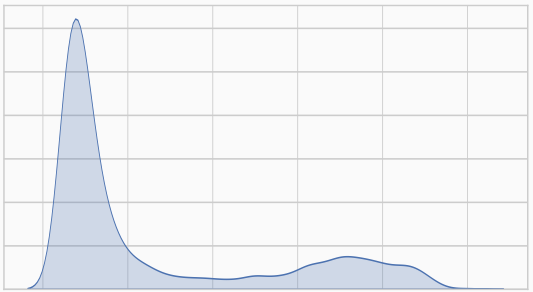


Reasoning Patterns [KJA+21]

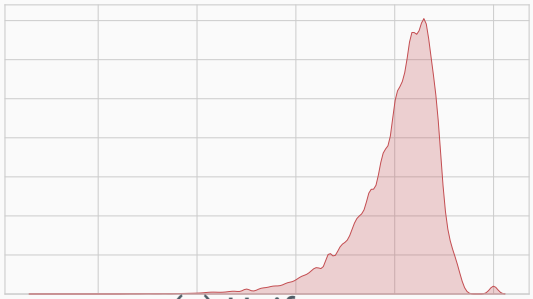




(a) Bimorph



(b) Dirac



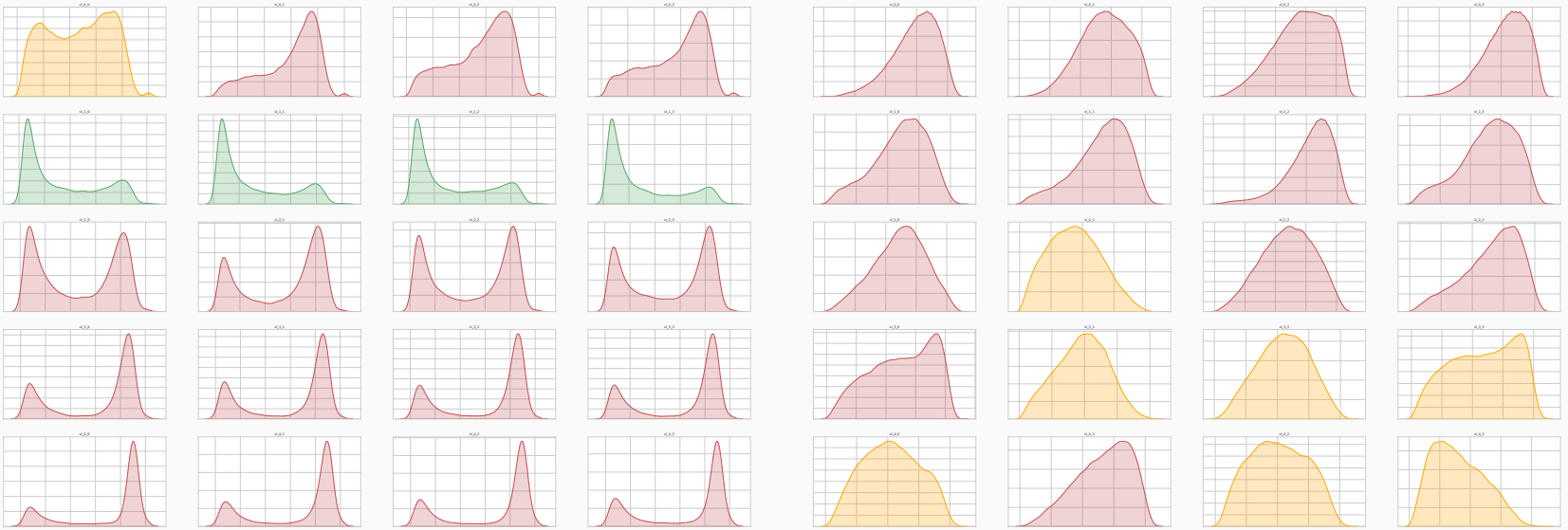
(c) Uniform

Figure: Attention modes learned by the oracle model.

Reasoning Patterns [KJA+21]

Attention modes: oracle vs. standard VQA

Higher *diversity* in visual oracle



(a) Oracle (b) Standard VQA

Figure: Attention modes of $t_{\times}^{L \leftarrow V}$ attention heads for oracle and standard VQA models. Rows indicates different $T_{\times}^{L \leftarrow V}$ layers.

Reasoning Patterns [KJA+21]

Attention modes vs. task functions

ex: `filter size`, `choose color`, `query name`, `relate`, `verify material`, *etc*...

Reasoning Patterns [KJA+21]

Attention modes vs. task functions

ex: `filter size, choose color, query name, relate, verify material`, *etc*...

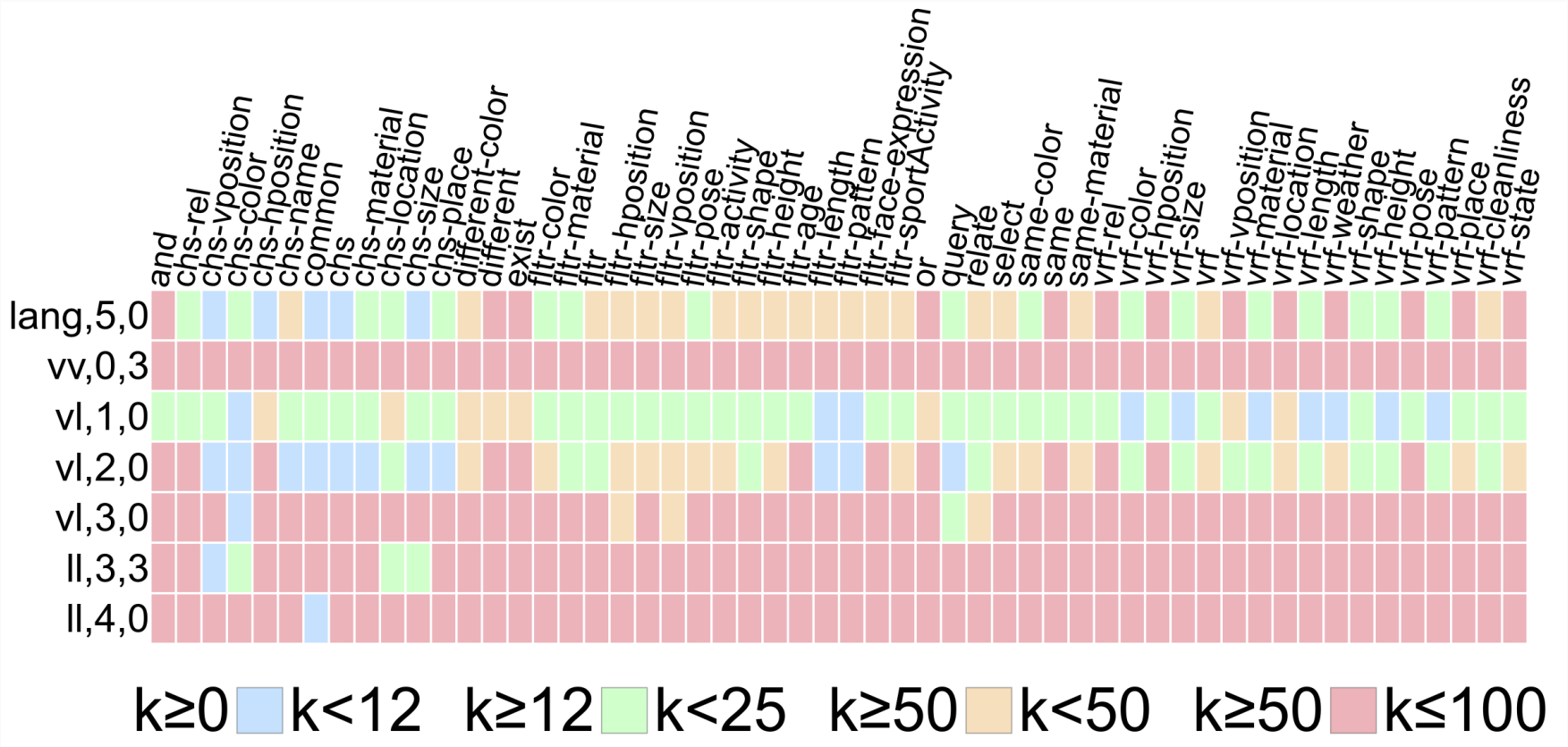


Figure: Oracle: attention modes for selected attention heads (rows) related to functions required to be solved to answer a question (columns).

Reasoning Patterns [KJA+21]

Cherry picked example: `choose color`

Impact of the function `choose color` on oracle attention modes.

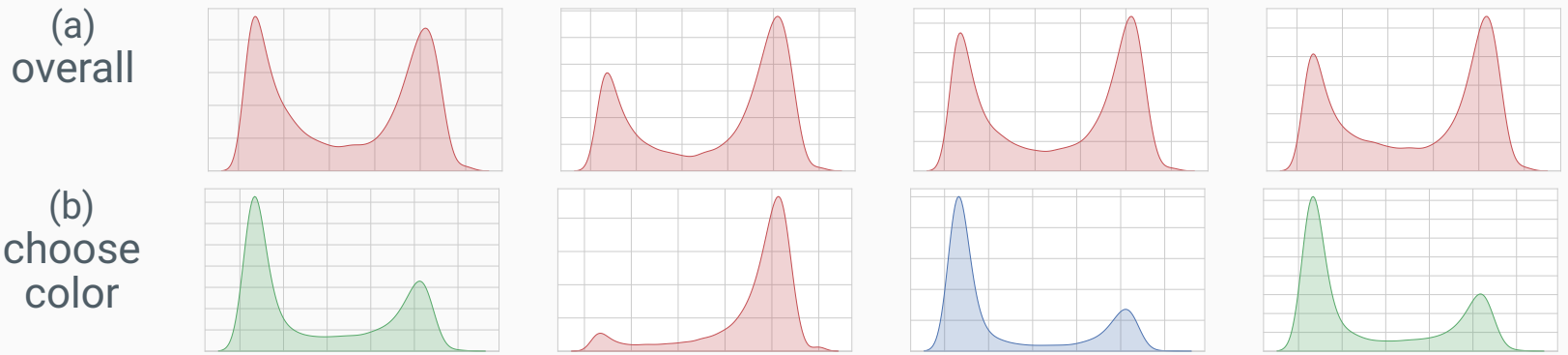
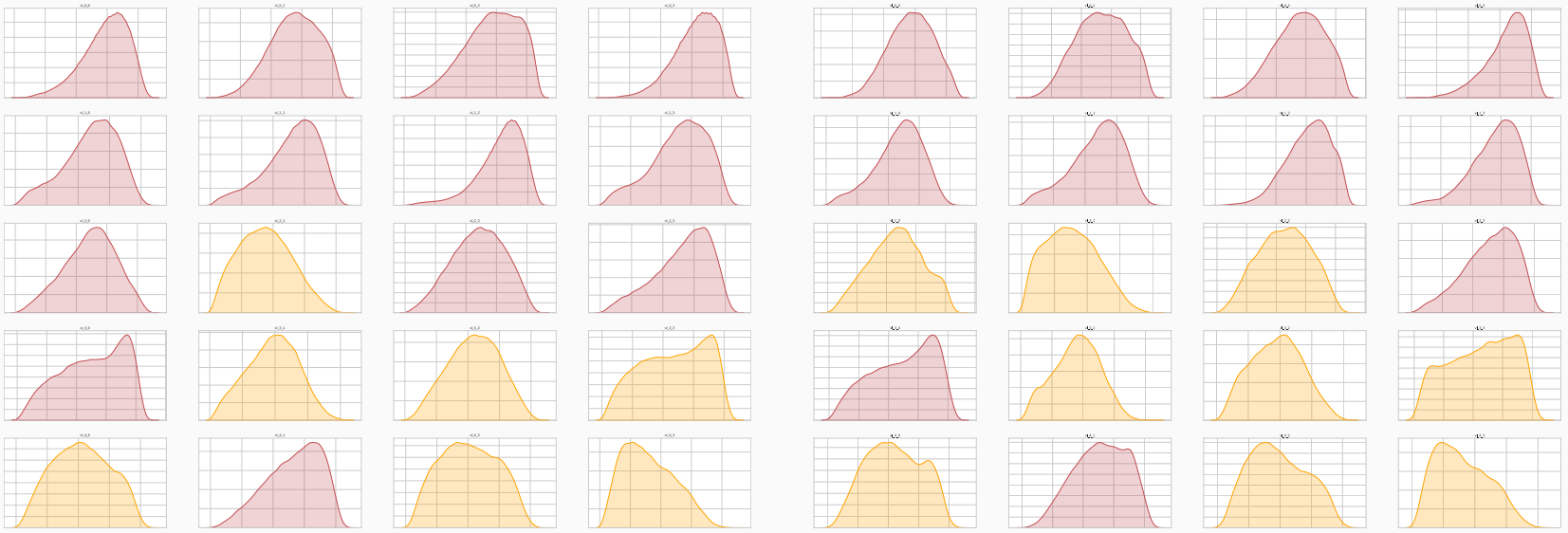


Figure: Visual oracle model

Reasoning Patterns [KJA+21]

Cherry picked example: `choose color`

Standard VQA attention modes remain the same



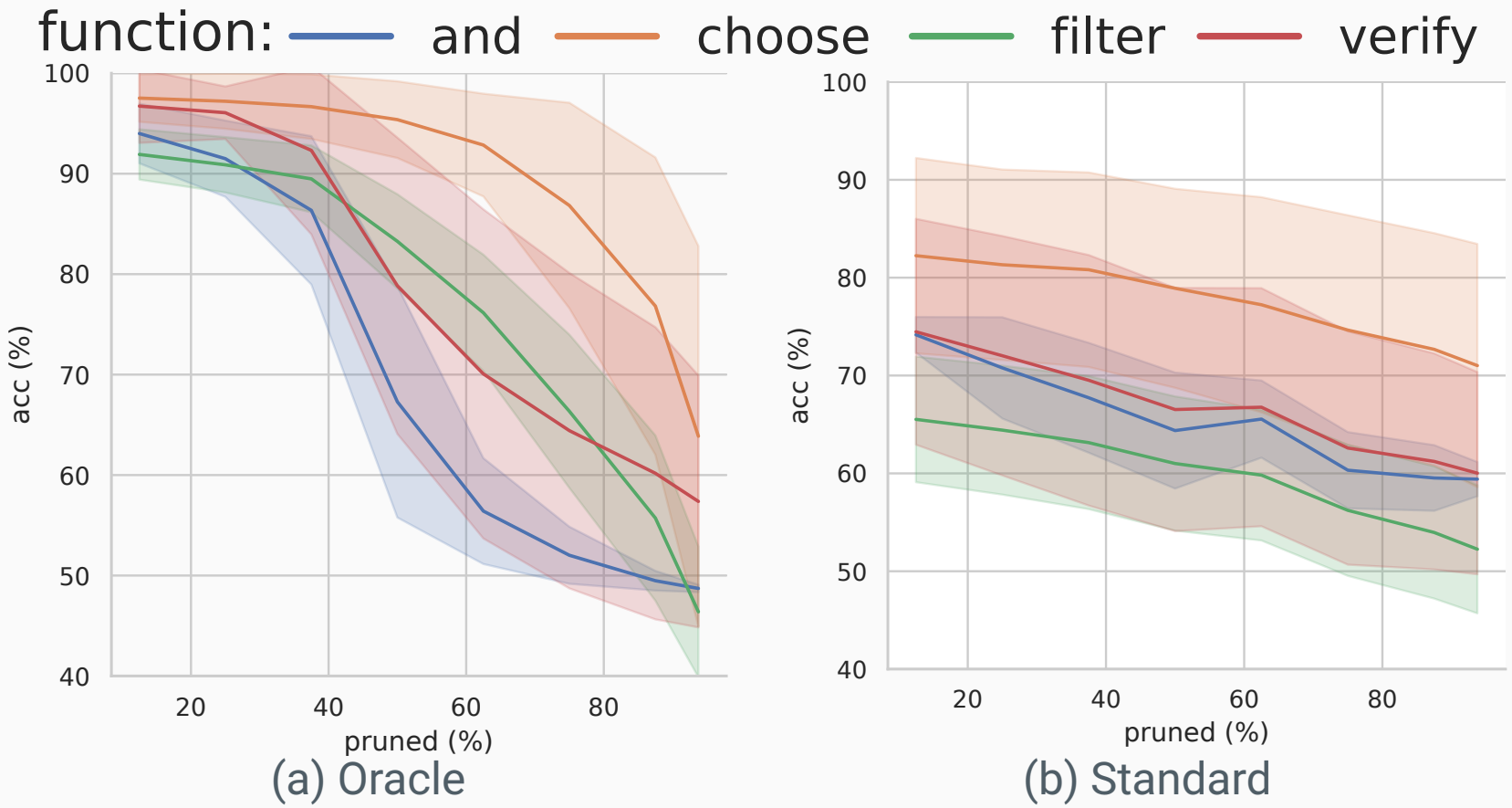
(a) overall (b) choose color

Figure: Standard VQA model

Reasoning Patterns [KJA+21]

Attention head pruning

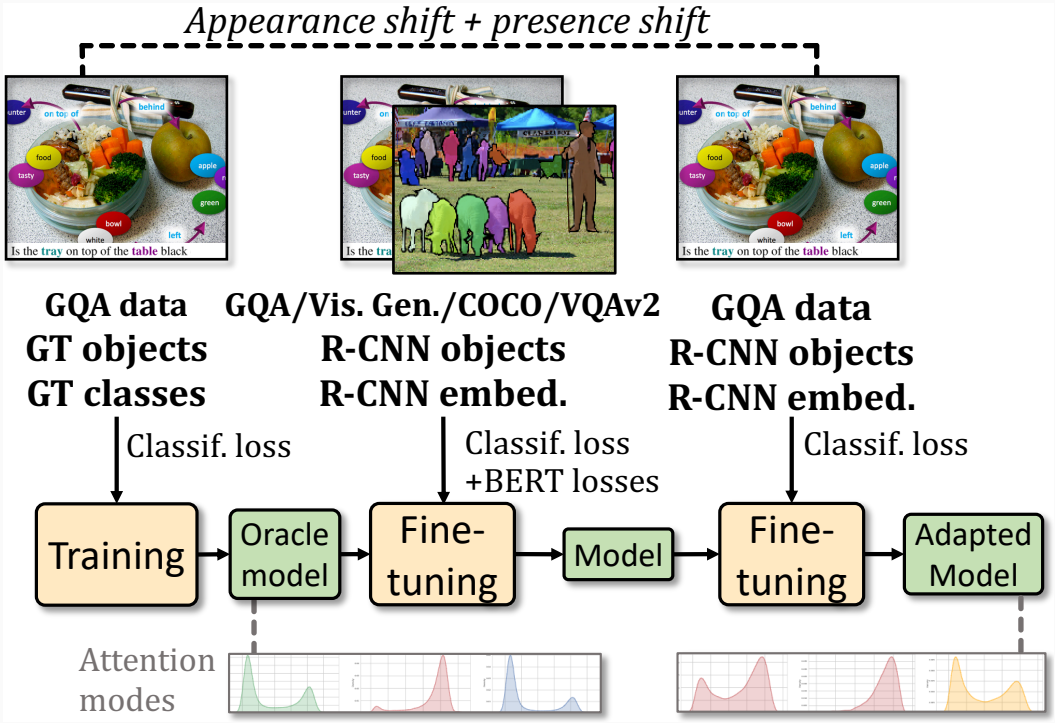
We randomly remove cross-modal attention head (replaced by average)



Reasoning Patterns [KJA+21]

Oracle transfer

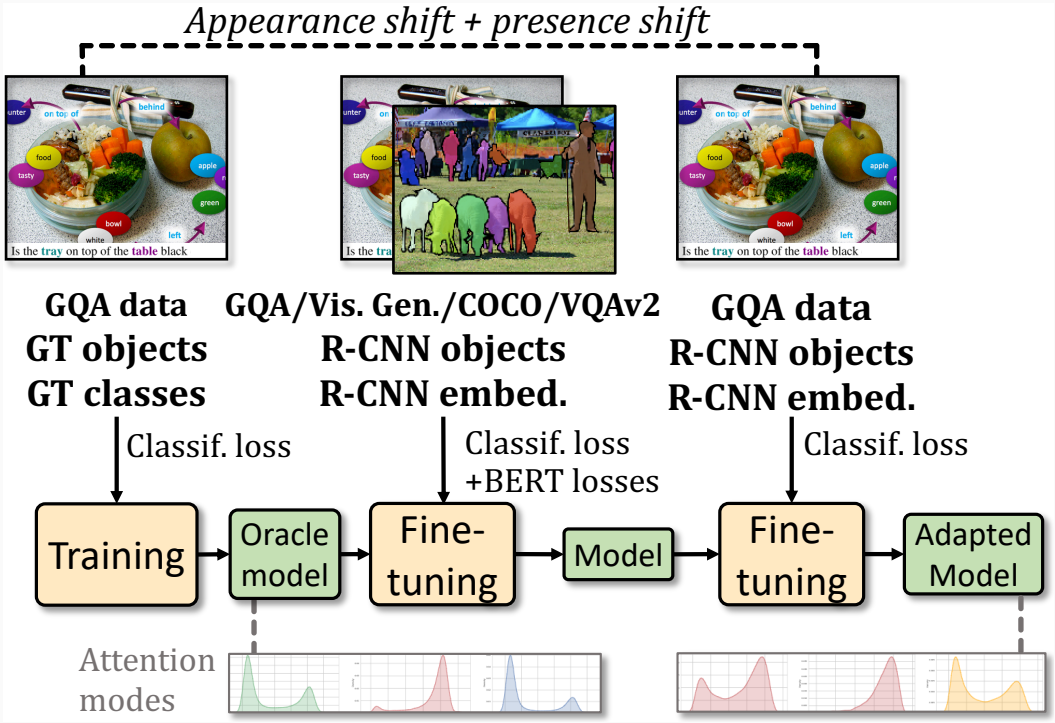
Using oracle and standard data:
(1) train the visual oracle;
(2) optionally, BERT-like pretraining;
(3) finetune on target dataset.



Reasoning Patterns [KJA+21]

Oracle transfer

Using oracle and standard data:
(1) train the visual oracle;
(2) optionally, BERT-like pretraining;
(3) finetune on target dataset.



Model	Pretraining Oracle	Pretraining BERT	GQA-OOD acc-tail	GQA-OOD acc-head	GQA overall	VQAv2 overall
(a) Baseline			42.9	49.5	52.4	-
(b) Ours	✓		**48.5**	**55.5**	**56.8**	-
(c) Baseline (+BERT)		✓	47.5	54.7	56.8	69.7
(d) Ours (+BERT)	✓	✓	**48.3**	**55.2**	**57.8**	**70.2**

Conclusion & Discussion

Contributions

* *GQA-OOD*: a benchmark for better evaluating biases in VQA
* A deep analysis of several aspects of VQA models linked to reasoning
* An *oracle transfer* method to reduce biases

Limitations

* Limited to the (partially) synthetic GQA [HM19] dataset
* The *oracle transfer* could be more efficient

Future work

* Extending OOD analysis to more natural settings
* Improving the *oracle transfer* with program prediction

Thanks

Thanks!
Any questions?

Roses are Red, Violets are Blue... But Should VQA expect Them To?
C. Kervadec, G. Antipov, M. Baccouche, C. Wolf @ *CVPR2021*

How Transferable are Reasoning Patterns in VQA?
C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, C. Wolf @ *CVPR2021*

More at `https://corentinkervadec.github.io/`
Twitter: `https://twitter.com/CorentK`

Bibliography I

[AAL+15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh.
Vqa: Visual question answering.
In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[ABPK18] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi.
Don't just assume; look and answer: Overcoming priors for visual question answering.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[AHB+18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang.
Bottom-up and top-down attention for image captioning and visual question answering.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[Bot14] Léon Bottou.
From machine learning to machine reasoning.
Machine learning, 94(2):133–149, 2014.

Bibliography II

[CDC+19] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al.
Rubi: Reducing unimodal biases for visual question answering.
In *Advances in Neural Information Processing Systems*, pages 839–850, 2019.

[CGL+21] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu.
Meta module network for compositional visual reasoning.
In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021.

[CYZ19] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer.
Don't take the easy way out: Ensemble based methods for avoiding known dataset biases.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4060–4073, 2019.

[GJM+20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann.
Shortcut learning in deep neural networks.
Nature Machine Intelligence, 2(11):665–673, 2020.

Bibliography III

[GKSS+17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh.
Making the v in vqa matter: Elevating the role of image understanding in visual question answering.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[HM19] Drew A Hudson and Christopher D Manning.
Gqa: A new dataset for real-world visual reasoning and compositional question answering.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[JHvdM+17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick.
Clevr: A diagnostic dataset for compositional language and elementary visual reasoning.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Bibliography IV

[KABW21] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf.
Roses are red, violets are blue... but should vqa expect them to?
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[KJA+21] Corentin Kervadec, Theo Jaunet, Grigory Antipov, Moez Baccouche, Romain Vuillemot, and Christian Wolf.
How transferable are reasoning patterns in vqa?
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[KJZ18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang.
Bilinear attention networks.
In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[TB19] Hao Tan and Mohit Bansal.
Lxmert: Learning cross-modality encoder representations from transformers.
In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114, 2019.

Bibliography V

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[YYC+19] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian.
Deep modular co-attention networks for visual question answering.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

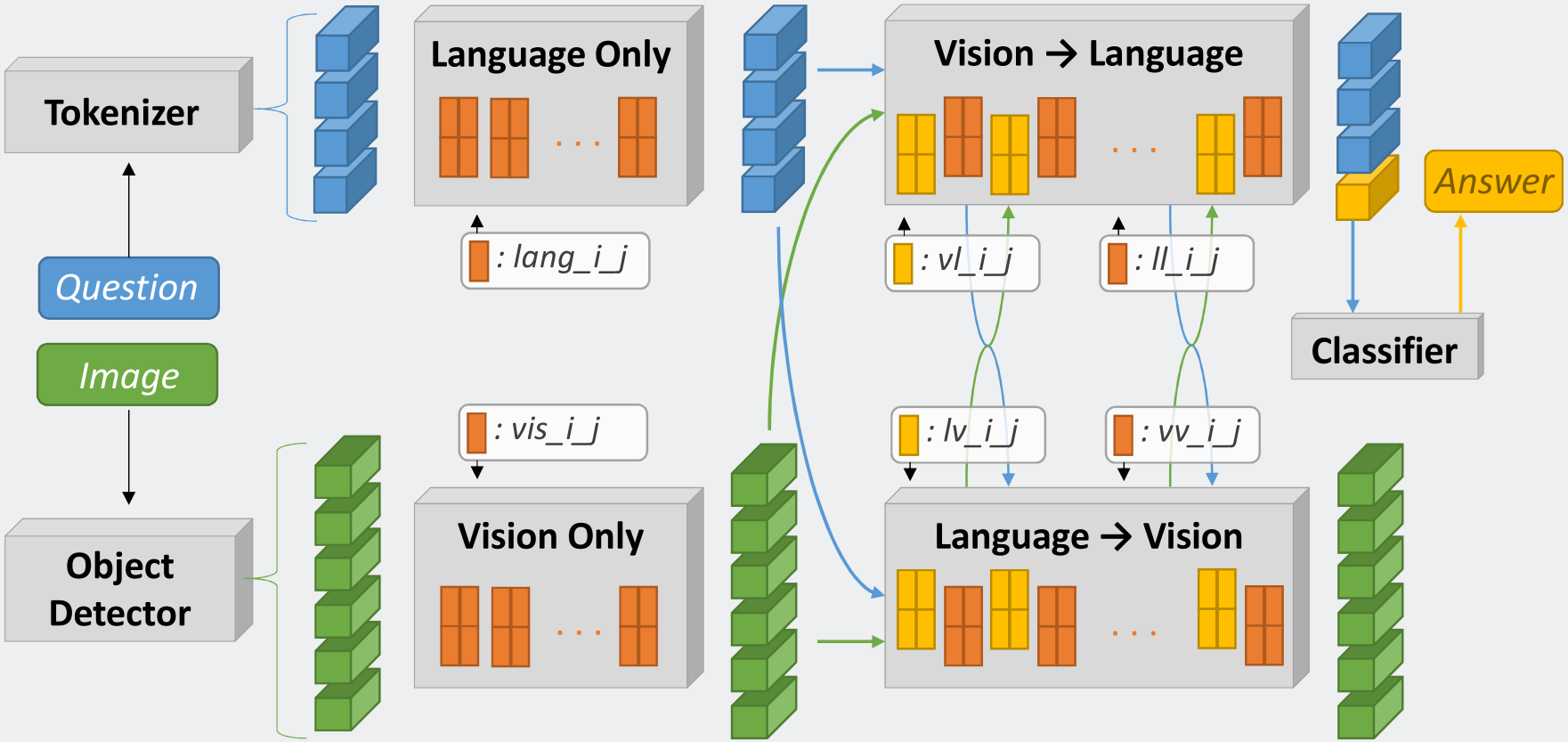
GQA-OOD: Measuring biases in VQA [KABW21]

Model	VQA2 overall	GQA overall	GQA dist.	VQA-CP2 overall	GQA-OOD acc-tail
Q. Prior	32.1	*27.0*	*55.6*	8.8	*17.8*
LSTM [AAL+15]	43.0	*39.1*	*3.6*	*22.1*	*24.0*
BUTD [AHB+18]	63.5	*51.6*±0.3	*1.8*	40.1	*42.1*±0.9
MCAN [YYC+19]	***66.1***	*56.3*±0.2	*1.6*	***42.5***	*46.5*±0.5
BAN4 [KJZ18]	65.9	*54.7*±0.4	*1.6*	40.7	*47.2*±0.5
MMN [CGL+21]	-	**59.6**	1.8	-	*48.0*
LXMERT [TB19]	**69.9**	***59.6***	***1.5***	-	***49.8***
BUTD [AHB+18]	**63.5**	*51.6*±0.3	*1.8*	40.1	***42.1***±0.9
+RUBi+QB	-	***51.9***±1.1	***1.7***	***47.6***±3.7	***42.1***±1.0
+RUBi [CDC+19]	61.2	*43.6*±2.0	*1.9*	44.2	*35.7*±2.3
+LM [CYZ19]	56.4	*39.7*±0.7	*2.1*	52.0	*32.2*±1.2
+BP [CYZ19]	63.2	*39.6*±0.3	*2.2*	39.9	*30.8*±1.0

Table: We compare the proposed *acc-tail* metric with other benchmarks. Results computed on the testdev split of *GQA-OOD* and GQA [HM19], the test split of VQA-CP2 [ABPK18] and the VQA2 [GKSS+17] validation split. Values in italic: trained and tested by ourselves.

Reasoning Patterns [KJA+21]

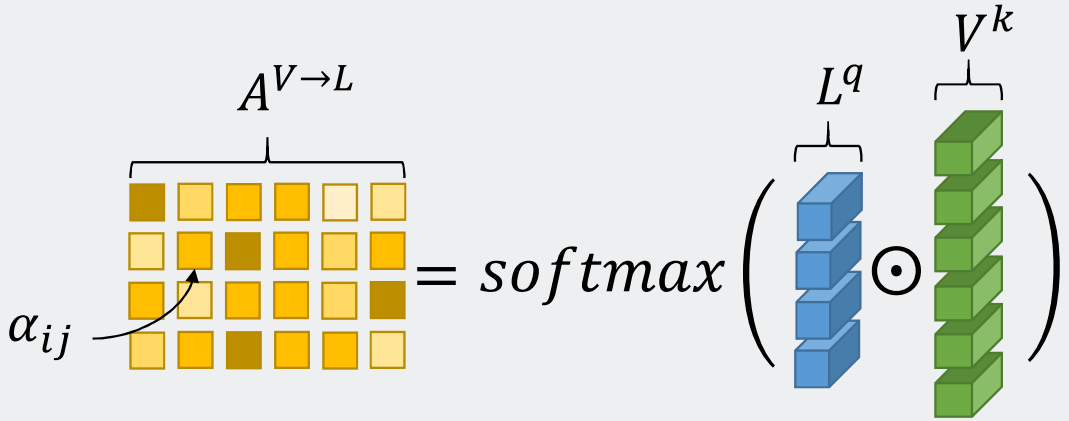
Vision Language (VL)-Transformer architecture [TB19]



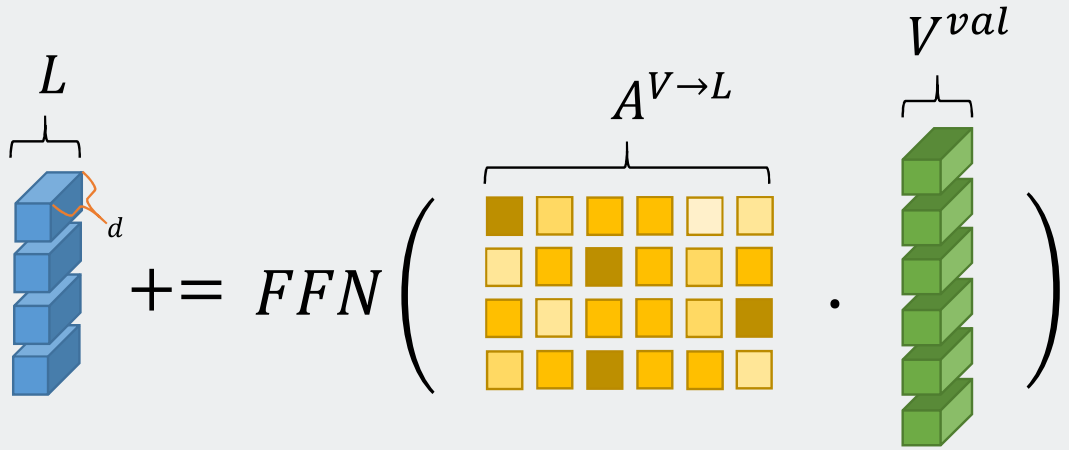
Reasoning Patterns [KJA+21]

Transformer's attention [VSP+17]

(a) Attention maps



(b) Contextualisation



GQA-OOD: Measuring biases in VQA [KABW21]

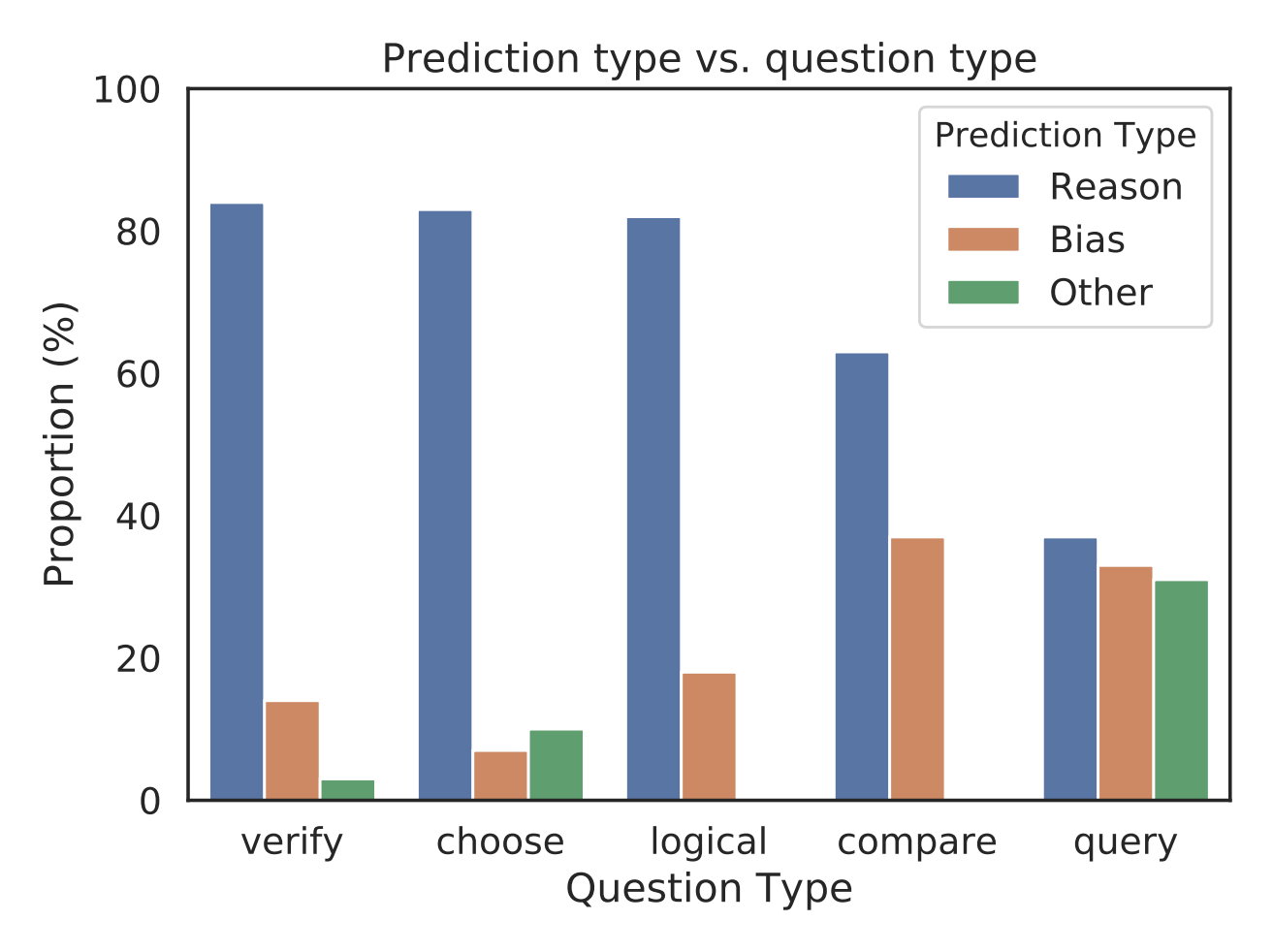


Figure: Distribution the estimated reasoning labels over the GQA [HM19] question types for the LXMERT [TB19] model.

GQA-OOD: Measuring biases in VQA [KABW21]

Multi-dimensional evaluation:

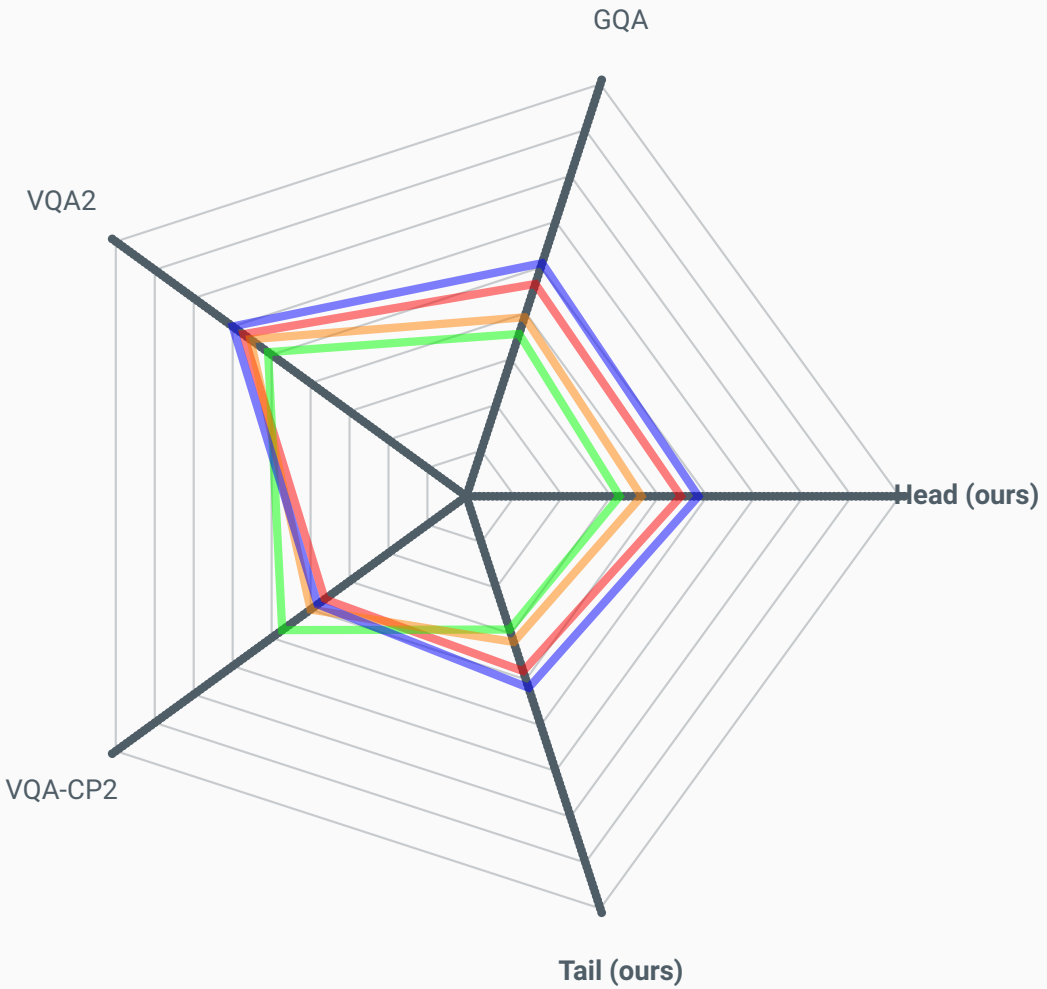


Figure: BUTD (baseline) – LM and RUBi (VQA-CP de-bias) – MCAN (sota)

Reasoning Patterns [KJA+21]

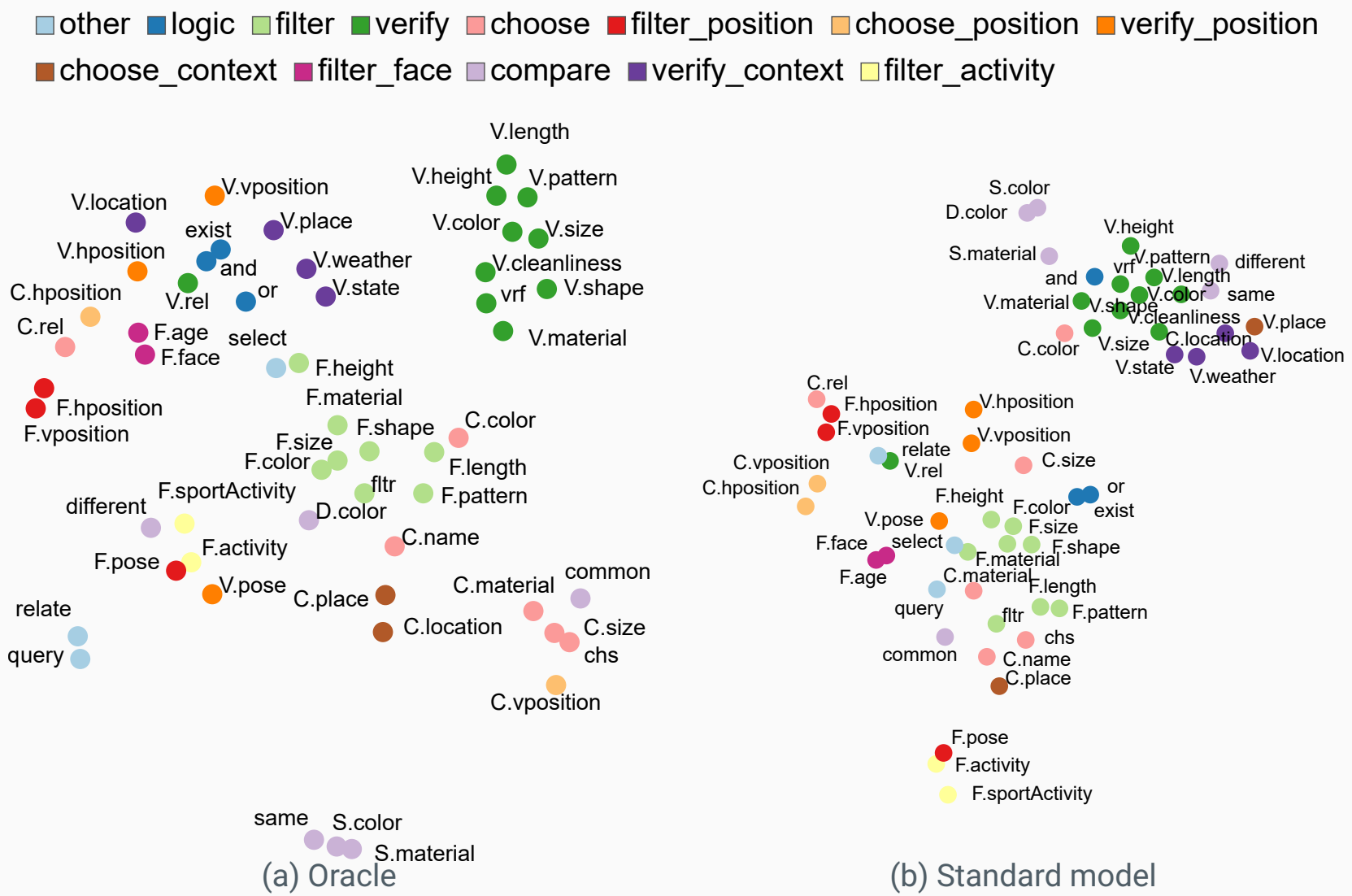
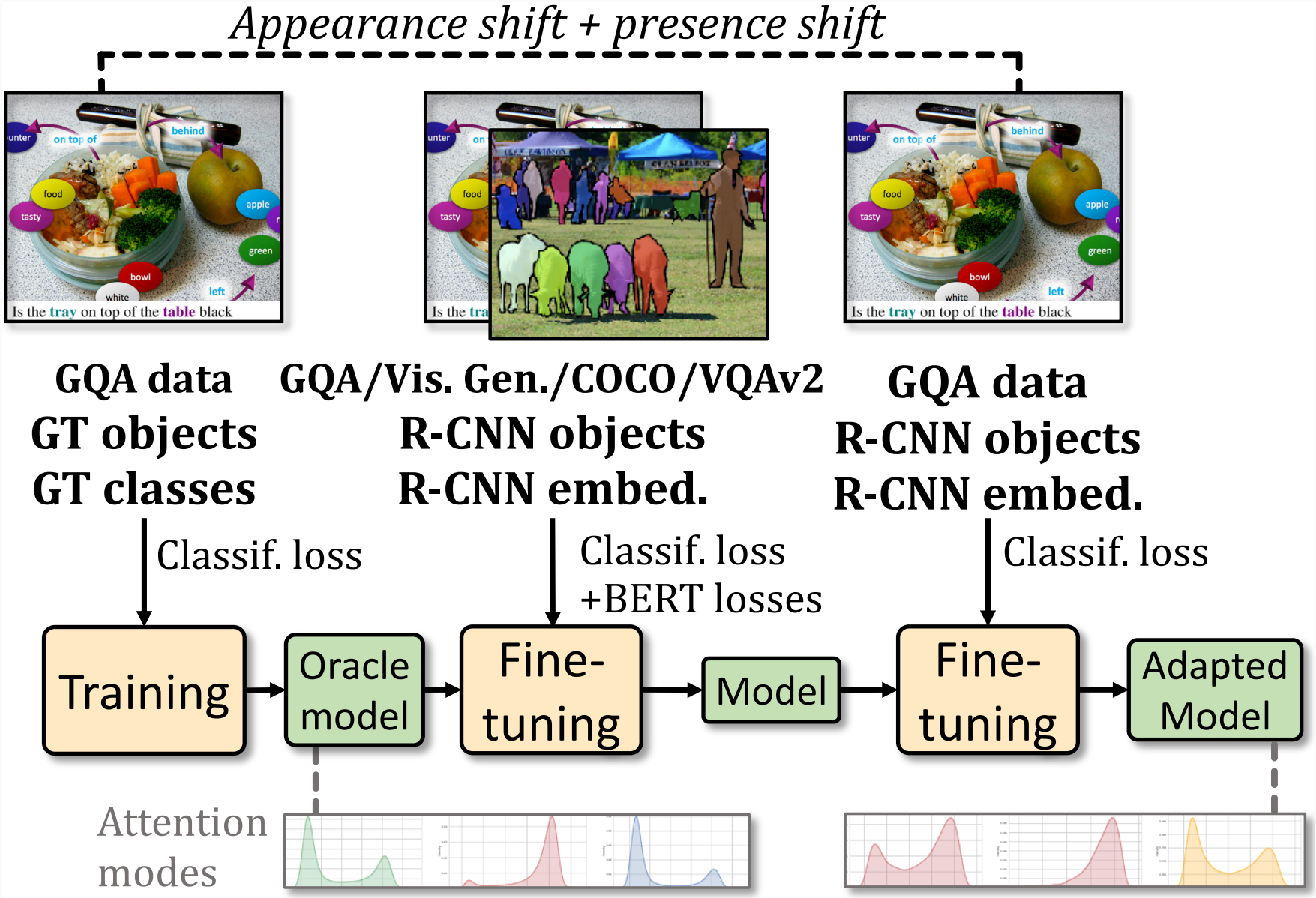


Figure: t-SNE projection of the attention mode space, *i.e.* the 80-dim representation median *k*-numbers, one per head of the model.

Reasoning Patterns [KJA+21]



GQA-OOD: Measuring biases in VQA [KABW21]

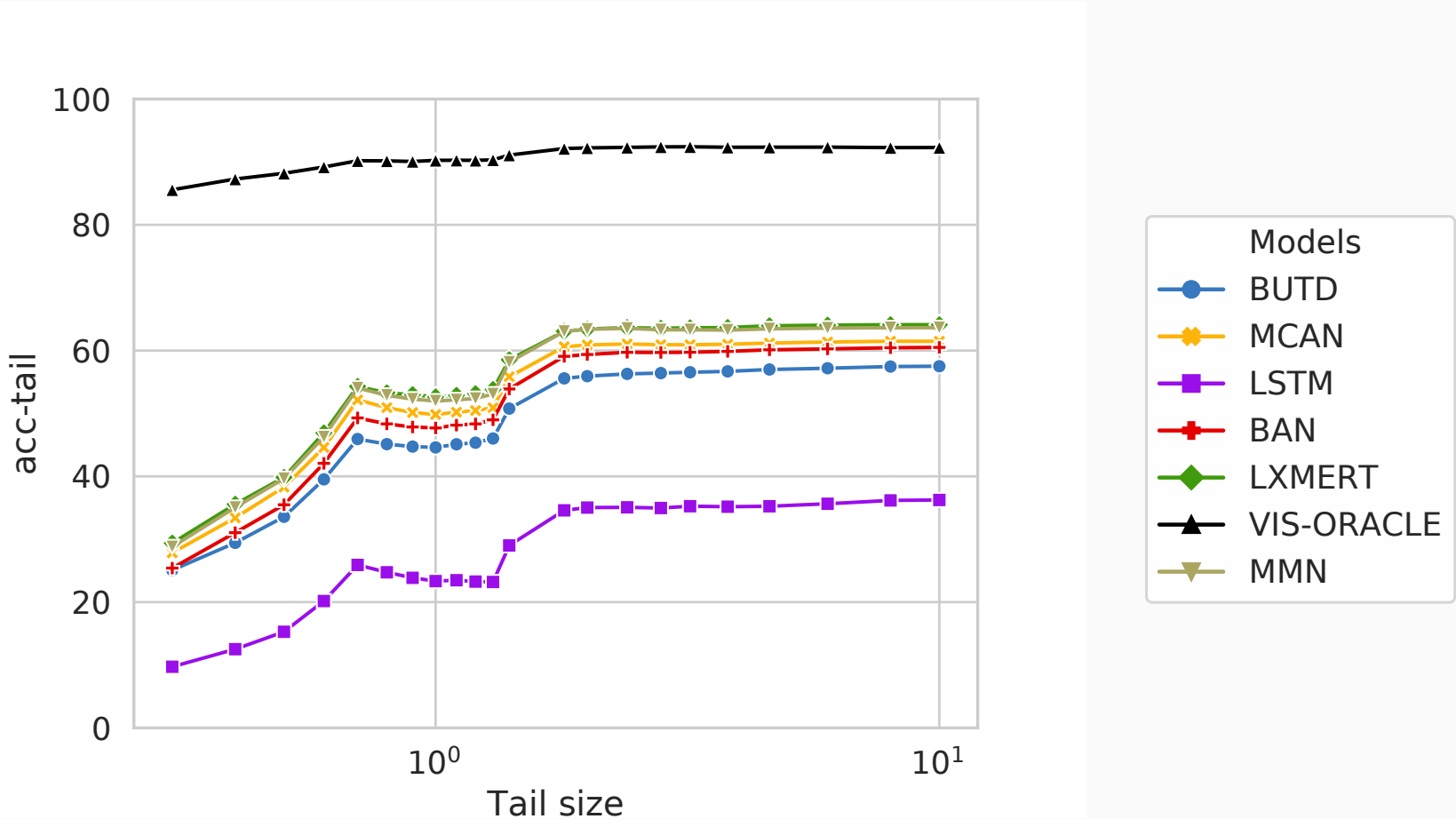
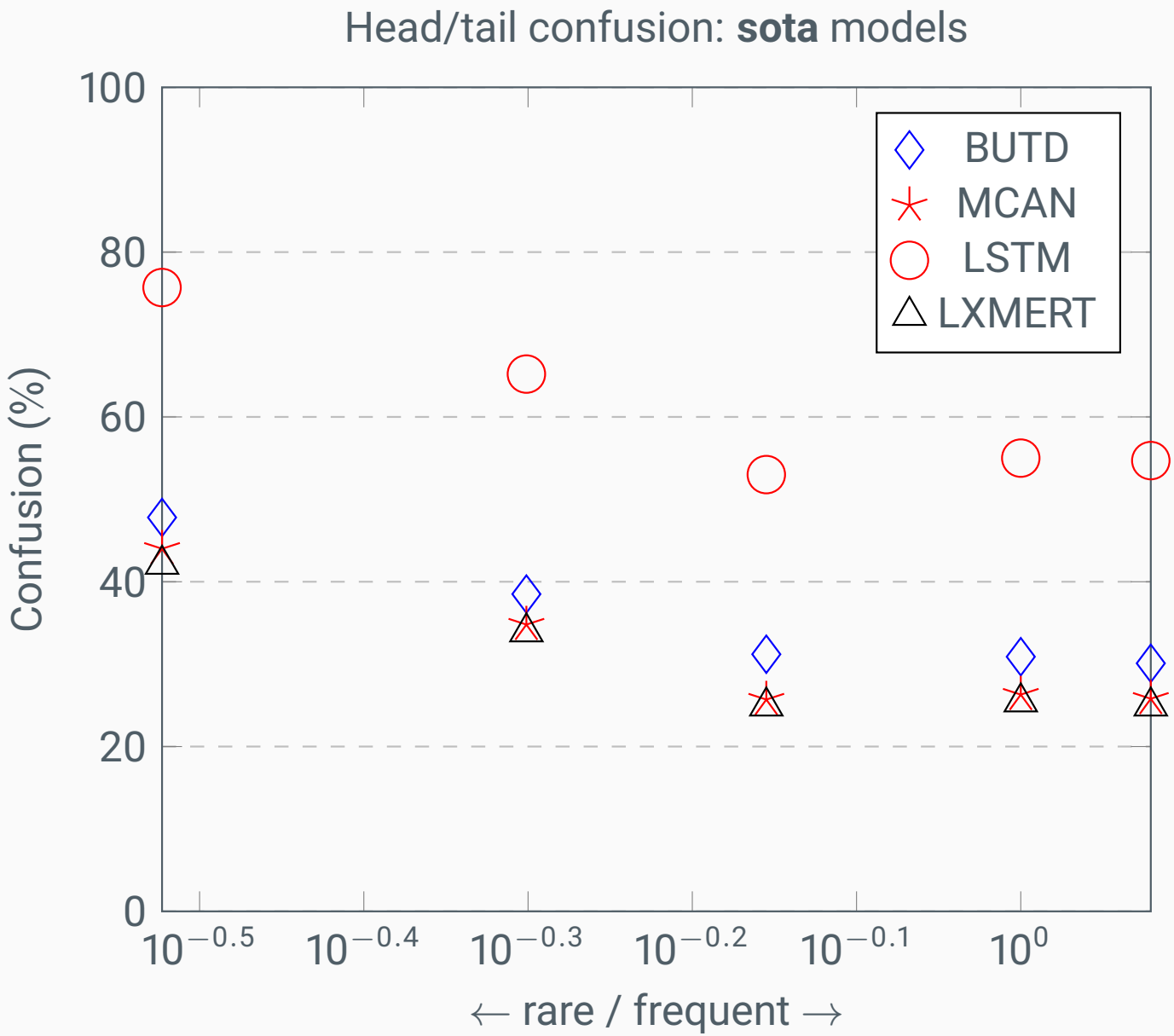


Figure: Performance (higher is better) for different definitions of the tail distribution (α parameter values) on the GQA-OOD benchmark. We compare several VQA **models**. The x-axis is in log-scale.

GQA-OOD: Measuring biases in VQA [KABW21]



GQA-OOD: Measuring biases in VQA [KABW21]