Advancing Multimodal Vision-Language Learning / Faire Progresser L'apprentissage Multimodal de la Vision et du Langage

Aishwarya Agrawal

Assistant Professor @ UdeM and Mila

Research Scientist @ DeepMind (one day a week)

Multimodal AI Research (MAIR) Lab

Oscar Mañas

Saba Ahmadi

Le Zhang

Sarvjeet Singh Ghotra

Rabiul Awal

Joining in Fall this year:

Kanishk Jain

Qian Yang

P Shravan Nayak

Vision-Language Tasks

Image Captioning

"A group of young people playing a game of Frisbee."

Visual Question Answering

Q: *"What is the mustache made of?"*

A: *"bananas"*

Vision-Language Tasks

Image Retrieval

“Grey haired man in black and yellow tie.”

Image Generation

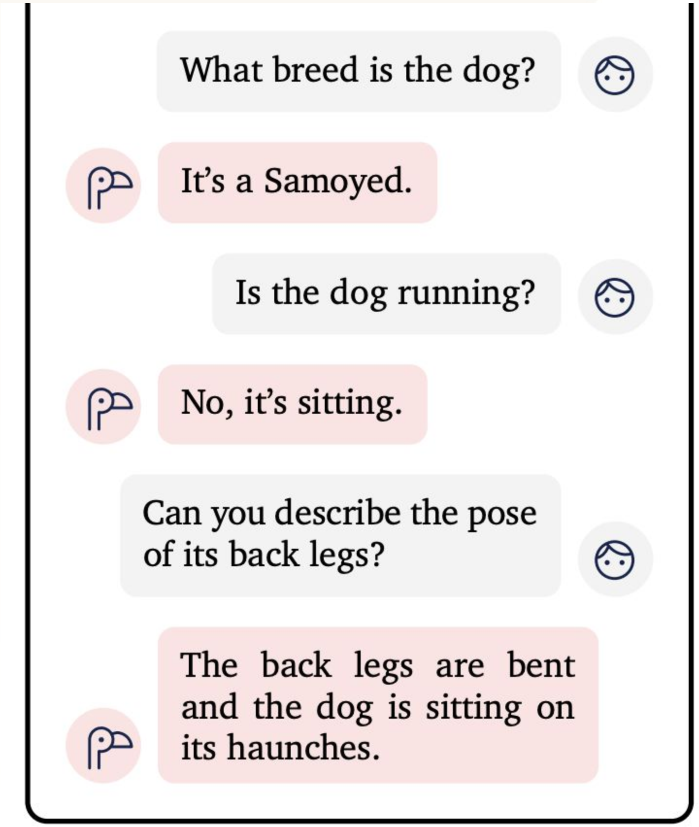
“Grey haired man in black and yellow tie.”

Why vision and language?

- Intuitive:
  - Humans learn in multimodal settings
- Applications:
  - Aid to visually impaired users
  - Online shopping and organizing photos
  - Grounded virtual assistants
- Scientific:
  - Visual recognition
  - Language understanding
  - Combining information across modalities
  - Visio-linguistic compositional reasoning
  - Commonsense and factual knowledge reasoning

Vision-Language Progress

DeepMind's Flamingo
[Link]



Vision-Language Progress

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

OpenAI's DALL.E 2

[Link]

Vision-Language Challenges

- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

DeepMind

Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization


Aishwarya Agrawal[*,$]

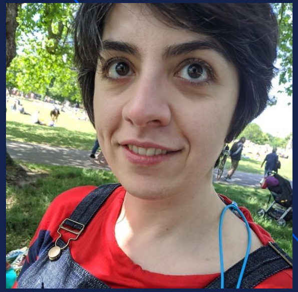

Ivana Kajić[*]


Emanuele Bugliarello[*]


Elnaz Davoodi[^]


Anita Gergely[^]

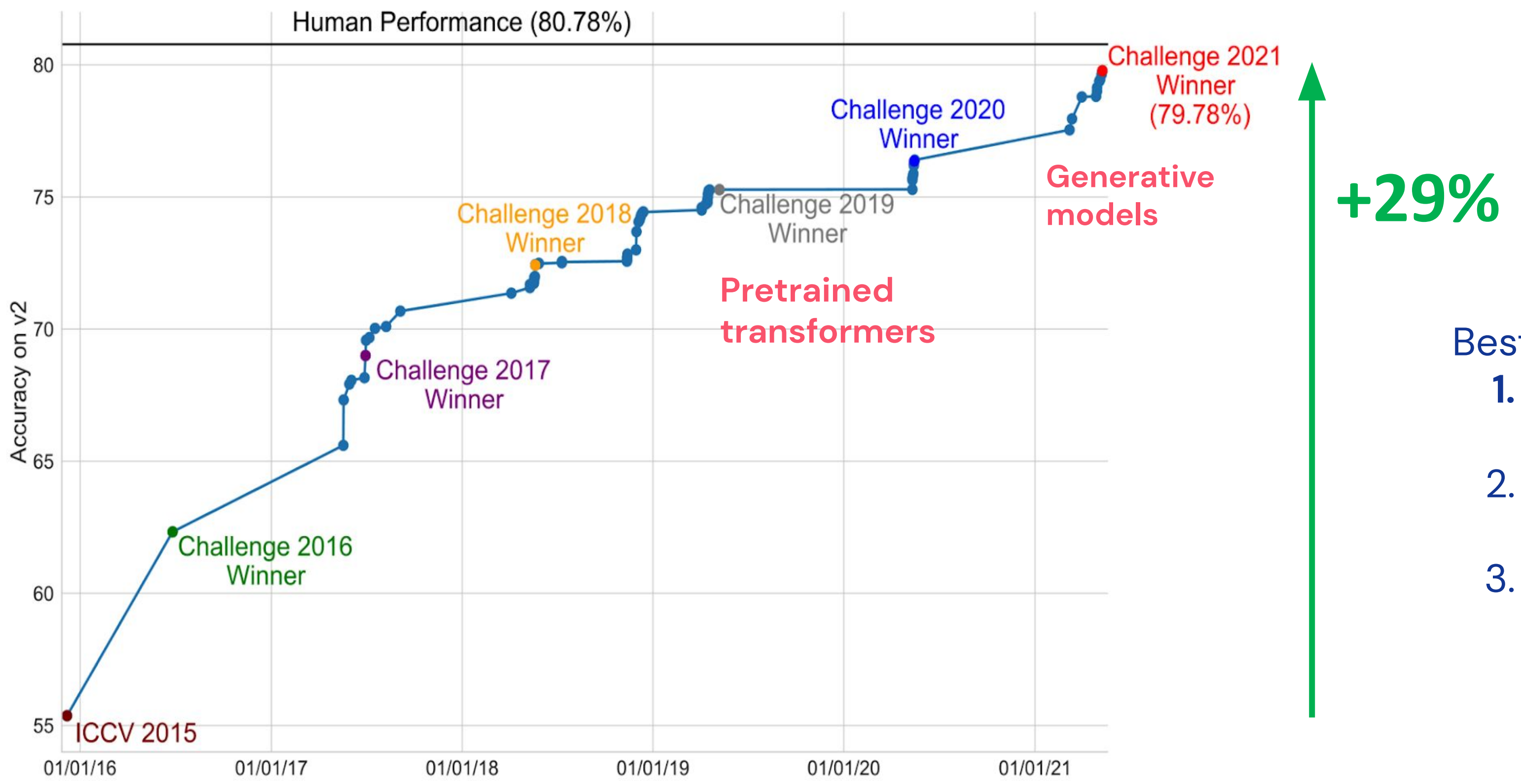

Phil Blunsom


Aida Nematzadeh[*,$]

EACL 2023

[*], [^] denote equal contribution, [$] denotes equal senior contribution

Progress on VQAv2 (Goyal et al., 2017)



Best performing models:*

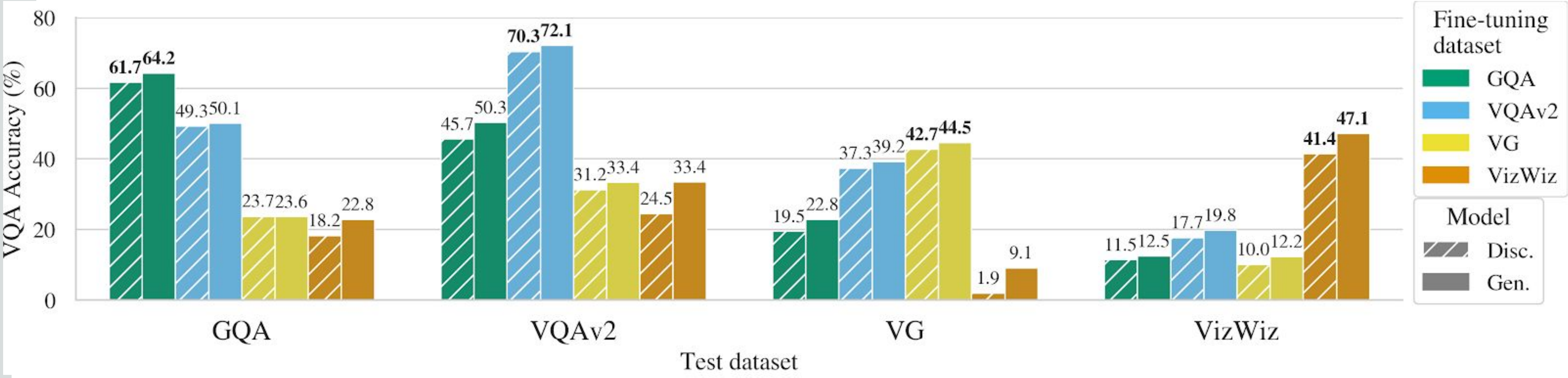
1. **PaLI: 84.3% (Chen et al., 2022)**
2. BEiT-3: 84.0% (Wang et al., 2022)
3. Flamingo: 82.0% (Alayrac et al., 2022)

* As of March 2023

- Is the VQA challenge solved?
  - No, we need to better evaluate our models
  - **Are models learning to solve the task of VQA or the dataset?**

Slide credits: Emanuele Bugliarello

IID vs OOD performance



- Large drop in performance for OOD evaluation

Slide credits: Ivana Kajić

Potential factors causing poor OOD generalization: A qualitative analysis

- Poor reasoning skills (logical, spatial, compositional)

  E.g., *"Is the cheese to the right or to the left of the empty plate?"*

- Overfitting to answer priors

  E.g., *"What is the skateboarder wearing to protect his head?"* → *"helmet"*

- Overfitting to question format

  E.g., *"What animal … ?", "What kind of animal … ?"* (GQA)

  45% accuracy drop

  "Who is … ?", "What is … ?" (VG)

Vision-Language Challenges

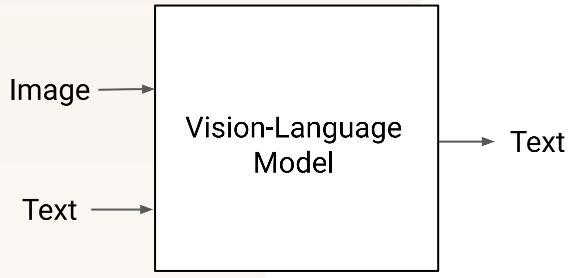
- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

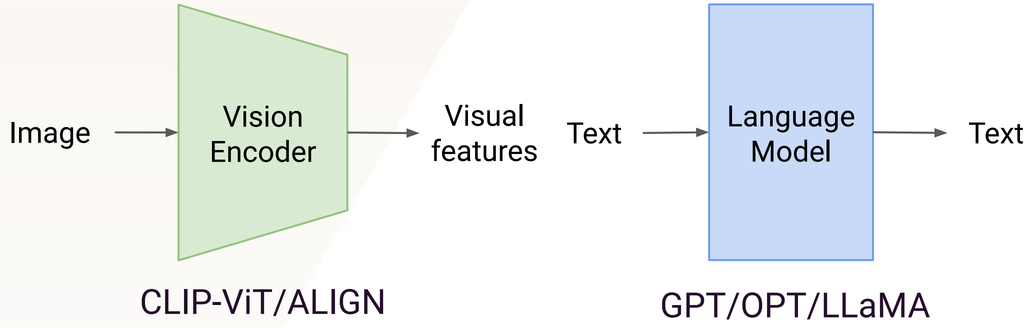
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Problem setup

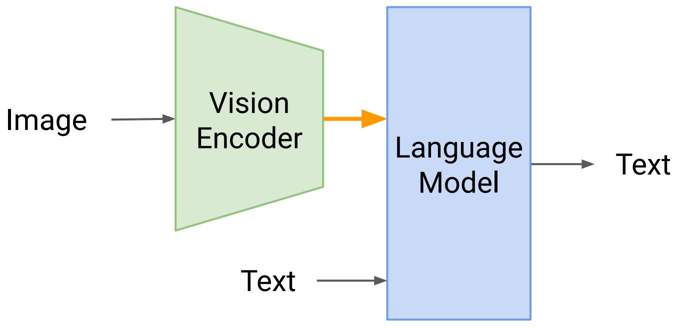
- **Goal**: build a vision-language model



- Impressive progress in unimodal models



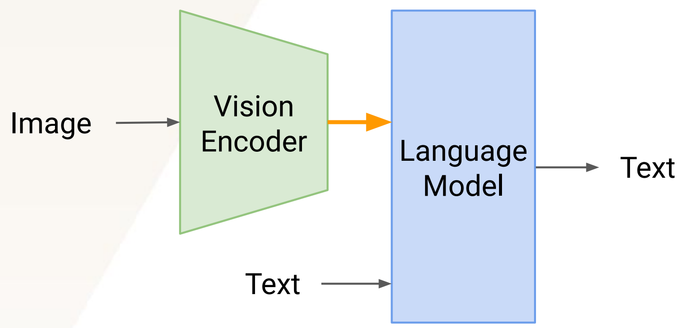
- **Research question**: Can we *efficiently adapt* unimodal models for multimodal tasks?



Slide adapted from Oscar Mañas

What existing approaches do

- Finetune the entire language model [Dai et al. 2022, Hao et al. 2022]
- Insert and train adapter layers in the language model [Eichenberg et al. 2021, Alayrac et al. 2022]
- Learn vision encoder from scratch [Tsimpoukelli et al. 2021]



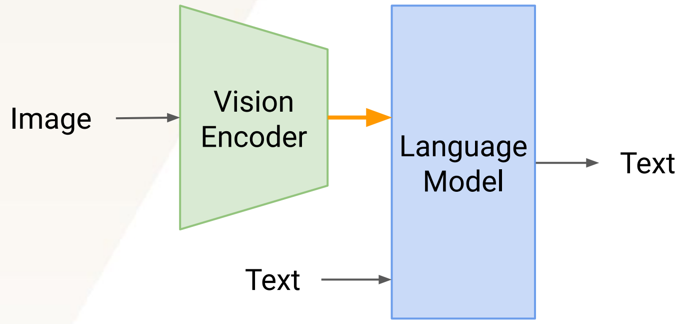
Issues with existing approaches:

- Large number of trainable parameters (~40M to ~10B)
- Inserting adapter layers is not straightforward
- Learning vision encoder from scratch does not scale well with larger vision encoders

Figure credits: Oscar Mañas

What we propose
(published at EACL 2023)

- **Reuse large** pre-trained unimodal **models** while keeping them **completely frozen** and free of adapter layers
- **Learn** a **lightweight mapping** between the representation spaces of pretrained unimodal models.



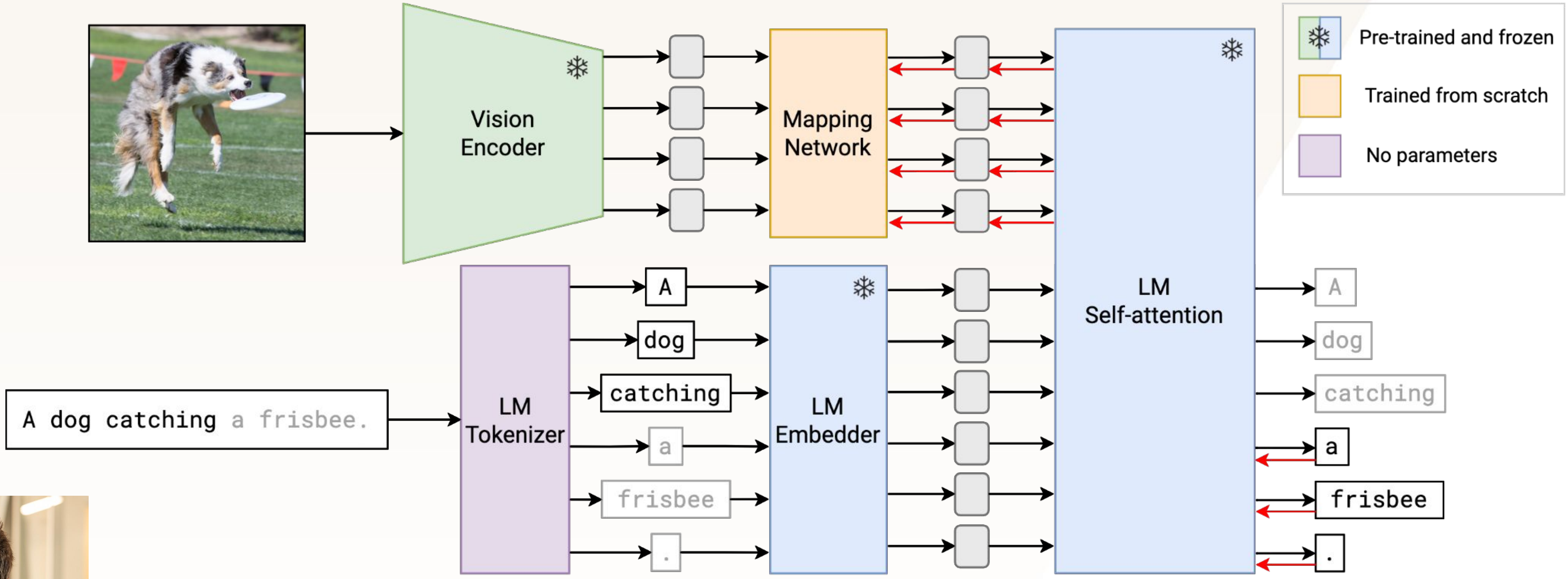
Benefits of our approach:

- Orders of magnitude fewer parameters
- Can be trained in just a few hours
- Uses modest computational resources and public datasets
- Modular, hence easily extensible to newer/better pretrained unimodal models

Oscar Mañas

Figure credits: Oscar Mañas

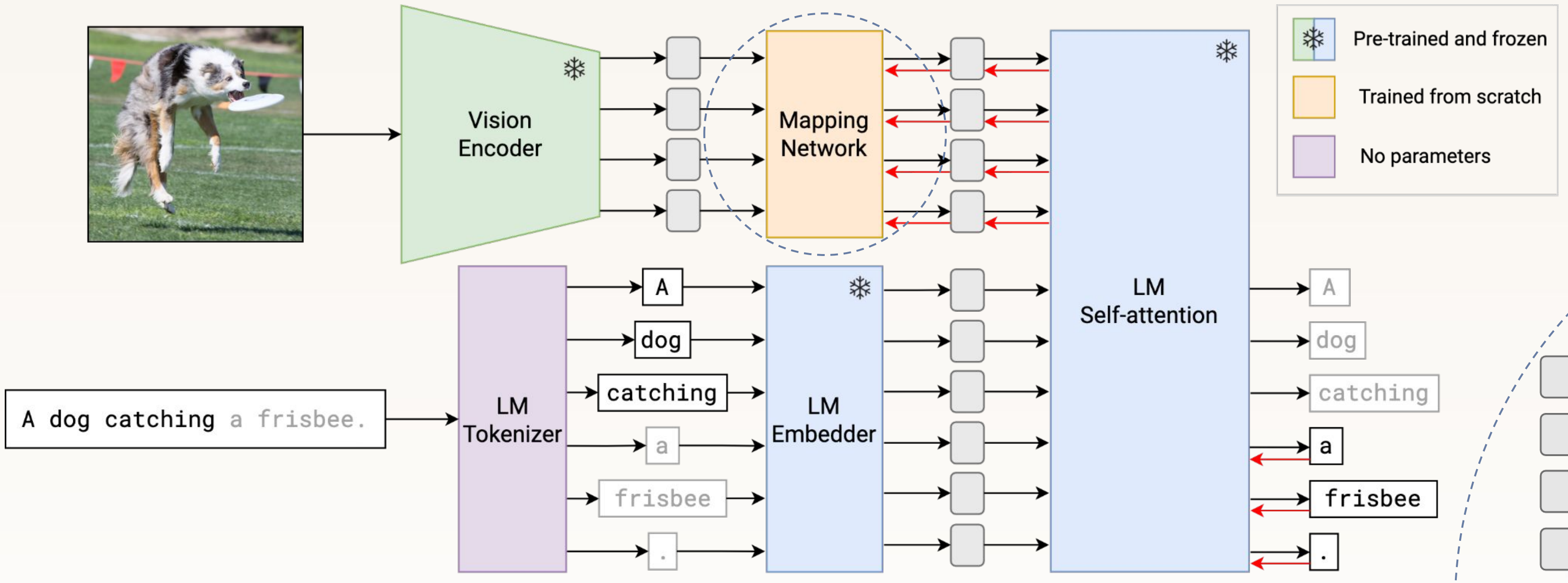
MAPL 🍁: method

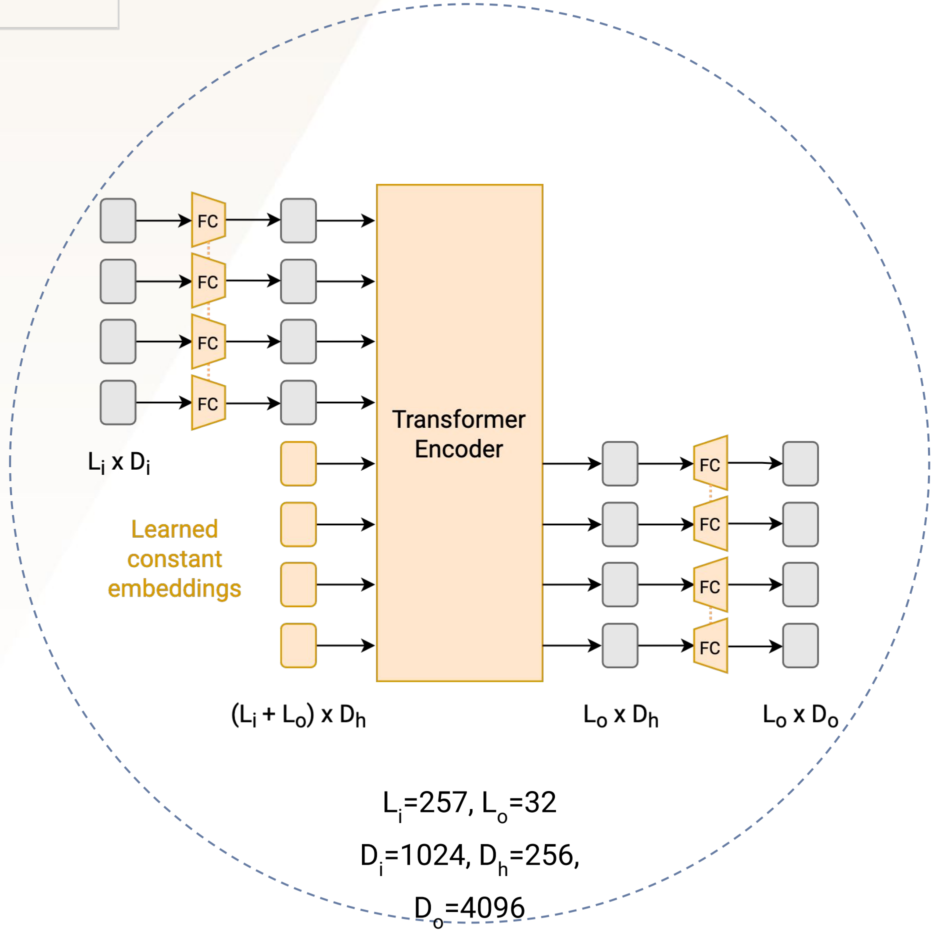


Oscar Mañas

Slide credits: Oscar Mañas

MAPL : method



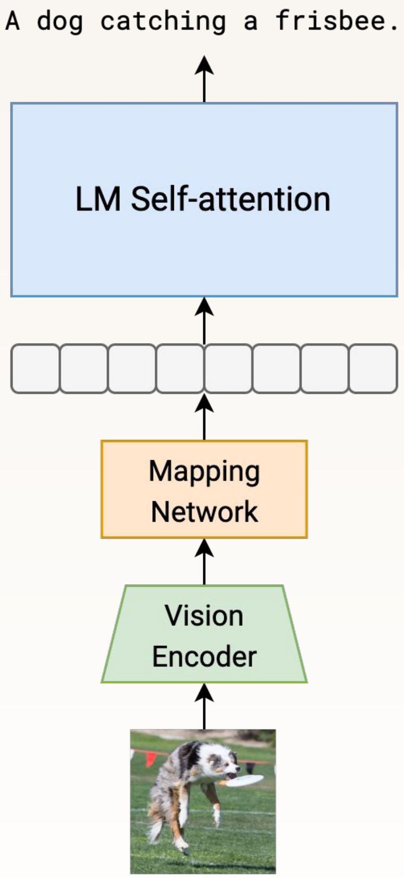


- Dimensionality bottleneck.
- Shared projection layers.
- Learned constant embeddings.

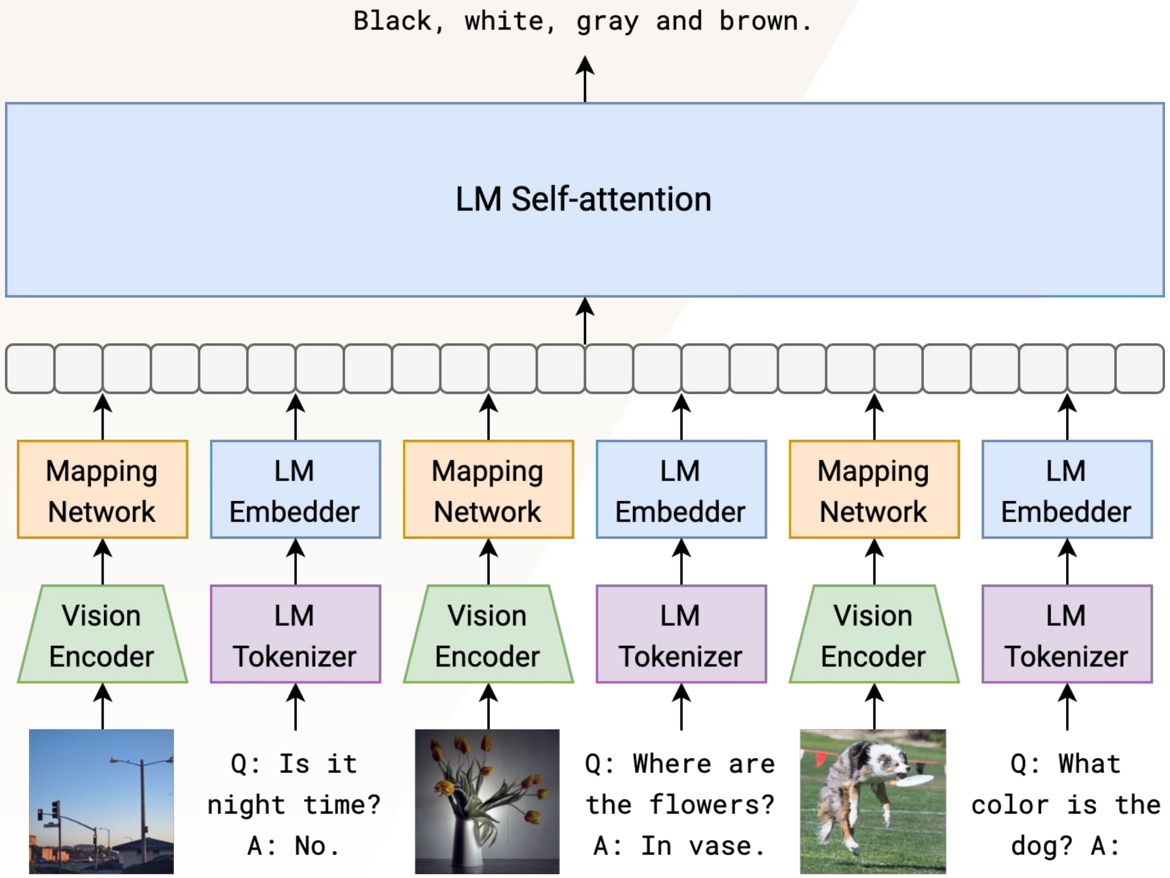
Oscar Mañas

Slide credits: Oscar Mañas

MAPL🍁: inference



0-shot image captioning.



2-shot VQA.

Oscar Mañas

Slide credits: Oscar Mañas

MAPL : experimental results

- MAPL achieves **superior or competitive** performance compared to similar methods while training orders of magnitude **fewer parameters**.
- MAPL is **more effective** than the baseline in **low-data** settings.
- MAPL is **more effective** than the baseline at **in-domain learning.**

	Trainable params	Training examples	n-shot VQAv2			n-shot OK-VQA			n-shot TextVQA			n-shot VizWiz-VQA			n-shot Overall		
			0	4	8	0	4	8	0	4	8	0	4	8	0	4	8
Existing methods using domain-agnostic training																	
Frozen	40.3M†	3.3M	29.50	38.20	-	5.90	12.60	-	-	-	-	-	-	-	-	-	-
MAGMA $_{CC12M}$	243M†	3.8M	36.90	45.40	-	13.90	23.40	-	-	-	-	5.60	10.60	-	-	-	-
VLKD $_{CC3M}$	406M	3.3M	38.60	-	-	10.50	-	-	-	-	-	-	-	-	-	-	-
Flamingo	10.2B	>2.1B	-	-	-	50.60	57.40	57.50	35.00	36.50	37.30	-	-	-	-	-	-
100% domain-agnostic training																	
MAPL-blind $_{CC\text{-}clean}$	3.4M	374K	20.62	35.01	35.11	4.84	14.68	14.28	3.68	5.43	5.82	3.18	8.65	9.55	8.08	15.94	16.19
Frozen* $_{CC\text{-}clean}$	40.3M	374K	25.98	37.80	38.52	5.51	18.86	19.91	5.11	6.15	6.30	4.33	11.28	16.68	10.23	18.52	20.35
MAPL $_{CC\text{-}clean}$	3.4M	374K	**33.54**	**45.13**	**45.21**	**13.84**	**24.25**	**23.93**	**8.26**	**8.88**	**8.77**	**11.72**	**18.46**	**19.52**	**16.84**	**24.18**	**24.36**
1% domain-agnostic training																	
Frozen* $_{CC\text{-}clean}$	40.3M	3.7K	26.22	36.69	37.41	5.50	**18.76**	**20.51**	5.71	**7.19**	7.53	3.83	**11.71**	**16.66**	10.31	**18.58**	**20.53**
MAPL $_{CC\text{-}clean}$	3.4M	3.7K	**30.80**	**37.38**	**37.95**	**8.77**	18.18	19.15	**6.40**	7.07	**7.74**	**5.68**	9.26	10.58	**12.91**	17.97	18.85
100% in-domain training																	
PICa*	0	0	20.61	46.86	47.80	11.84	**31.28**	**33.07**	-	-	-	-	-	-	-	-	-
Frozen* $_{COCO}$	40.3M	414K	32.09	38.90	39.42	9.81	20.72	21.83	7.54	6.82	6.74	5.87	12.07	17.35	13.82	19.63	21.33
Frozen* $_{TextCaps}$	40.3M	103K	32.49	37.39	38.03	11.34	19.87	20.82	8.83	7.33	7.51	6.25	12.26	16.86	14.73	19.21	20.80
Frozen* $_{VizWiz}$	40.3M	110K	26.93	37.38	37.91	5.85	19.12	20.64	6.38	7.44	7.47	5.57	13.06	18.06	11.18	19.25	21.02
MAPL $_{COCO}$	3.4M	414K	**43.51**	**48.75**	**48.44**	**18.27**	31.13	31.63	10.99	11.10	11.08	**14.05**	17.72	19.18	21.70	**27.17**	**27.58**
MAPL $_{TextCaps}$	3.4M	103K	38.83	43.34	43.43	16.33	25.07	25.92	**22.27**	**19.53**	**19.75**	12.31	16.69	18.18	**22.43**	26.15	26.82
MAPL $_{VizWiz}$	3.4M	110K	32.80	42.94	43.20	11.70	24.91	25.73	9.27	10.36	10.23	10.42	**20.63**	**23.10**	16.05	24.71	25.56
1% in-domain training																	
Frozen* $_{COCO}$	40.3M	4.1K	30.18	37.23	37.89	9.33	19.60	20.71	7.43	7.65	7.67	4.37	12.00	16.48	12.83	19.12	20.69
Frozen* $_{TextCaps}$	40.3M	1.0K	32.09	36.72	37.25	10.75	18.85	19.51	8.17	7.57	7.28	5.39	11.79	16.20	14.10	18.73	20.06
Frozen* $_{VizWiz}$	40.3M	1.1K	29.62	37.30	37.87	7.57	19.36	20.60	7.16	7.17	7.25	4.53	**12.51**	**17.56**	12.22	19.08	**20.82**
MAPL $_{COCO}$	3.4M	4.1K	**37.69**	**40.42**	**40.84**	**13.92**	**21.66**	**22.41**	8.30	6.96	6.84	**6.94**	10.72	12.43	**16.71**	**19.94**	20.63
MAPL $_{TextCaps}$	3.4M	1.0K	33.57	36.70	36.87	12.46	17.45	18.21	**9.34**	**8.29**	**8.62**	6.54	9.58	11.62	15.48	18.00	18.83
MAPL $_{VizWiz}$	3.4M	1.1K	31.88	36.81	37.04	9.59	17.64	17.64	7.25	5.99	6.04	4.73	9.48	11.33	13.36	17.48	18.01

Oscar Mañas

ArXiv: https://arxiv.org/abs/2210.07179

Vision-Language Challenges

- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Visio-Linguistic Compositional Reasoning

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

[Thrush et al. CVPR 2022]

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

[Hudson et al. CVPR 2019]

Visual Genome Relation

Assessing relational understanding (23,937 test cases)

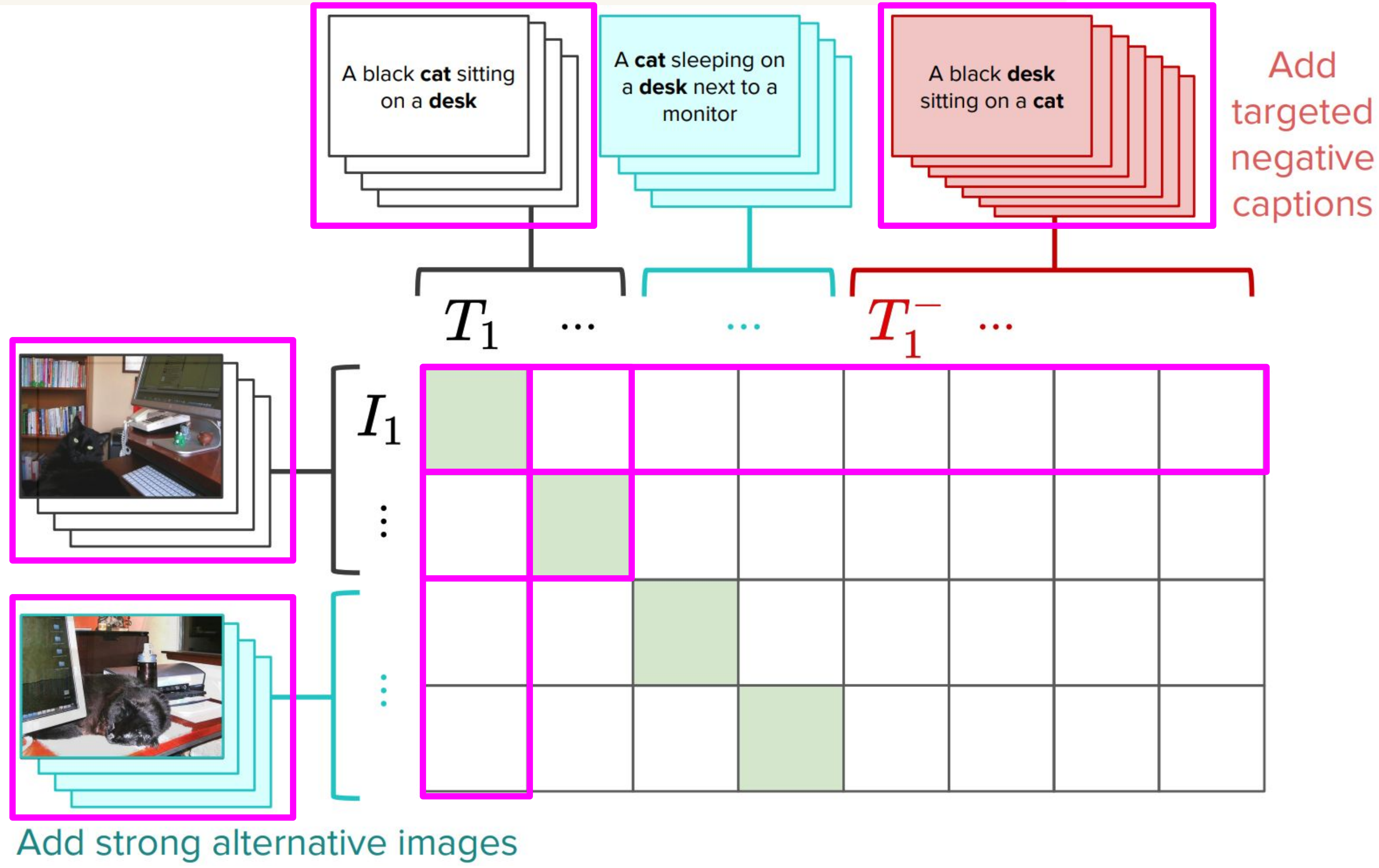
✔ *the horse is eating the grass*
X *the grass is eating the horse*

[Yuksekgonul et al. ICLR 2023]

What existing approaches do

1. **Create** hard-negative sentences and images

2. **Contrast** them against correct image-caption pairs



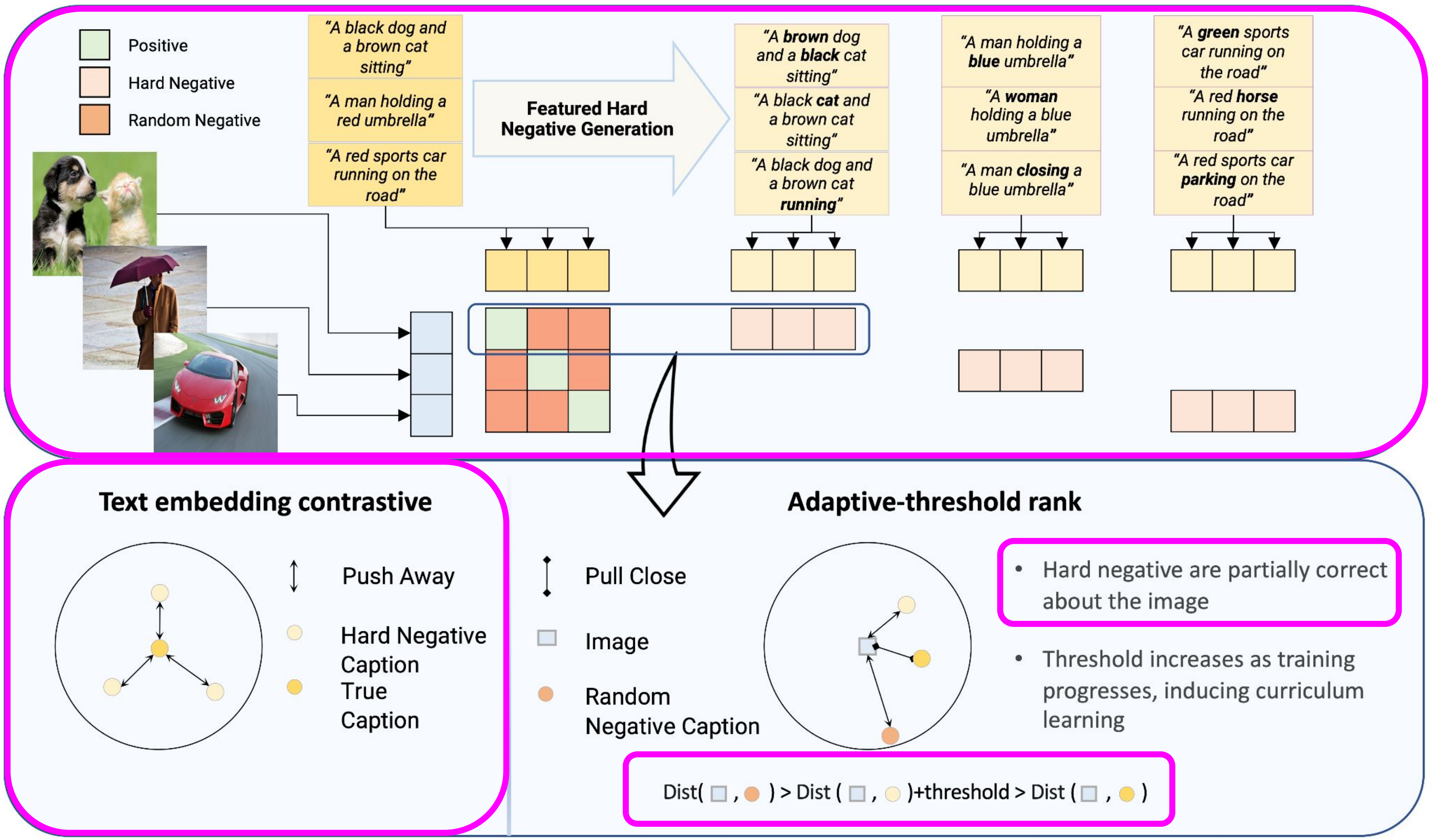
NegCLIP approach from [Yuksekgonul et al. ICLR 2023](#)

What we propose – training time approach

(work in progress)

3. Additionally **contrast** hard-negative **sentences** against correct **sentences**

4. Add **rank loss** between correct and hard-negative image-text pairs



Le Zhang

Figure credits: Le Zhang

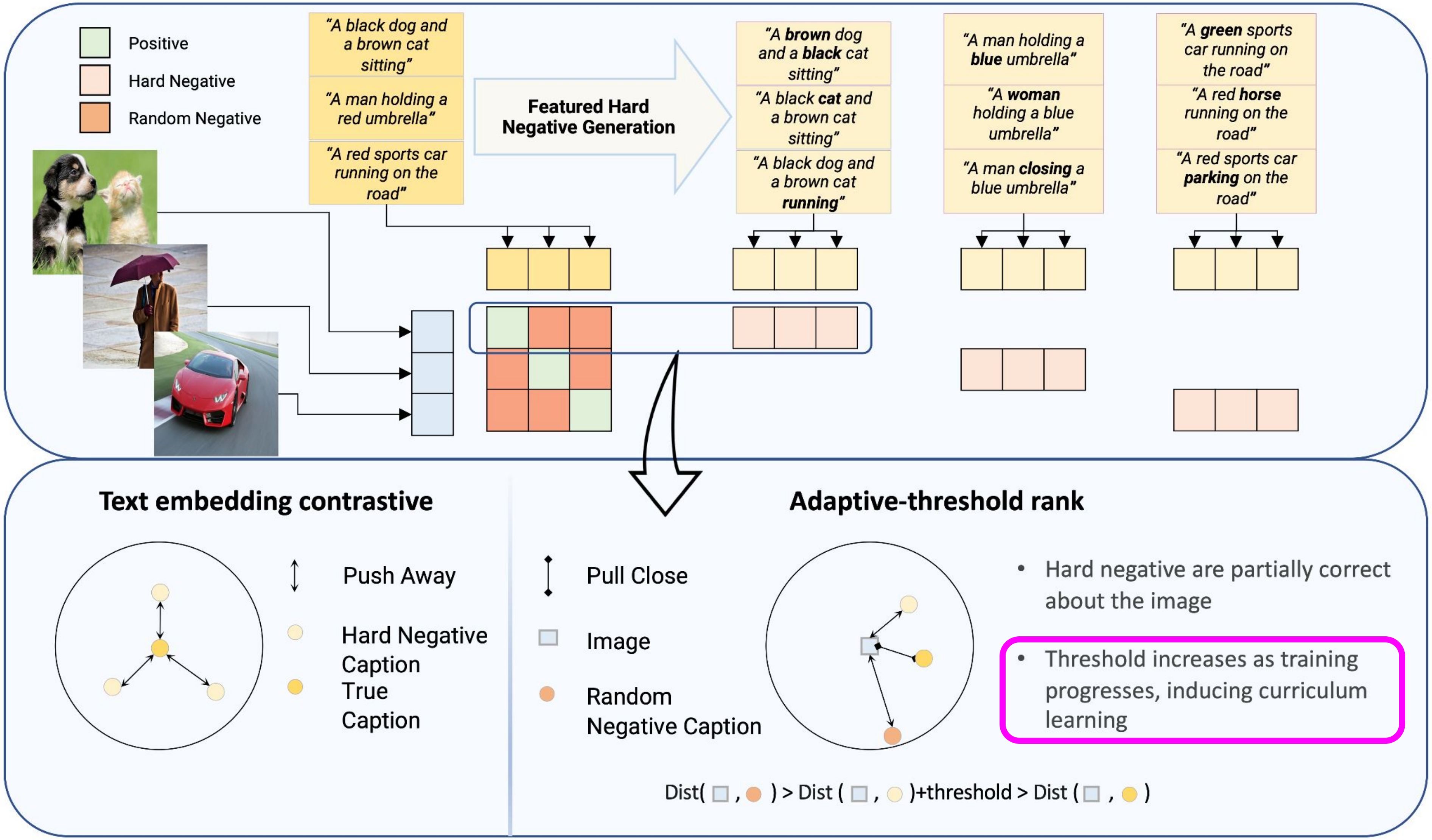
What we propose – training time approach

(work in progress)

3. Additionally **contrast** hard-negative **sentences** against correct **sentences**

4. Add **rank loss** between correct and hard-negative image-text pairs

5. Use **adaptive margin** for the rank loss – **curriculum learning**



Le Zhang

Figure credits: Le Zhang

Experimental Results

- We **outperform** existing methods **significantly** in both relation and attribution understanding.

Model	VG-Relation	VG-Attribution
Random Chance	50.00	50.00
CLIP	59.28	62.86
NegCLIP [22]	80.22	70.46
Ours *itc*	61.46	65.80
Ours *itc(hn)*	79.28	70.26
Ours *itc+rank*	79.59	69.23
Ours *itc+tec*	81.11	71.70
Ours *itc(hn)+rank*	81.34	72.24
Ours *itc(hn)+tec*	81.92	74.54
Ours *itc(hn)+tec+rank*	**82.70**	**76.31**

+2.5%

+5.9%

Table 1. Results on the ARO dataset across the combination of losses. *itc* represents finetune model on COCO dataset without generated negatives, *itc(hn)* represents finetune with generated negatives using $\mathcal{L}_{itm(hn)}$

Le Zhang

Experimental Results

- We **outperform** existing methods **significantly** in both relation and attribution understanding.

- Ablation studies show that **both proposed losses are effective** for learning compositionality

Model	VG-Relation	VG-Attribution
Random Chance	50.00	50.00
CLIP	59.28	62.86
NegCLIP [22]	80.22	70.46
Ours *itc*	61.46	65.80
Ours *itc(hn)*	79.28	70.26
Ours *itc+rank*	79.59	69.23
Ours *itc+tec*	81.11	71.70
Ours *itc(hn)+rank*	81.34	72.24
Ours *itc(hn)+tec*	81.92	74.54
Ours *itc(hn)+tec+rank*	**82.70**	**76.31**

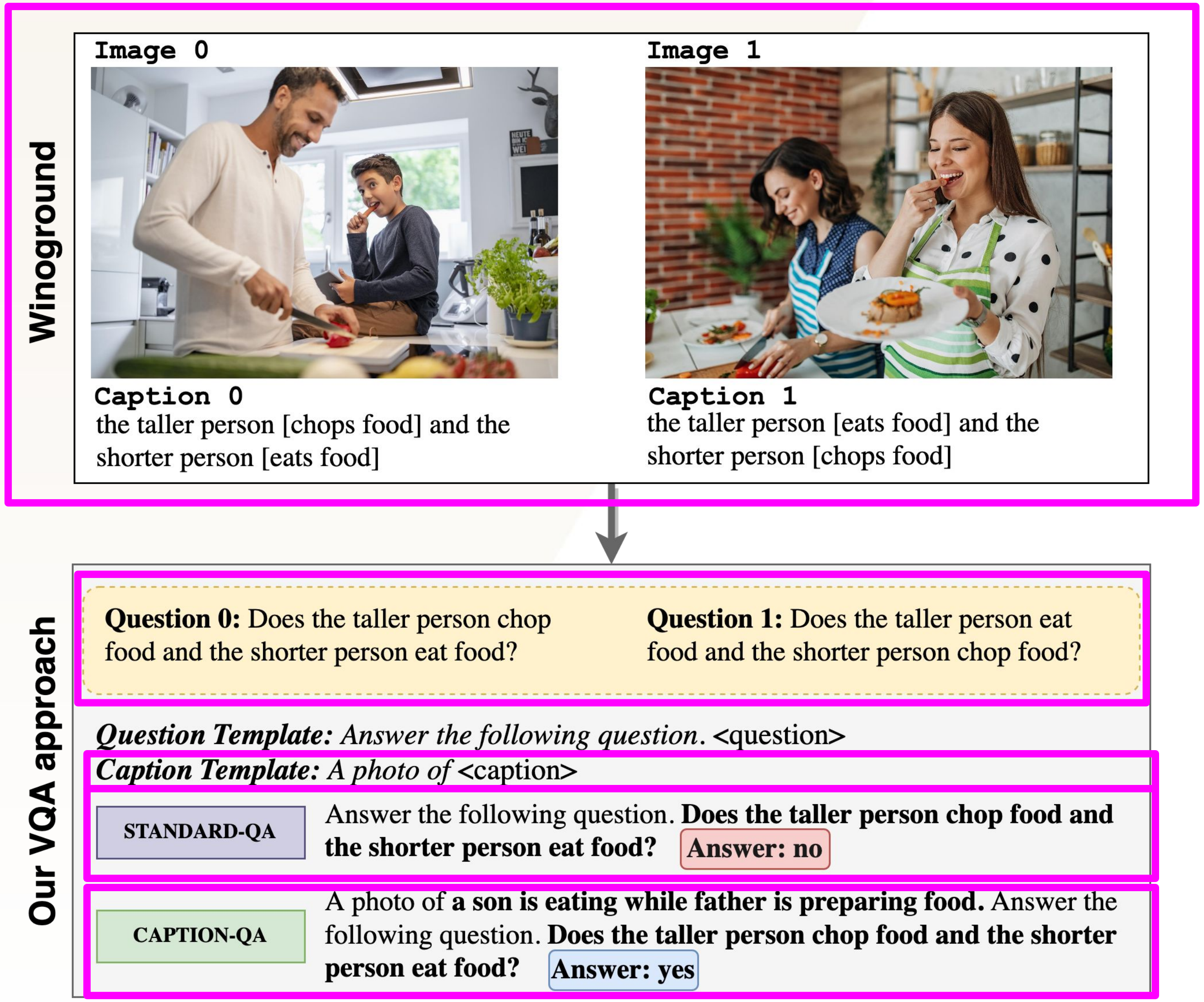
Table 1. Results on the ARO dataset across the combination of losses. *itc* represents finetune model on COCO dataset without generated negatives, *itc(hn)* represents finetune with generated negatives using $\mathcal{L}_{itm(hn)}$

Le Zhang

What we propose – inference time approach

(work in progress)

- Prompt the model to generate a caption for the image.
- Feed the generated caption along with the question.



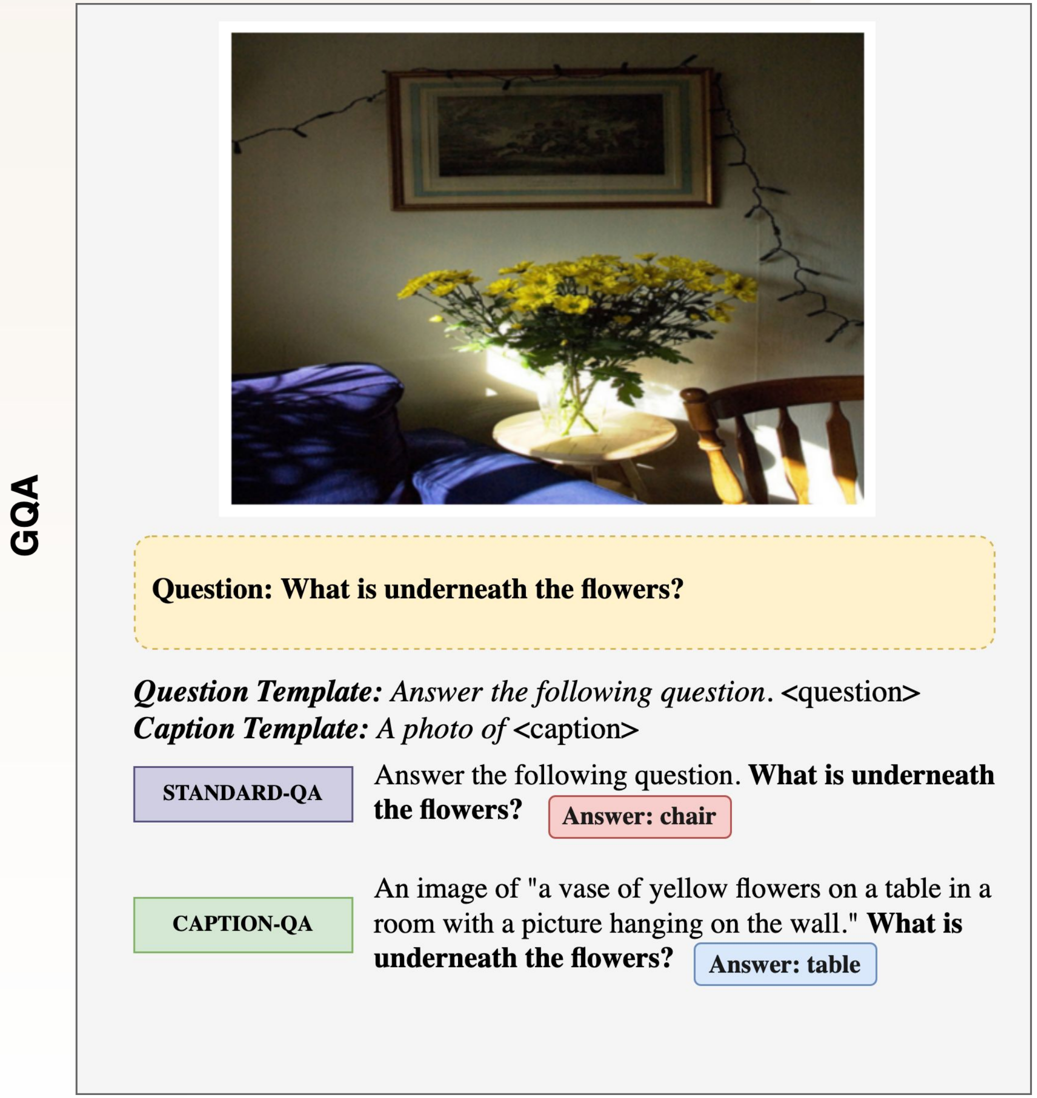
Rabiul Awal

Le Zhang

Figure credits: Rabiul Awal

Improving compositional reasoning through effective prompting

- Zero-shot Prompting for VLMs is underexplored
- Effective prompts can improve eliciting proper response on a given question
- An image description e.g. caption can provide additional visual cues in the text-encoder (chain-of-thought prompting)




Rabiul Awal


Le Zhang

Slide credits: Rabiul Awal

What we propose – inference time approach

(work in progress)

- Image-caption prompting significantly improves the performance on Winoground-QA.
- The caption should contain information relevant to the question.

Model Name	Prompt Name	STANDARD-QA			CAPTION-QA		
		Group	$Q_0 2I$	$Q_1 2I$	Group	$Q_0 2I$	$Q_1 2I$
BLIP2 FLAN-$T5_{XL}$	AnswerFollowingYNQuestion / describeTheImage	4.0	17.5	22.0	4.25	16.5	21.5
	doesItDescribeImage / inThisImage	5.25	21.75	22.75	8.25	25.5	27.2
	isTrueAboutImageAnswerYN / describeTheScene	**5.75**	19.75	22.0	**8.75**	27.75	25.5
BLIP2 FLAN-$T5_{XXL}$	AnswerFollowingYNQuestion / describeTheScene	**7.25**	22.5	24.25	**10.25**	28.75	28.0
	doesItDescribeImage / aPhotoOf	5.25	18.5	23.7	8.75	26.25	27.7
	isTrueAboutImageAnswerYN / describeTheImage	6.0	20.0	22.75	9.5	24.75	28.5
BLIP-VQA	-	6.75	25.5	22.0	-	-	-
Random chance	-	6.25	25	25	6.25	25	25

Rabiul Awal **Le Zhang**

Vision-Language Challenges

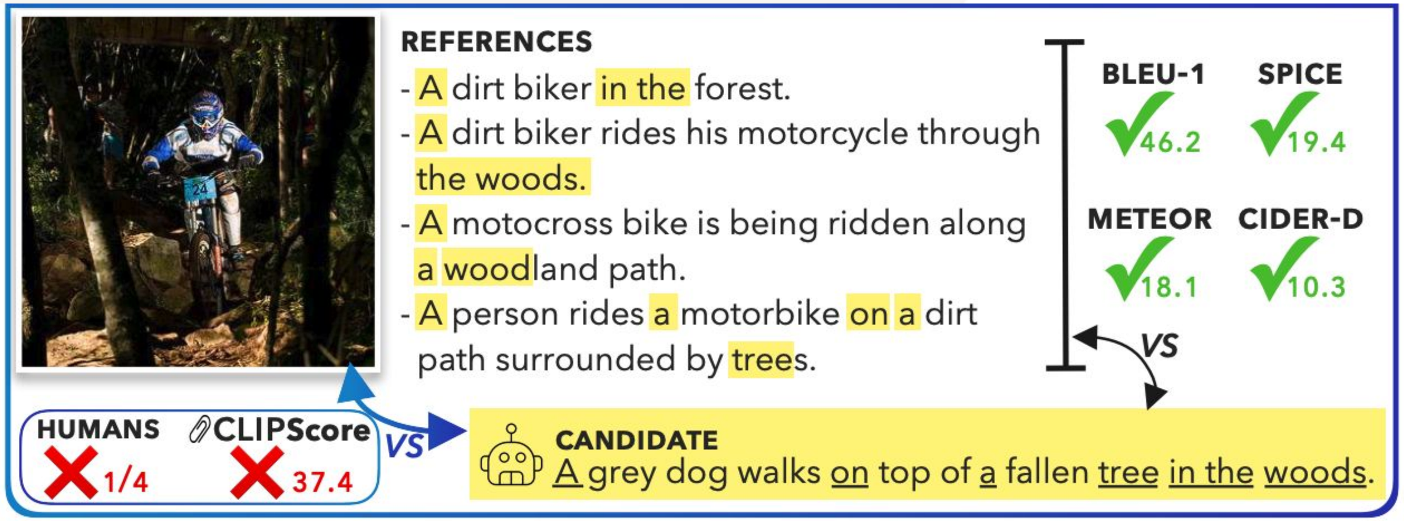
- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Vision-Language Challenges

- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Robust Automatic Evaluation

- Evaluating image captioning is difficult!
- Existing metrics rely on **n-gram matches** between candidate and reference captions.
- Recently, **reference-free** metrics have been proposed – CLIPScore, UMIC.



[Hessel et al. EMNLP 2021]

Robust Automatic Evaluation

(work in progress)

Recently proposed reference-free image-captioning metrics are not robust enough!

- They fail to recognize **fine-grained differences** between correct and incorrect captions.
- They have poor understanding of **negation**.
- They are biased by the **length** of the captions.

Captions	CLIPScore	UMIC
The title of the book is topology.	0.62	0.19
The title of the book is muffin.	0.74	0.62

Saba Ahmadi

Figure credits: Saba Ahmadi

Robust Automatic Evaluation

(work in progress)

Recently proposed reference-free image-captioning metrics are not robust enough!

- CLIPScore is more sensitive (than UMIC) to the **number and size of objects** mentioned in the caption.

Captions	CLIPScore	UMIC
Small Object: There is a knife.	0.62	0.36
Big Object: There is a pizza.	0.72	0.34

Saba Ahmadi

Figure credits: Saba Ahmadi

Robust Automatic Evaluation

(work in progress)

Recently proposed reference-free image-captioning metrics are not robust enough!

- CLIPScore is more sensitive (than UMIC) to the **number and size of objects** mentioned in the caption.
- CLIPScore is indifferent to the **sentence structure**.

Captions	CLIPScore	UMIC
Small Object: There is a knife.	0.62	0.36
Big Object: There is a pizza.	0.72	0.34
Shuffled Small Object: A there knife is.	0.63	0.19
Shuffled Big Object: A there pizza is.	0.74	0.18

Saba Ahmadi

Figure credits: Saba Ahmadi

Robust Automatic Evaluation

- CLIPScore shows high sensitivity to the number of image-relevant objects mentioned in the caption while UMIC is not notably sensitive to it.
- Both metrics are sensitive to the size of image-relevant objects mentioned in the caption; however, CLIPScore increases with size while UMIC decreases.
- While visual grounding remained the same and we shuffled the sentences, interestingly UMIC was sensitive to sentence structure, whereas CLIPScore was not.

Captions	CLIPScore [Hessel et al. EMNLP 2021]	UMIC [Lee et al. ACL 2021]
There is a person.	0.536	0.143
There is a person and a sports ball.	0.639	0.156
There is a person, a sports ball and a baseball bat.	0.746	0.150

Number of Objects

Captions	CLIPScore [Hessel et al. EMNLP 2021]	UMIC [Lee et al. ACL 2021]
Small Object: There is a knife.	0.619	0.363
Big Object: There is a pizza.	0.721	0.336
Shuffled Small Object: A there knife is.	0.631	0.193
Shuffled Big Object: A there pizza is.	0.740	0.180

Object-size and sentence structure

Figure credits: Saba Ahmadi

Vision-Language Challenges

- Out-of-distribution generalization
- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Thanks!

Questions?

Vision-Language Challenges

- Parameter-efficient learning
- Visio-linguistic compositional reasoning
- Robust automatic evaluation
- Common sense and factual knowledge reasoning
- Interpretability and explainability
- Overcoming spurious correlations and biases in data
- …

Thanks!
Questions?