Language + Vision (1)

Mateusz Malinowski

Plan

- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - Early Visual Question Answering systems
  - Non-local computations (Relation Nets, Transformer)
  - Graph Neural Networks
  - Soft-Attention & Hard-Attention in Computer Vision
  - Bias
- Vision + Language (2)
  - Self-supervised learning
  - Generation and AI Art
  - Prompting & recent advances
  - “Vision as a Language”
  - Scalability

Why Vision + Language?

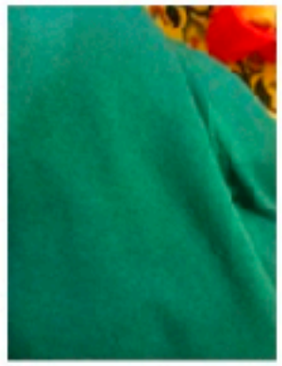
Blade Runner 2049

Why Vision + Language?

Answer Visual Questions from People Who Are Blind

Q: Does this foundation have any sunscreen?
A: yes

Q: What is this?
A: 10 euros

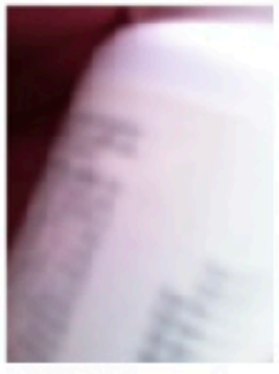


Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

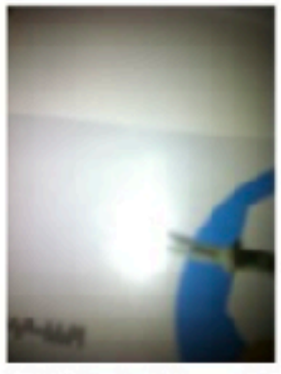
Q: Is it sunny outside?
A: yes

Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning



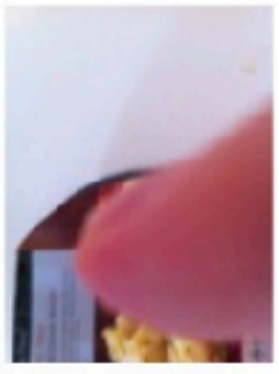
Q: What type of pills are these?
A: unsuitable image

Q: What type of soup is this?
A: unsuitable image



Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable

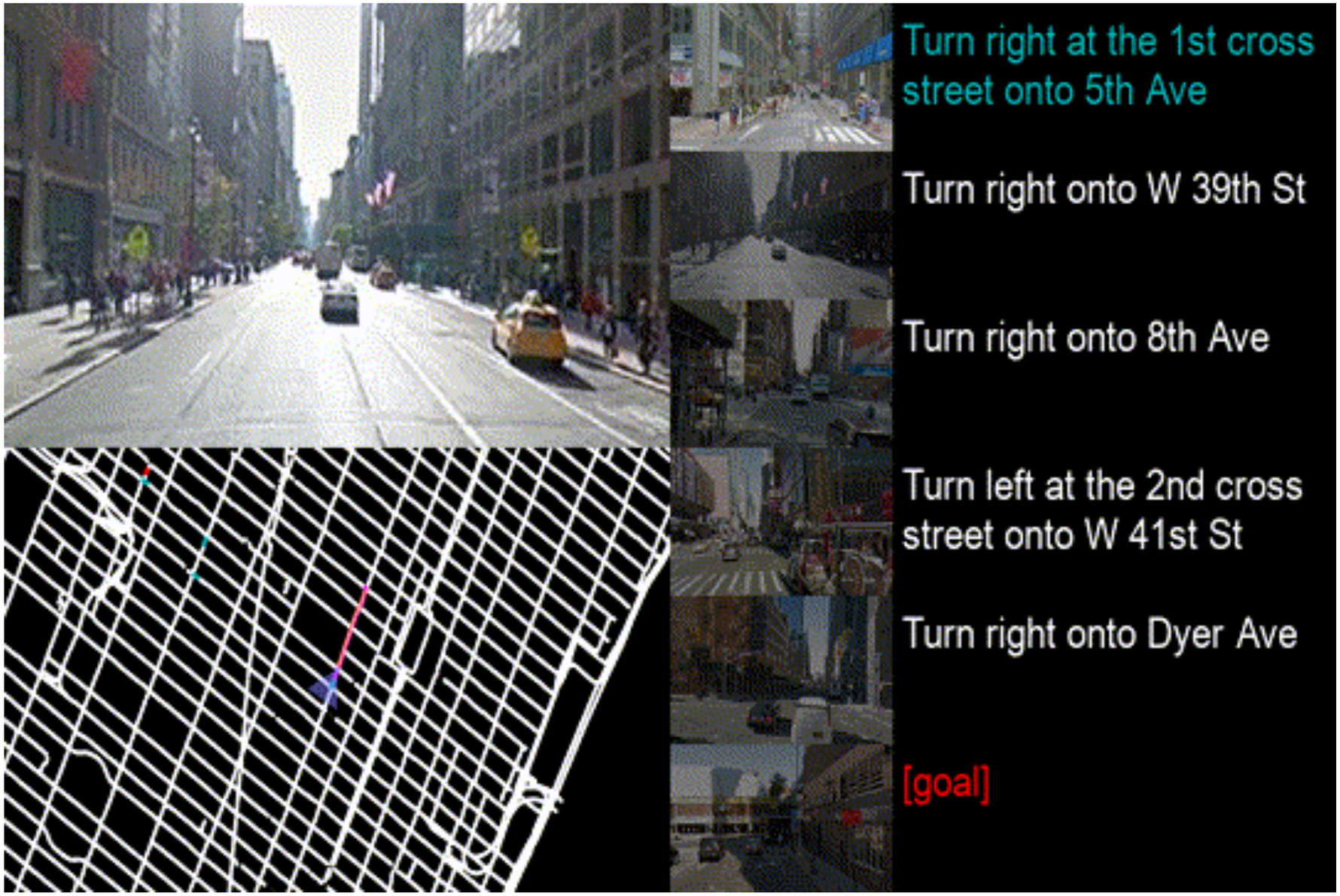


Q: What is this?
A: unanswerable

Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

"VizWiz" D. Gurari et al.

Why Vision + Language?



“StreetLearn” P. Mirowski, M. Malinowski, K. Hermann et al.

Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?

Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?
  - Is reasoning linguistic?

Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?
  - Is reasoning linguistic?
  - The role of (grounded) communication?

Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?
  - Is reasoning linguistic?
  - The role of (grounded) communication?
  - Do we need language? Do we need vision?

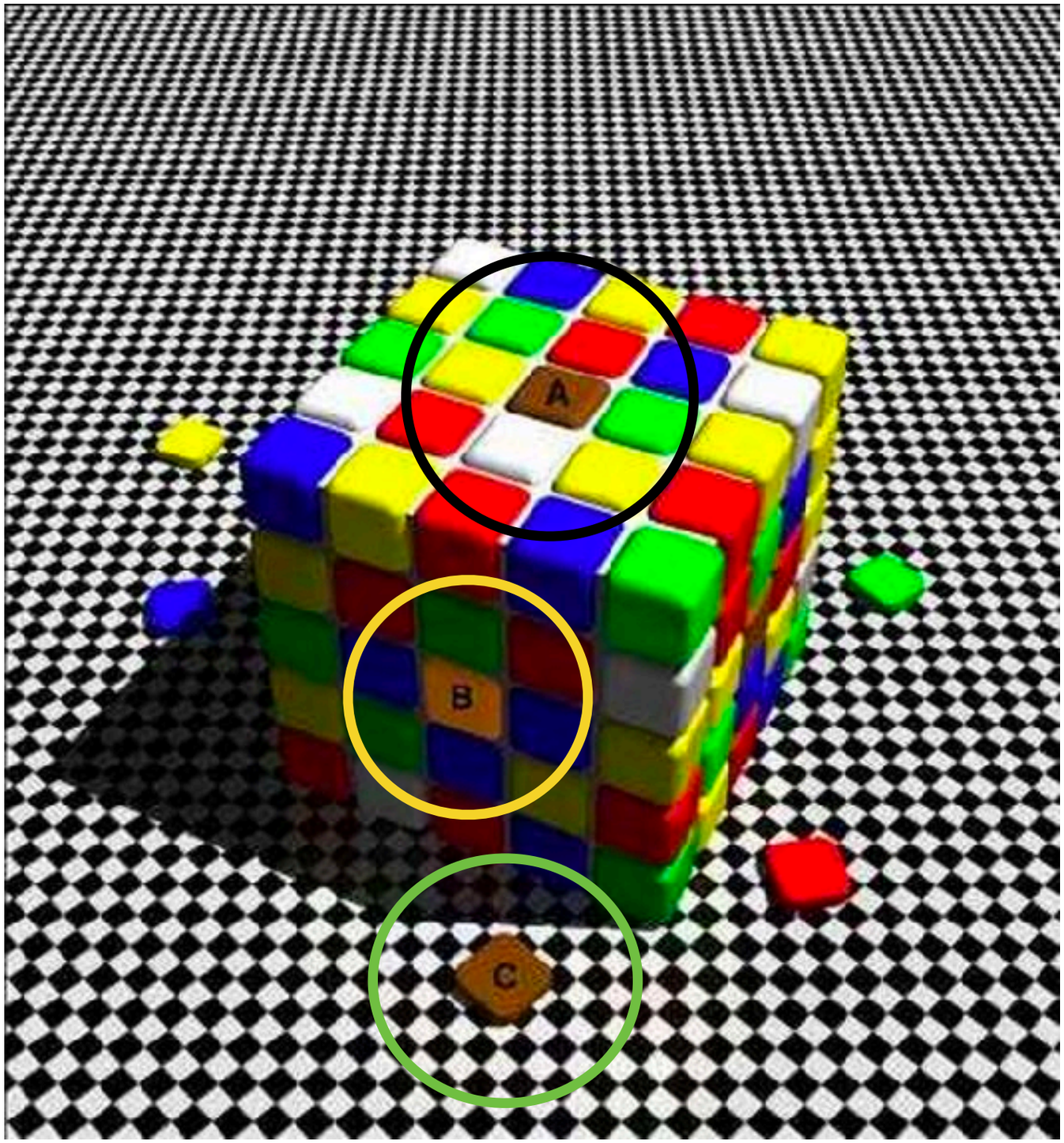
Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?
  - Is reasoning linguistic?
  - The role of (grounded) communication?
  - Do we need language? Do we need vision?
  - How to deal with ambiguities (in language; in vision)?

Why Vision + Language?

- Research curiosity
  - How does categorisation emerge?
  - Is reasoning linguistic?
  - The role of (grounded) communication?
  - Do we need language? Do we need vision?
  - How to deal with ambiguities (in language; in vision)?
  - How to evaluate visual systems?

Can we trust our vision?



Beau Lotto

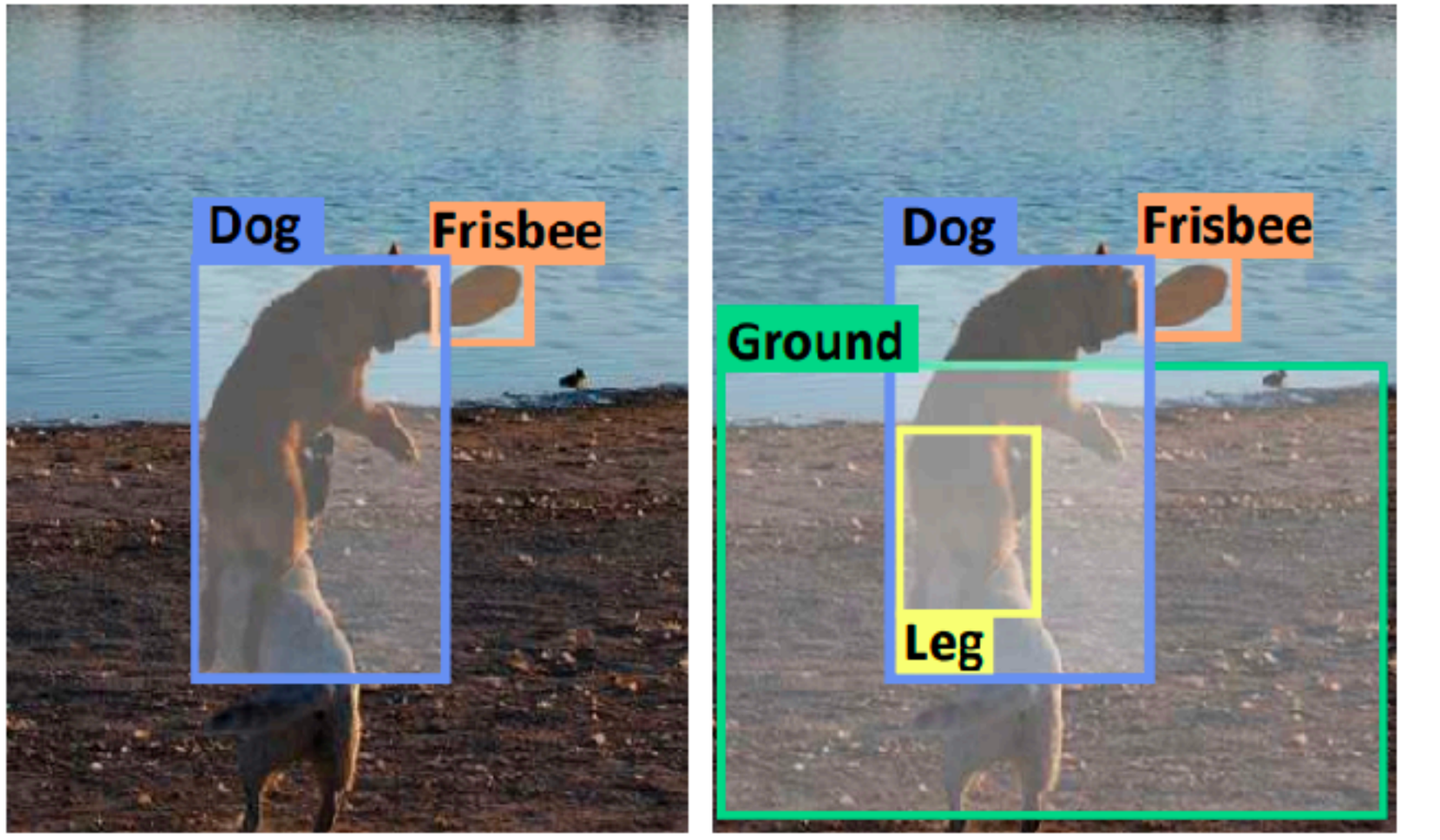
Can we trust our vision?

- We live in two realities
  - Subjective psychophysical experience
  - Objective physical reality
- And they are not the same

Beau Lotto

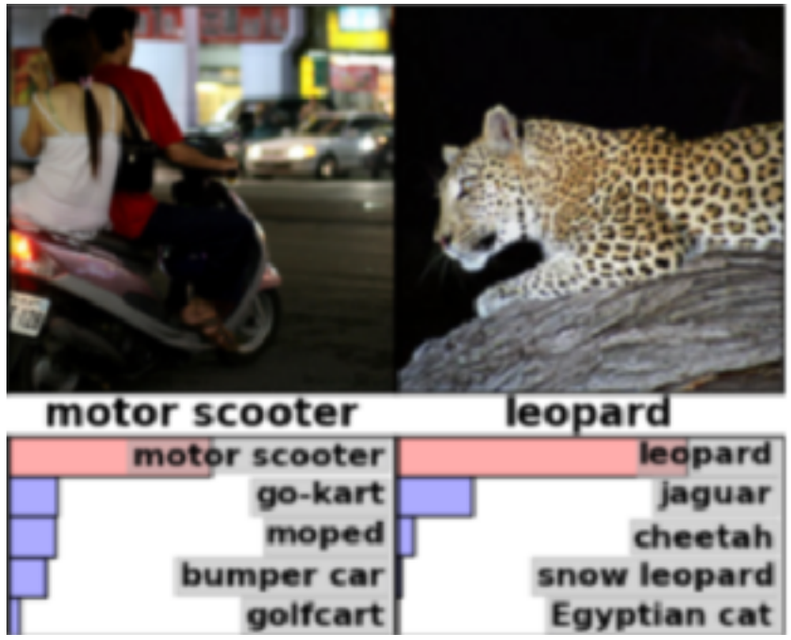
How to evaluate vision? Bounding boxes?

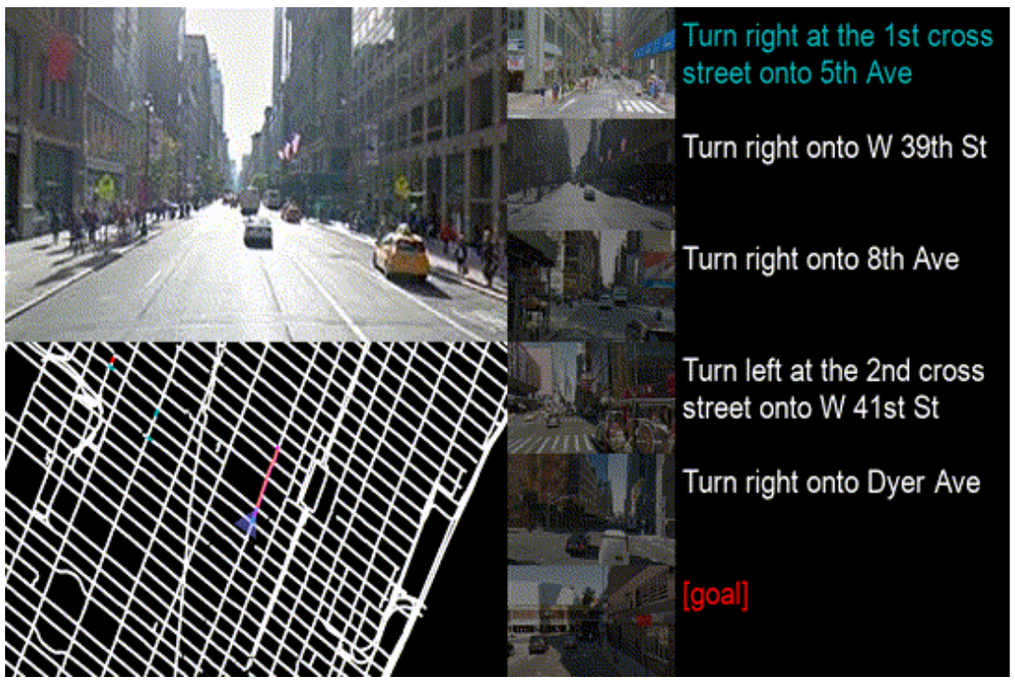
Captioning



T. Wang et al. “Visual Commonsense R-CNN”

Intelligence is NOT about a single task

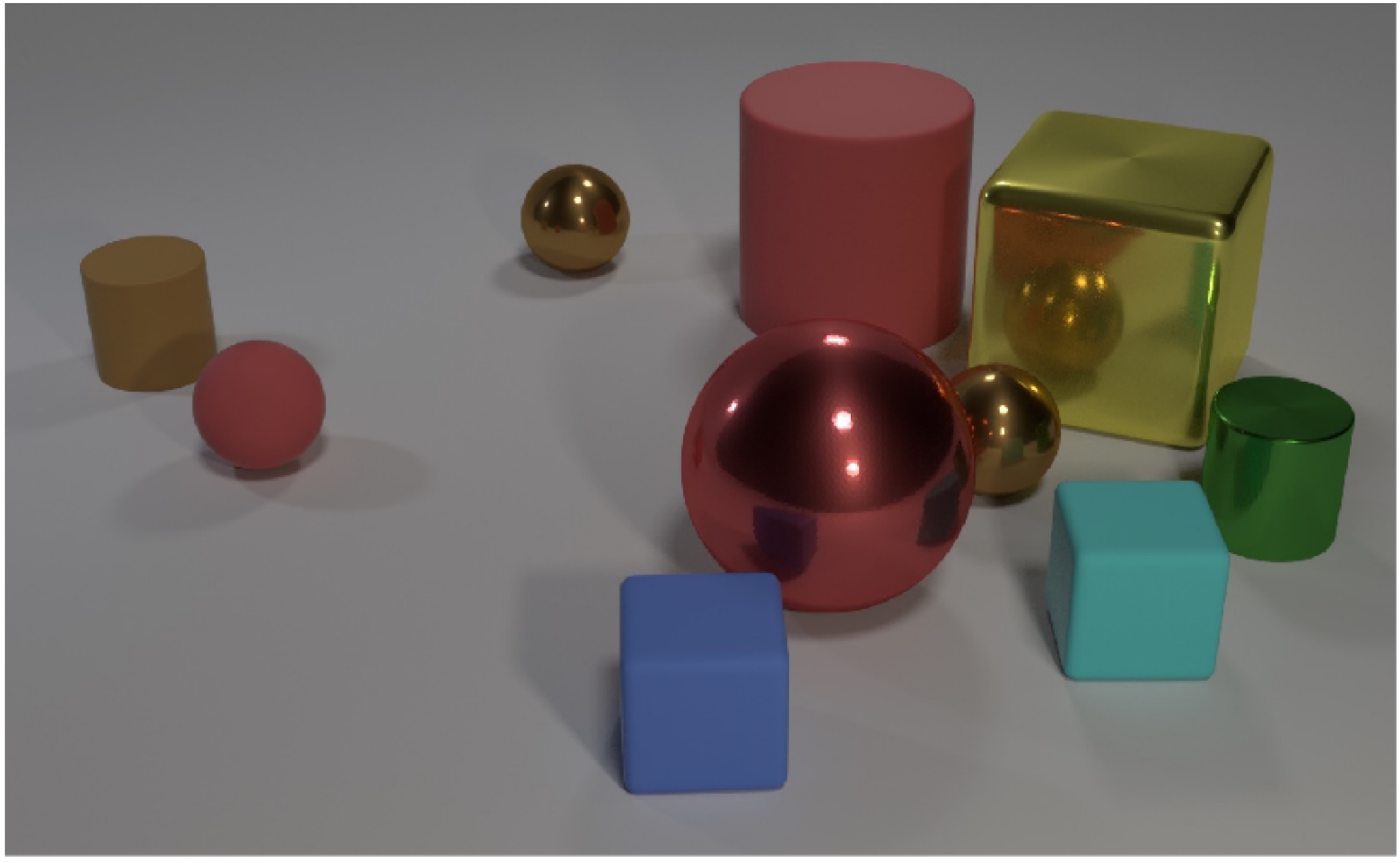




Visual Question Answering

M. Malinowski et al. “A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input”. NeurIPS 2014.

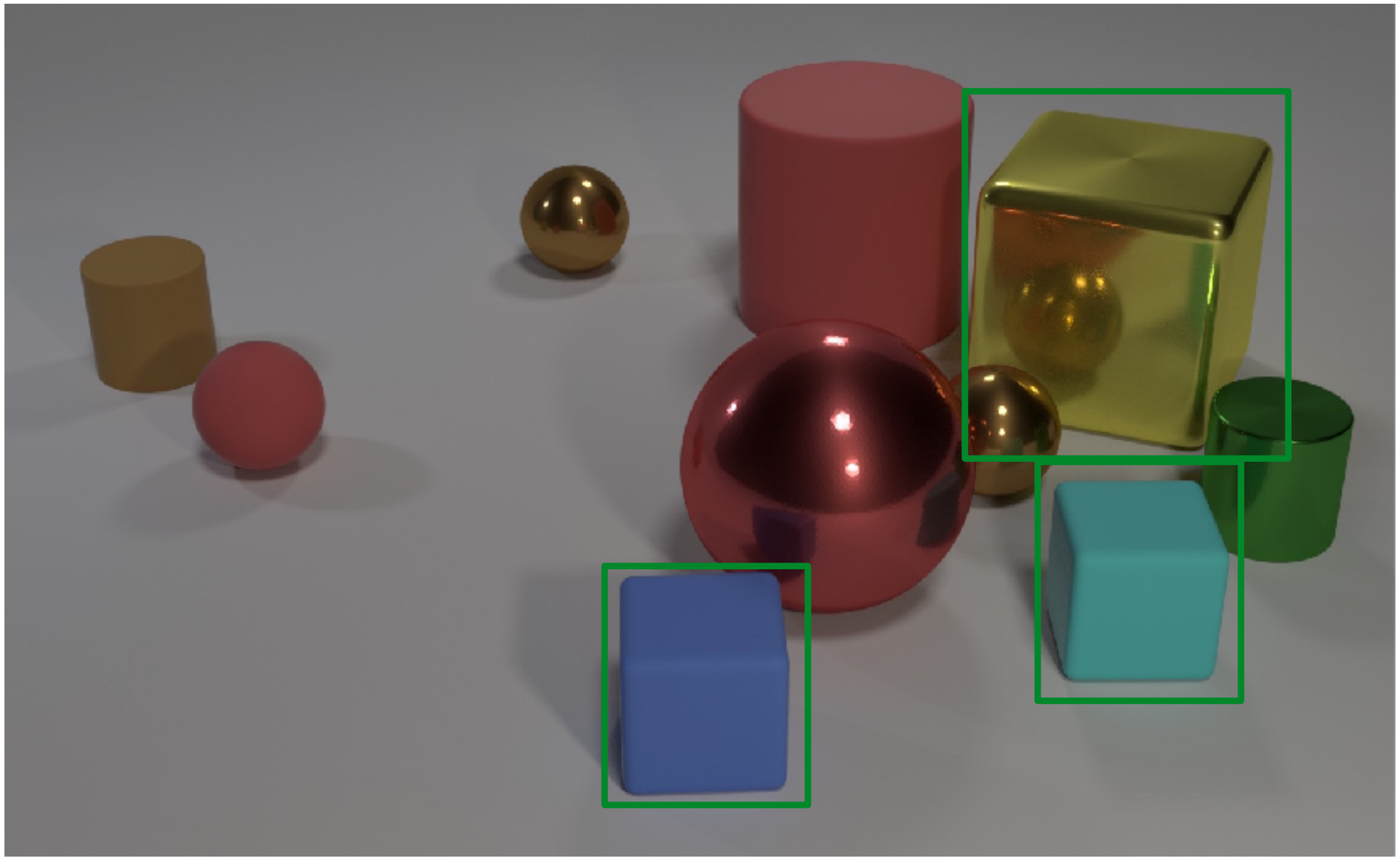
Visual Reasoning



What is the color of the cube to the left of the big metallic sphere?

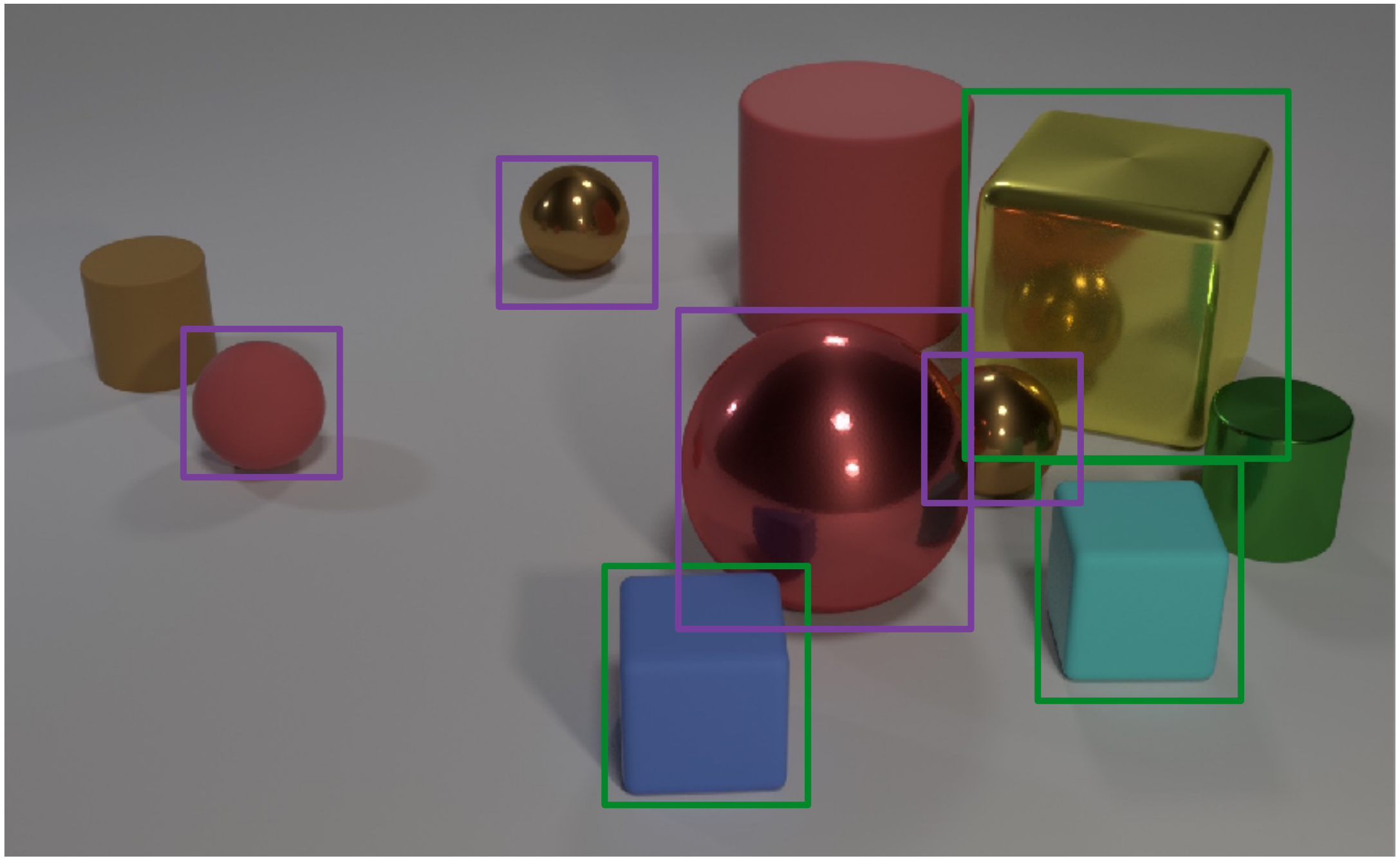
J. Johnson et. al. “CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning ”

Visual Reasoning



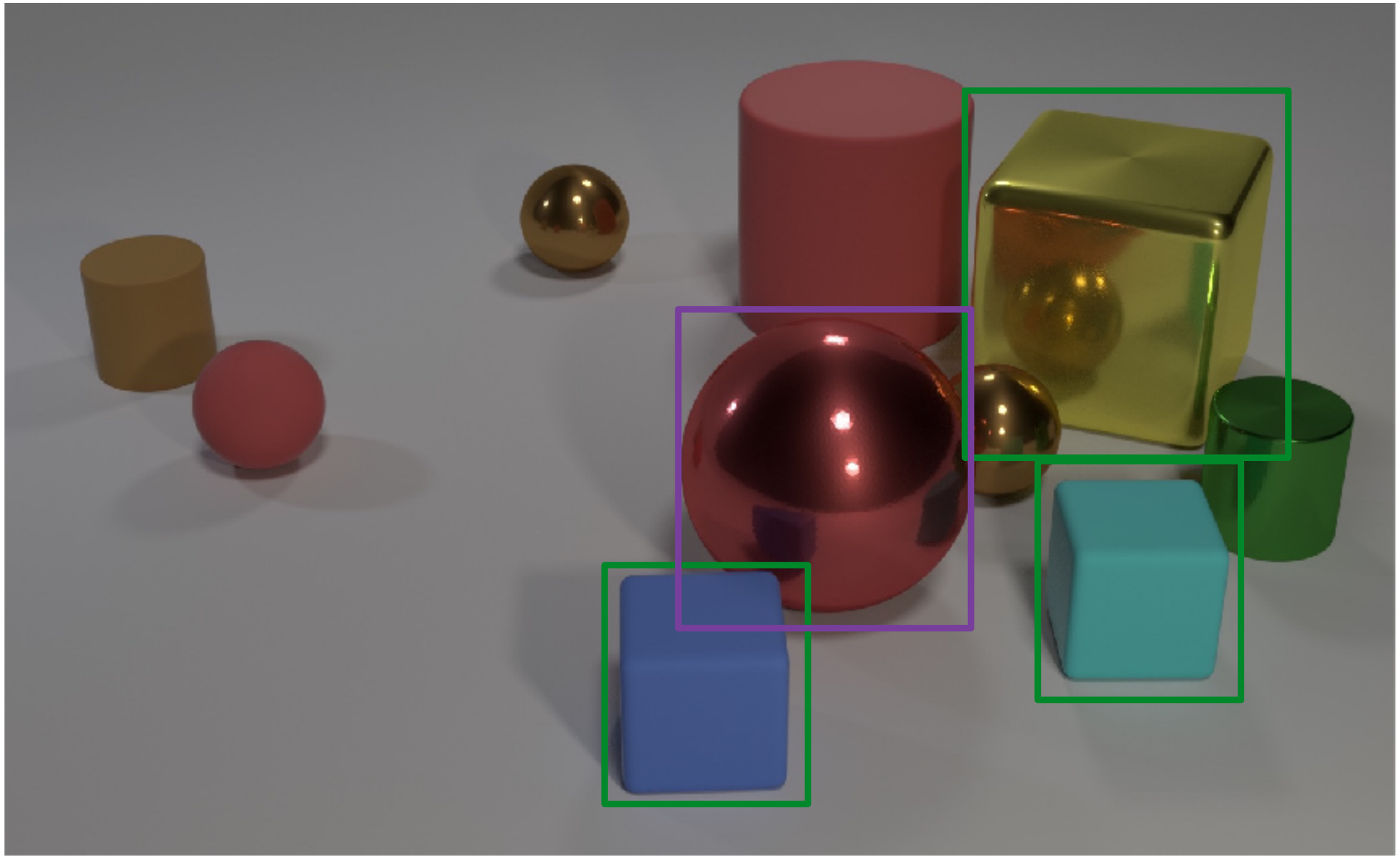
"What is the color of the cube to the left of the big metallic sphere?"

Visual Reasoning



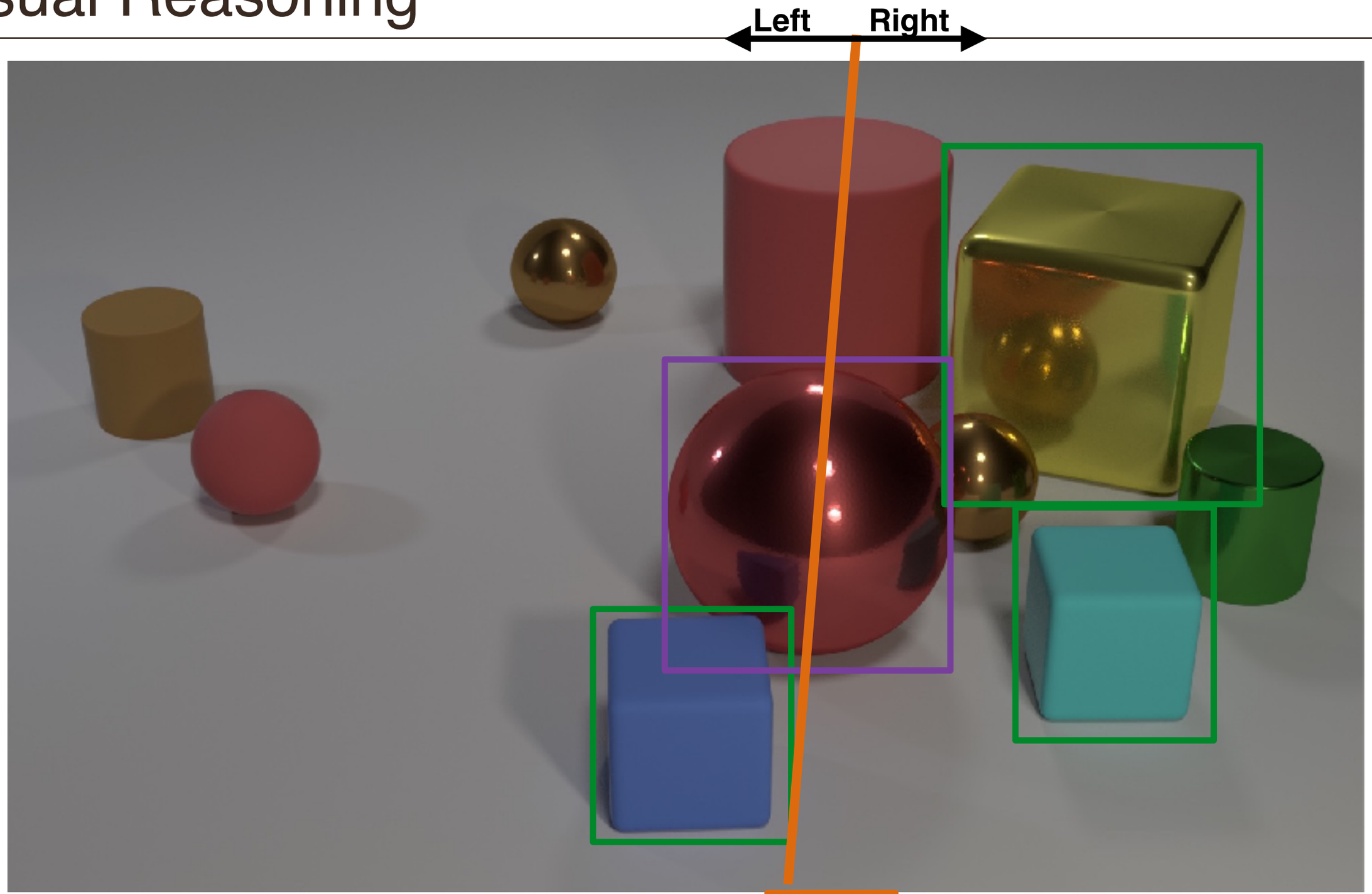
"What is the color of the cube to the left of the big metallic sphere?"

Visual Reasoning



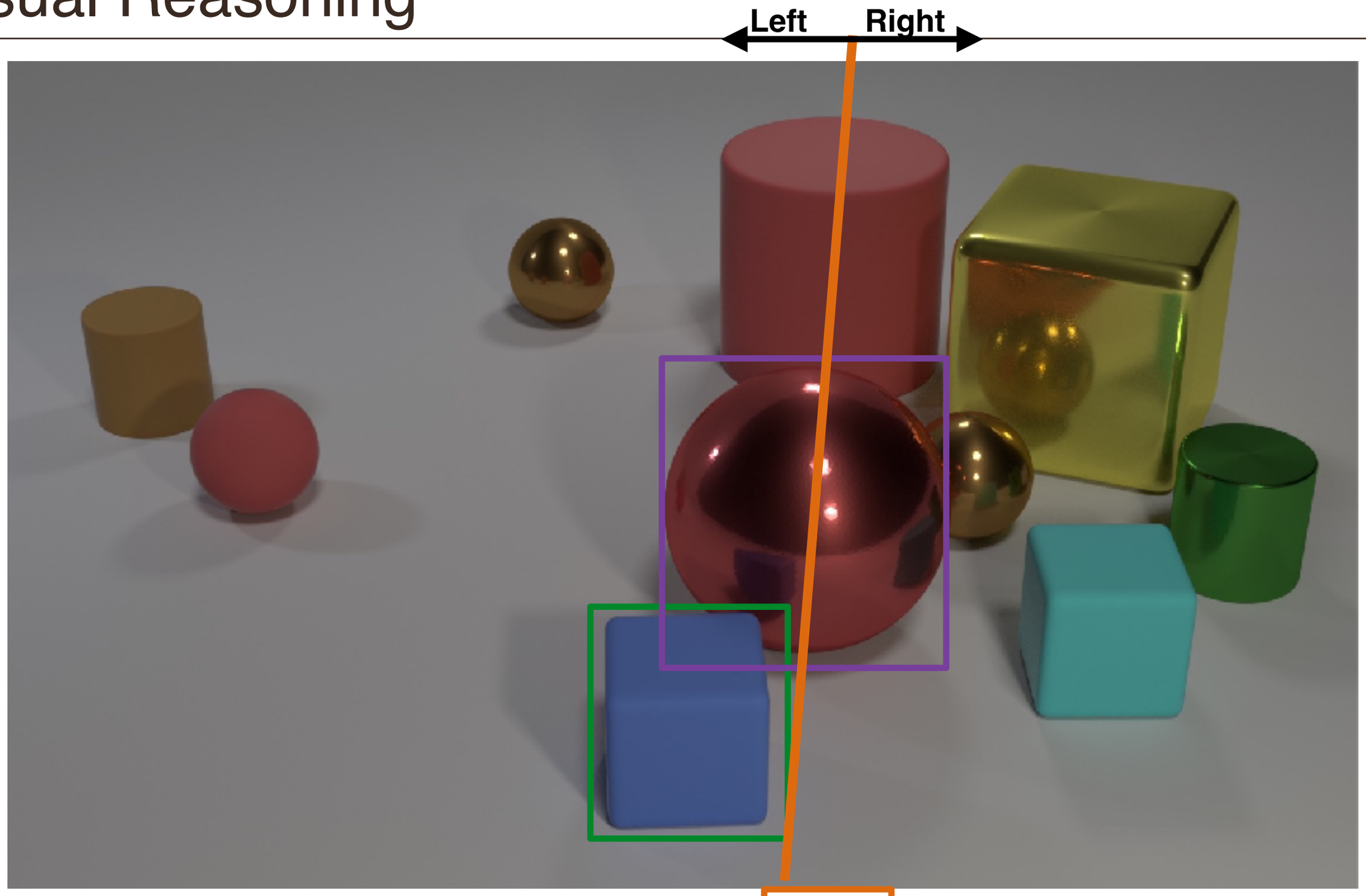
"What is the color of the cube to the left of the big metallic sphere?"

Visual Reasoning



"What is the color of the cube to the left of the big metallic sphere?"

Visual Reasoning

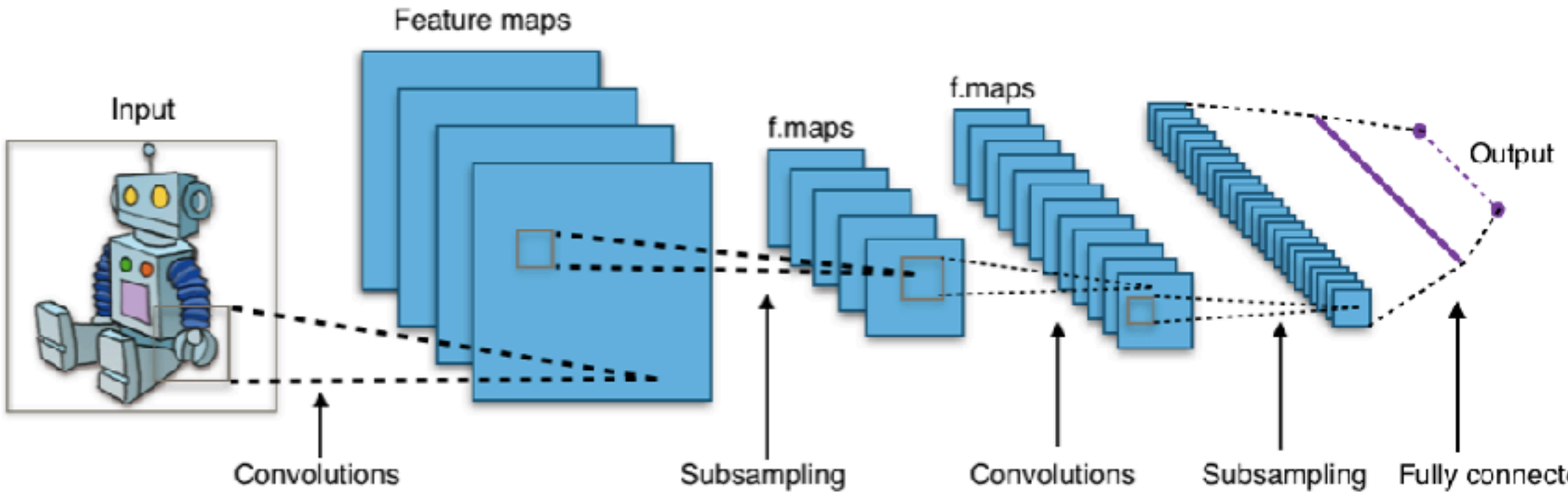


"What is the color of the cube to the left of the big metallic sphere?"

Plan

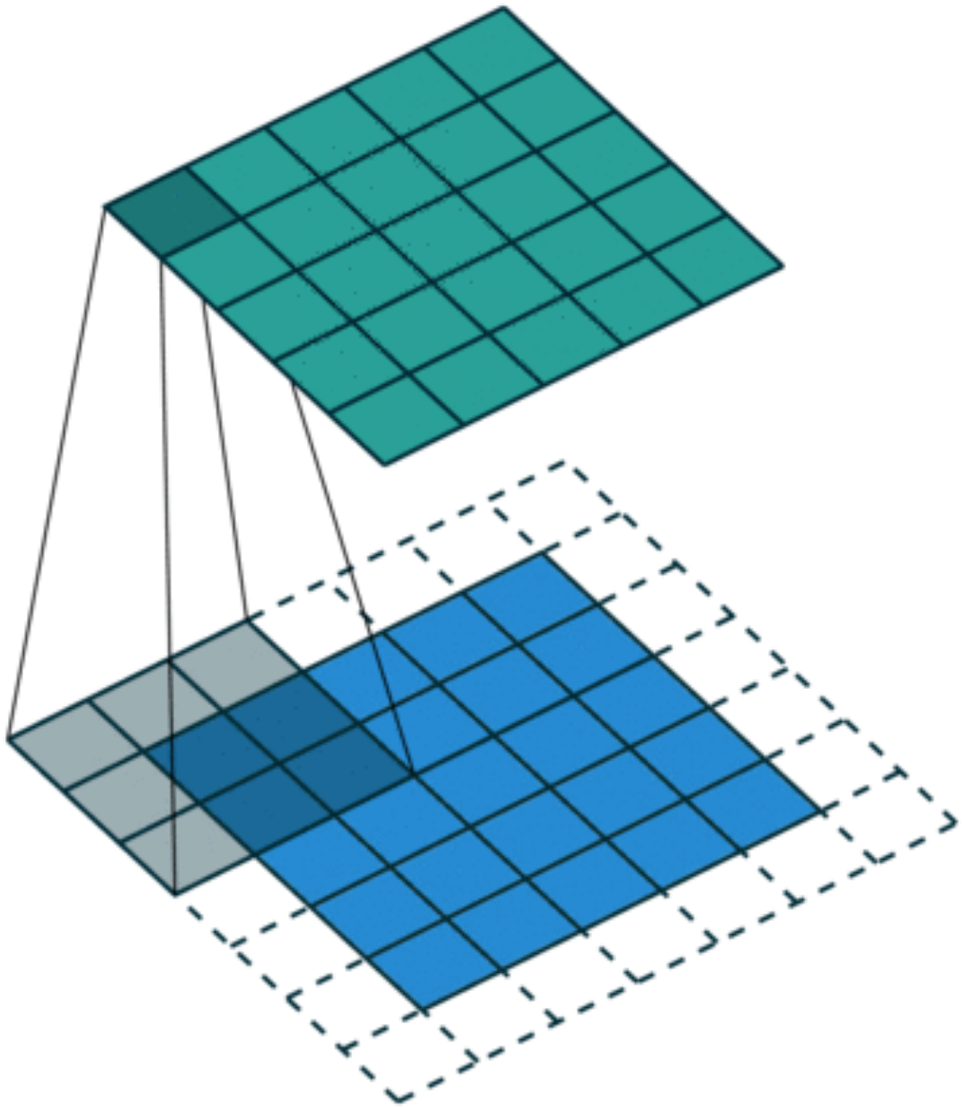
- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - **Early Visual Question Answering systems**
  - Non-local computations (Relation Nets, Transformer)
  - Graph Neural Networks
  - Soft-Attention & Hard-Attention in Computer Vision
  - Bias

CNNs



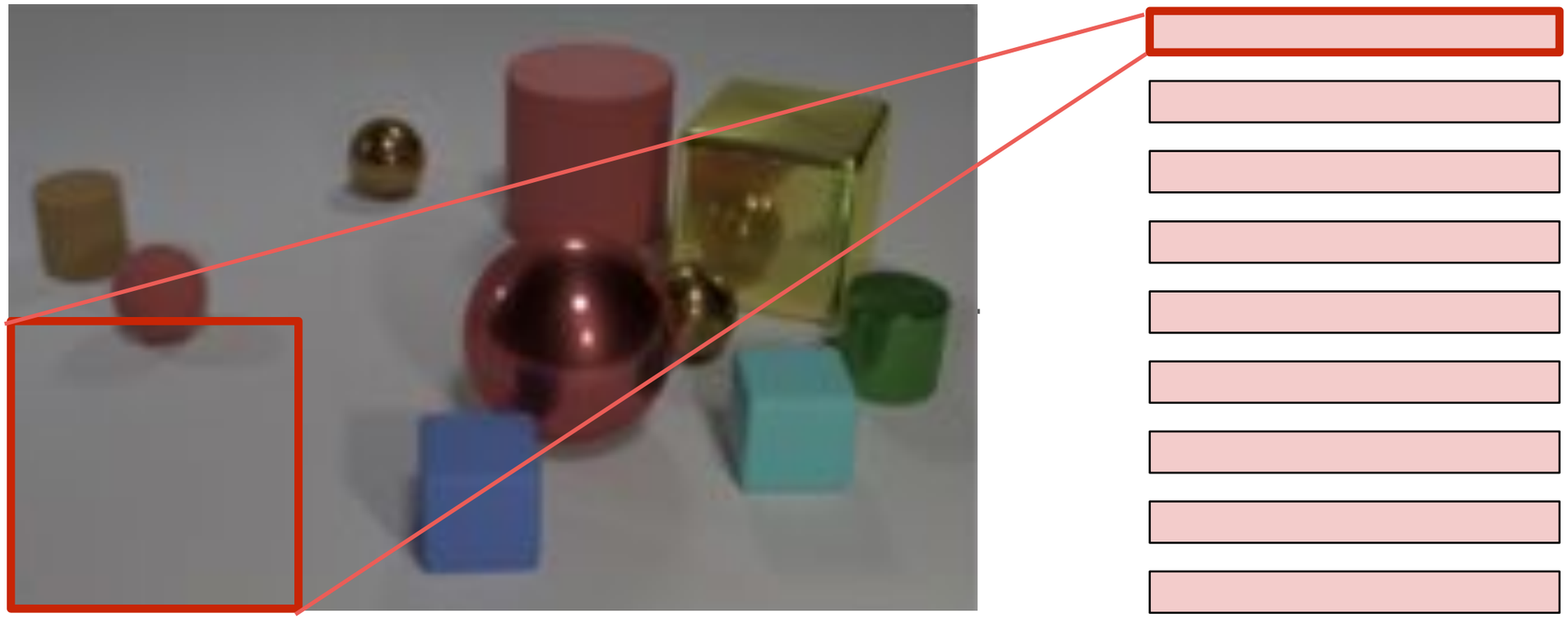
Wikipedia

CNNs

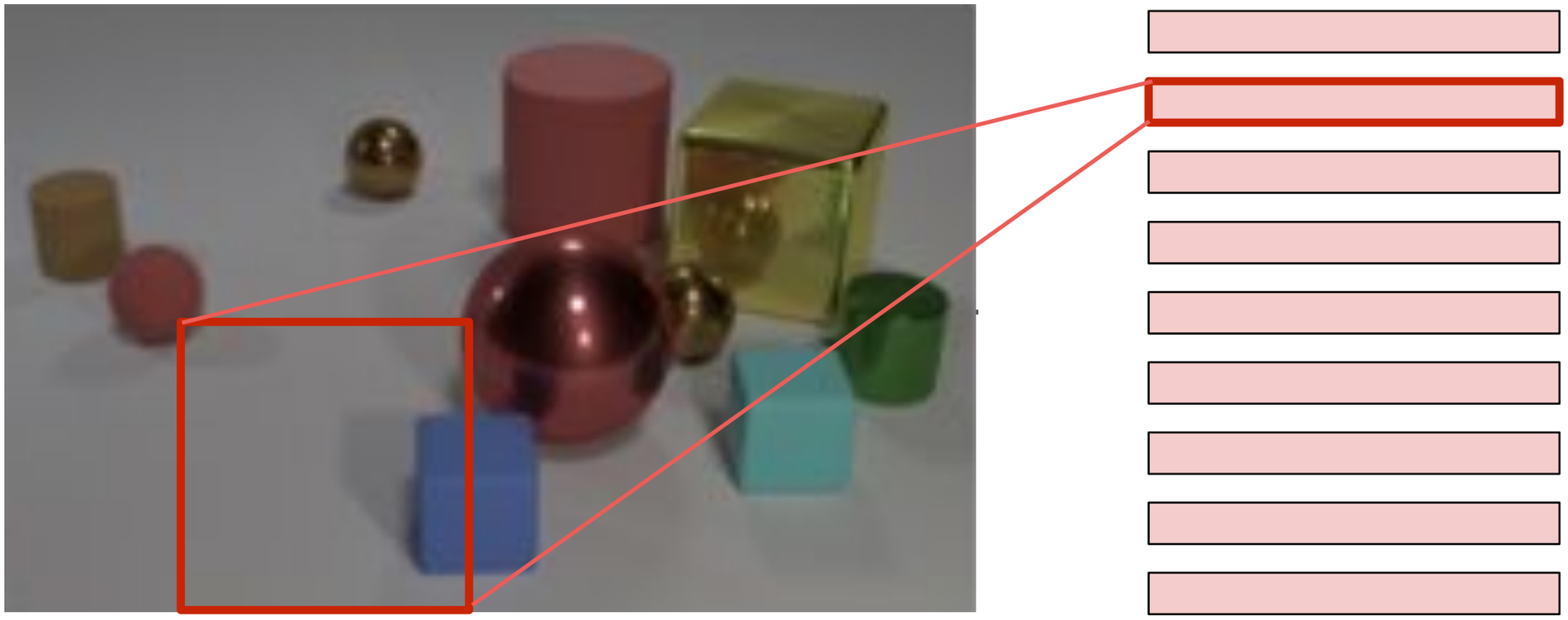


https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

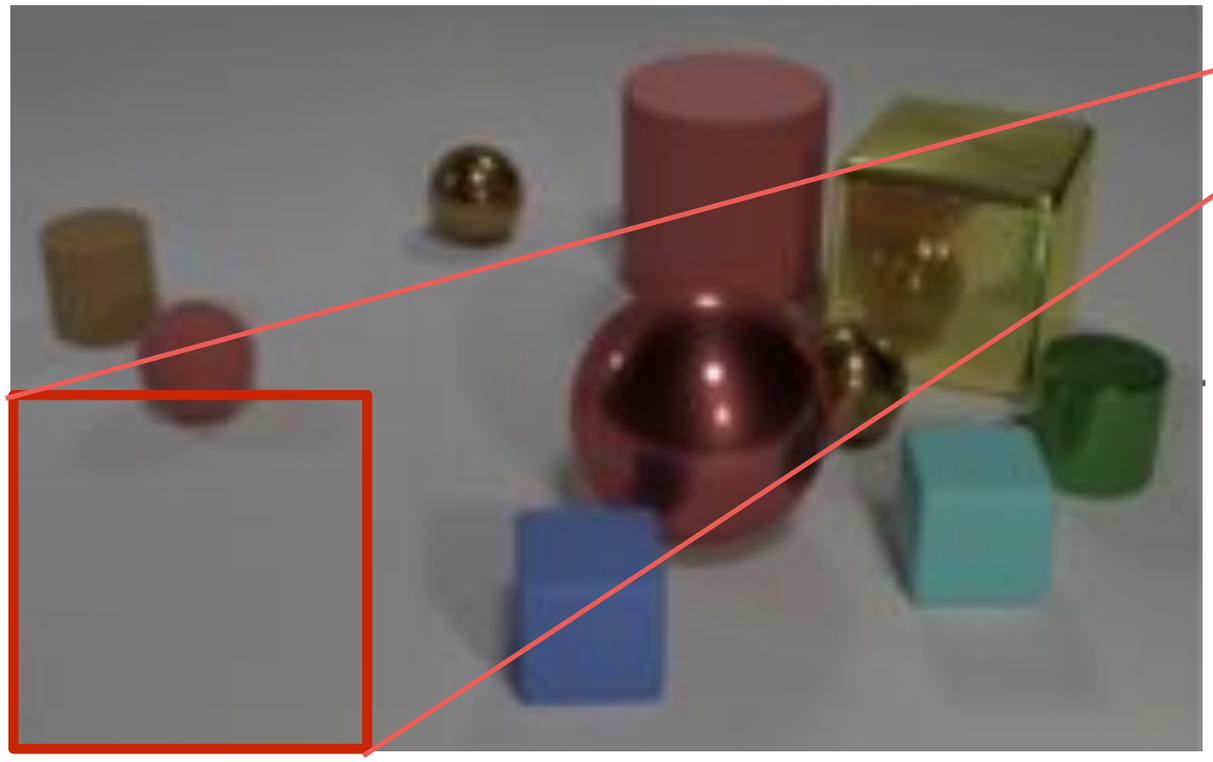
CNNs



CNNs



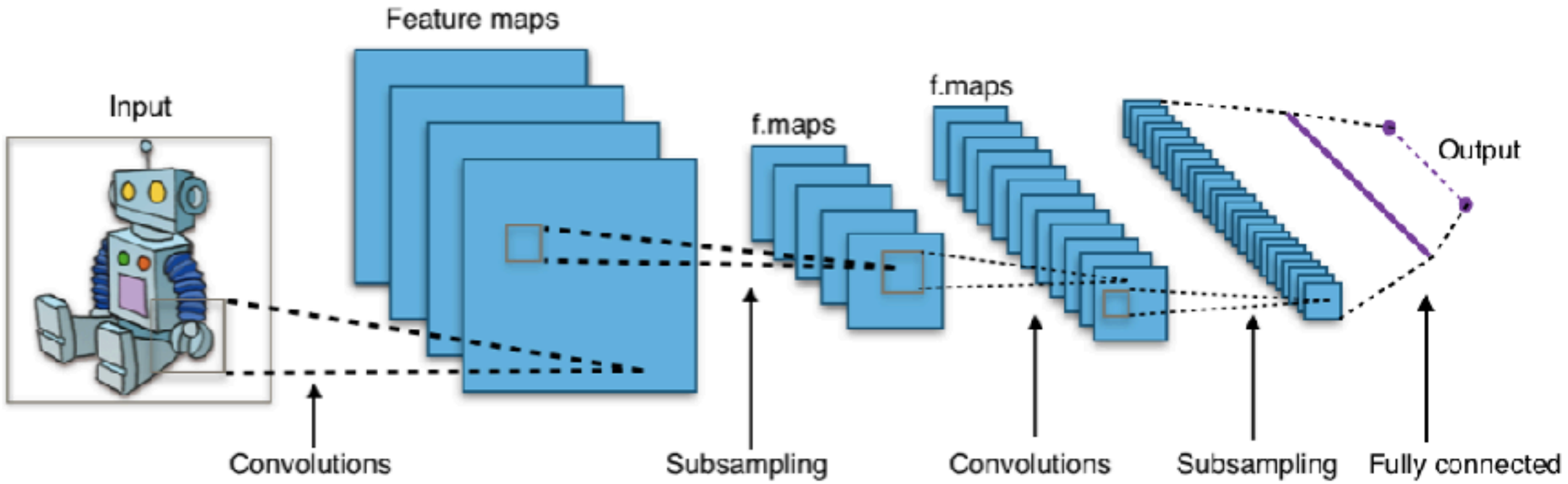
CNNs



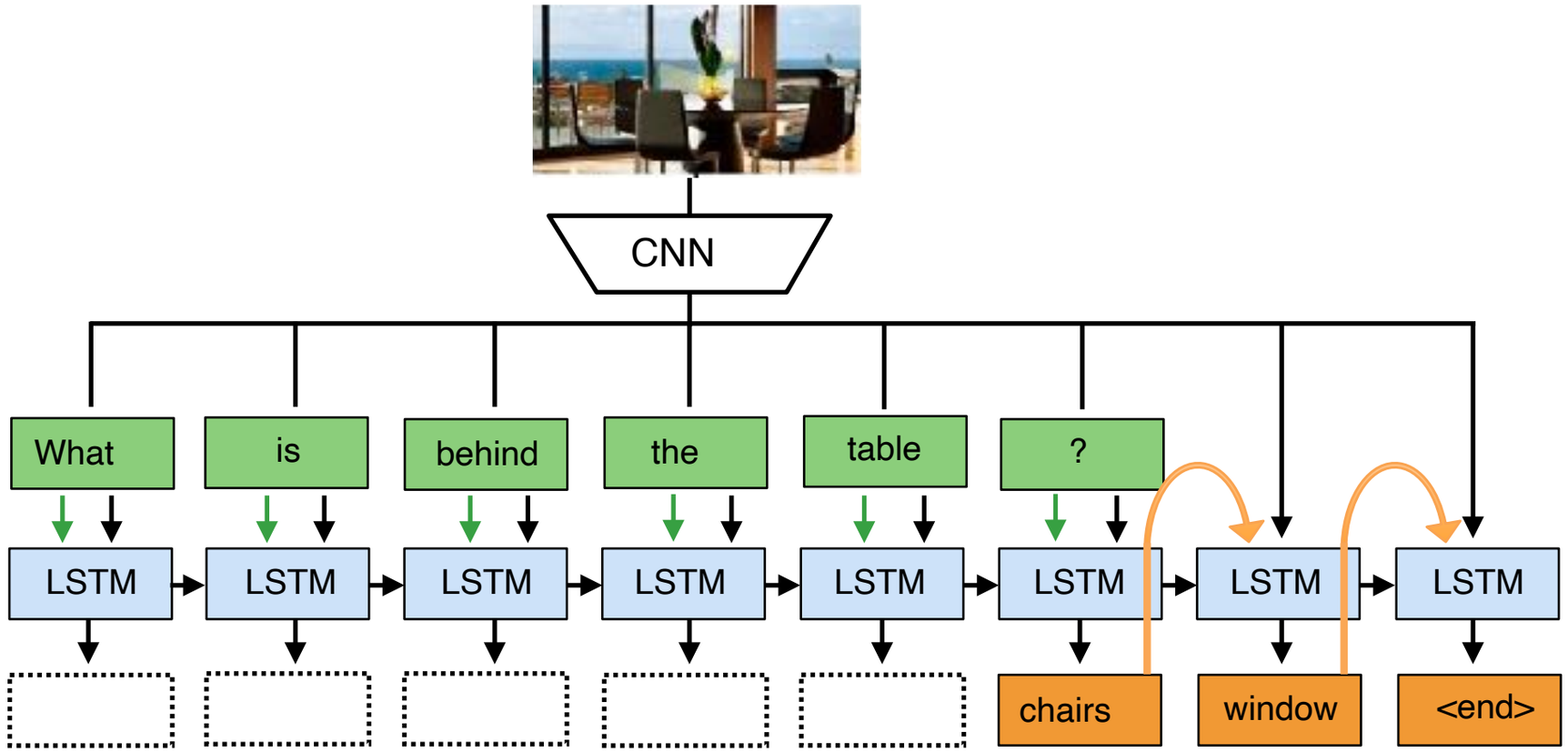
f_θ
f_θ
f_θ
f_θ
f_θ
f_θ
f_θ
f_θ
f_θ

CNNs

- Statistical efficiency
  - Weight sharing -> fewer parameters to learn
- Computational efficiency
  - Tiny kernels (weights) -> faster computations
- Efficient capacity
  - Hierarchical representations

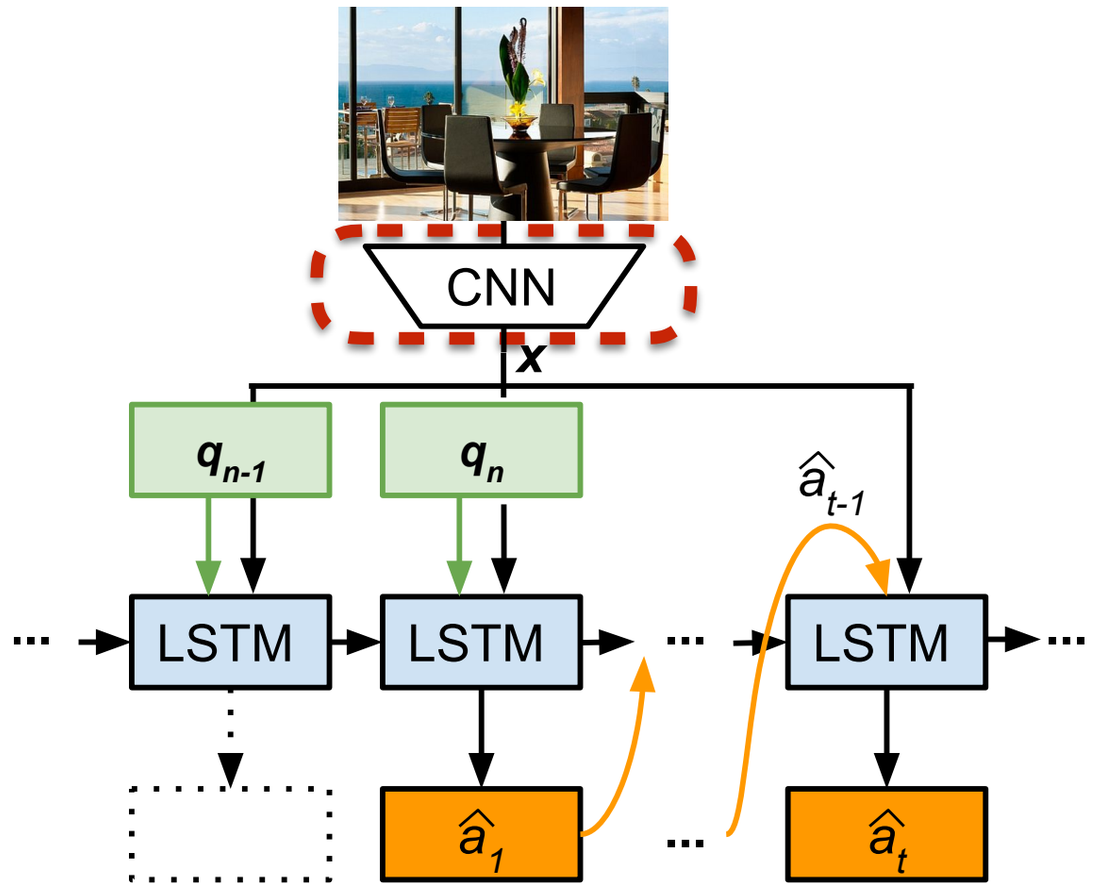


Early Visual Question Answering systems



M. Malinowski et. al. “Ask Your Neurons: A Neural-based Approach to Answering Questions about Images”

Early Visual Question Answering systems



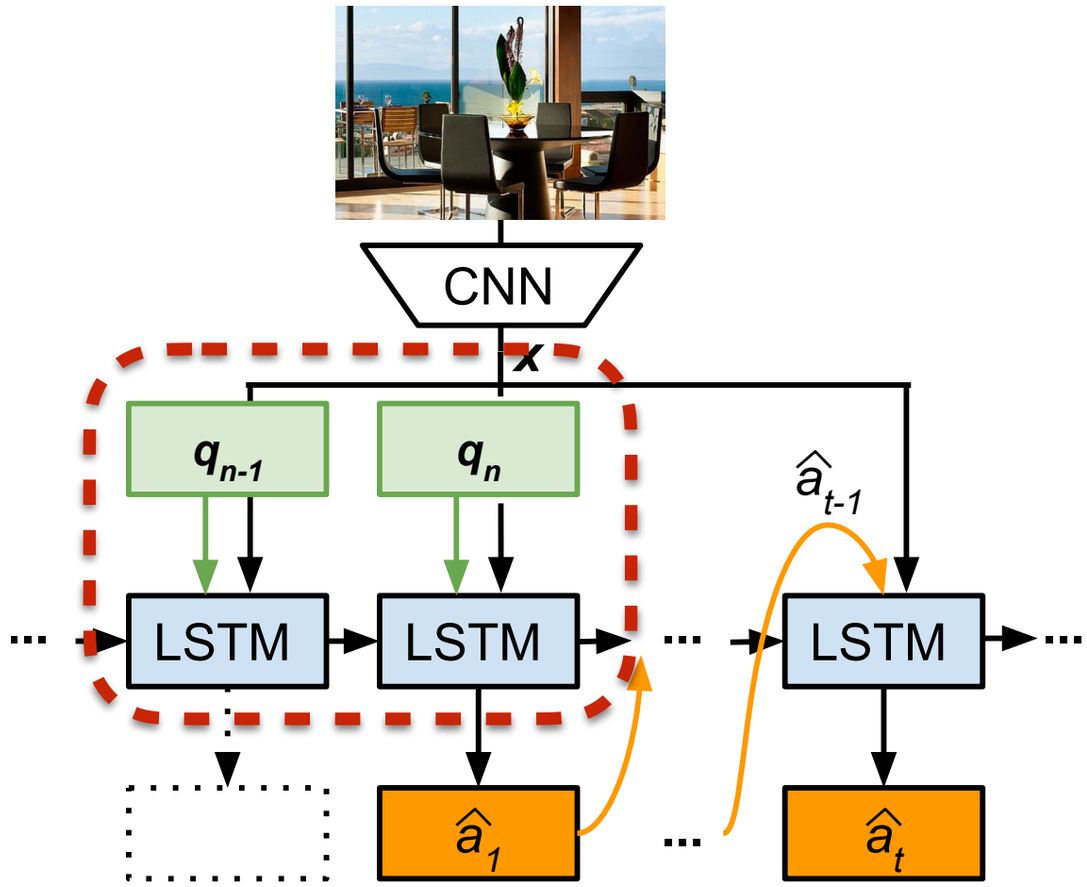
- Predicting answer sequence
  - Recursive formulation

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Early Visual Question Answering systems



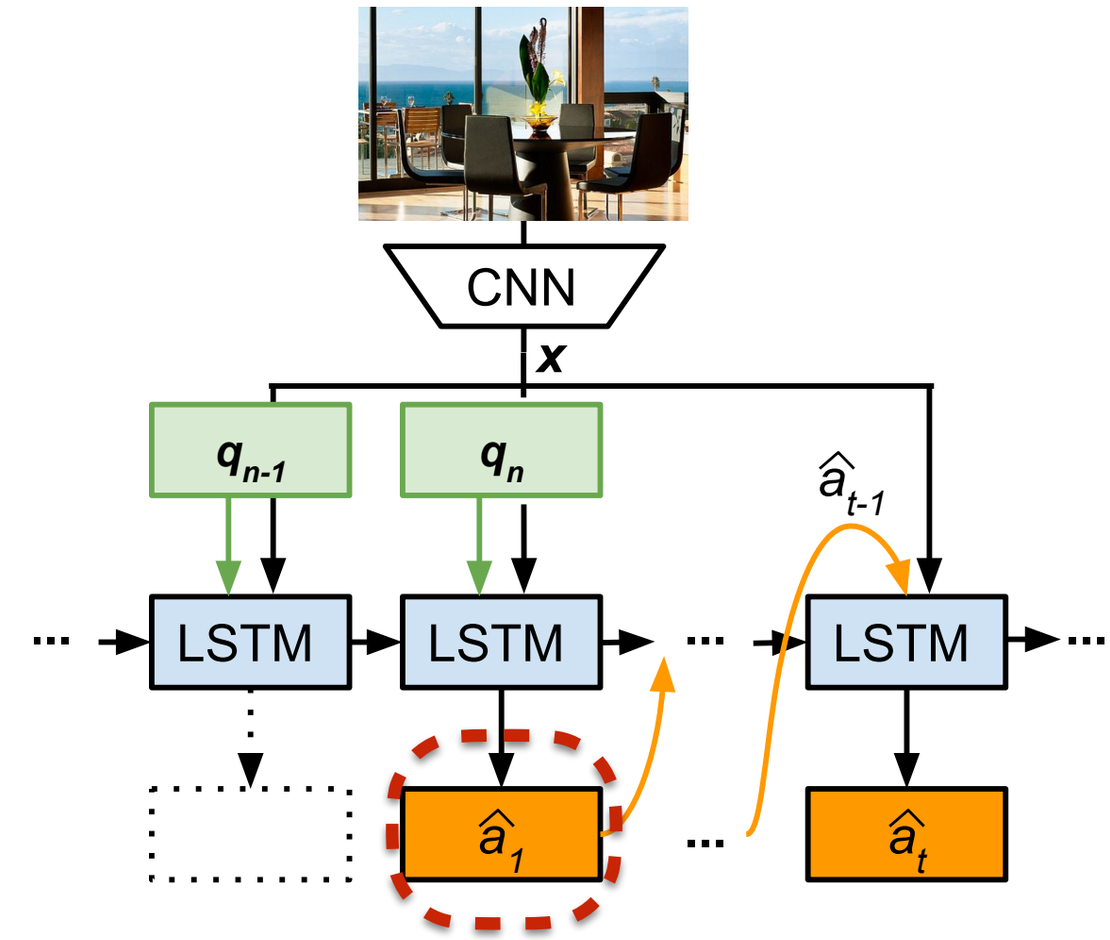
- Predicting answer sequence
  - Recursive formulation

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Early Visual Question Answering systems



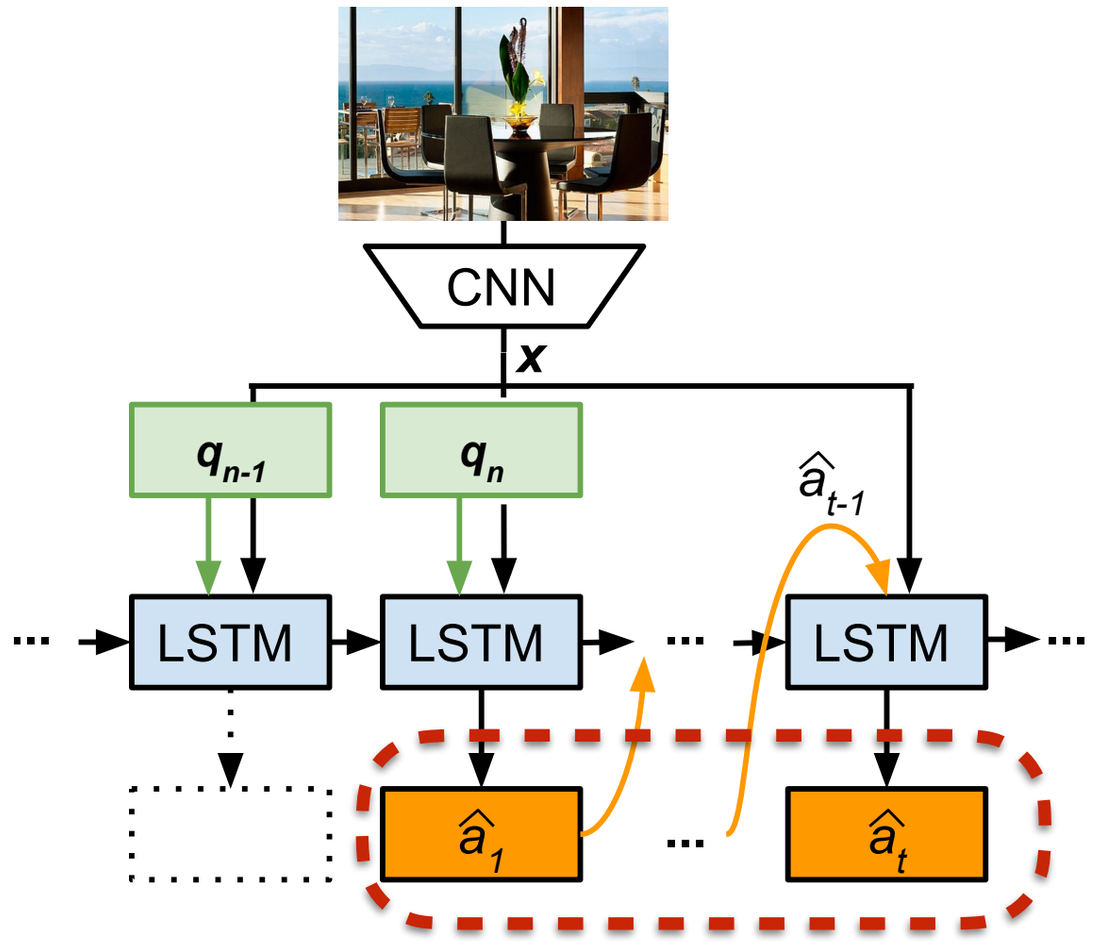
- Predicting answer sequence
  - Recursive formulation

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

Early Visual Question Answering systems



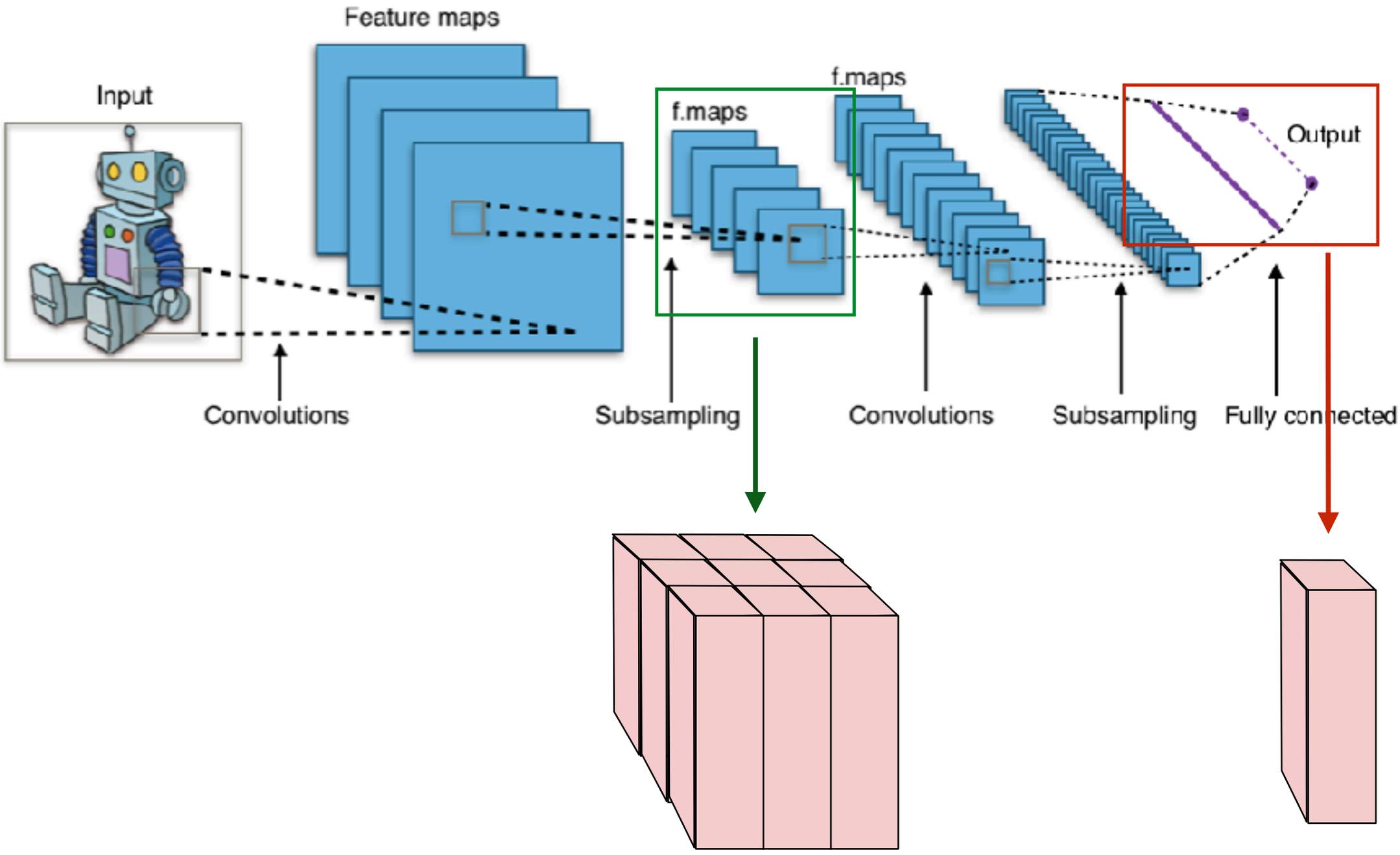
- Predicting answer sequence
  - Recursive formulation

$$\hat{\boldsymbol{a}}_t = \arg\max_{\boldsymbol{a}\in\mathcal{V}} p(\boldsymbol{a}|\boldsymbol{x}, \boldsymbol{q}, \hat{A}_{t-1}; \boldsymbol{\theta}),$$ $\boldsymbol{x}$ - image representation

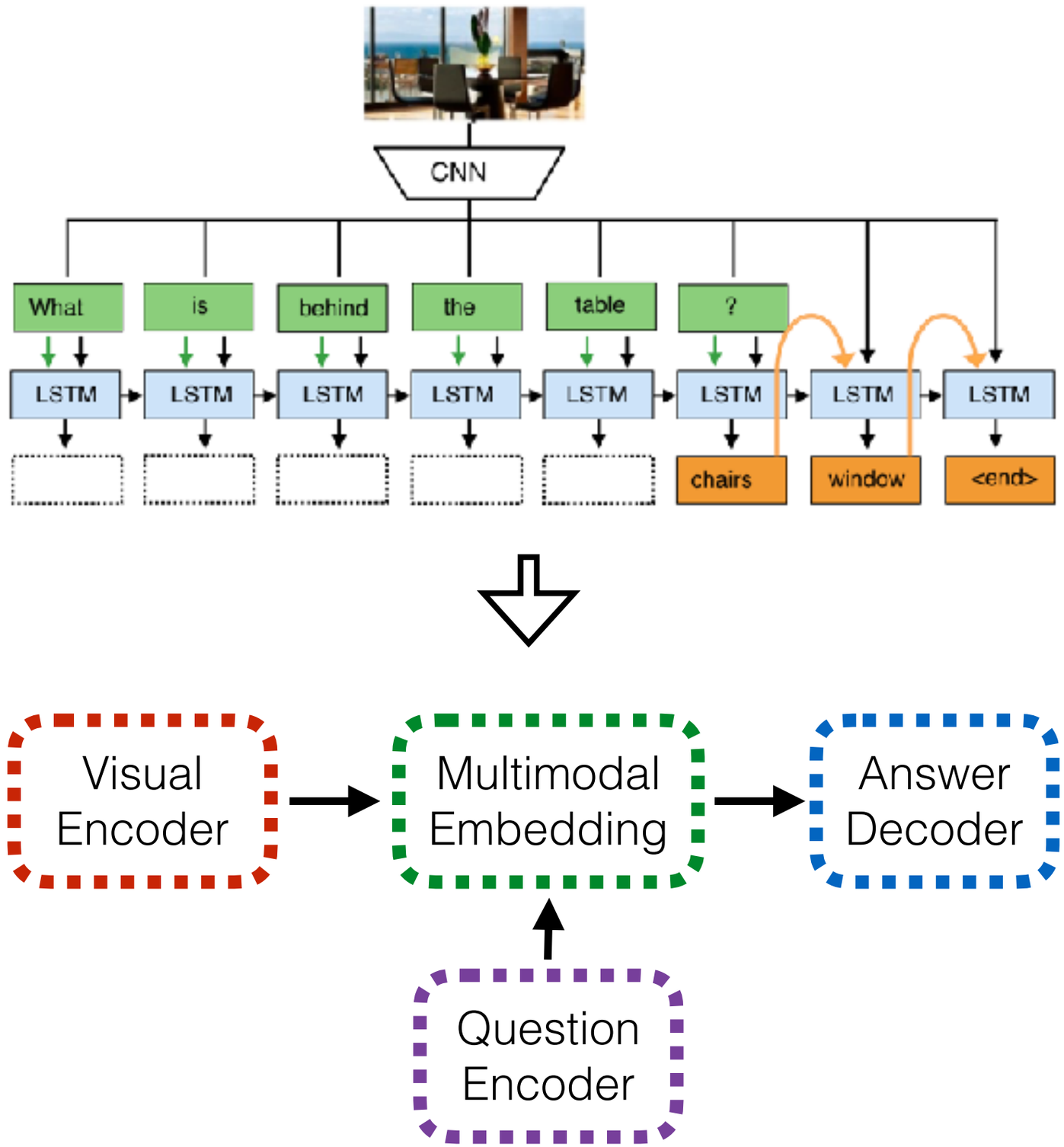
$$\boldsymbol{q} = \left[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}, [\![?]\!]\right],$$ $\boldsymbol{q}_j$ - question word index

$\mathcal{V}$ - vocabulary, $\hat{A}_{t-1} = \{\hat{\boldsymbol{a}}_1, \ldots, \hat{\boldsymbol{a}}_{\boldsymbol{t-1}}\}$ - previous answer words

What is the output of the CNNs?

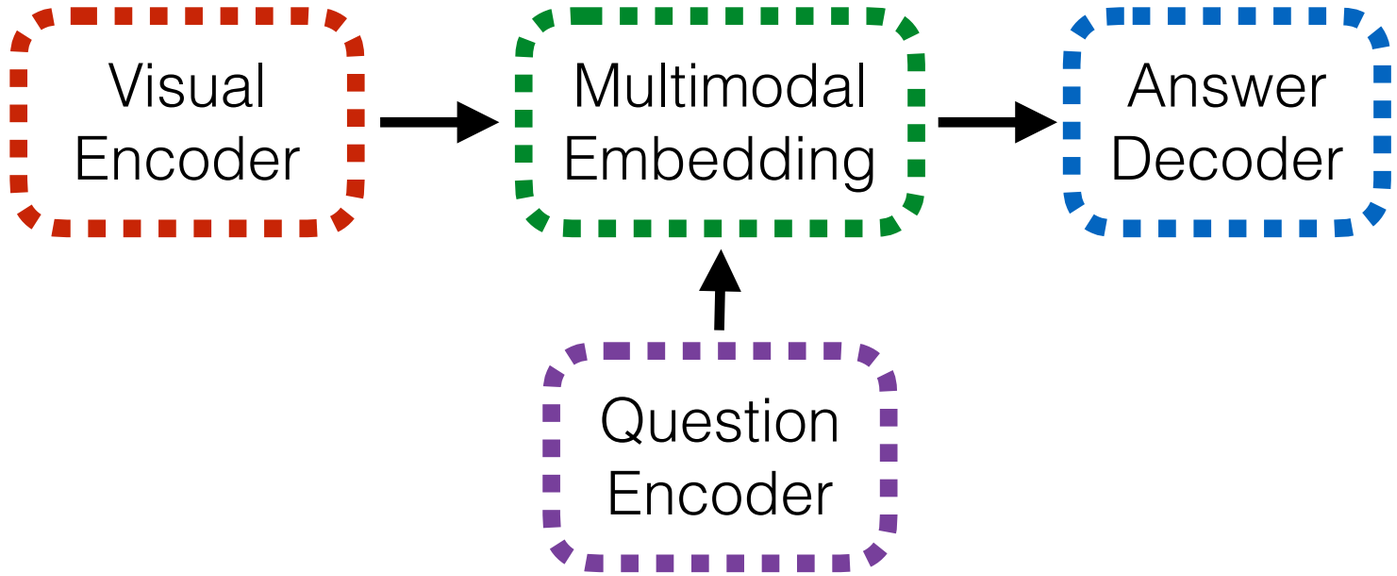


Early Visual Question Answering systems



Early Visual Question Answering systems

- AlexNet
- GoogLeNet
- VGG-19
- ResNet-152
- VIT



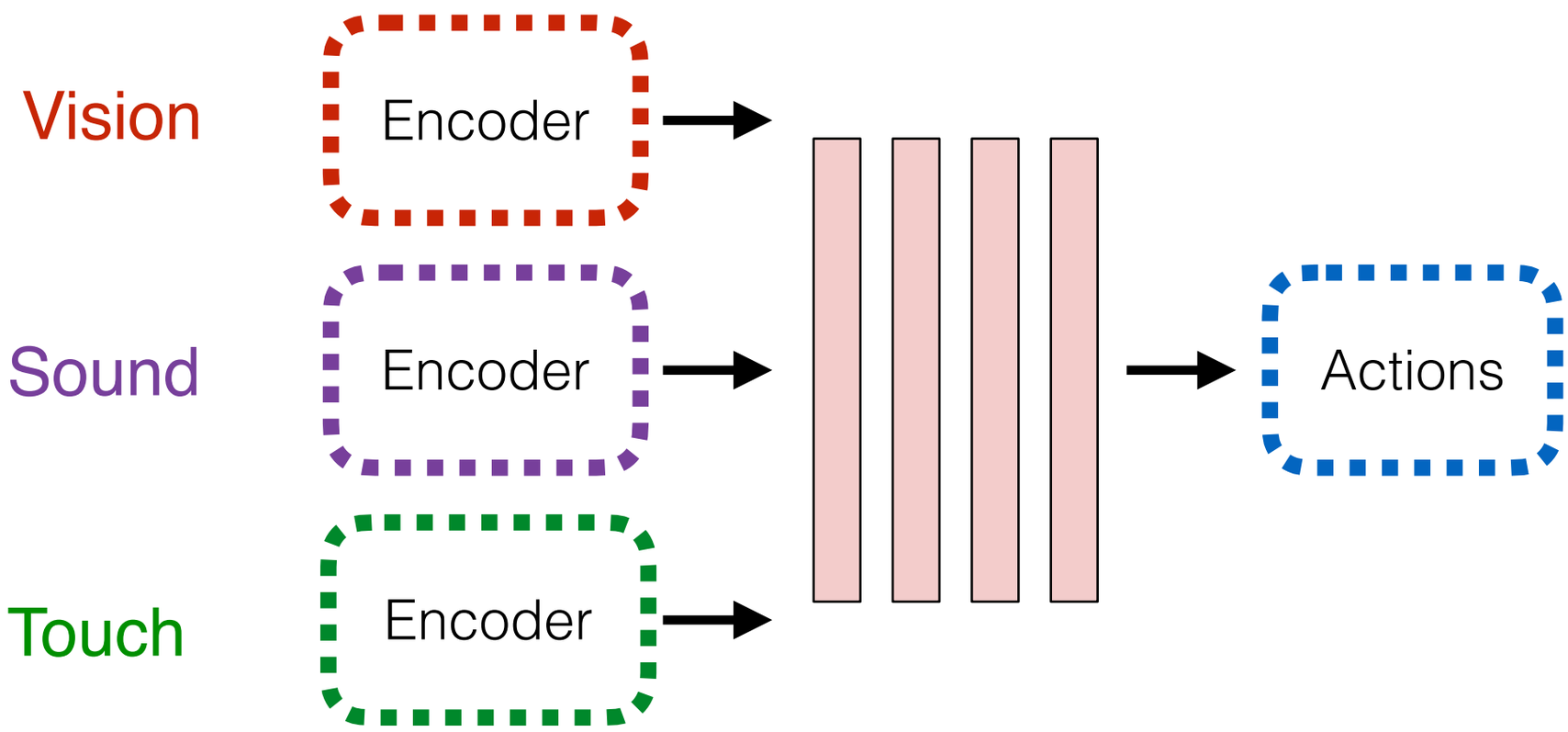
- Multiply
- Sum
- Concat
- Cross-Attention

- BOW
- CNN
- LSTM
- GRU
- Transformer

- Generation
- Classification with top frequent answers

Vectors as the “common currency”

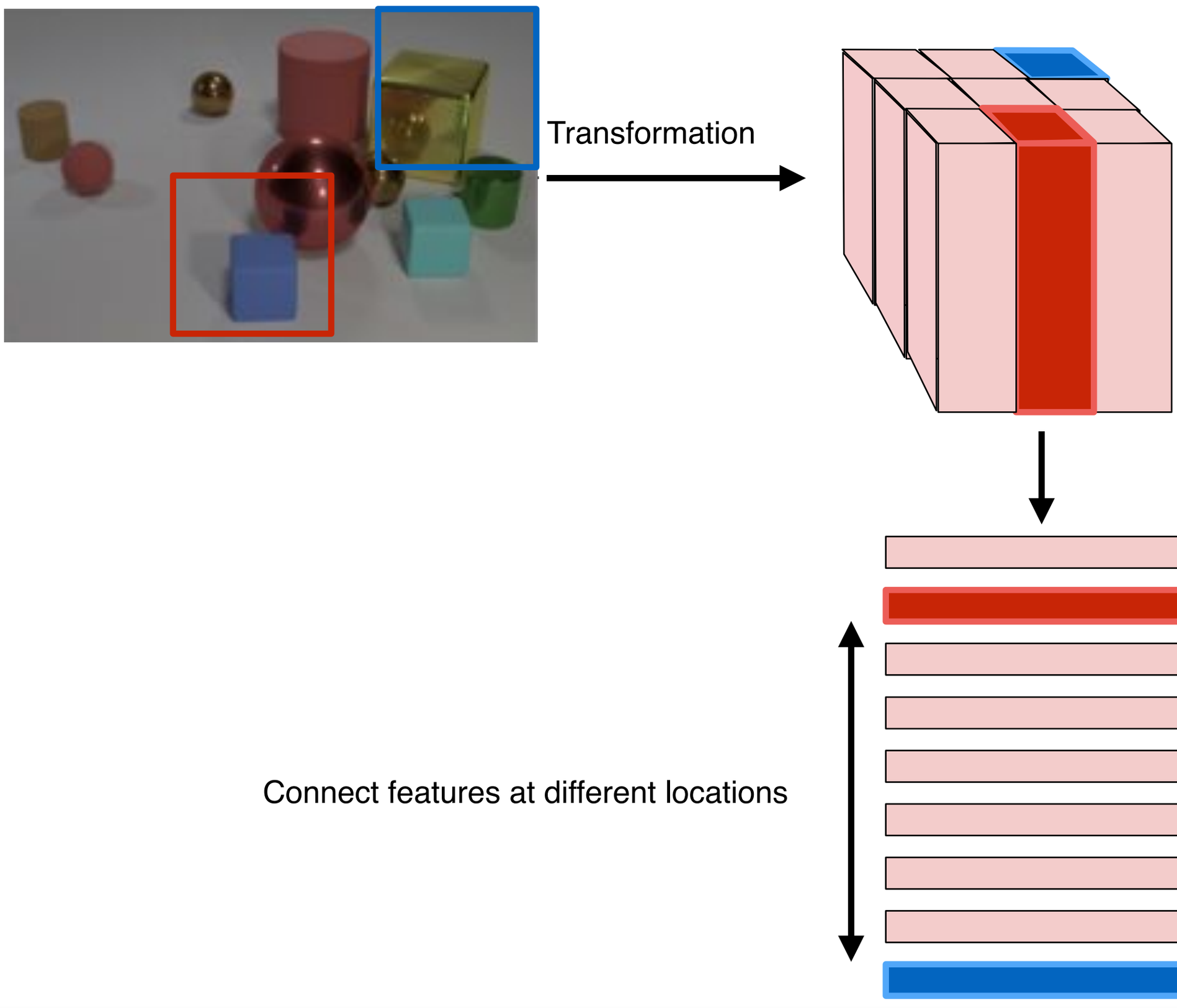
- Deep Learning
  - Transforms inputs into vectors
- Brain
  - Transforms inputs into electrochemical signals



Plan

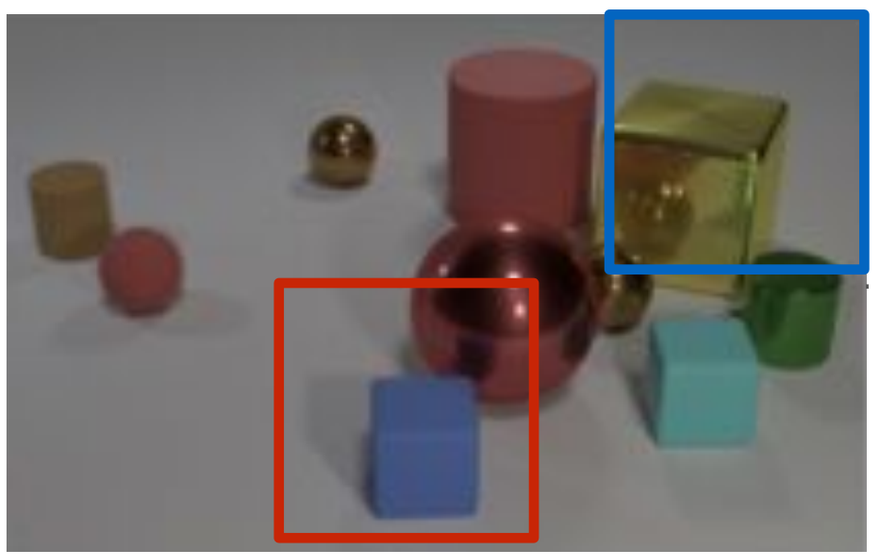
- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - Early Visual Question Answering systems
  - **Non-local computations (Relation Nets, Transformer)**
  - Graph Neural Networks
  - Soft-Attention & Hard-Attention in Computer Vision
  - Bias

Why non-local computations?

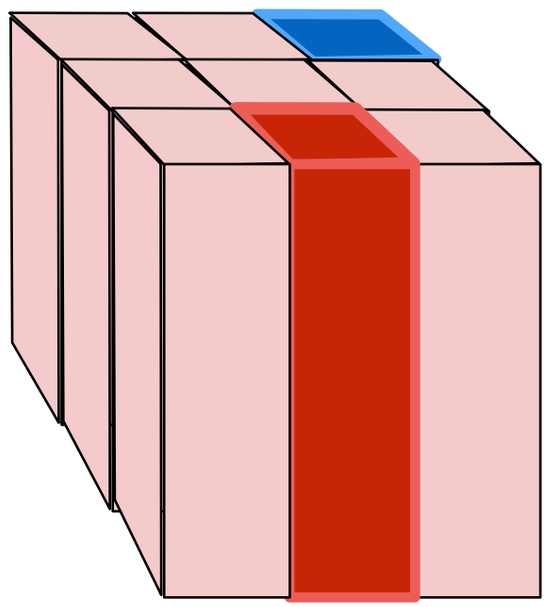


Connect features at different locations

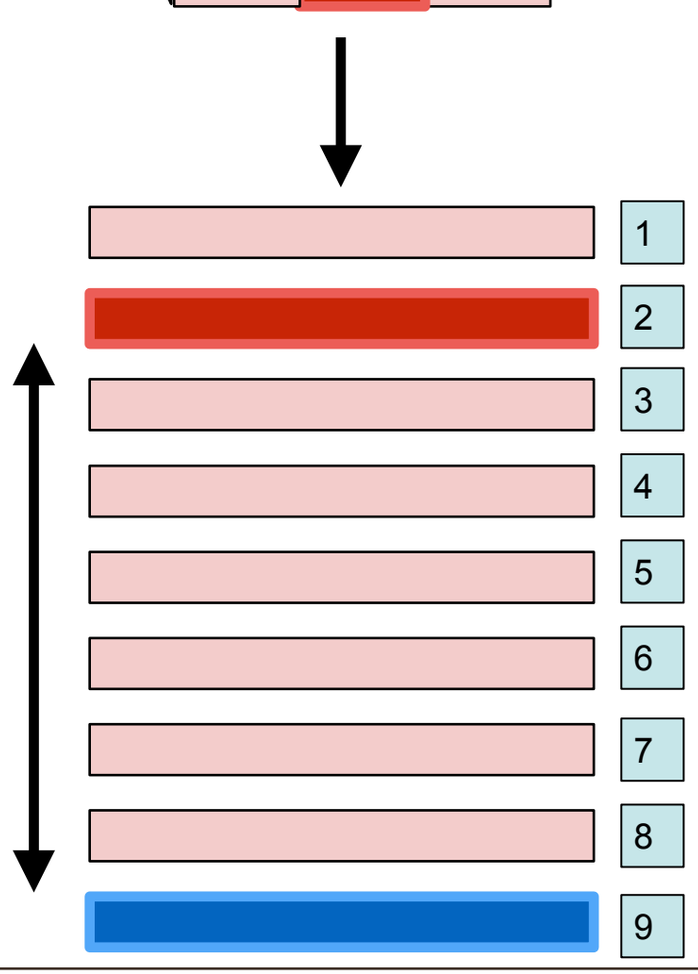
Why non-local computations?



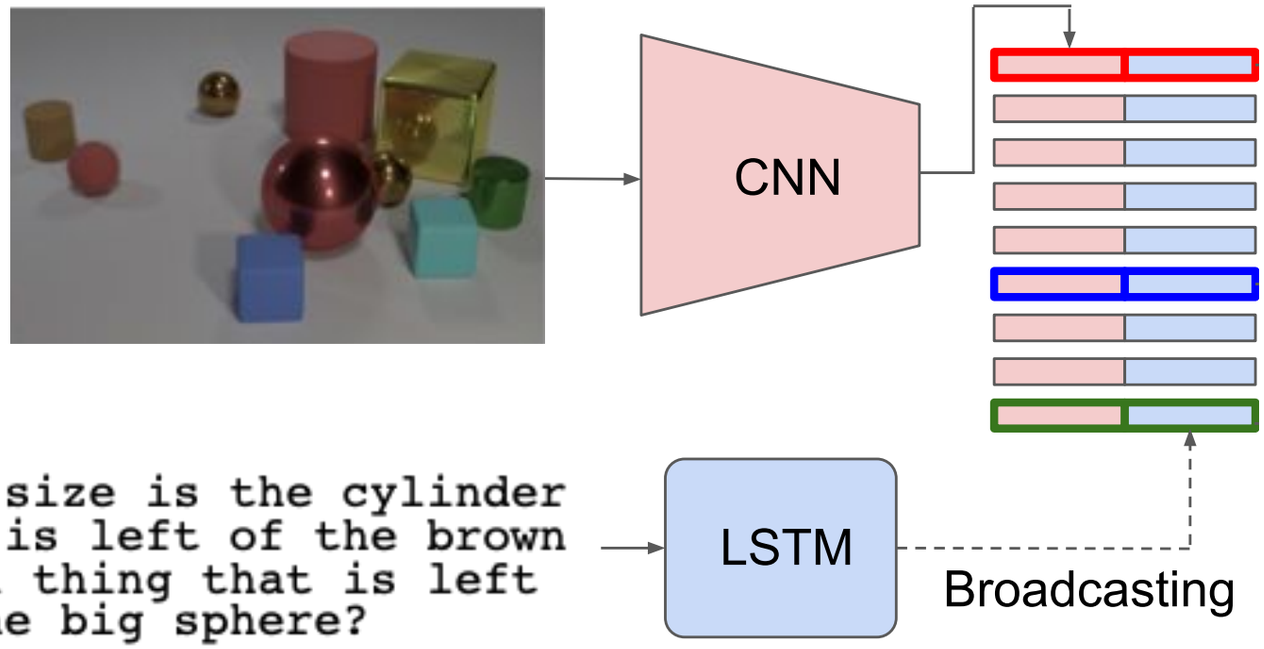
Transformation



Connect features at different locations

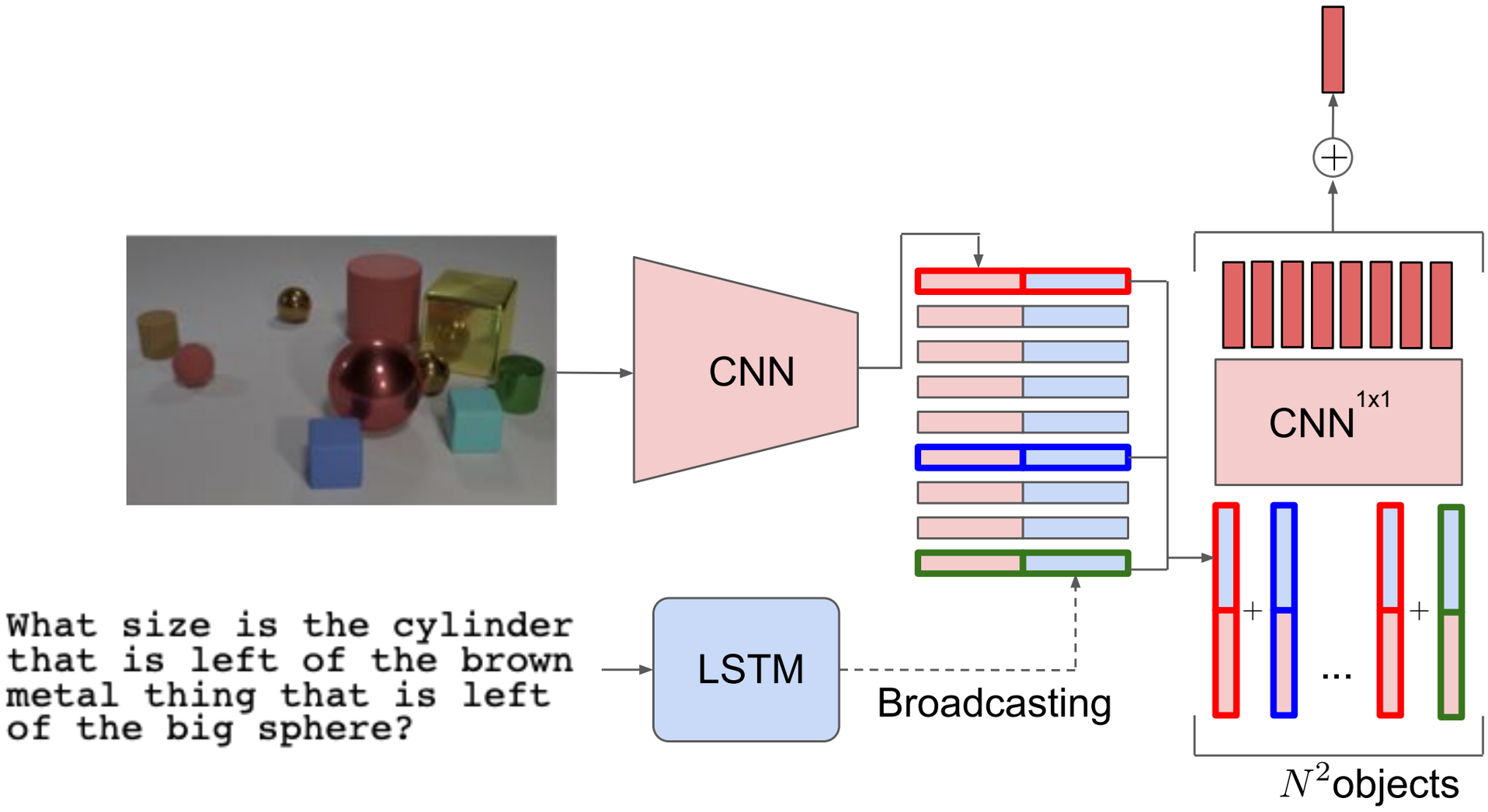


Relation Nets



A. Santoro, D. Raposo et. al. “A simple neural network module for relational reasoning”. NeurIPS’17

Relation Nets



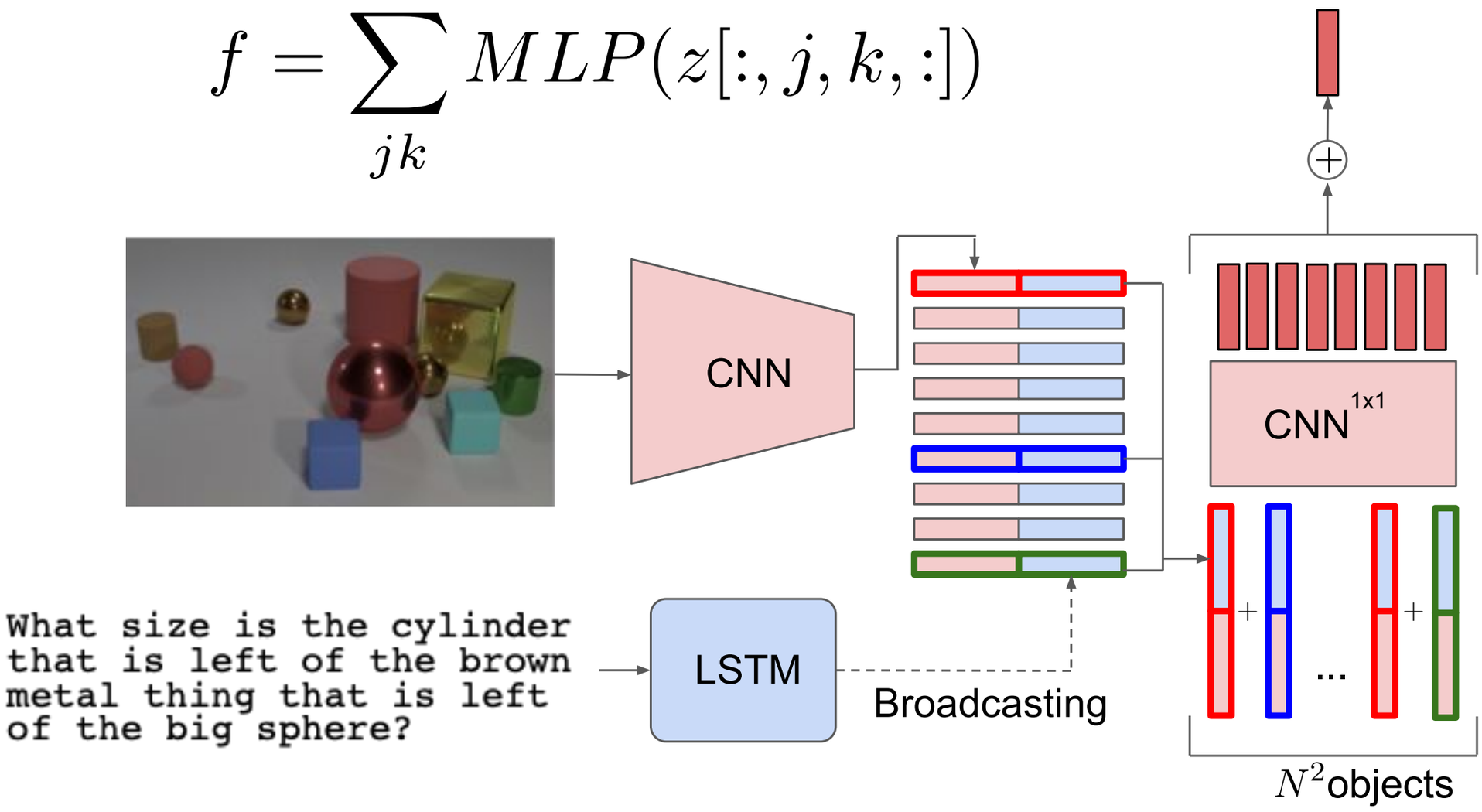
A. Santoro, D. Raposo et. al. “A simple neural network module for relational reasoning”. NeurIPS’17

Relation Nets

$$object \in \mathbb{R}[b, s, f]$$

$$z = object[:, :, None, :] + object[:, None, :, :]$$

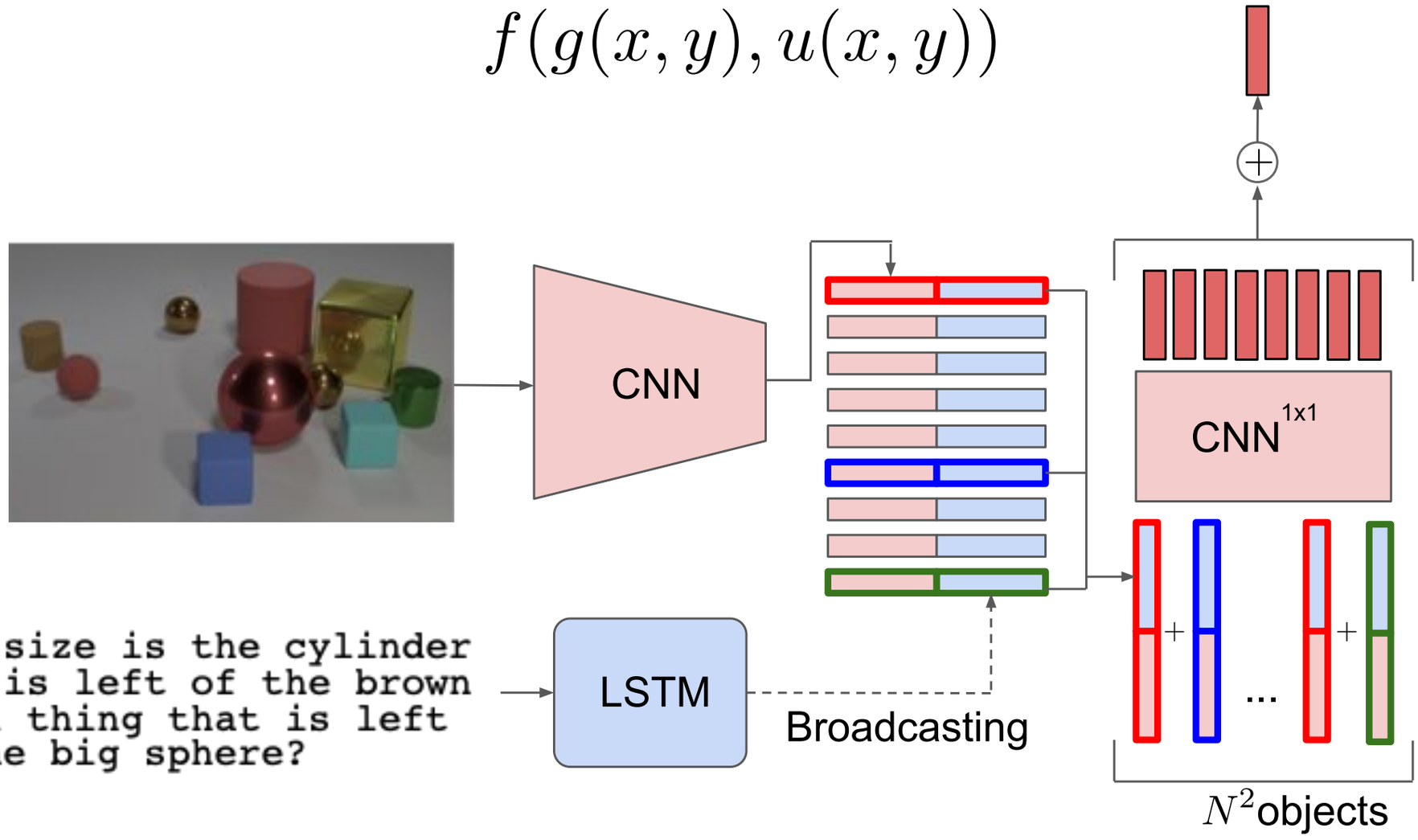
$$f = \sum_{jk} MLP(z[:, j, k, :])$$



A. Santoro, D. Raposo et. al. “A simple neural network module for relational reasoning”. NeurIPS’17

Relation Nets

$$f(g(x,y),u(x,y))$$



A. Santoro, D. Raposo et. al. “A simple neural network module for relational reasoning”. NeurIPS’17

Relation Nets and Transformer

$$f(g(x,y),u(x,y))$$

Relation Networks

$$g(x,y) = \mathrm{MLP}(x,y;\theta)$$

Transformer

$$g(x,y) = \mathrm{softmax}((W^q x)^T (W^k y))$$
$$u(x,y) = W^v y$$
$$f(\alpha,x) = \alpha x$$

[1] A. Santoro, D. Raposo et. al. “A simple neural network module for relational reasoning”. NeurIPS’17
[2] A. Vaswani et al. “Attention Is All You Need”. NeurIPS’17

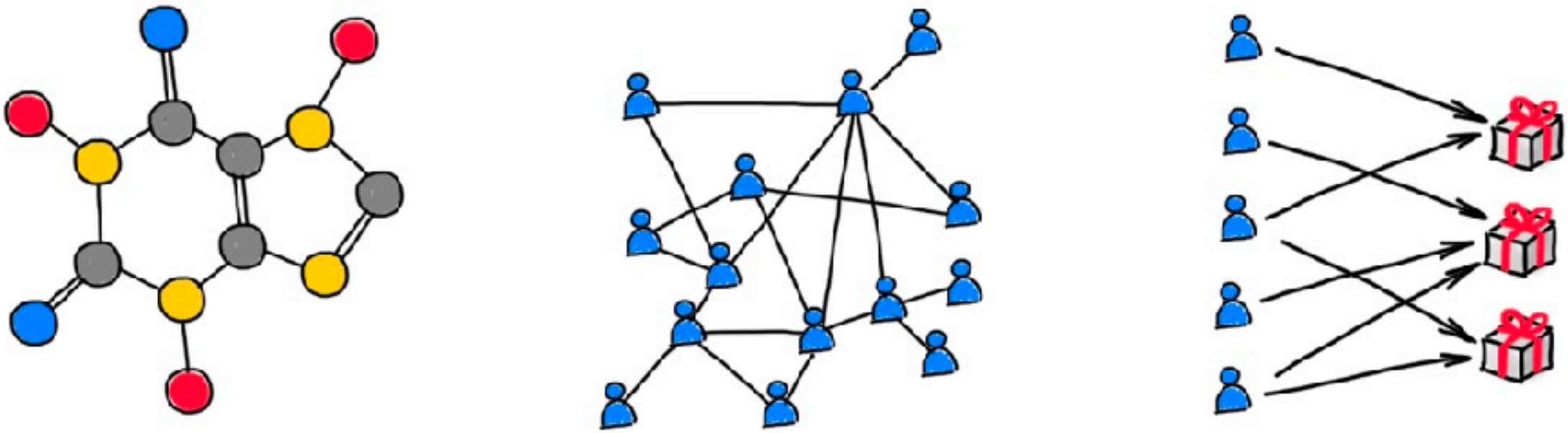
Relation Nets: Results

RN [1]	95.5
Human [2]	92.6

Plan

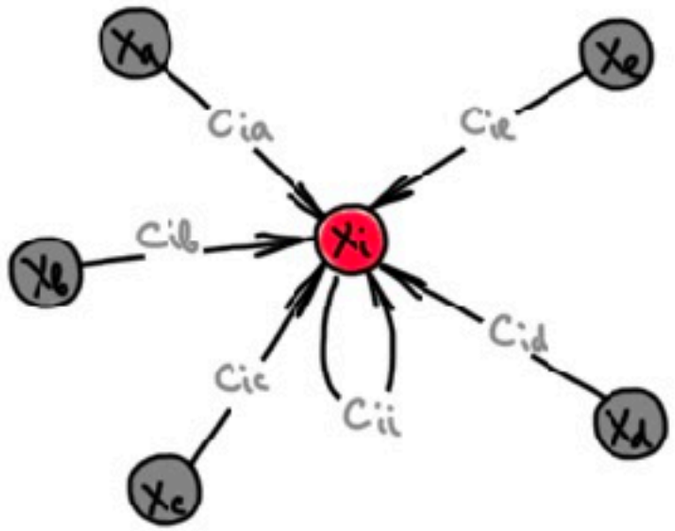
- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - Early Visual Question Answering systems
  - Non-local computations (Relation Nets, Transformer)
  - **Graph Neural Networks**
  - Soft-Attention & Hard-Attention in Computer Vision
  - Bias

Learning on graphs



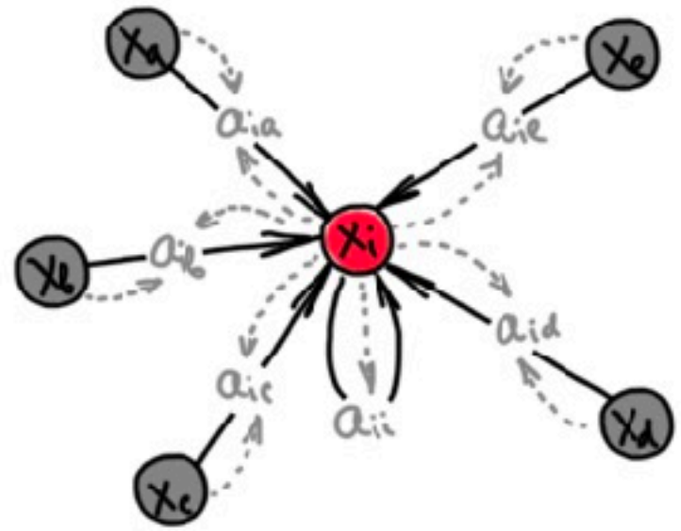
Michael Bronstein

Three flavours of Graph Nets



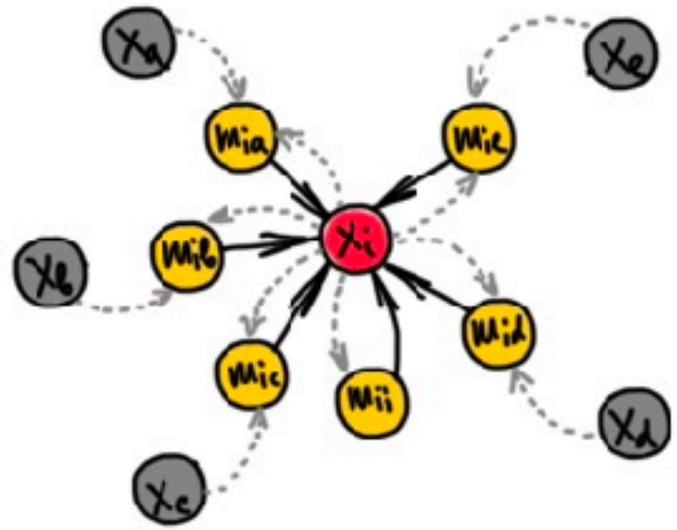
Michael Bronstein

Three flavours of Graph Nets



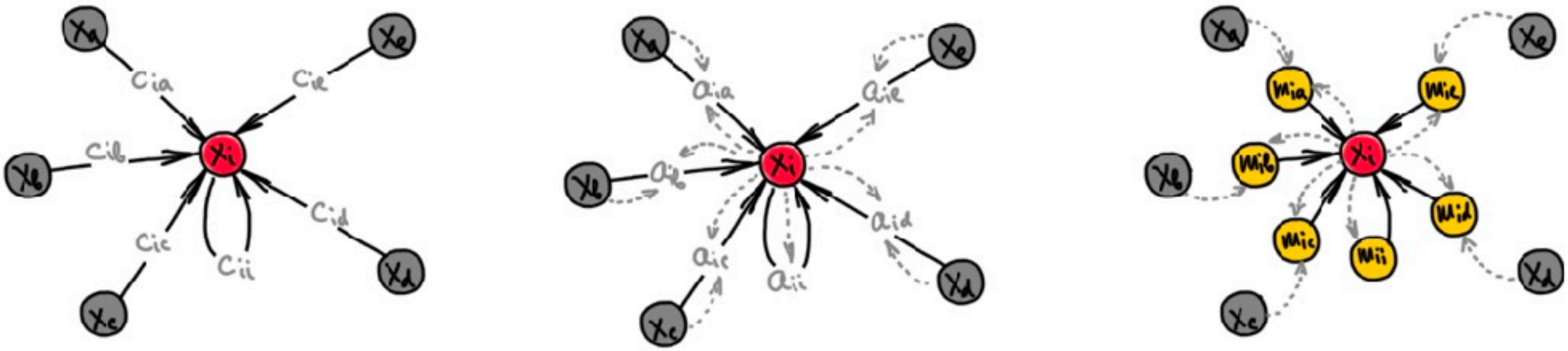
Michael Bronstein

Three flavours of Graph Nets



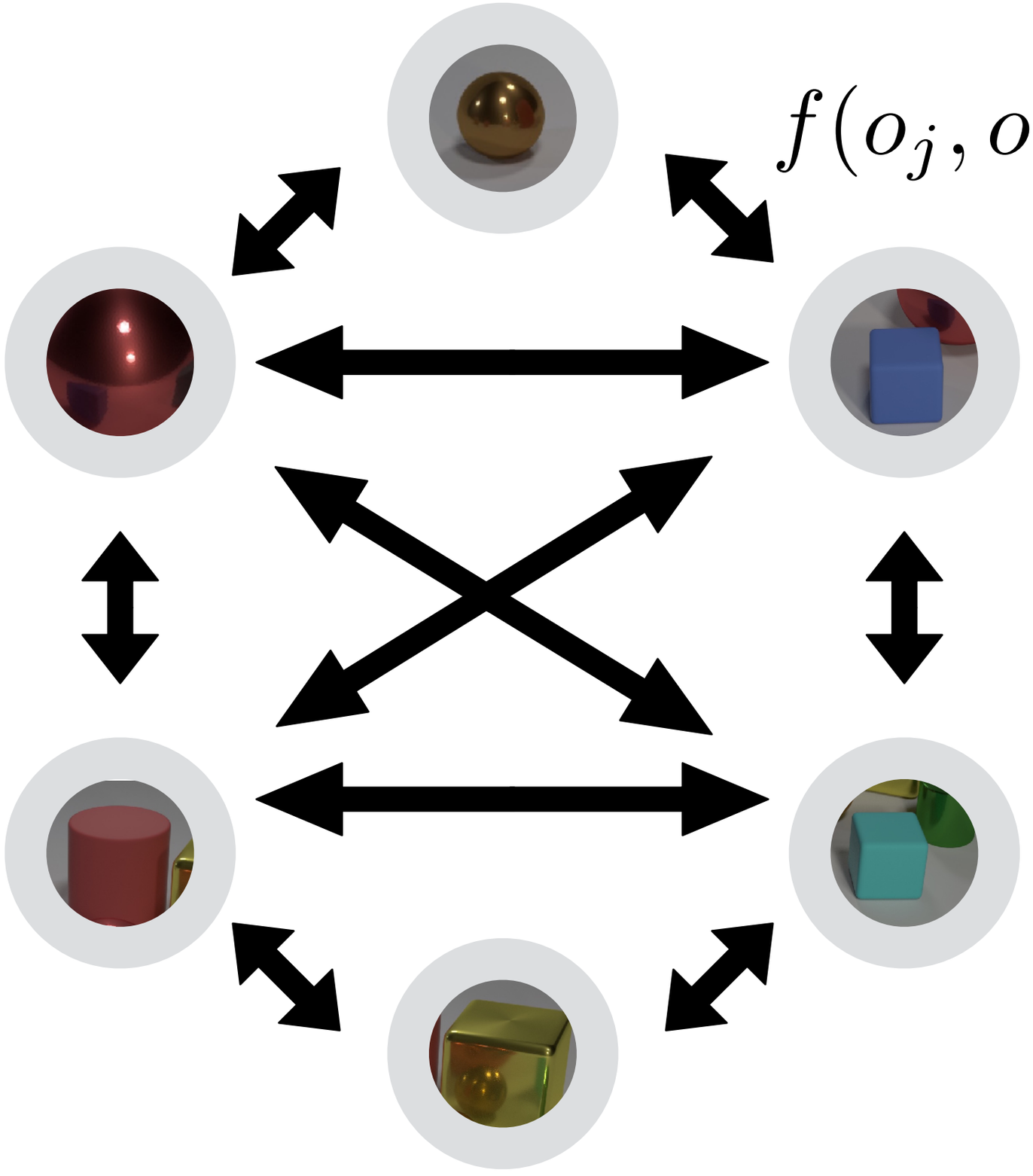
Michael Bronstein

Three flavours of Graph Nets

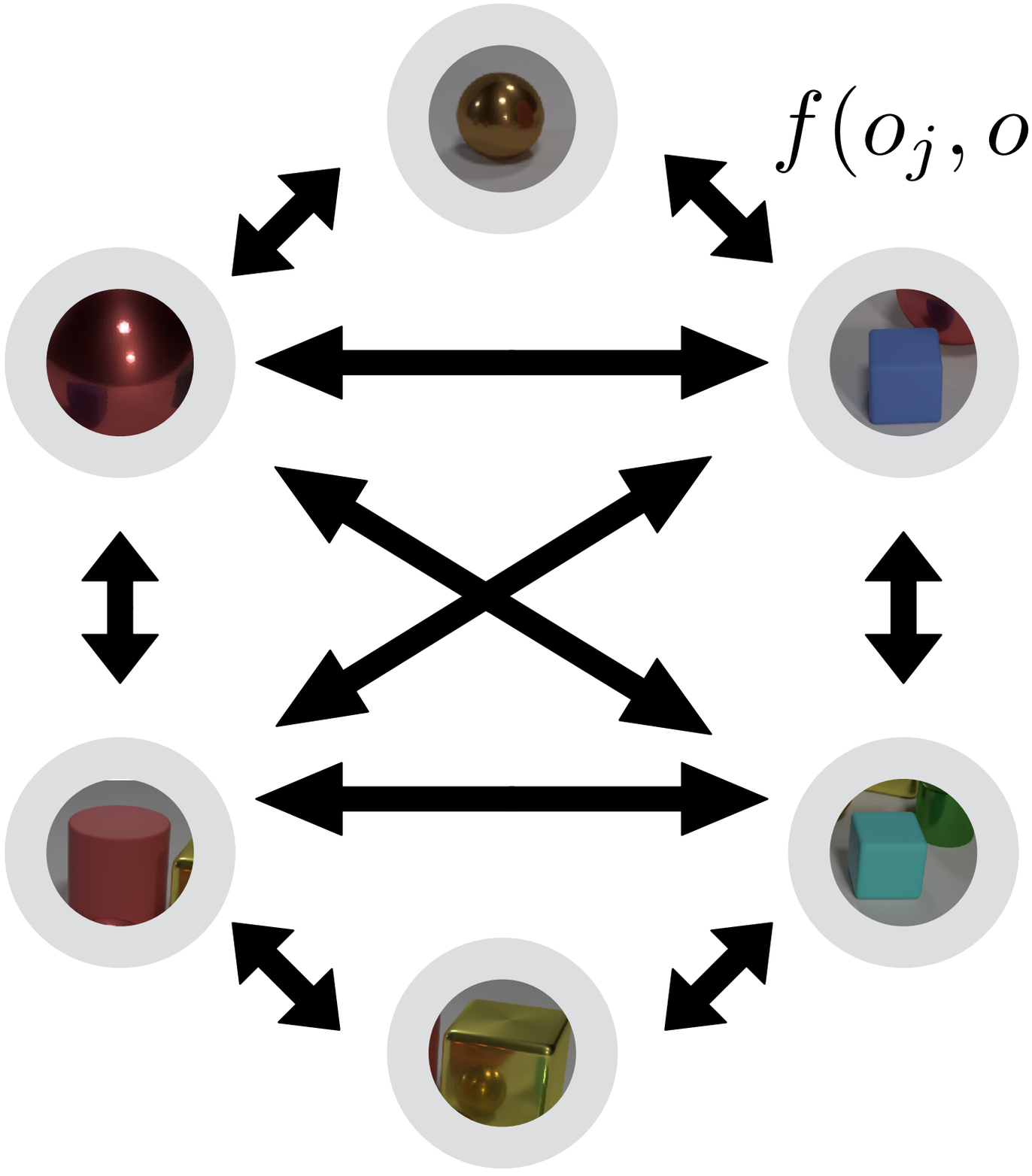


Michael Bronstein

Transformer & Relation Nets as GraphNets



Transformer & Relation Nets as GraphNets



$$f(o_j, o_k)$$

- Bi-linear
  - Convolutional GraphNet
  - Transformer
- A neural network
  - Message Passing GraphNet
  - Relation Net

Plan

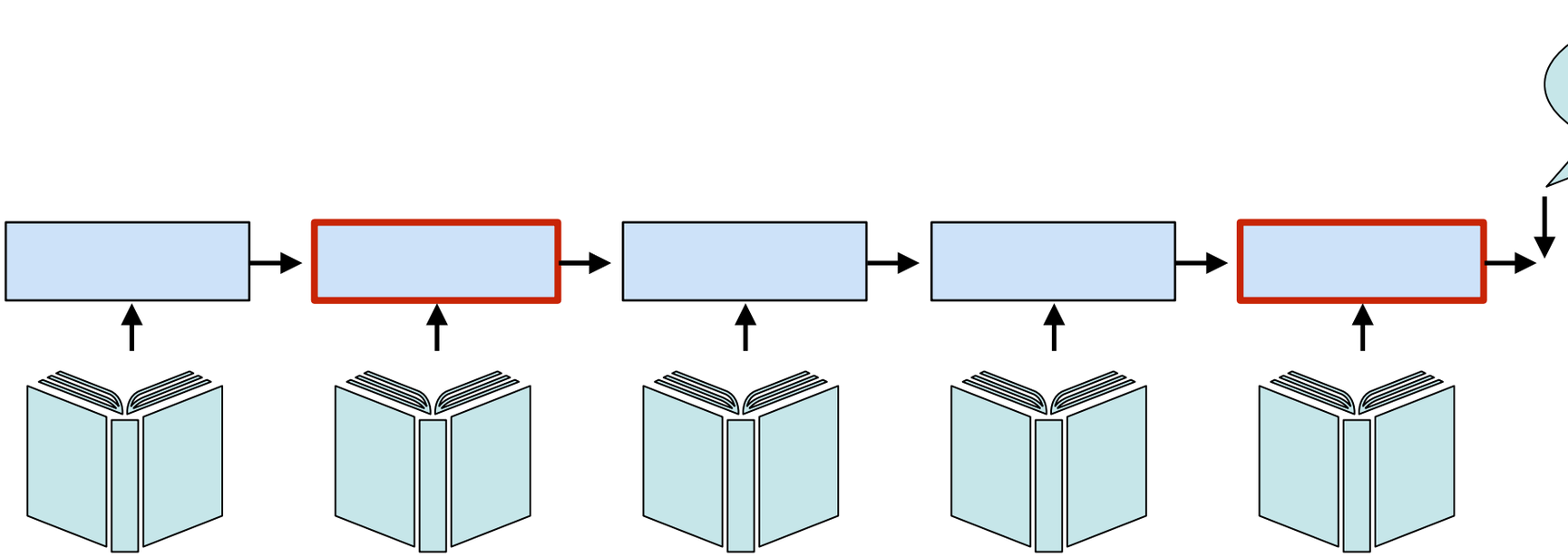
- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - Early Visual Question Answering systems
  - Non-local computations (Relation Nets, Transformer)
  - Graph Neural Networks
  - **Soft-Attention & Hard-Attention in Computer Vision**
  - Bias

Selective Attention

http://www.theinvisiblegorilla.com

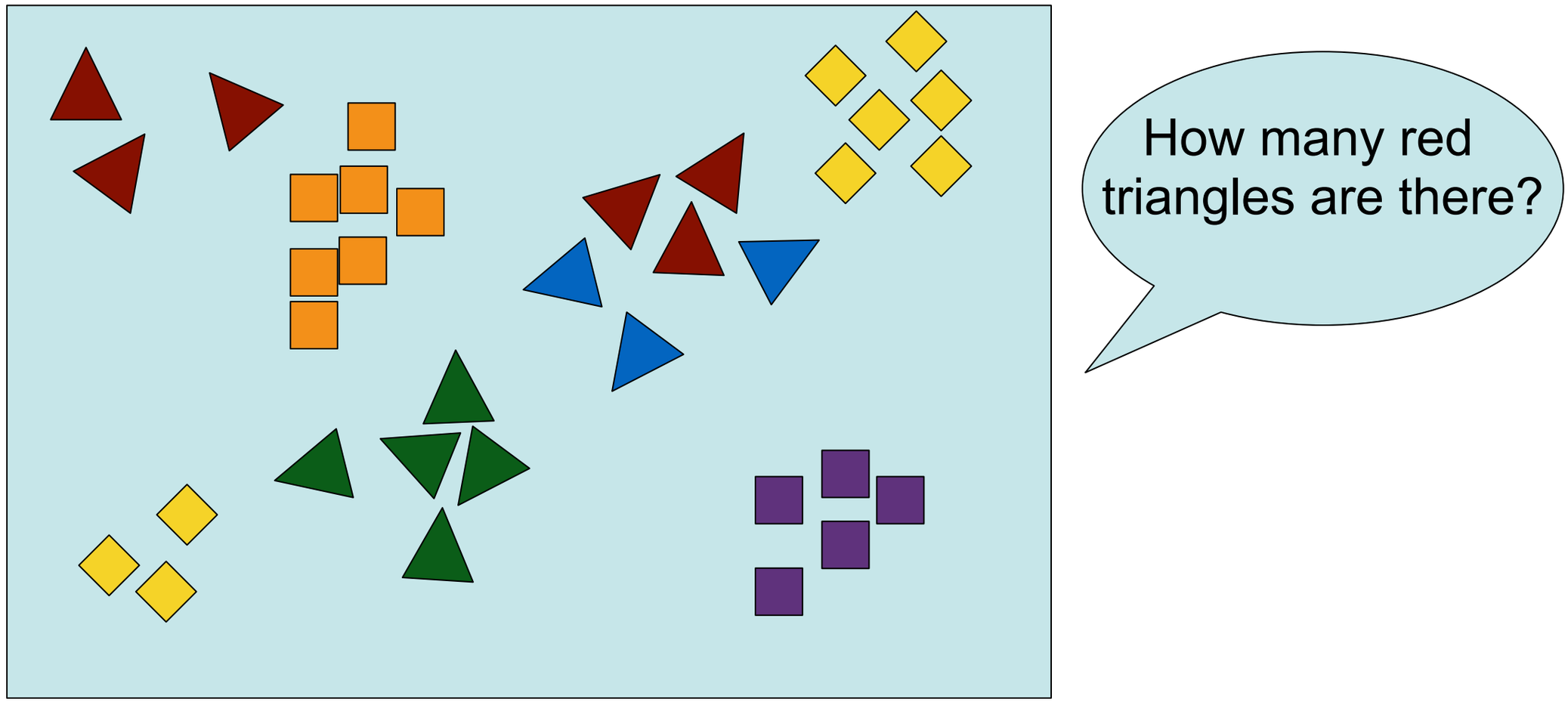
Why attention?

- Long term memories - attending to memories
  - Dealing with gradient vanishing problem



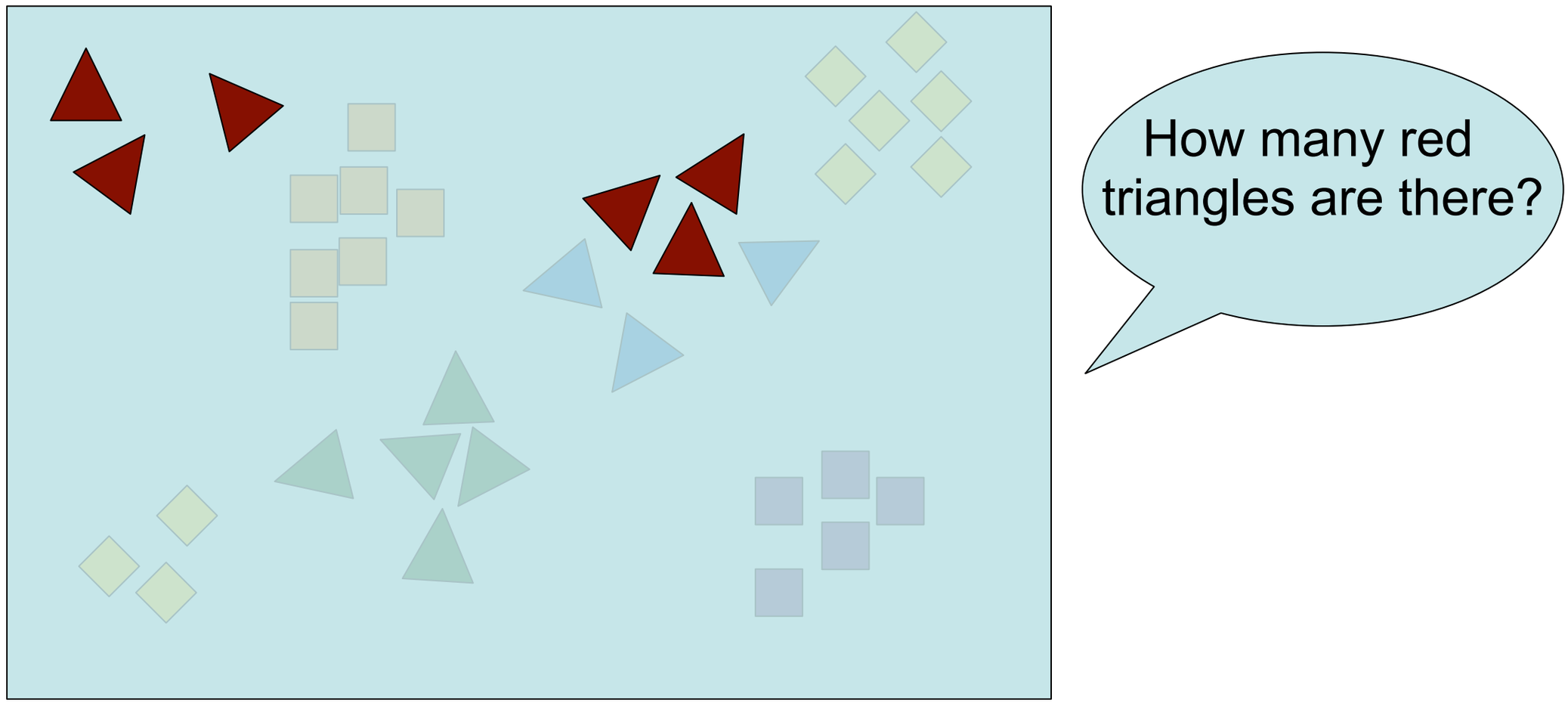
Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
  - Attending/focusing to smaller parts of data
    - patches in images
    - words or phrases in sentences



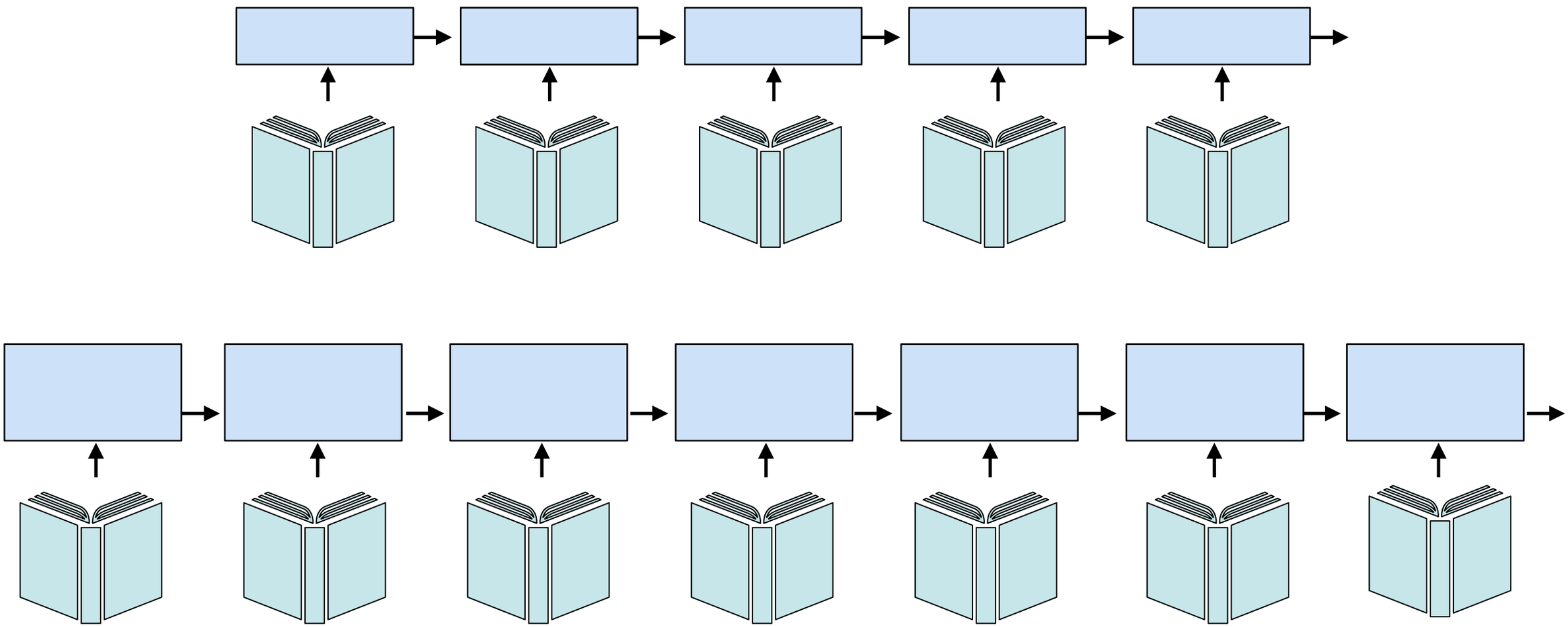
Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
  - Attending/focusing to smaller parts of data
    - patches in images
    - words or phrases in sentences



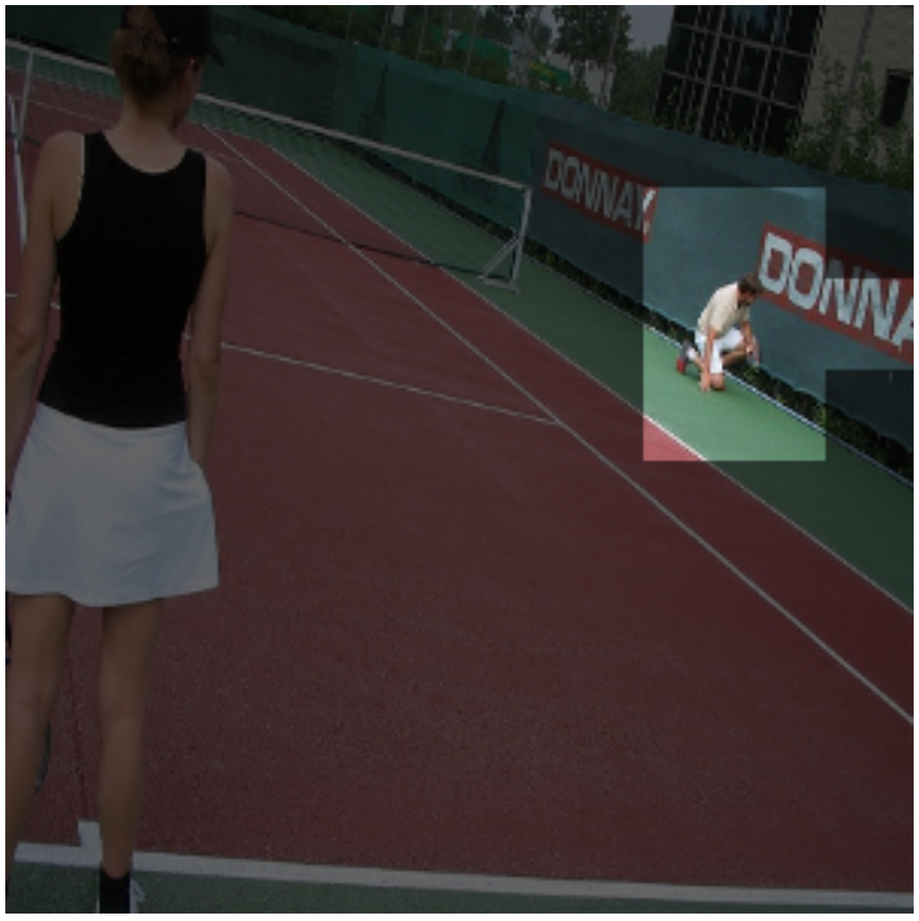
Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
- Decoupling representation from a problem
  - LSTM with longer sentences requires larger vectors



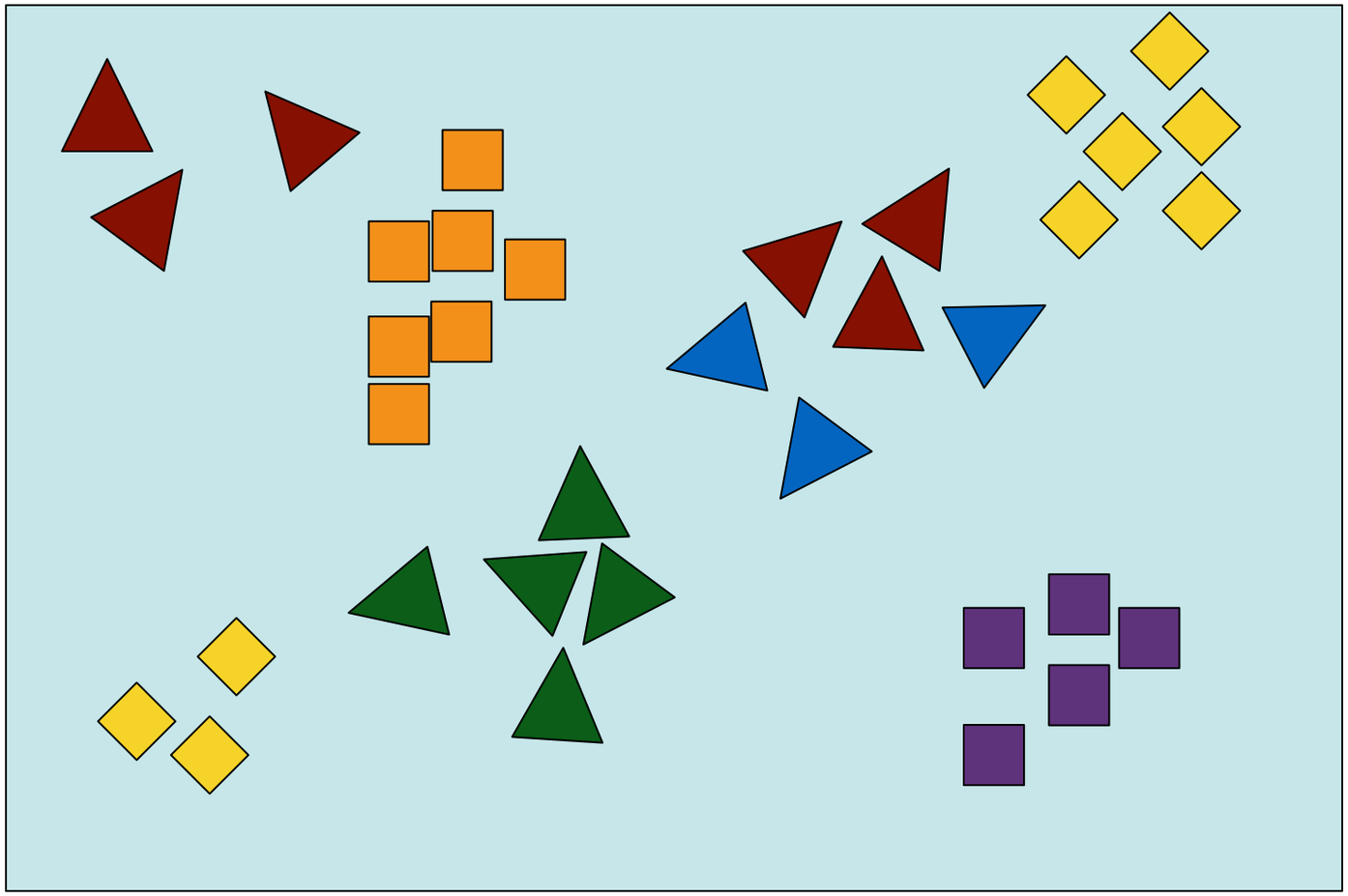
Why attention?

- Long term memories - attending to memories
- Exceeding limitations of a global representation
- Decoupling representation from a problem
  - LSTM with longer sentences requires larger vectors
- Adds some interpretability to the models (error inspection)



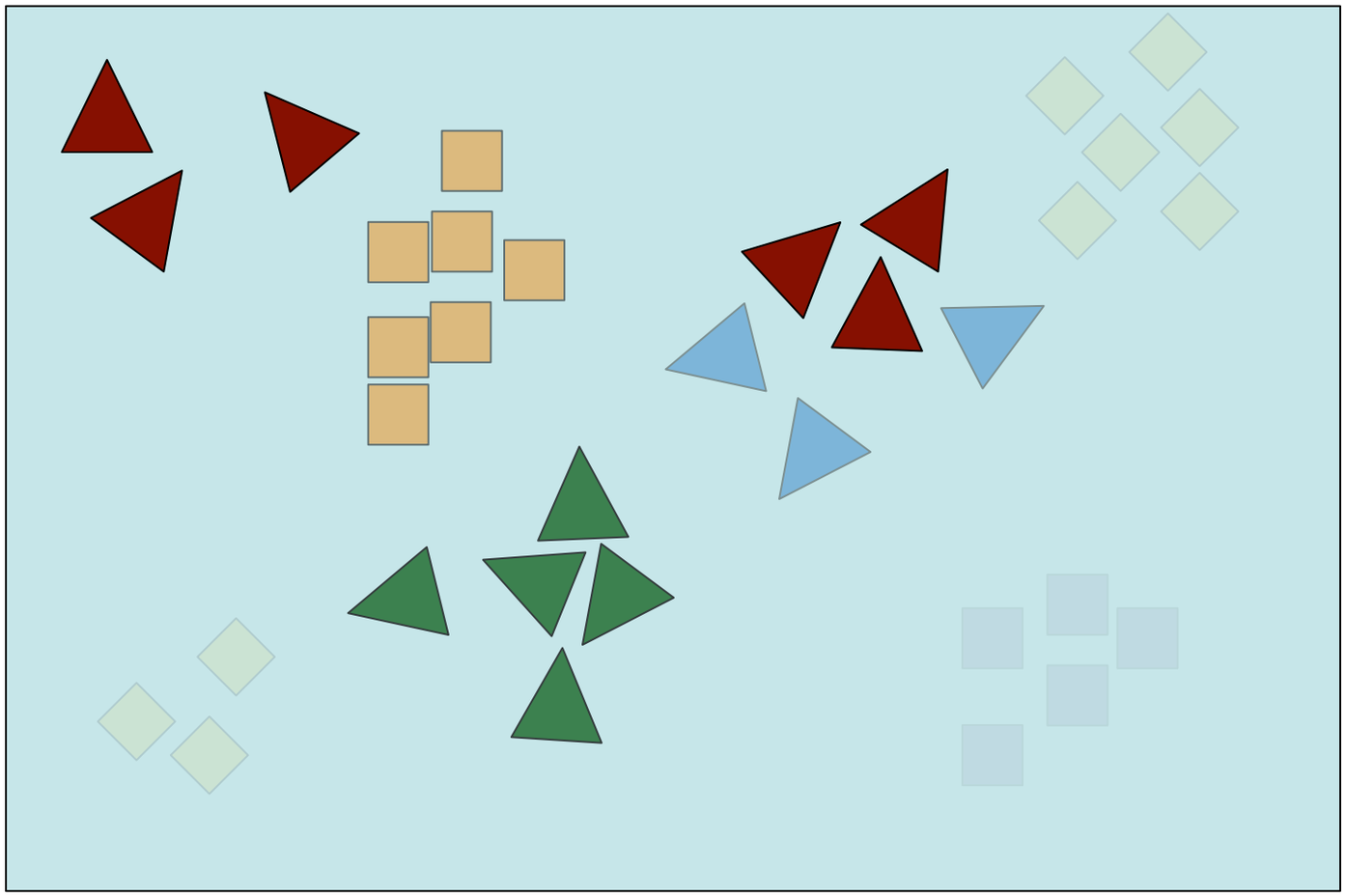
What color is her skirt? White

Two flavours of “Attention”



How many red triangles are there?

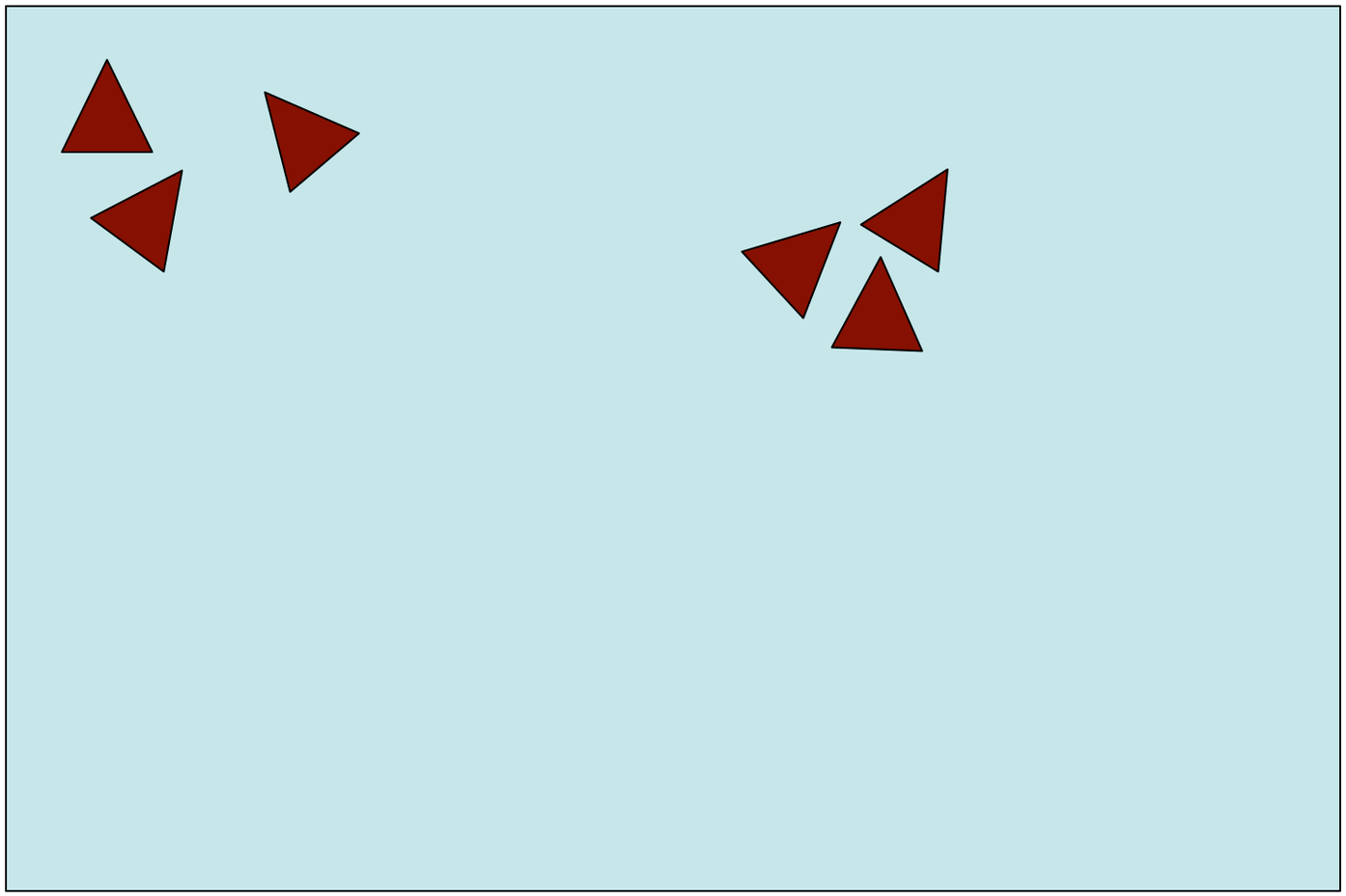
Two flavours of “Attention”: Soft



How many red triangles are there?

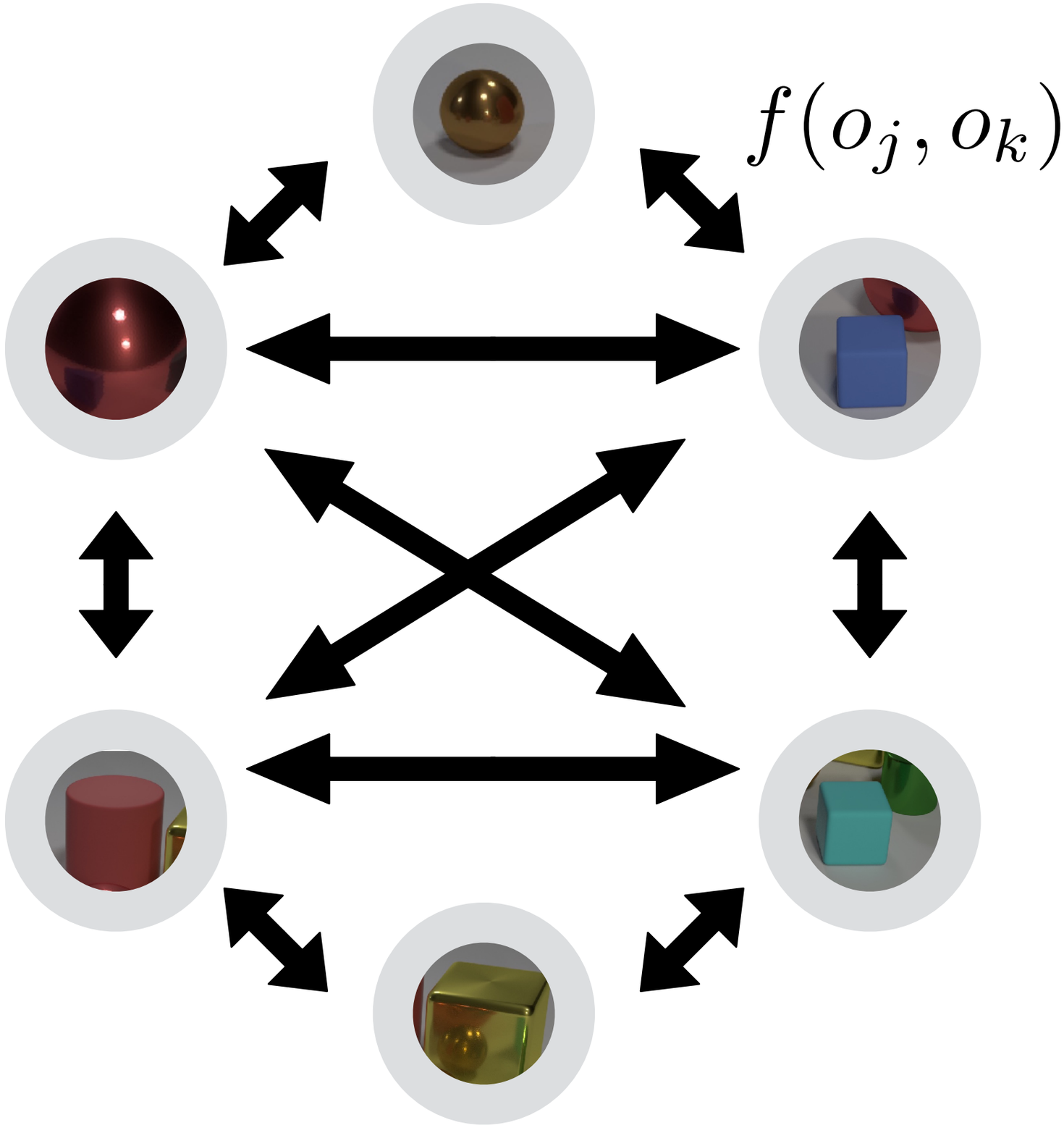
- Pros
  - Differentiable
  - Statistical efficiency
- Cons
  - No computational benefits

Two flavours of “Attention”: Hard

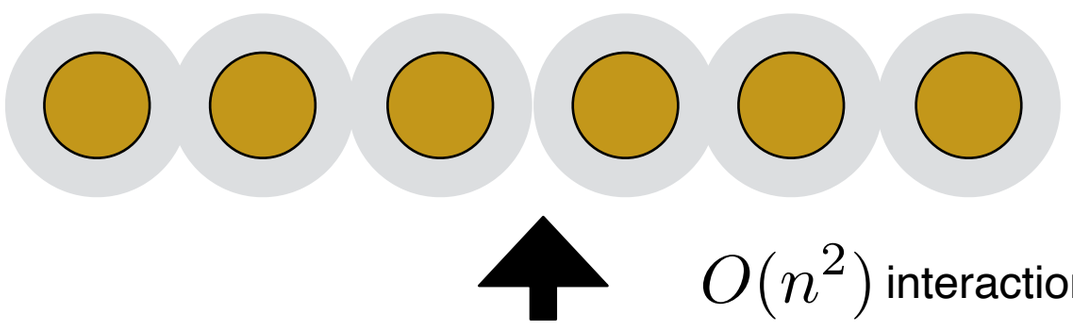


- Pros
  - Computational efficiency
  - Statistical efficiency
- Cons
  - Non-differentiable

Taming the quadratic complexity

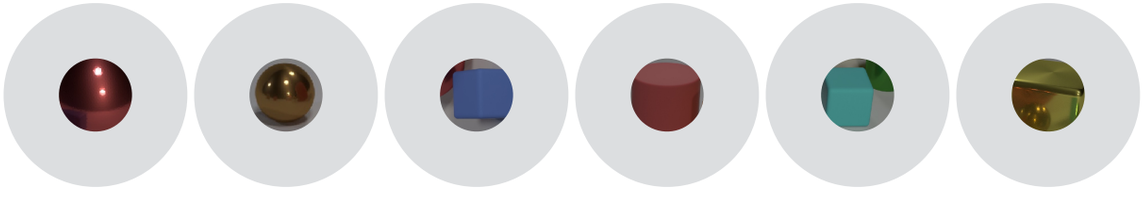


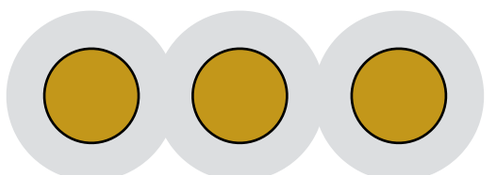
Taming the quadratic complexity



$O(n^2)$ interactions

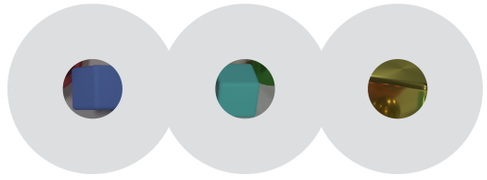
Self-Attention (all pairs)



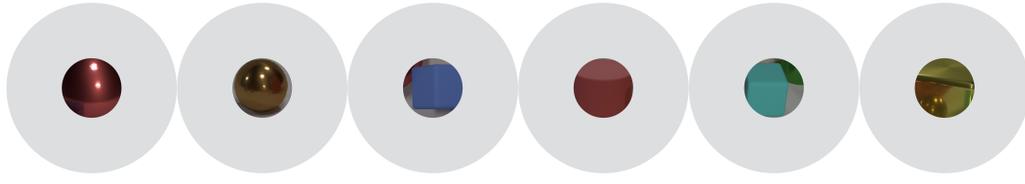


$O(k^2)$ interactions

Self-Attention (all pairs)



Hard-Attention (per-token)



M. Malinowski et al “Learning Visual Question Answering by Bootstrapping Hard Attention”. ECCV’18

Straight through estimator

- Divide a non-differentiable function into two steps
  - “Hard-“ forward pass; e.g. select inputs based on the attention mask and a threshold
  - “Soft-“ backward pass; e.g. use the whole attention mask during the backprop
- Easy to implement

$$f_{\text{soft}}(x) + \cancel{\nabla}(f_{\text{hard}}(x) - f_{\text{soft}}(x))$$

- Drawbacks
  - Still computationally demanding (backward-pass)
  - No guarantees
    - E.g. $f_{\text{soft}}(x)$ and $f_{\text{hard}}(x)$ might be arbitrarily different

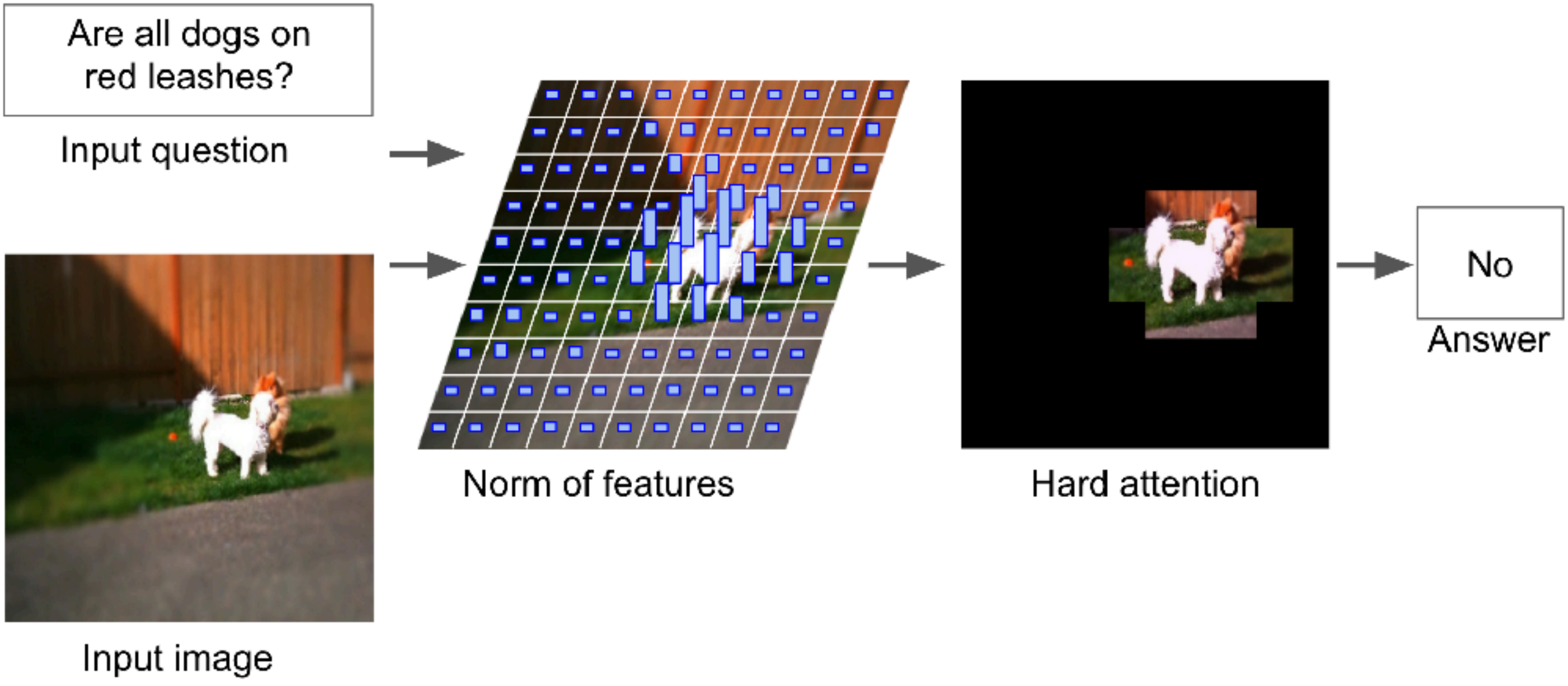
Using top-k with a measure

- Use a sub-differentiable function argmax
- Easy to implement

$$\text{argmax}_k(x, \mathcal{M}(x))$$

- Efficient
- Drawbacks
  - No guarantees
    - e.g. gradient might not flow through some of the inputs

Hard-Attention



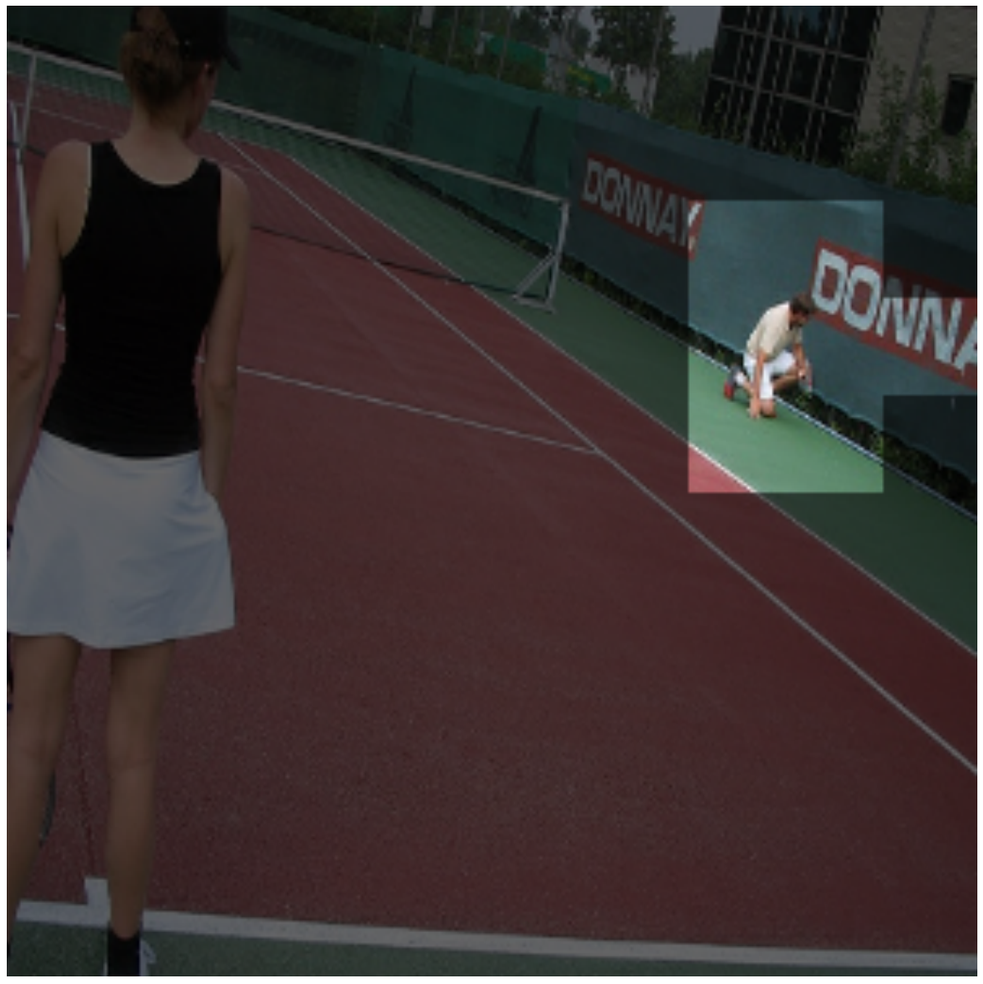
Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Results on CLEVR

HAN + RN, 25%	95.2
HAN + RN (+), 25%	96.9
HAN + RN (++), 33% (64 cells [same as RN])	**98.8**
RN [1]	95.5
Human [2]	92.6

Hard-Attention

Are this lions? No



What color is her skirt? White

Malinowski et al. “Learning Visual Question Answering by Bootstrapping Hard-Attention”

Plan

- Vision + Language (1)
  - Why Vision and Language?
  - Captioning, Visual Question Answering, Visual Reasoning
  - Early Visual Question Answering systems
  - Non-local computations (Relation Nets, Transformer)
  - Graph Neural Networks
  - Soft-Attention & Hard-Attention in Computer Vision
  - **Bias**

Let’s build a wonderful VQA system together

How many eyes are in the picture?

Let’s build a wonderful VQA system together

How many eyes are in the picture? -> 2

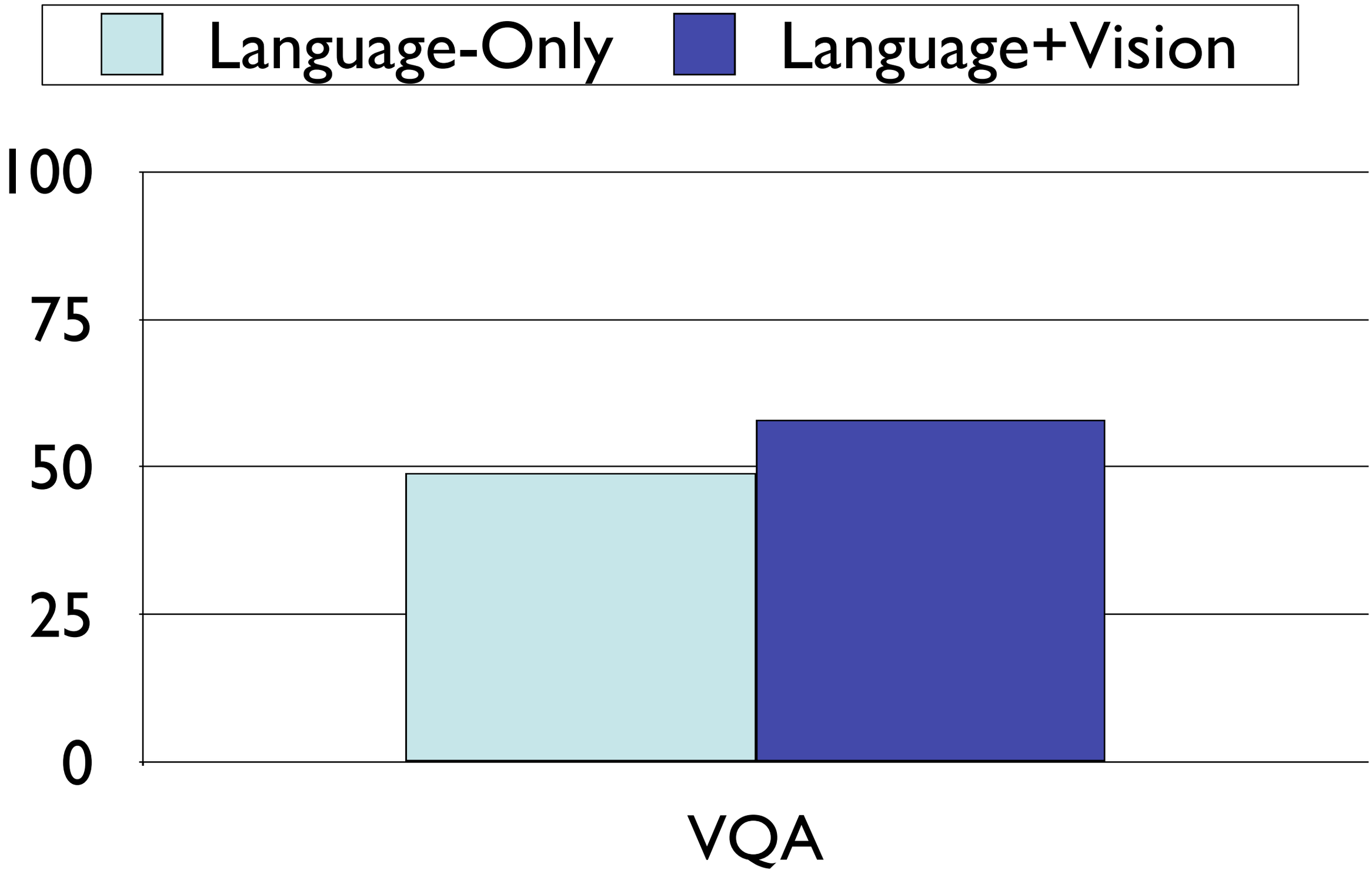
Let’s build a wonderful VQA system together

How many eyes are in the picture?

Let’s build a wonderful VQA system together

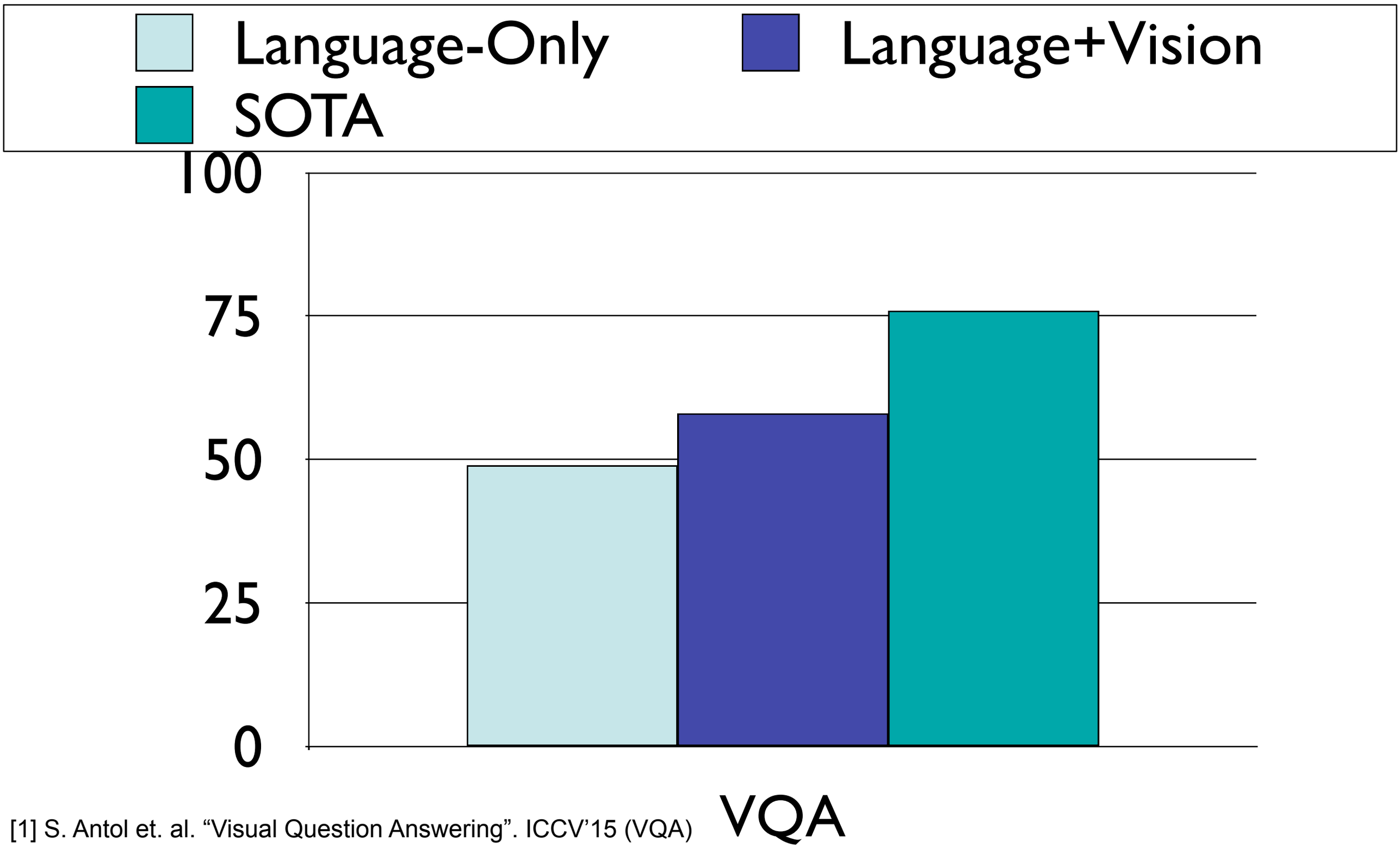
How many eyes are in the picture? -> 2

Performance on VQA



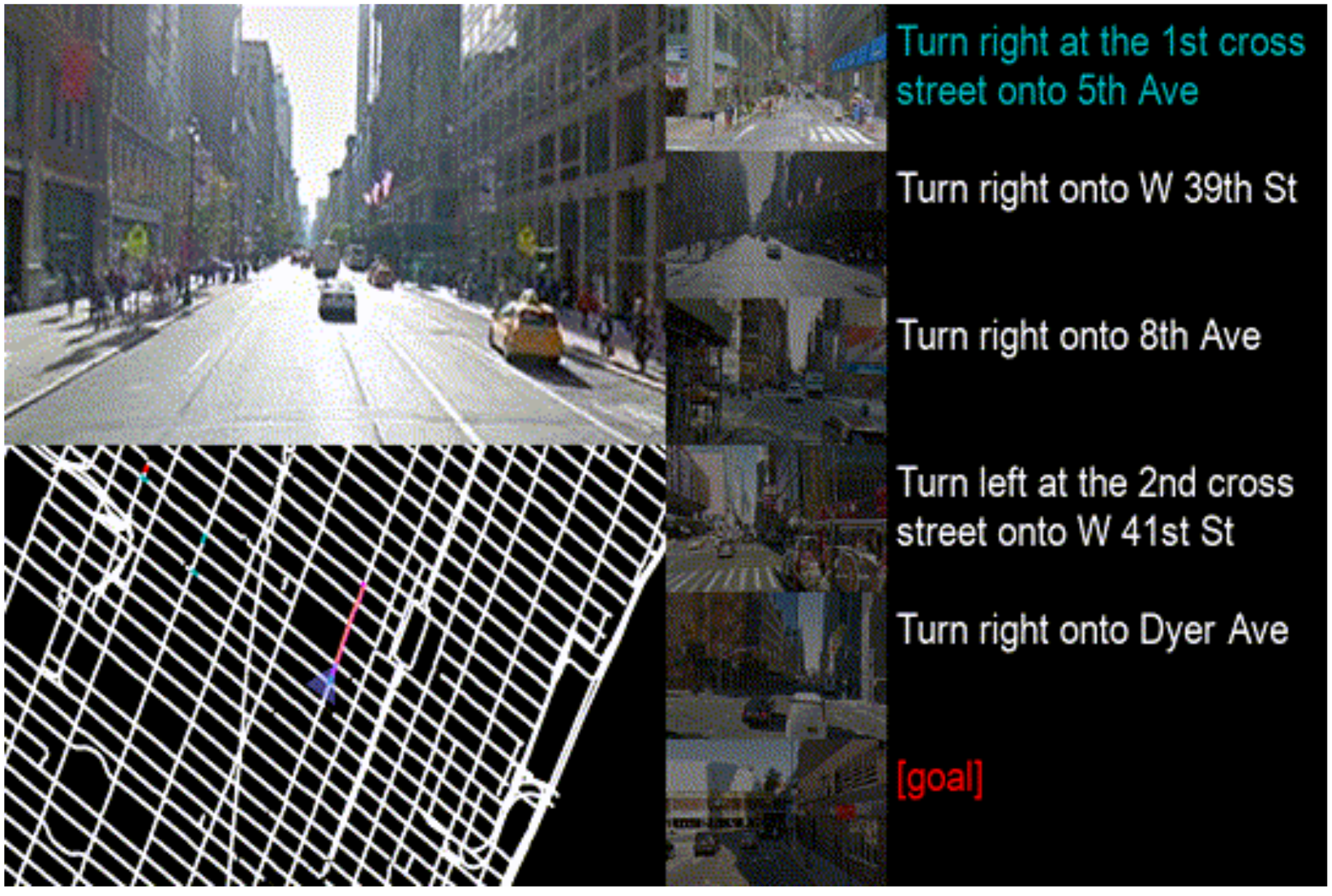
[1] S. Antol et. al. “Visual Question Answering”. ICCV’15 (VQA)
[2] M. Malinowski et. al. “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering”. NeurIPS’14 (DAQUAR)

Performance on VQA

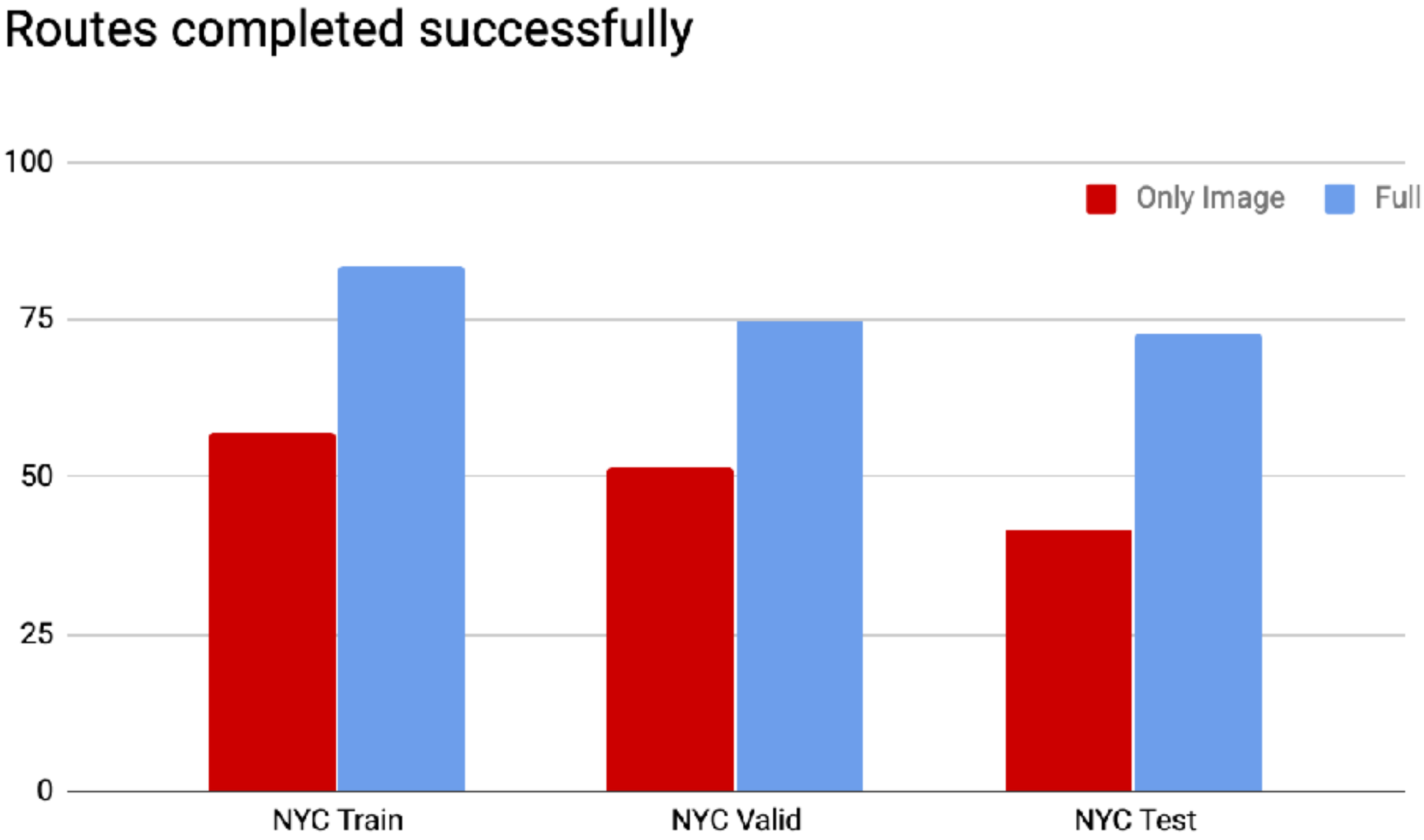


[1] S. Antol et. al. “Visual Question Answering”. ICCV’15 (VQA)
[2] M. Malinowski et. al. “Ask Your Neurons: A Deep Learning Approach to Visual Question Answering”. NeurIPS’14 (DAQUAR)
[3] D. Nguyen et al. Grid Features + MoVie (SOTA)

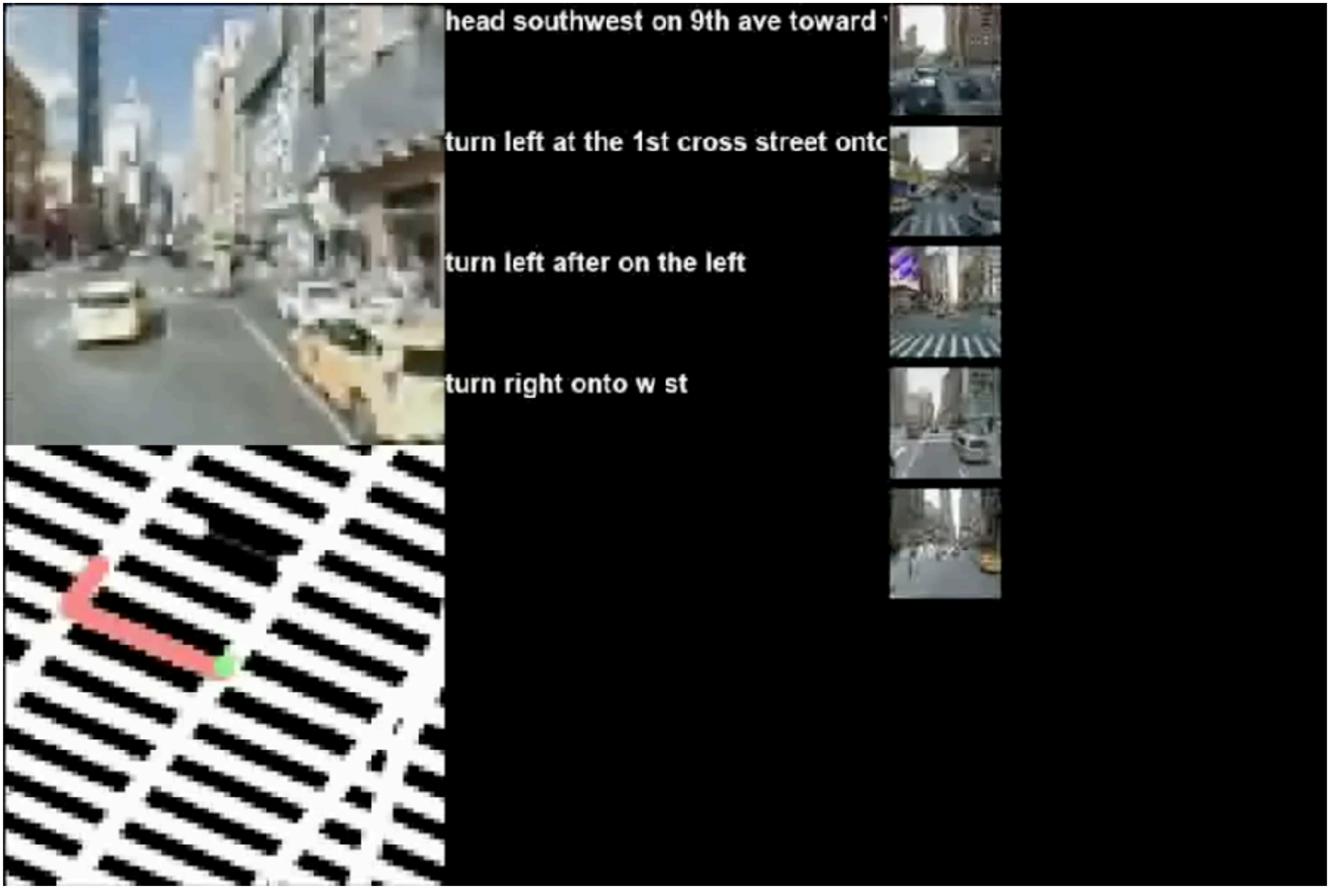
Bias in StreetLearn



Bias in StreetLearn



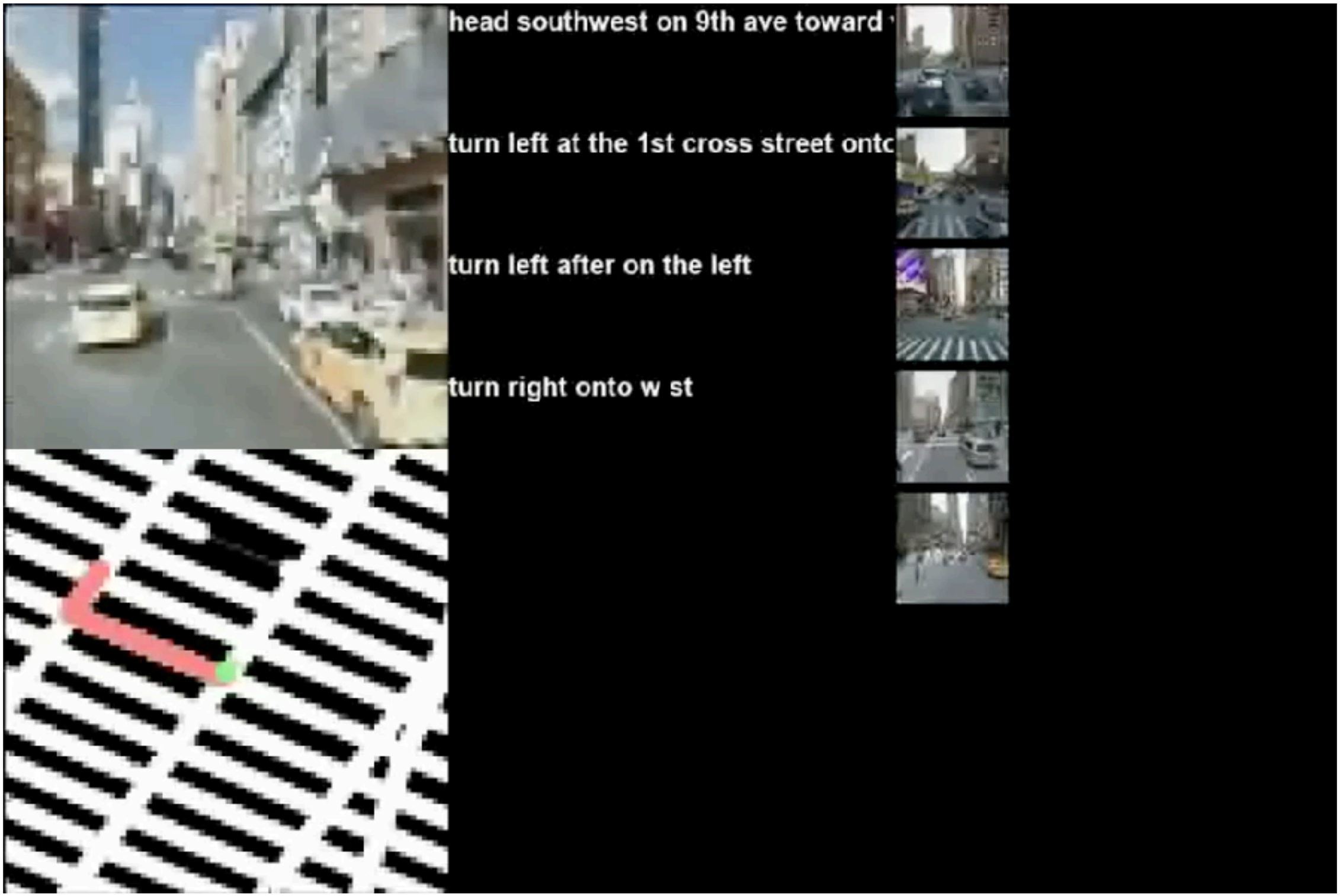
Bias in StreetLearn



Language + Vision (2)

Mateusz Malinowski

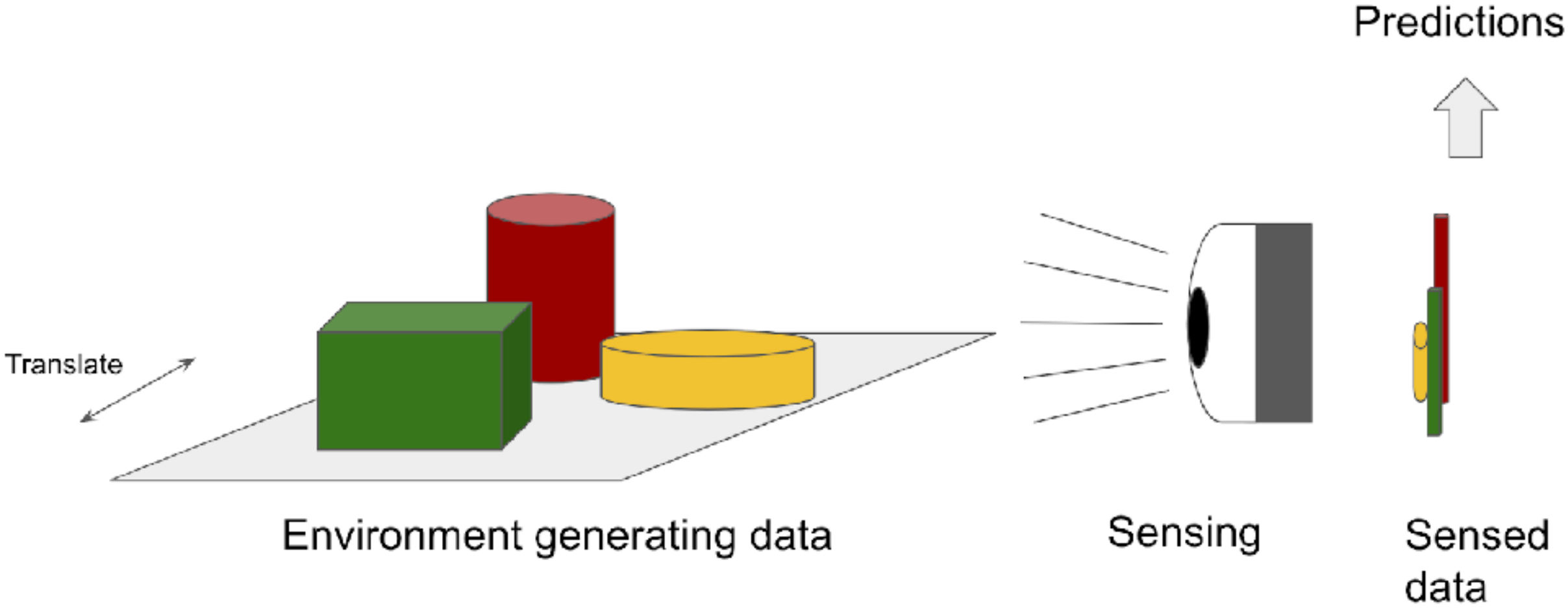
Bias in StreetLearn



“Super-human” performance on CLEVR

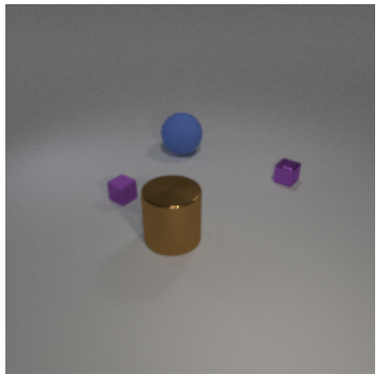
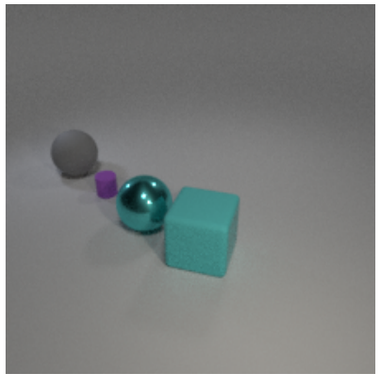
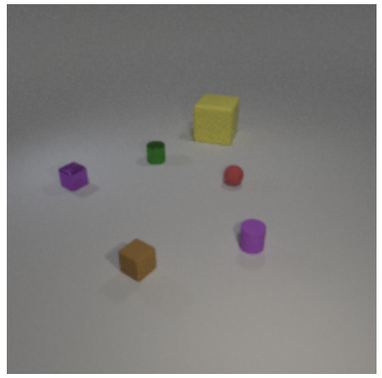
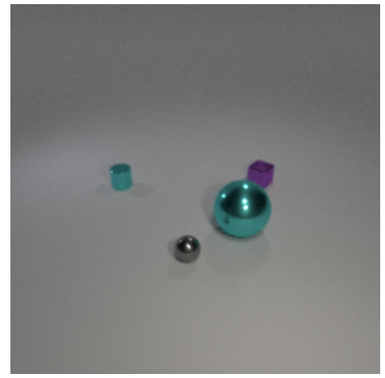
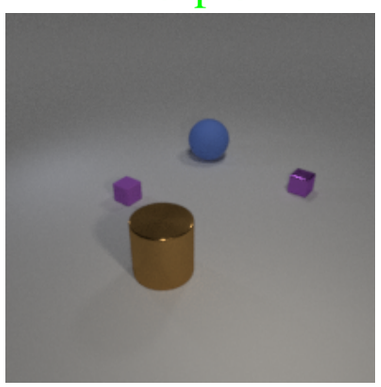
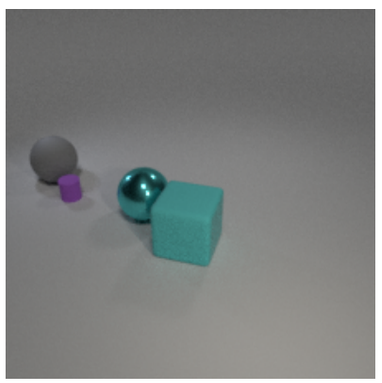
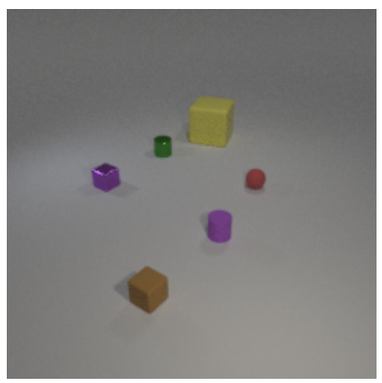
SAN (Yang et al.)	72.1	Raw Signals
FILM (Perez et al.)	96.2	
RN (ours)	95.5	
IEP (Johnson et al.)	96.9	Program
TbD (Mascharka et al.)	99.1	
MDetr (Kamath et al.)	99.7	States

Bias in CLEVR



S. Mouselinos et al. “Measuring CLEVRness”

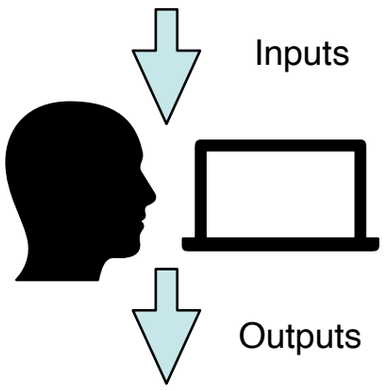
Representation should be invariant under “meaningless changes”

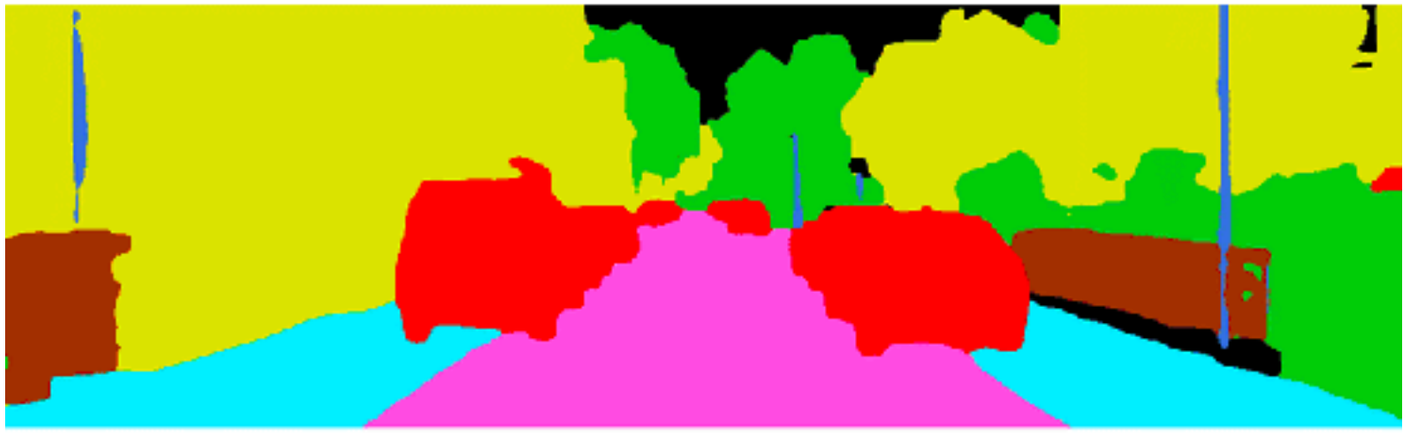
	What is the color of the block that is left of the big blue matte ball?	How many other objects are the same shape as the big gray matte object?	There is a tiny block behind the purple cylinder; what number of red matte spheres are in front of it?	How many other things are the same shape as the big metallic thing?
FiLM - Before				
	Purple	1	0	1
FiLM - After				
	Brown	0	1	2

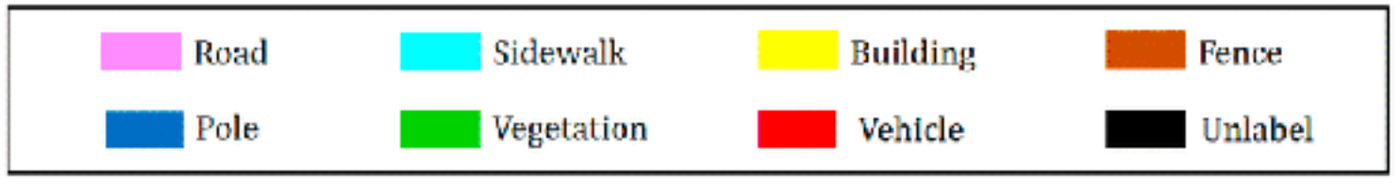
Plan

- Vision + Language (2)
  - Self-supervised learning
  - “Vision as a Language”
  - Generation and AI Art
  - Scalability

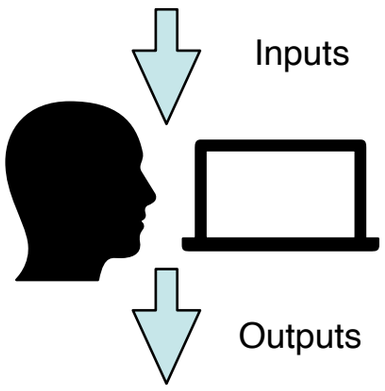
Supervised training

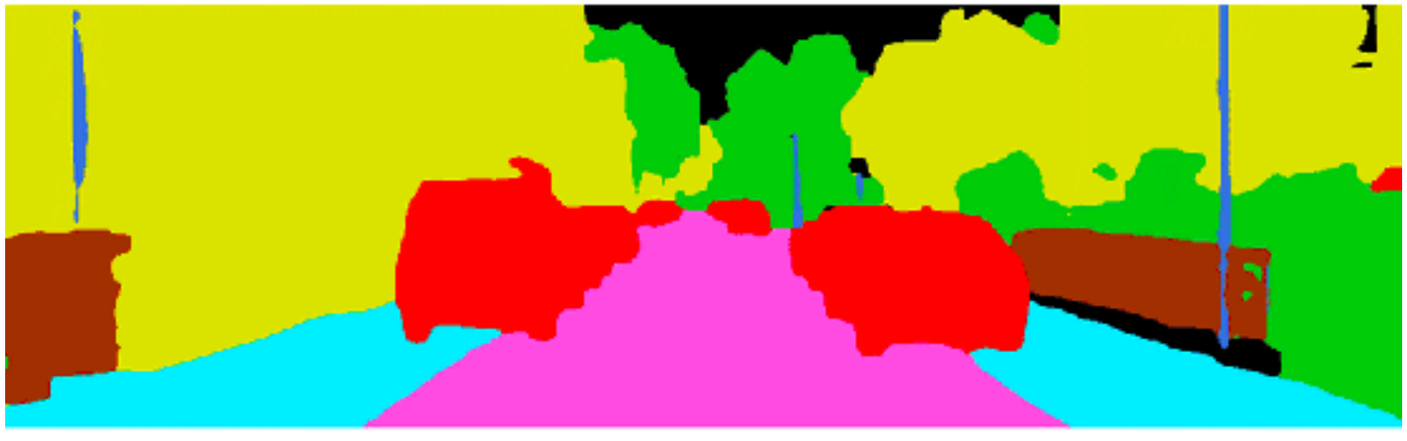


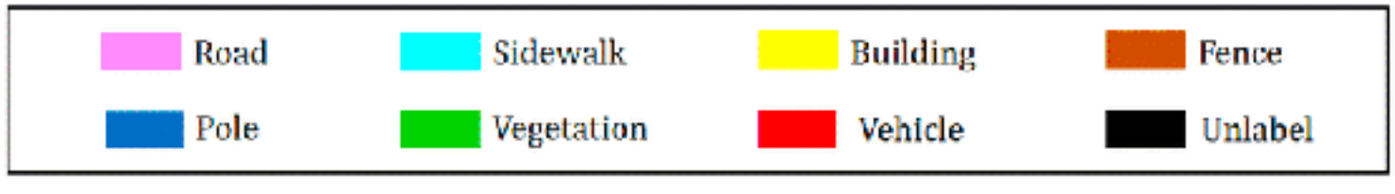


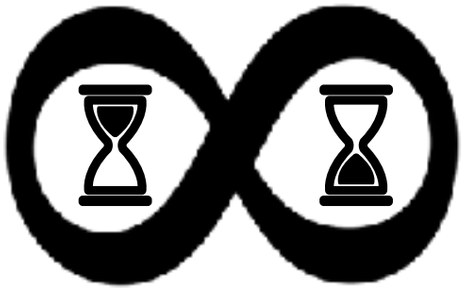


Supervised training



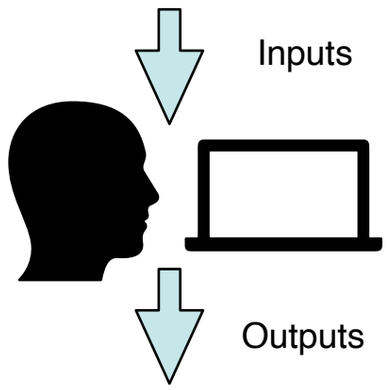


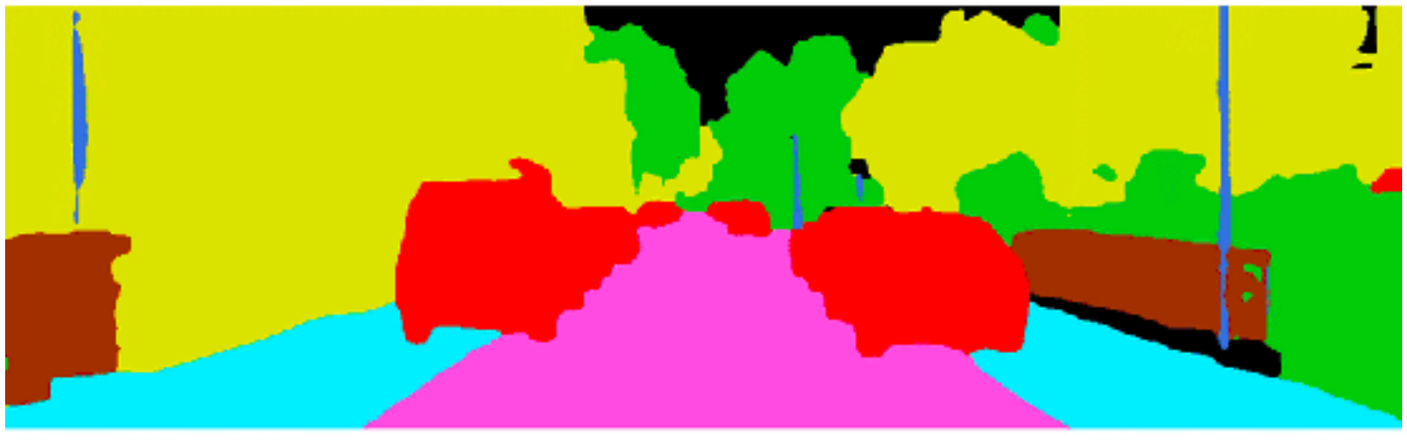


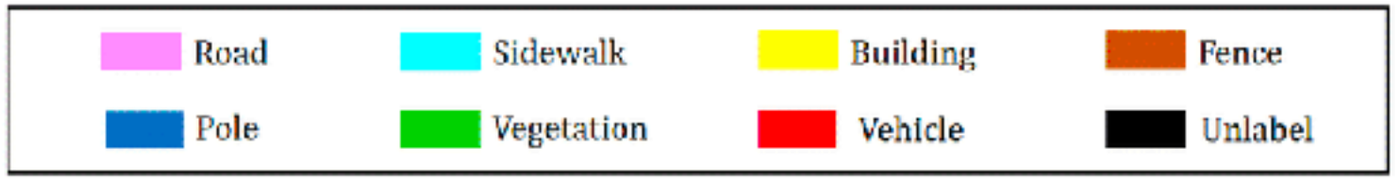


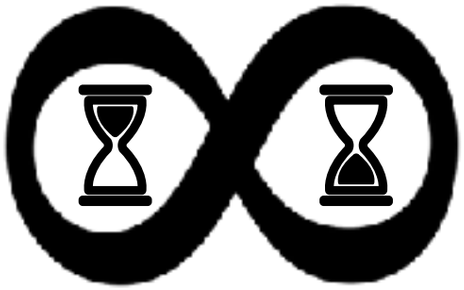
Time-consuming

Supervised training

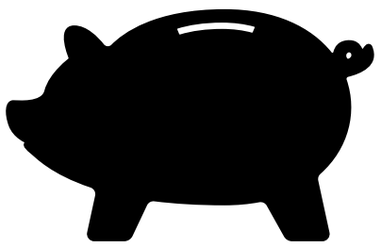






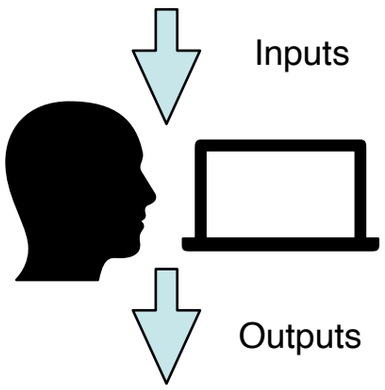


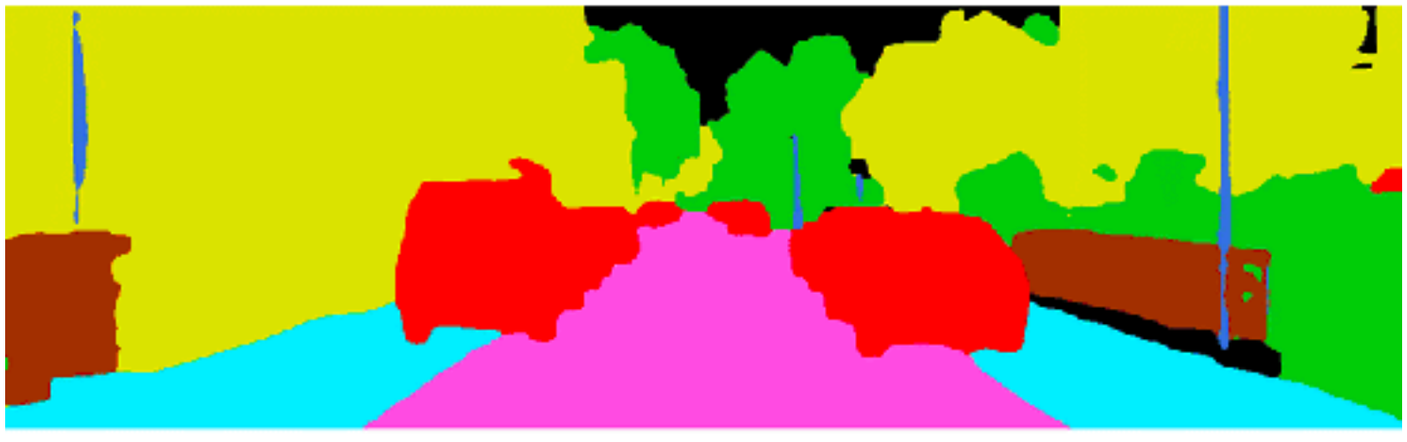
Time-consuming

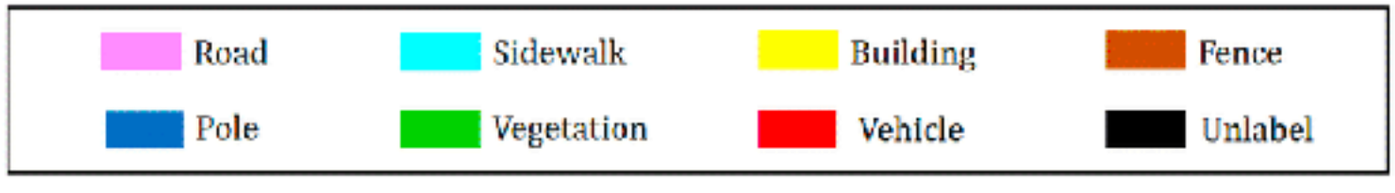


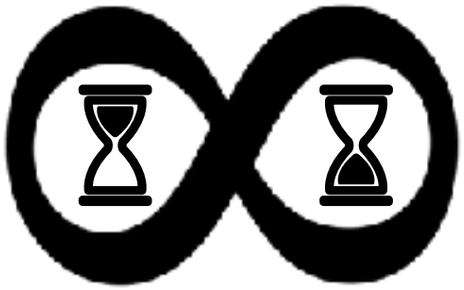
High costs

Supervised training

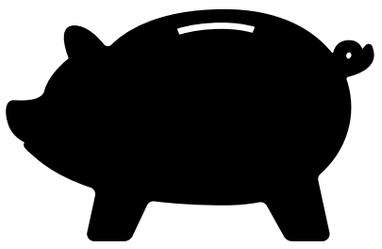




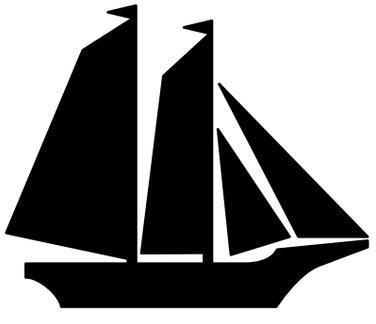




Time-consuming

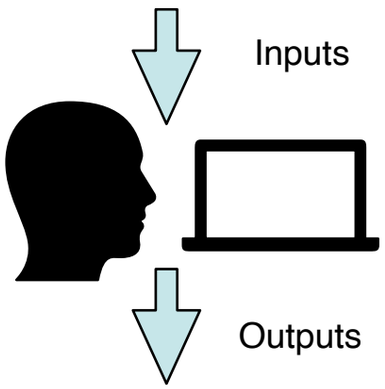


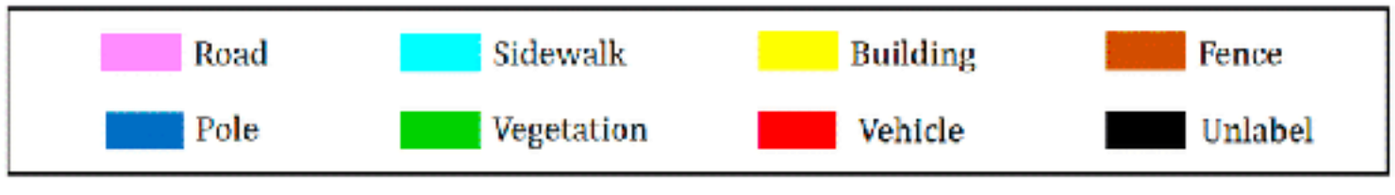
High costs

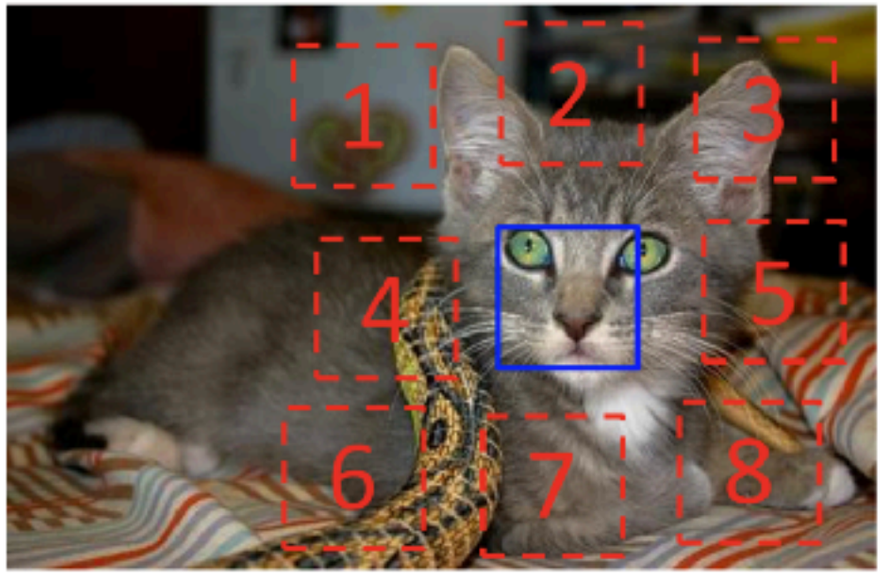


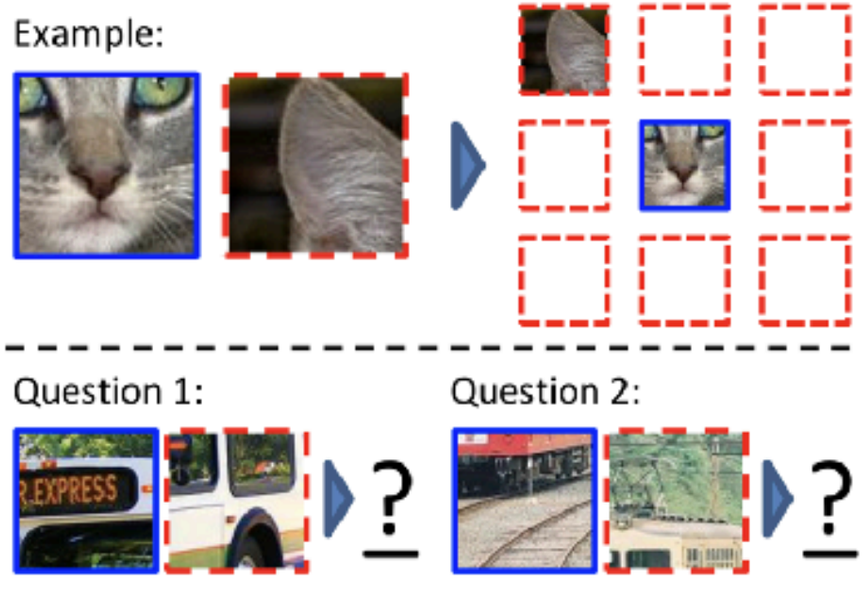
Often sparse (exploration)

Supervised training —> Self-supervision



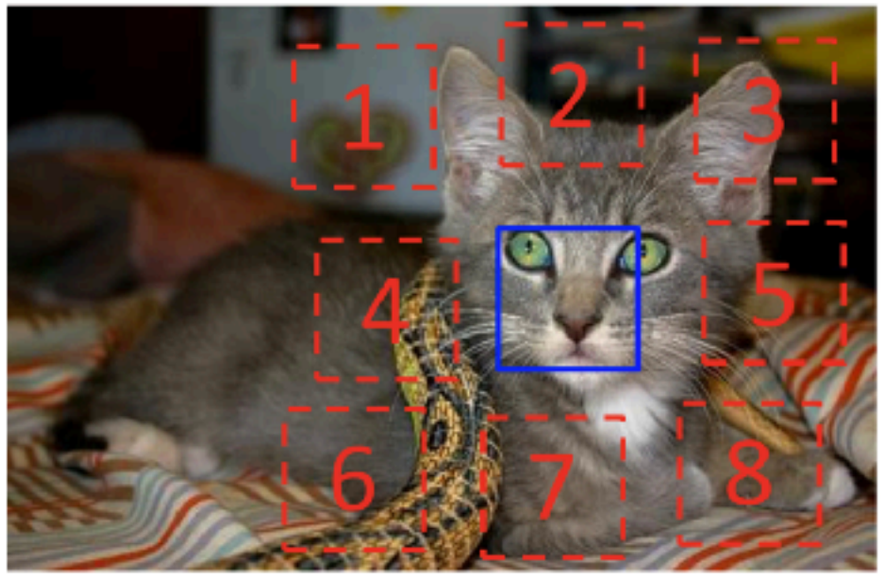






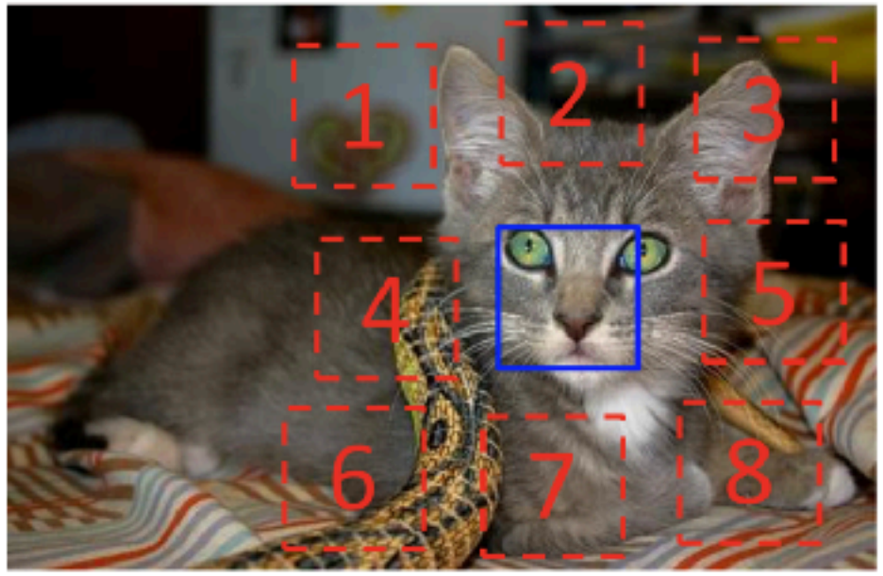
C. Doersch et al. “Unsupervised Visual Representation Learning by Context Prediction”. ICCV’15

Self-supervision



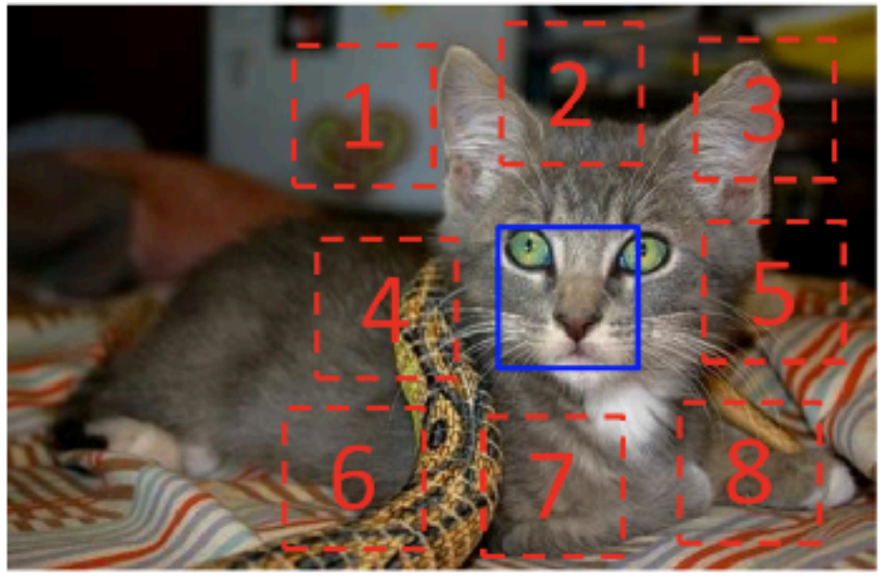
C. Doersch et al. “Unsupervised Visual Representation Learning by Context Prediction”. ICCV’15

Self-supervision



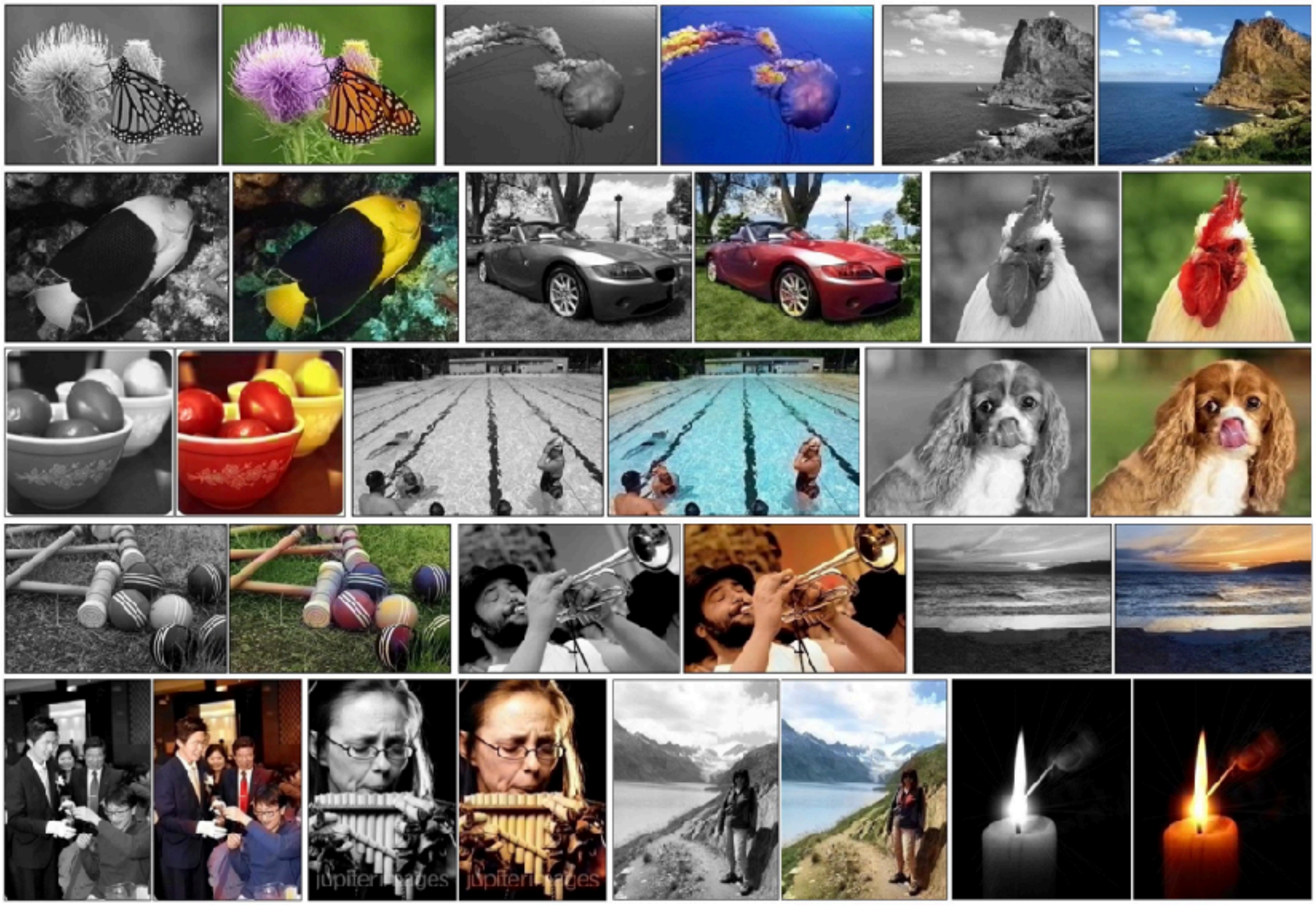
C. Doersch et al. “Unsupervised Visual Representation Learning by Context Prediction”. ICCV’15

Self-supervision



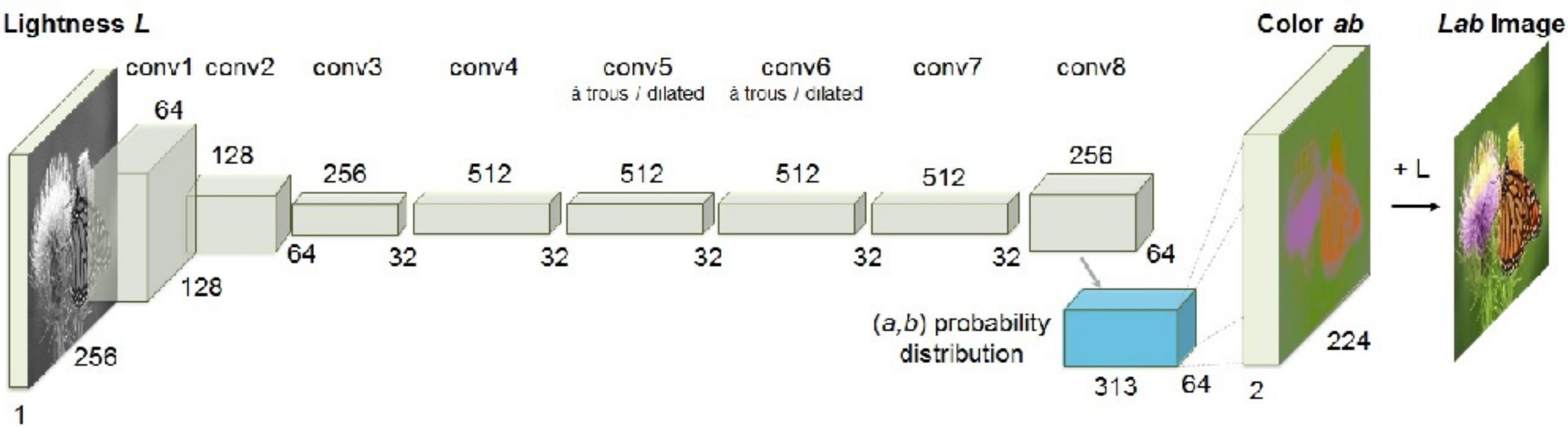
C. Doersch et al. “Unsupervised Visual Representation Learning by Context Prediction”. ICCV’15

Self-supervision: colorisation



R. Zhang et al. “Colorful Image Colorization”. CVPR’16

Self-supervision: colorisation



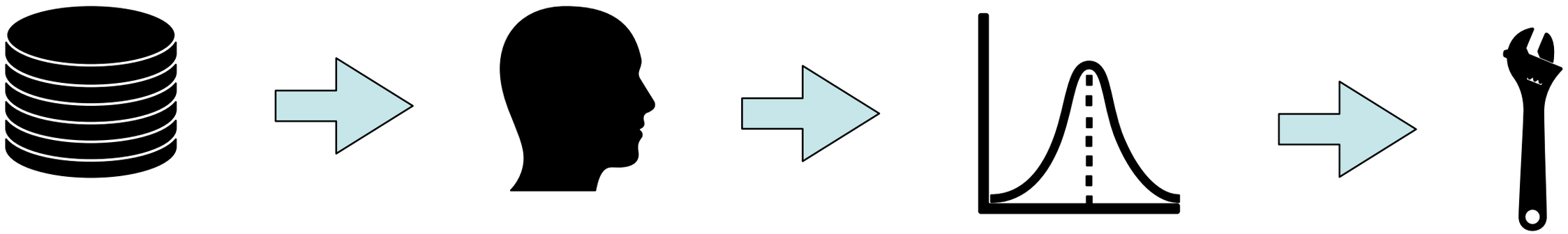
R. Zhang et al. “Colorful Image Colorization”. CVPR’16

Self-supervised learning

VOLO [1] (no extra data)	87.1	Supervised
Colorization [2] (2016)	35.2	Self-supervised
EsVIT [3] (2021)	81.3	

[1] L. Yuan et al. “VOLO: Vision Outlooker for Visual Recognition”. CVPR’21
[2] R. Zhang et al. “Colorful Image Colorization”. CVPR’16
[3] C. Li et al. “Efficient Self-supervised Vision Transformers for Representation Learning”. CVPR’21

Self-supervised learning: recipe



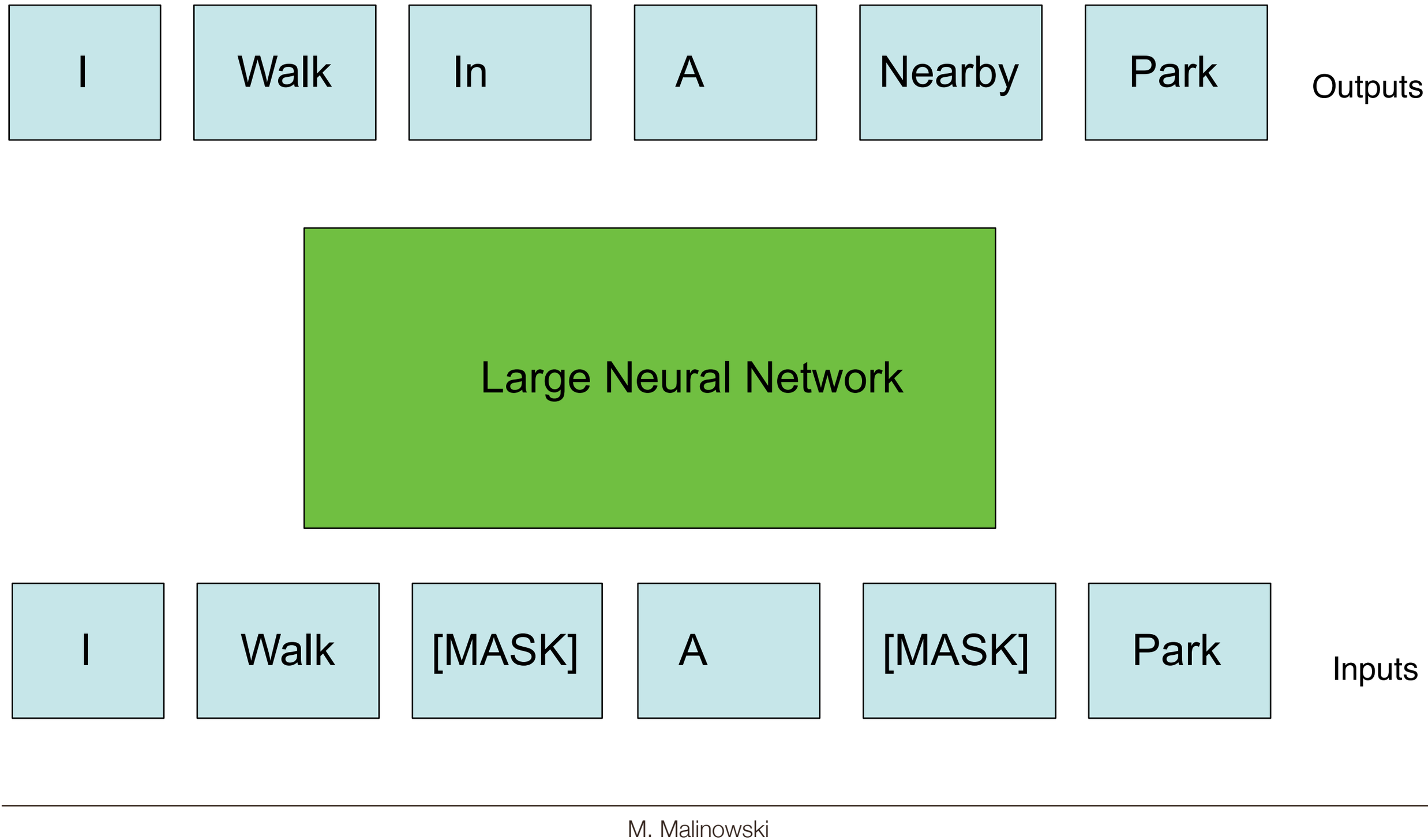
Dataset

Pre-text Task

Pre-training

Fine-tuning or
Probing on the
downstream task

Self-supervised leaning in NLP



Self-supervision in reasoning

Question: *A starts a business with Rs.40,000. After 2 months, B joined him with Rs.60,000. C joined them after some more time with Rs.120,000. At the end of the year, out of a total profit of Rs.375,000, C gets Rs.150,000 as his share. How many months after B joined the business, did C join?*
Options: *A) 30, B) 32, C) 35, D) 36, E) 40*
Rationale:

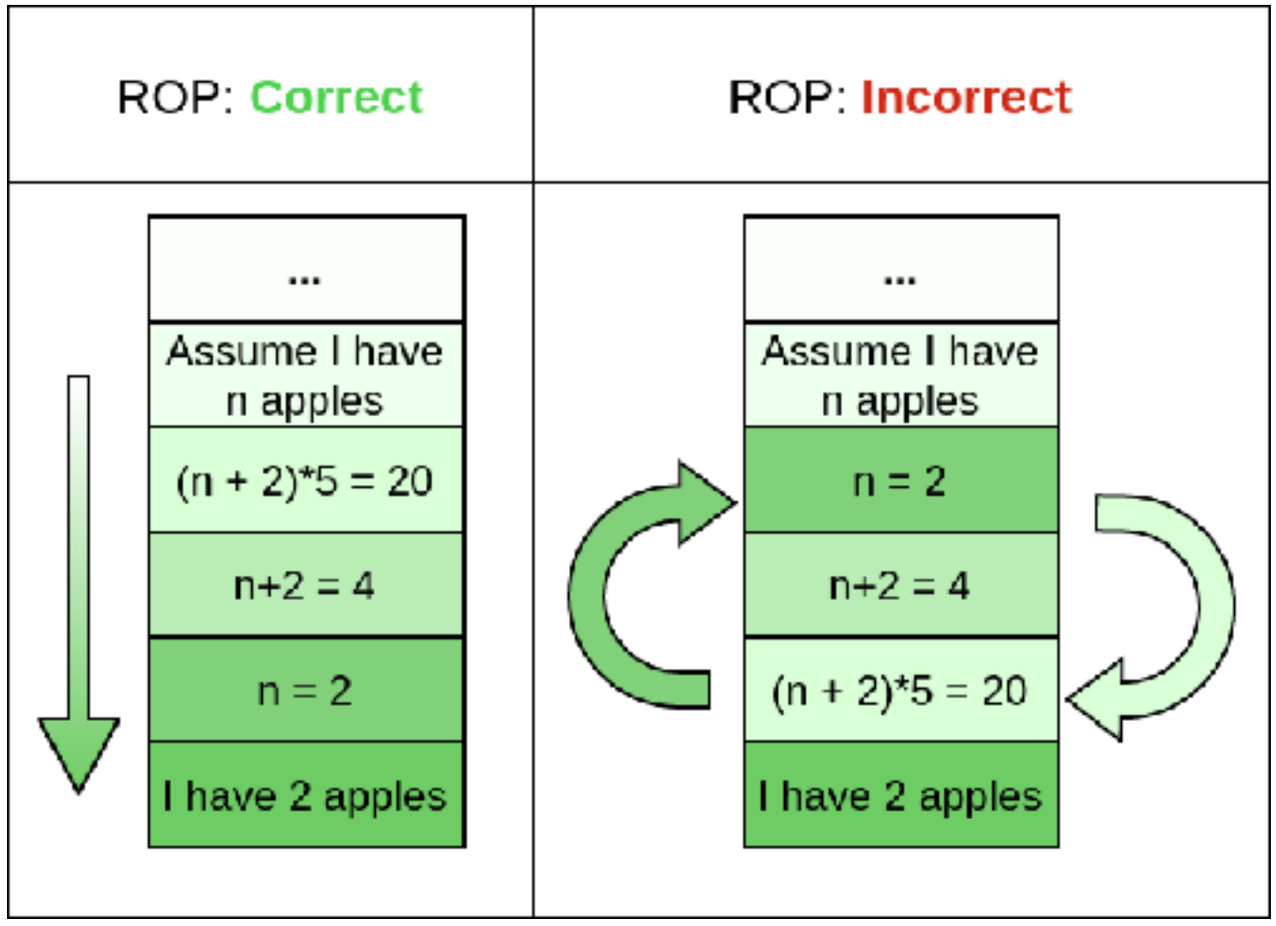
Assume that C was there in the business for x months

$$A : B : C = 40000 * 12 : 60000 * 10 : 120000 * x$$
$$= 40 * 12 : 60 * 10 : 120x = 40 : 5 * 10 : 10x$$
$$= 8 : 10 : 2x$$
$$= 4 : 5 : x$$
$$C\text{'s share} = 375000 * x/(9 + x) = 150000$$
$$=> 375x/(9 + x) = 150$$
$$=> 15x = 6(9 + x)$$
$$=> 5x = 18 + 2x$$
$$=> 3x = 18$$
$$=> x = 18/3 = 6$$

It means C was there in the business for 6 months. Given that B joined the business after 2 months. Hence C joined after 4 months after B joined
Answer is B

P. Piekos et al. “Measuring and Improving BERT's Mathematical Abilities by Predicting the Order of Reasoning”. ACL’21

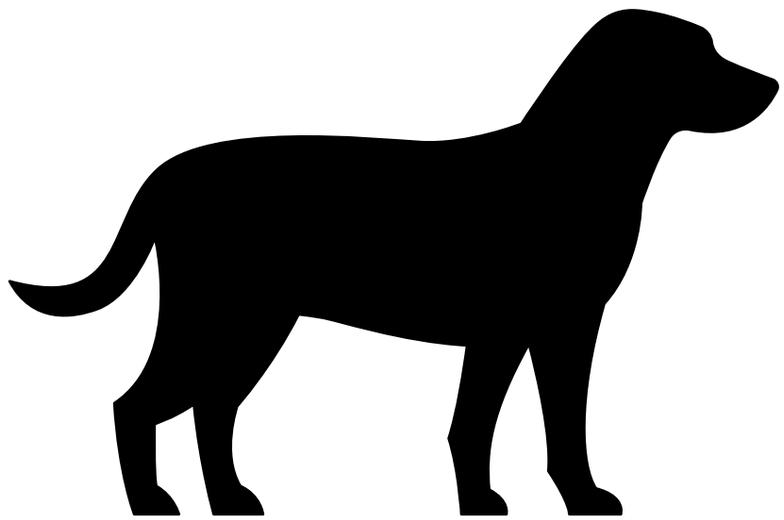
Self-supervision in reasoning



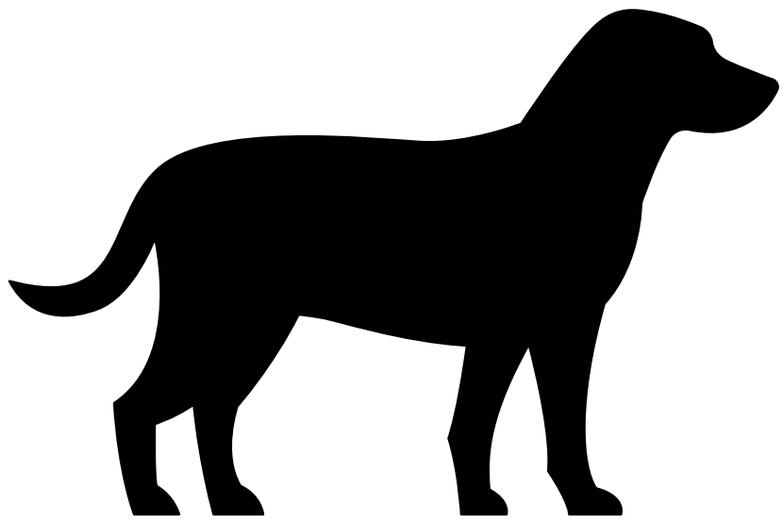
BERT	32.3
BERT + ROP	35.4
BERT + NROP	**37.0**

P. Piekos et al. “Measuring and Improving BERT's Mathematical Abilities by Predicting the Order of Reasoning”. ACL’21

But the world is multimodal

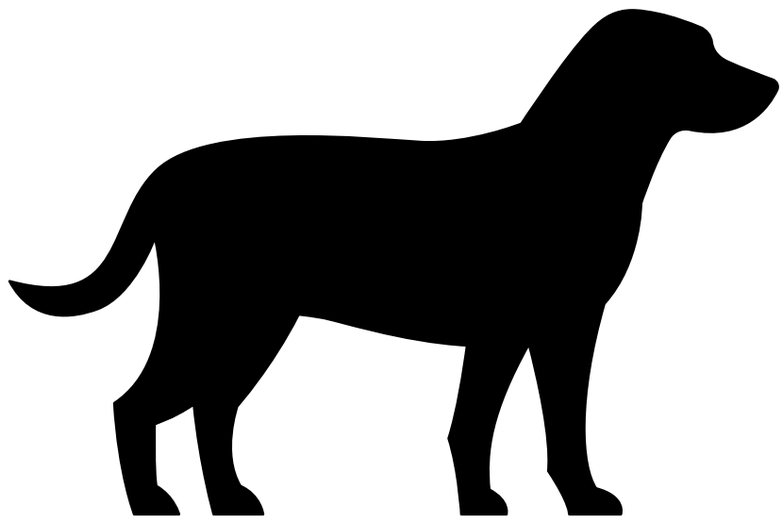


But the world is multimodal

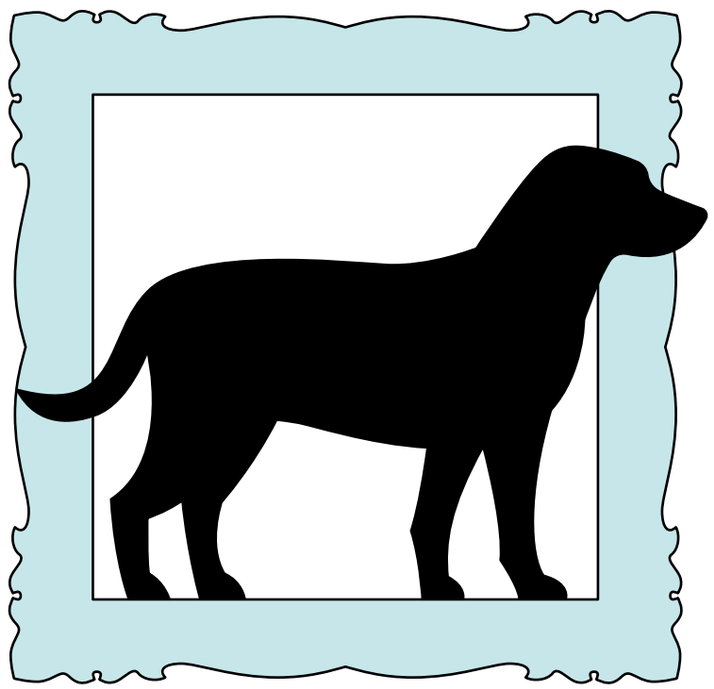


Dog

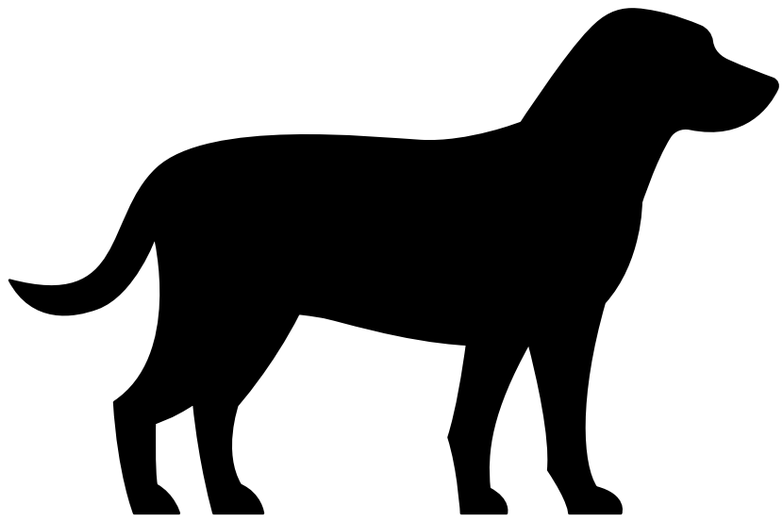
But the world is multimodal

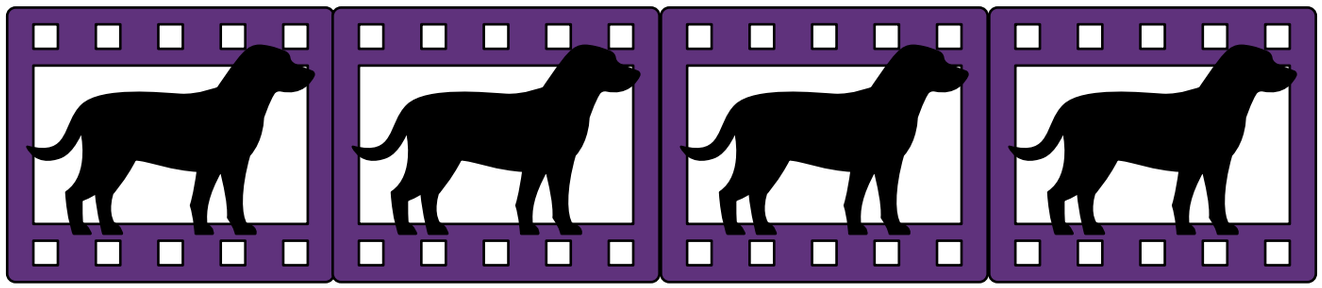


Dog

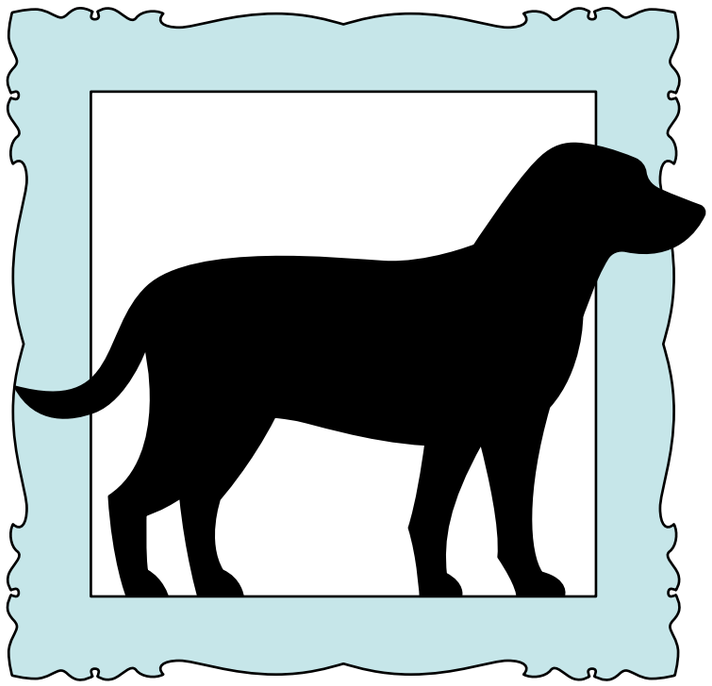


But the world is multimodal

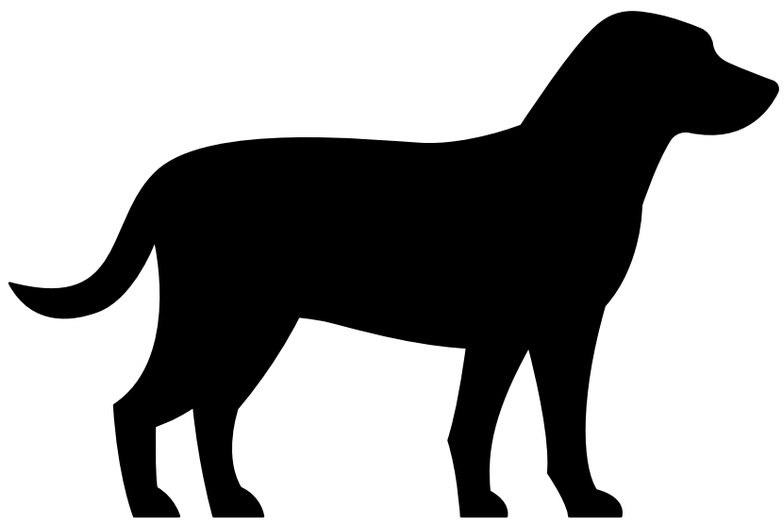


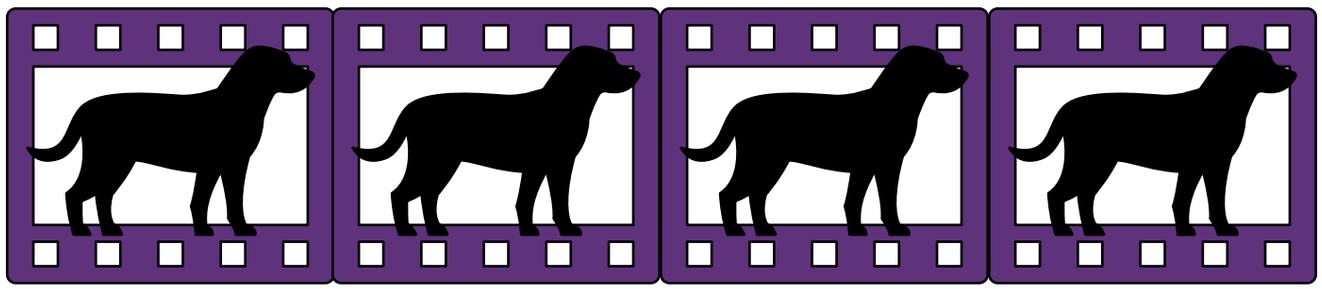


Dog

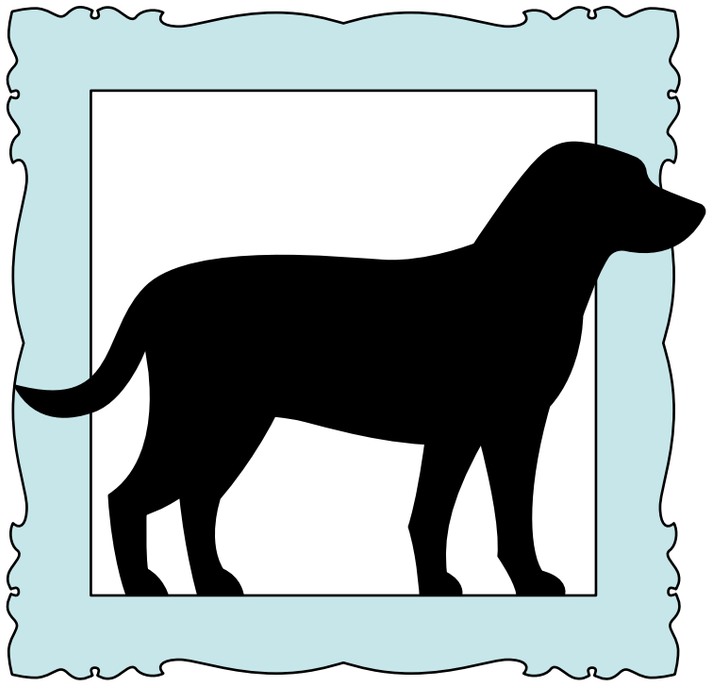


But the world is multimodal

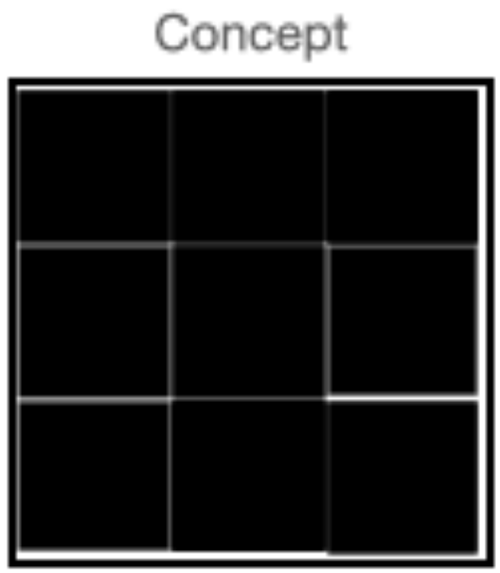




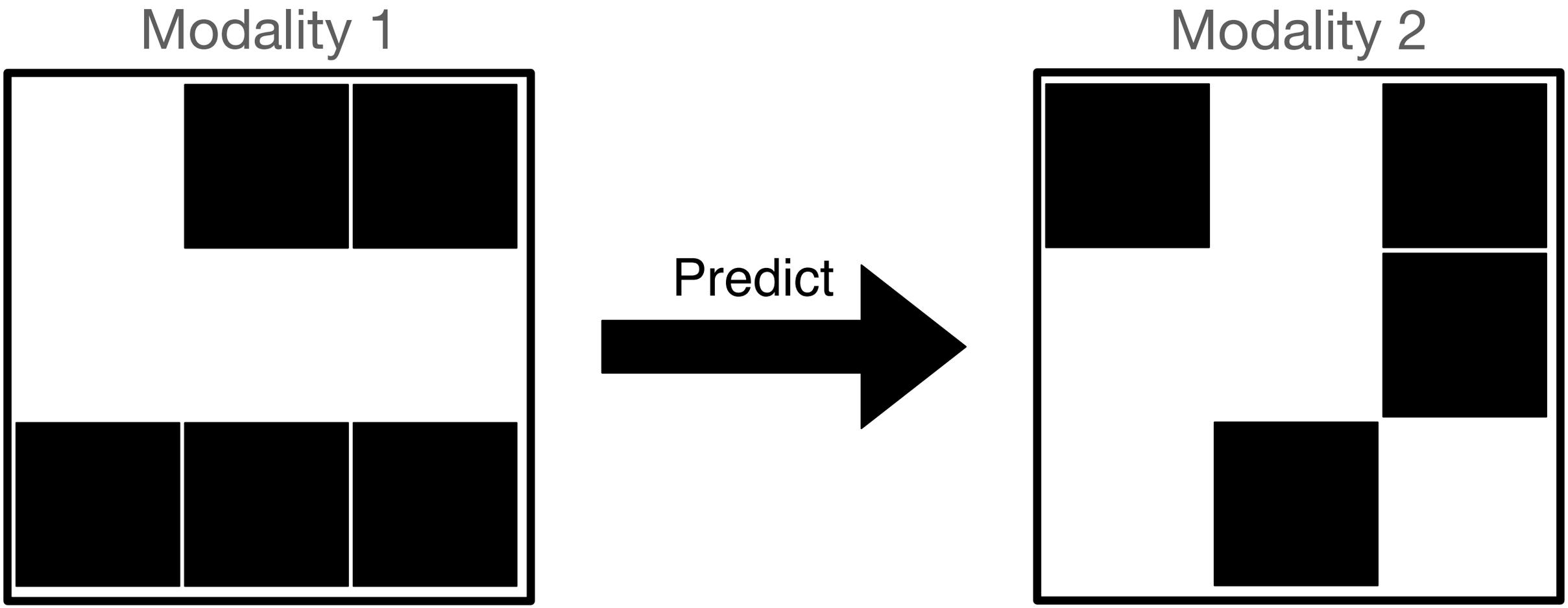
Dog



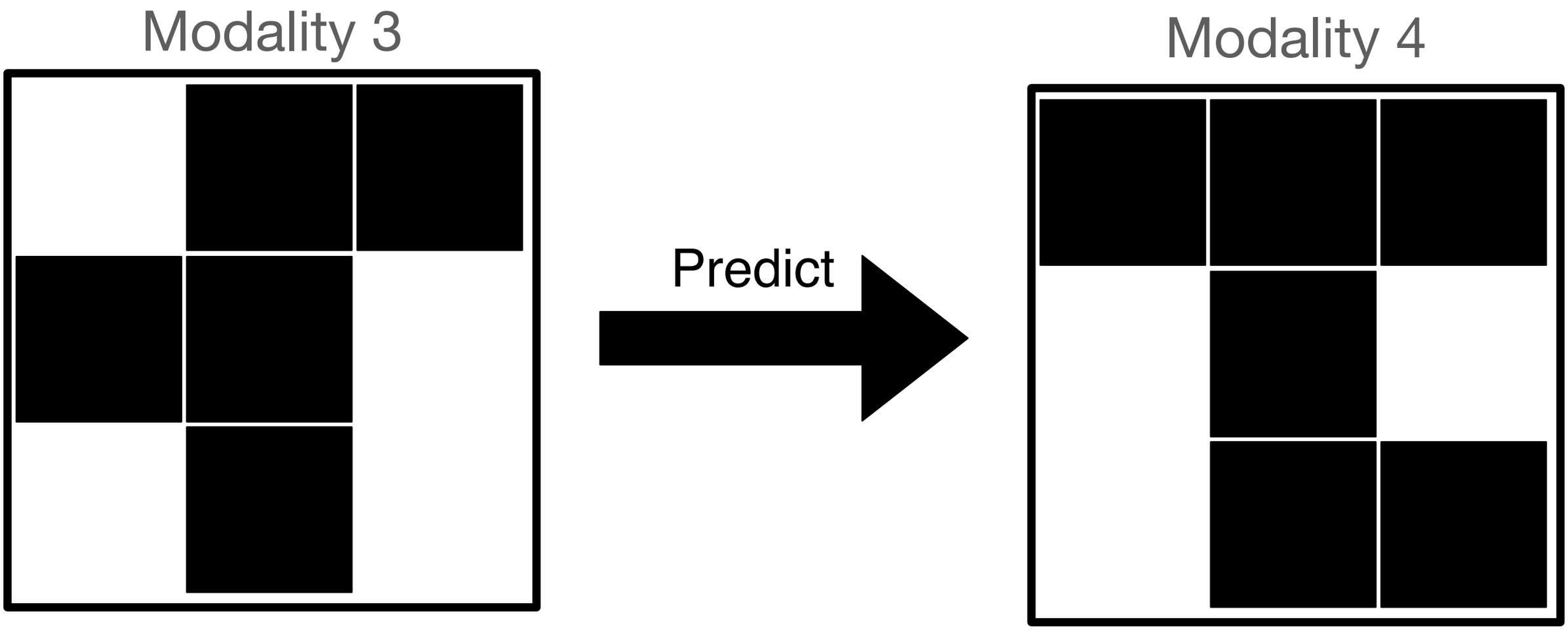
Every modality is a piece of a larger puzzle



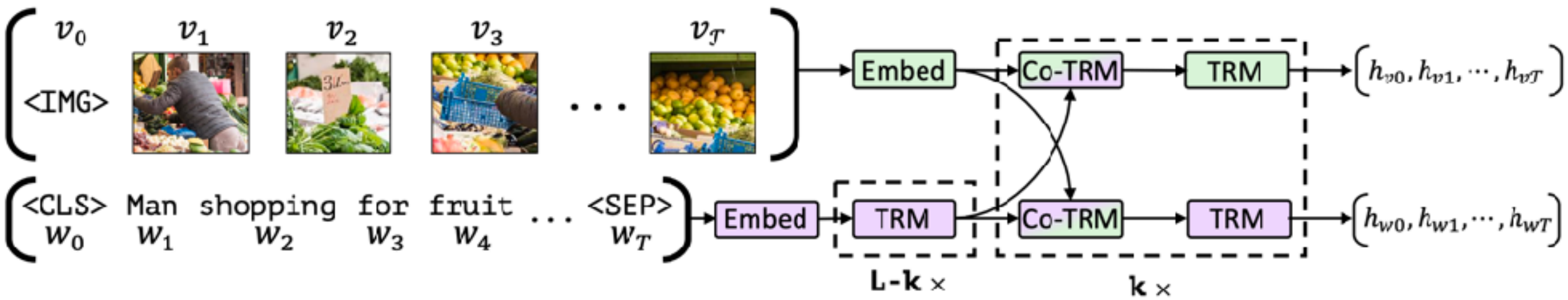
Modalities are supporting each other



Modalities are supporting each other

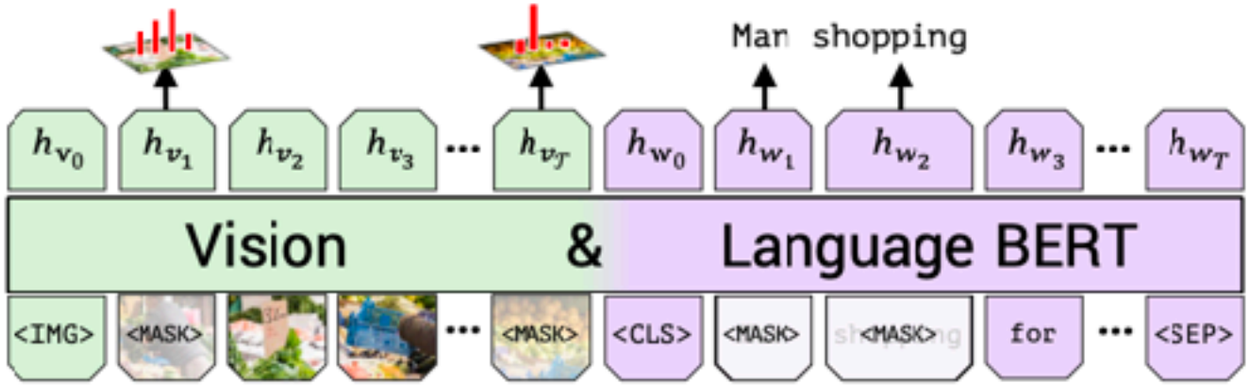


Self-supervised learning in Vision + Language

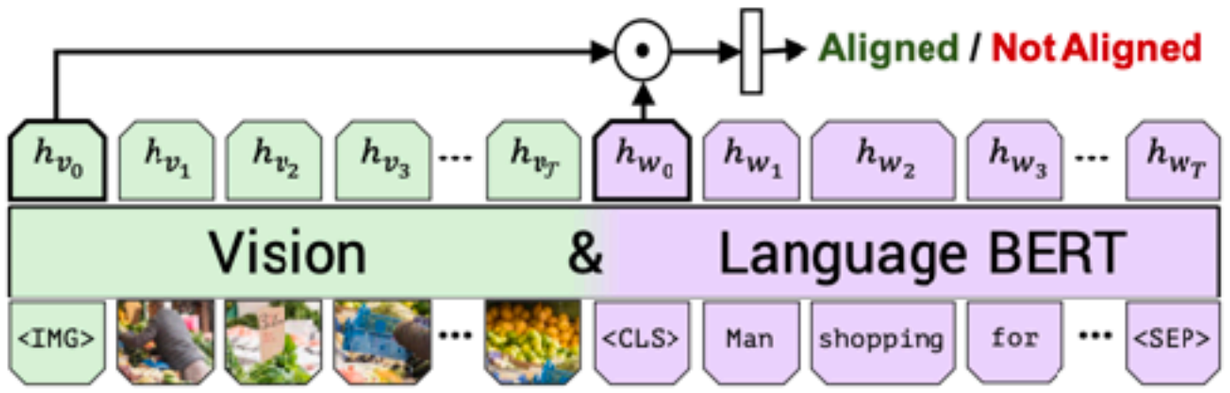


J. Lue et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”

Self-supervised learning in Vision + Language



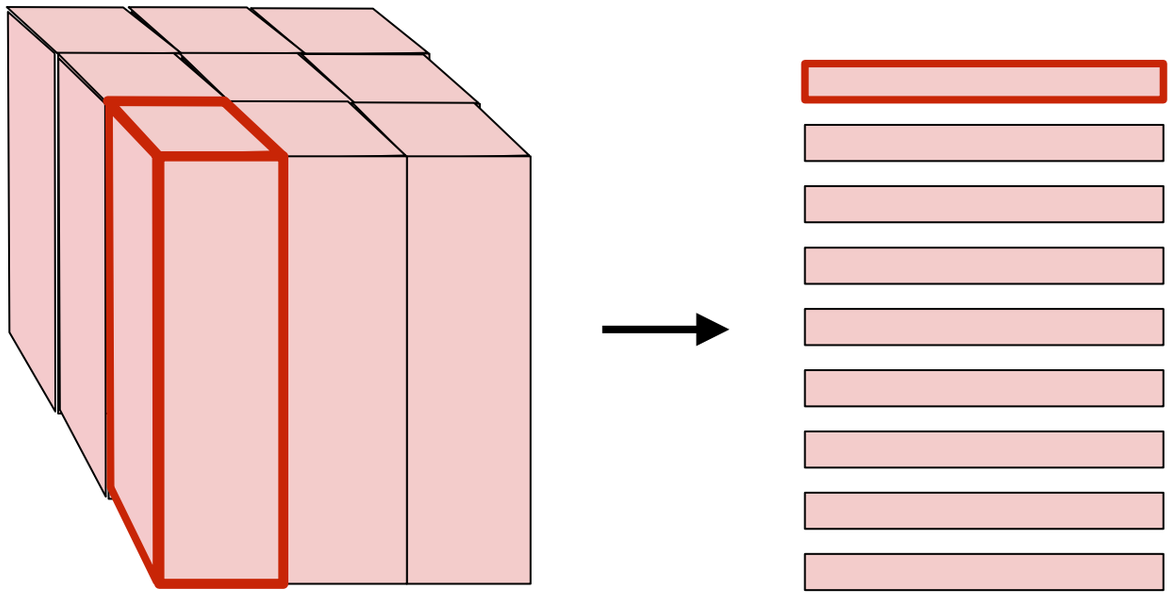
(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

J. Lue et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”

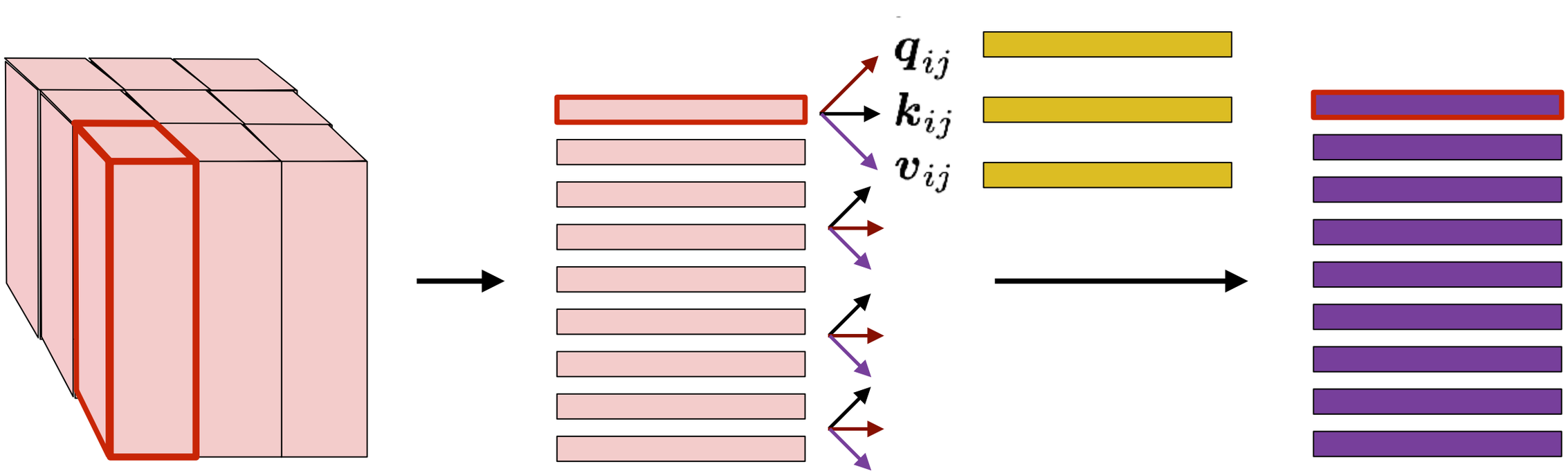
Self-attention (Transformer)



Self-attention (Transformer)

$$\boldsymbol{q}_{ij} := \boldsymbol{W}_q \boldsymbol{x}_{ij}$$
$$\boldsymbol{k}_{ij} := \boldsymbol{W}_k \boldsymbol{x}_{ij}$$
$$\boldsymbol{v}_{ij} := \boldsymbol{W}_v \boldsymbol{x}_{ij}$$

$$\tilde{\boldsymbol{x}}_{lk} := \sum_{ij} \text{softmax}\left(\boldsymbol{q}_{lk}^T \boldsymbol{k}_{ij}\right) \boldsymbol{v}_{ij}$$

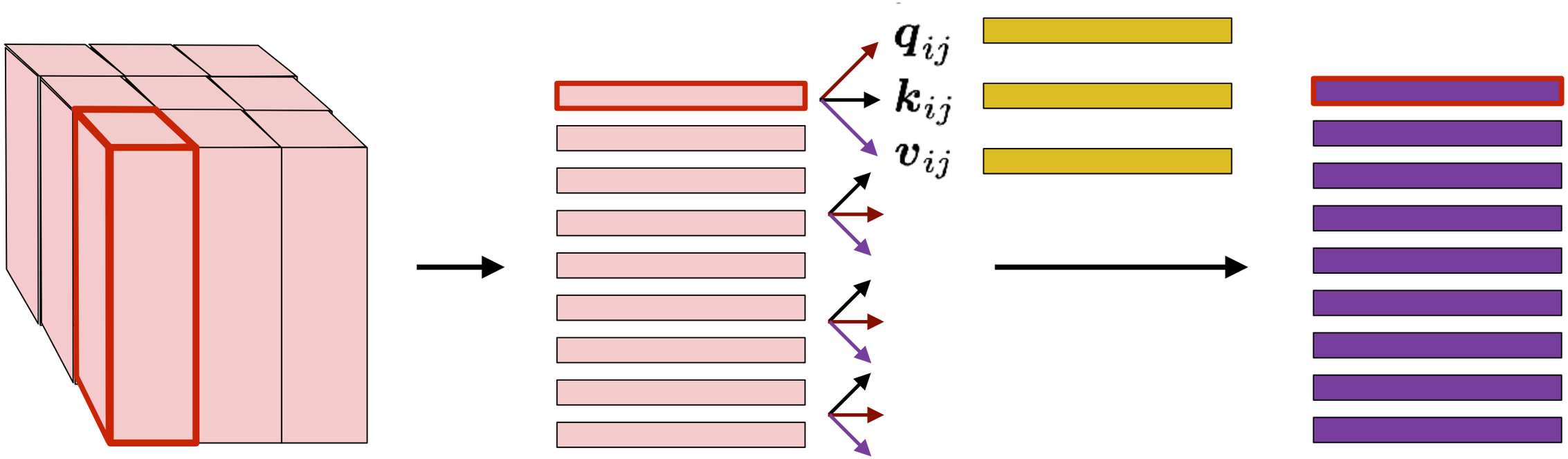


Self-attention (Transformer)

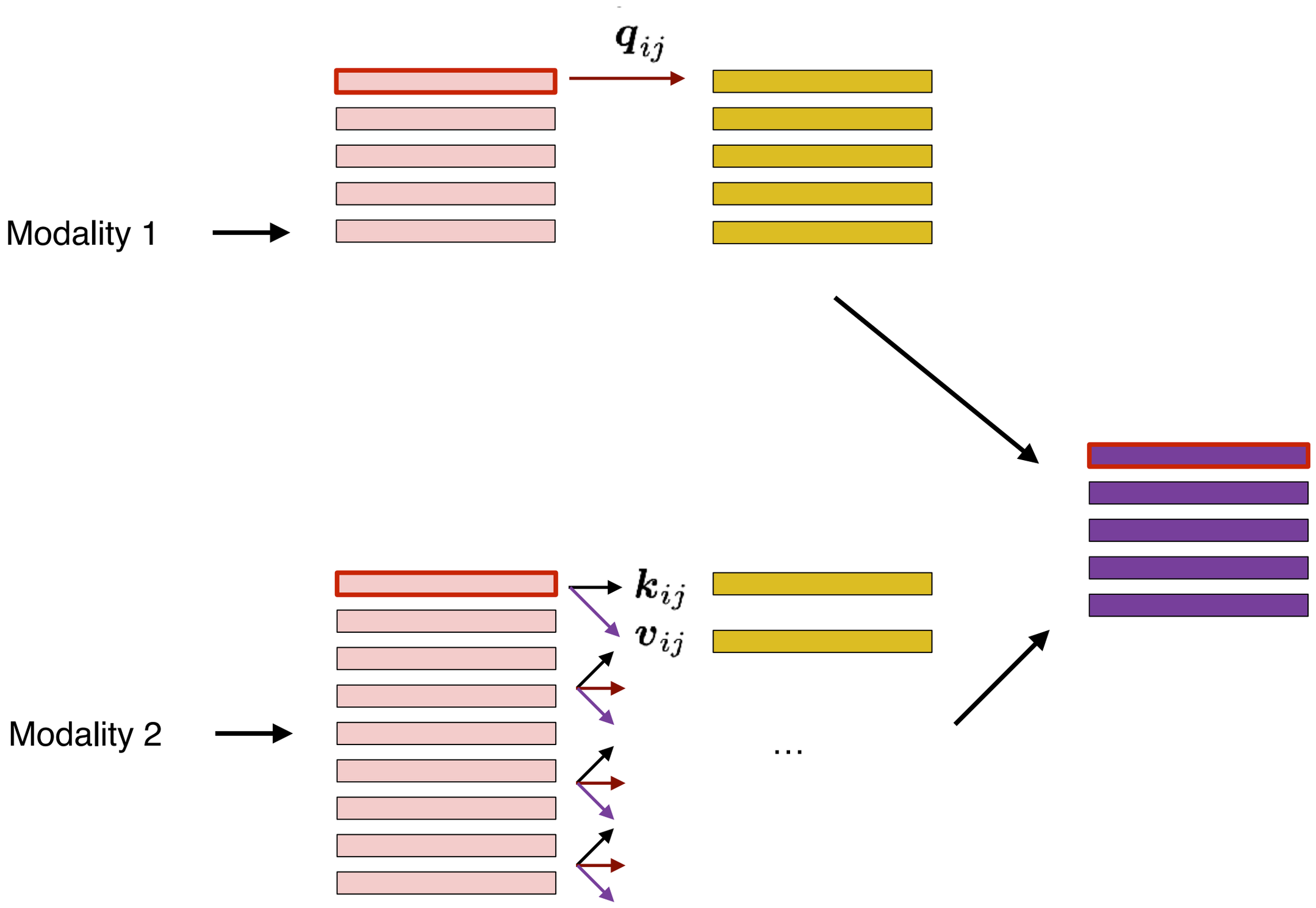
$$\boldsymbol{q}_{ij} := \boldsymbol{W}_q \boldsymbol{x}_{ij}$$
$$\boldsymbol{k}_{ij} := \boldsymbol{W}_k \boldsymbol{x}_{ij}$$
$$\boldsymbol{v}_{ij} := \boldsymbol{W}_v \boldsymbol{x}_{ij}$$

$$\tilde{\boldsymbol{x}}_{lk} := \sum_{ij} \text{softmax}\left(\boldsymbol{q}_{lk}^T \boldsymbol{k}_{ij}\right) \boldsymbol{v}_{ij}$$

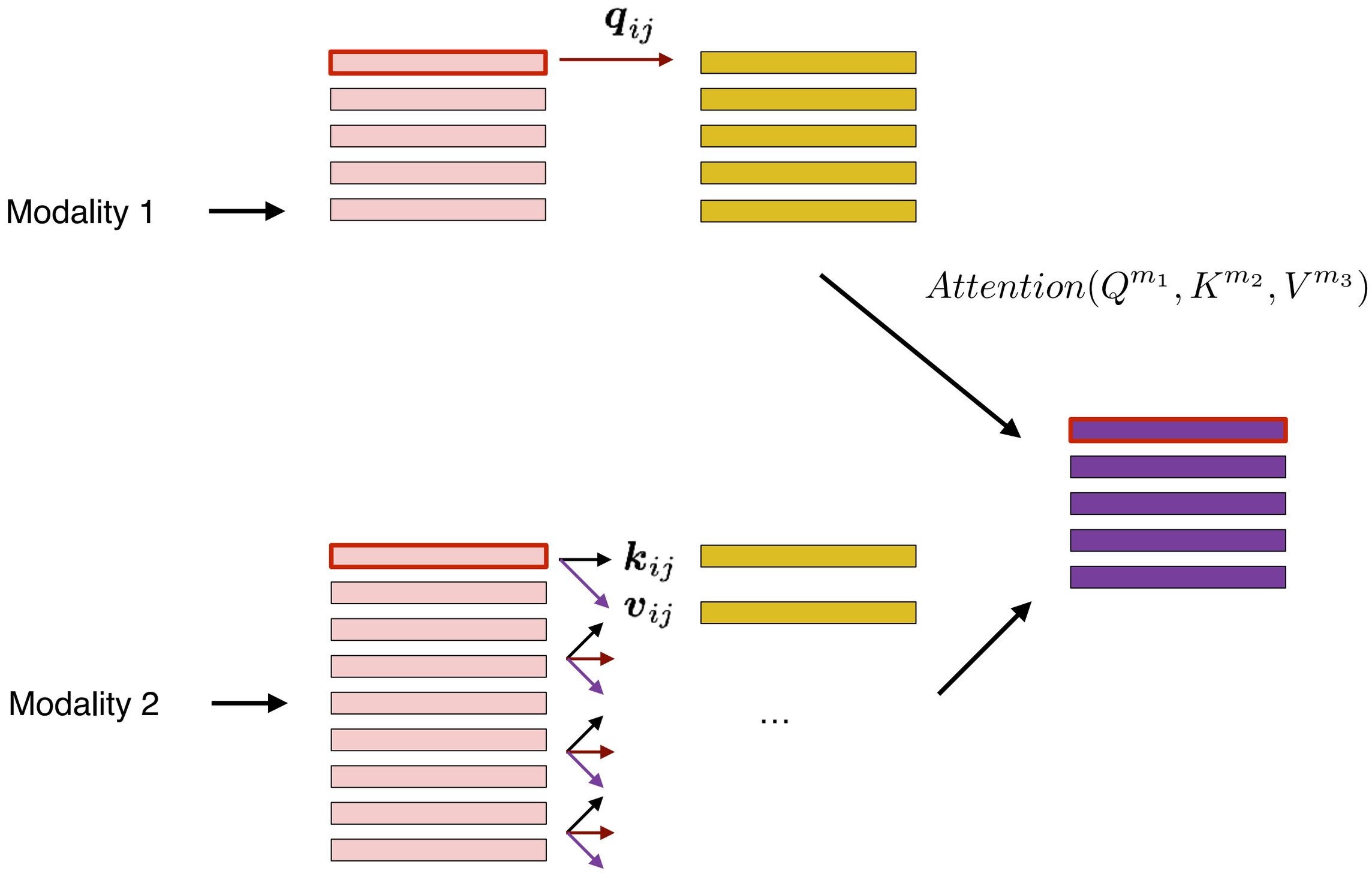
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



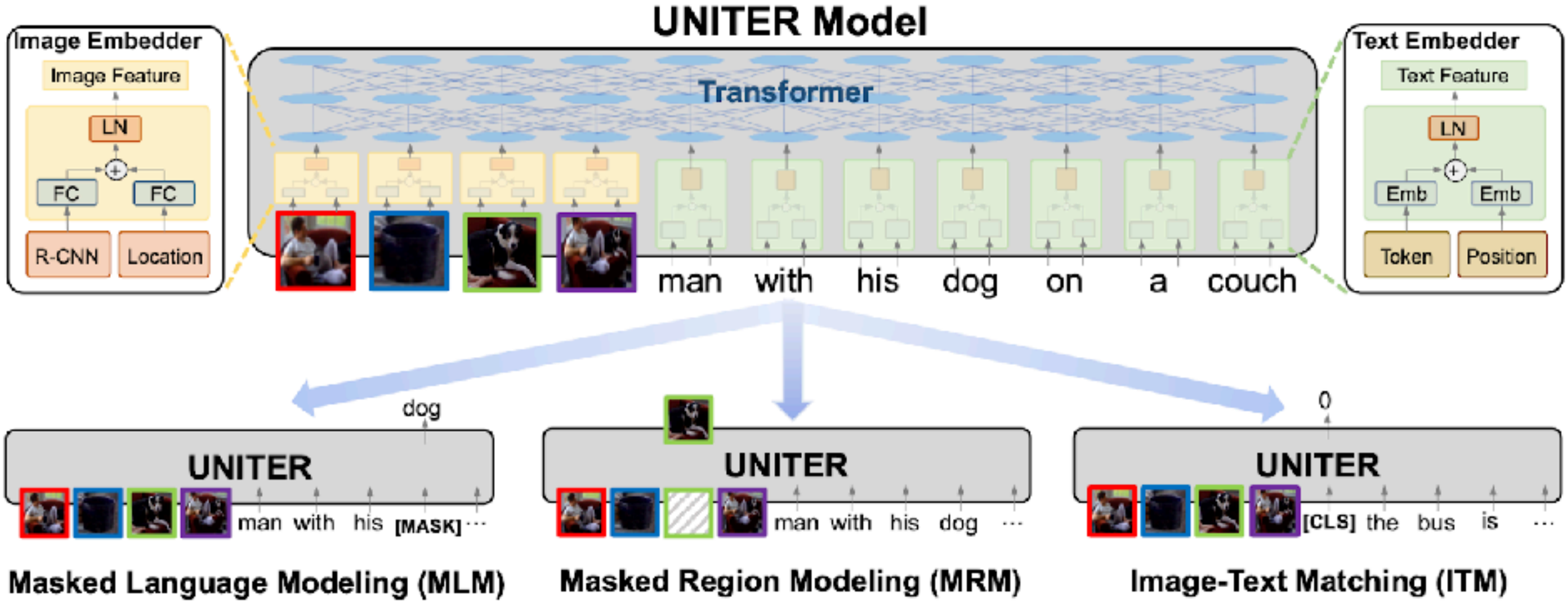
Cross-attention



Cross-attention

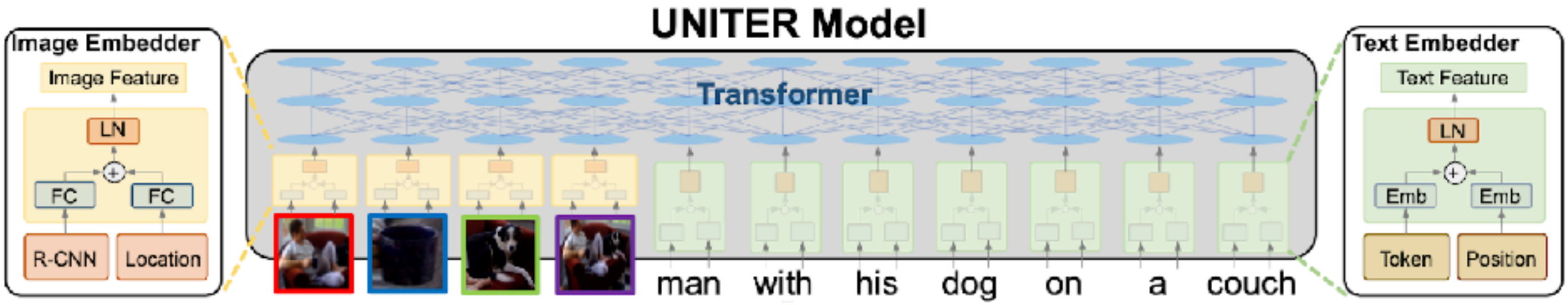


Uniter


UNITER Model
Image Embedder
Image Feature
LN
FC
FC
R-CNN
Location
Transformer
man with his dog on a couch
Text Embedder
Text Feature
LN
Emb
Emb
Token
Position
dog
UNITER
man with his [MASK] ···
Masked Language Modeling (MLM)
UNITER
man with his dog ···
Masked Region Modeling (MRM)
0
UNITER
[CLS] the bus is ···
Image-Text Matching (ITM)

Uniter

- Consider image regions (Faster R-CNN)
- Consider words (WordPieces)
- Transformer that combines both representation

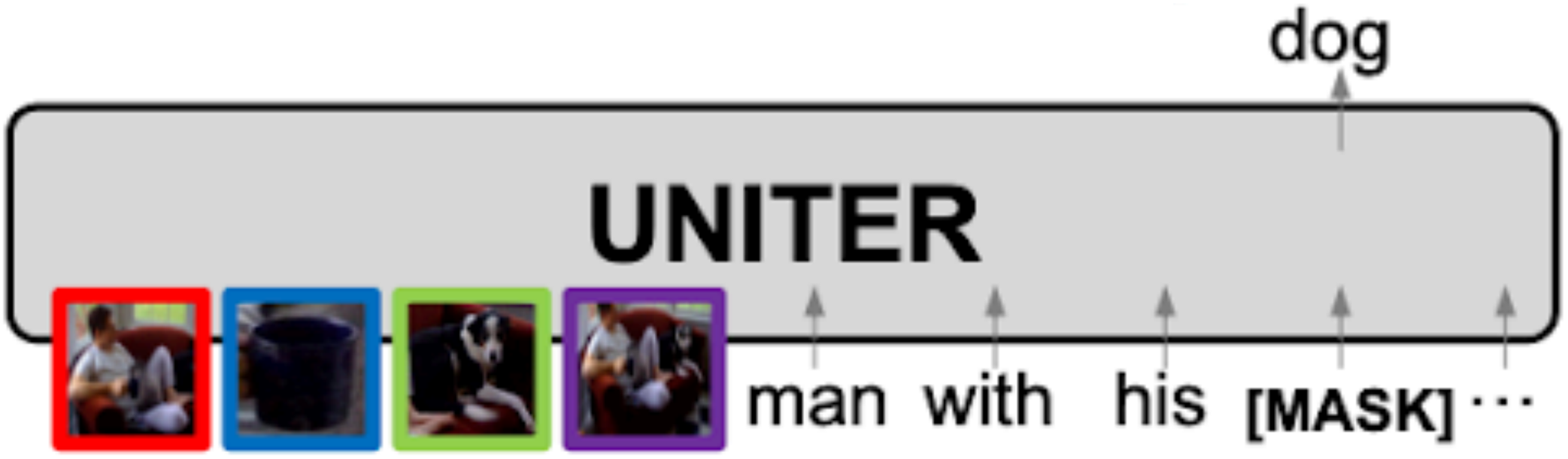


Y. Chen, et al. “UNITER: Learning UNiversal Image-TExt Representations”

Uniter

- Mask words randomly (mask = token)
- Signal of supervision: predict the masked words given context

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m}|\mathbf{w}_{\backslash\mathbf{m}}, \mathbf{v})$$

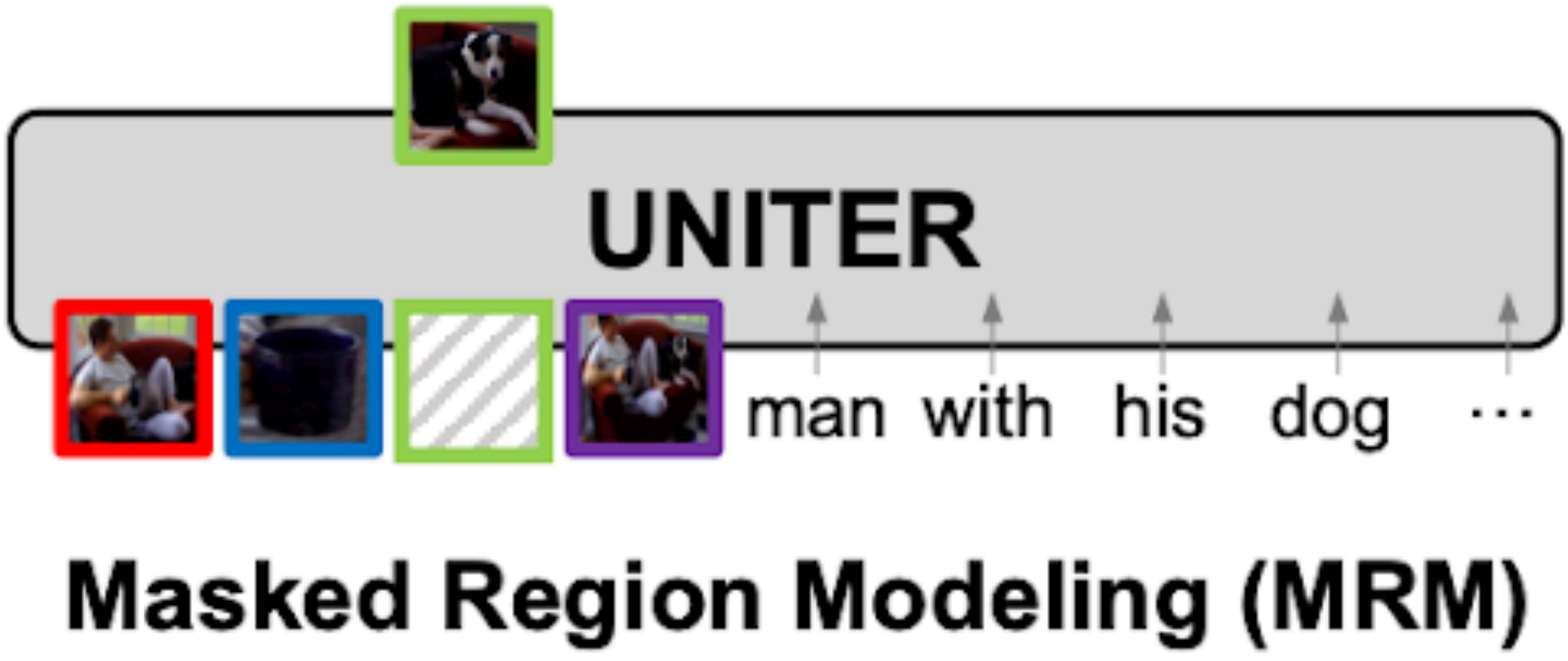


Y. Chen, et al. “UNITER: Learning UNiversal Image-TExt Representations”

Uniter

- Mask words randomly (mask = ‘zero image’)
- Signal of supervision: predict the masked regions given context

$$\mathcal{L}_{\mathrm{MRM}}(\theta) = E_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w})$$



Y. Chen, et al. “UNITER: Learning UNiversal Image-TExt Representations”

Masked Region Modelling

- Unlike words (classes), regions are high dimensional
  - Feature regression $f_\theta(\mathbf{v_m}|\mathbf{v}_{\setminus\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \|h_\theta(\mathbf{v_m}^{(i)}) - r(\mathbf{v_m}^{(i)})\|_2^2$
  - Region classification
    - output of the object detection as ground truth class

$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\setminus\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \mathrm{CE}(c(\mathbf{v_m}^{(i)}), g_\theta(\mathbf{v_m}^{(i)}))$$

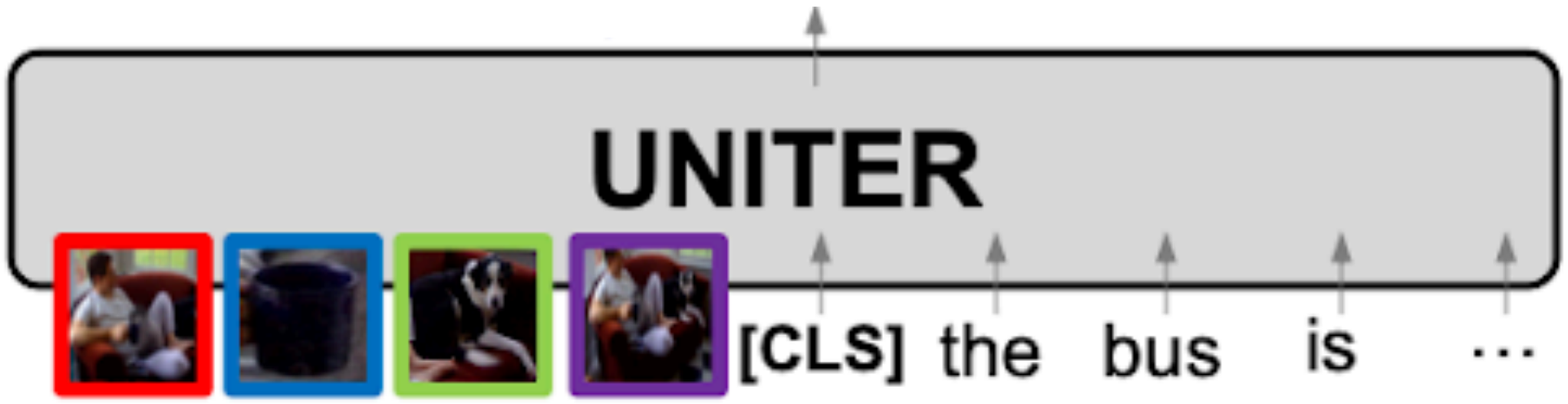
  - Region class regression
    - Regress to the object detection distribution rather than to individual classes (distillation)

$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\setminus\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\mathbf{v_m}^{(i)})||g_\theta(\mathbf{v_m}^{(i)}))$$

Y. Chen, et al. “UNITER: Learning UNiversal Image-TExt Representations”

Uniter

- [CLS] token ‘separates’ modalities
- [CLS] as the joint representation
- Sample positive/negative pairs (negative: image or text replaced by a random data point) — binary classification with cross-entropy

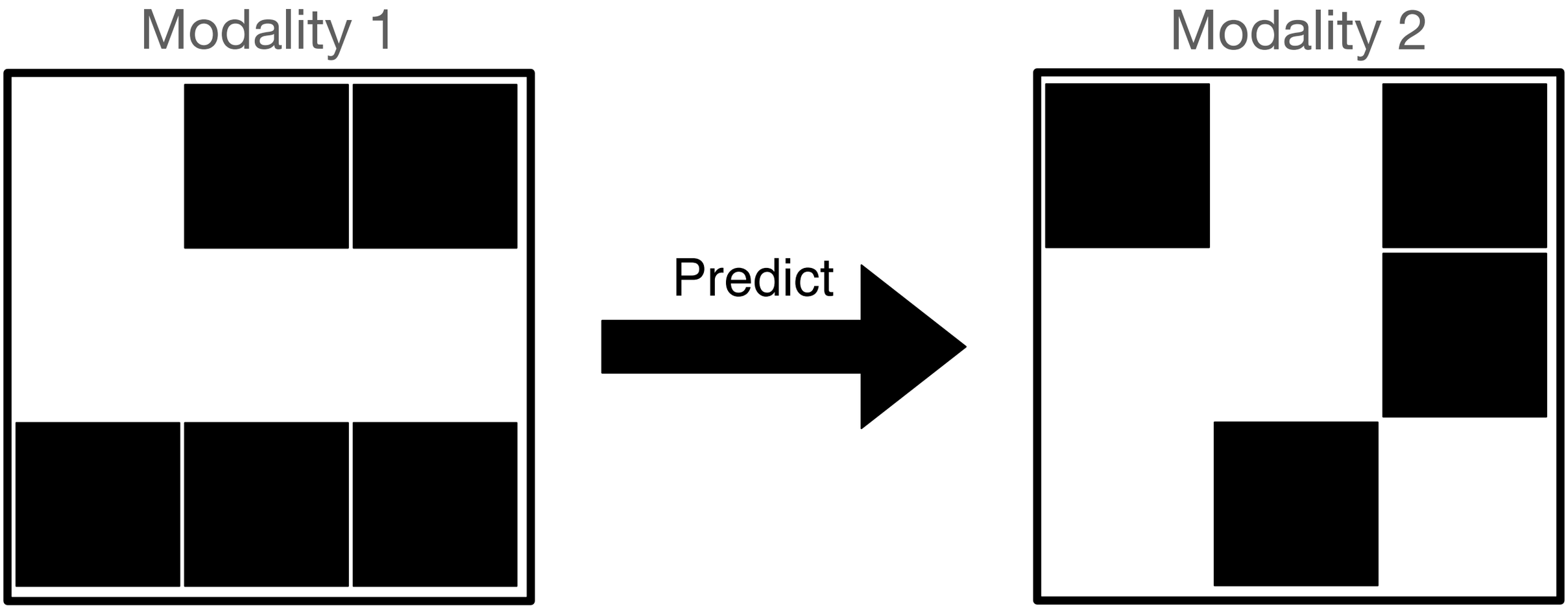


Y. Chen, et al. “UNITER: Learning UNiversal Image-TExt Representations”

Results: Train on captions; test on downstream tasks

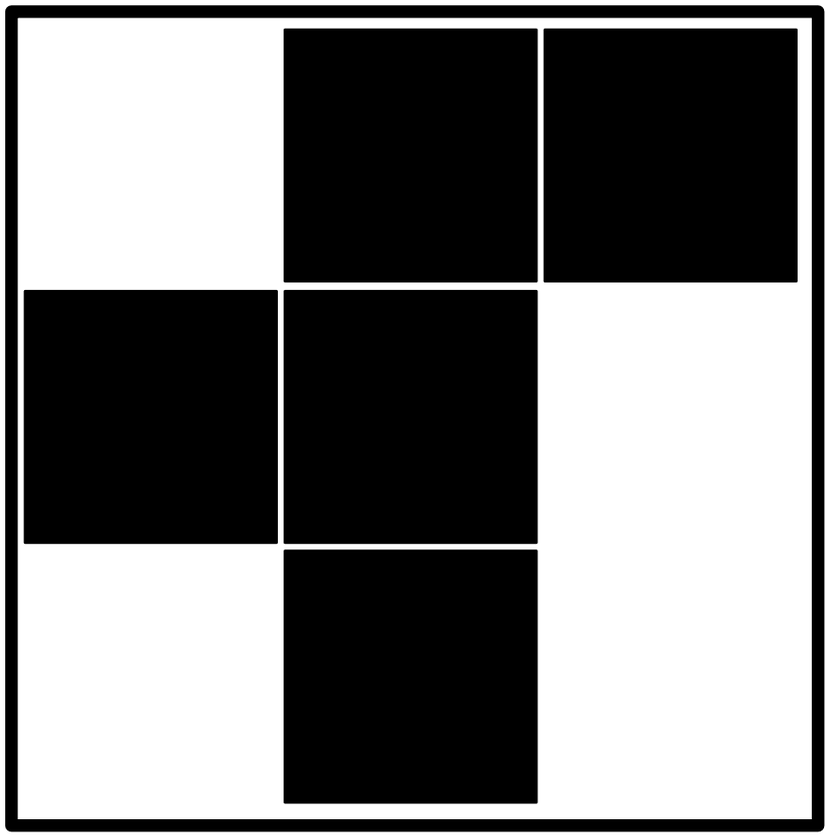
Tasks		SOTA	ViLBERT	VLBERT	Unicoder -VL	VisualBERT	LXMERT	UNITER BASE	UNITER LARGE
VQA	test-dev	70.63	70.55	70.50	-	70.80	72.42	72.27	**73.24**
	test-std	70.90	70.92	70.83	-	71.00	72.54	72.46	**73.40**
VCR	Q→A	72.60	73.30	74.00	-	71.60	-	75.00	**77.30**
	QA→R	75.70	74.60	74.80	-	73.20	-	77.20	**80.80**
	Q→AR	55.00	54.80	55.50	-	52.40	-	58.20	**62.80**
NLVR^2	dev	54.80	-	-	-	67.40	74.90	77.14	**78.40**
	test-P	53.50	-	-	-	67.00	74.50	77.87	**79.50**
SNLI-VE	val	71.56	-	-	-	-	-	78.56	**79.28**
	test	71.16	-	-	-	-	-	78.02	**78.98**
ZS IR (Flickr)	R@1	-	31.86	-	42.40	-	-	62.34	**65.82**
	R@5	-	61.12	-	71.80	-	-	85.62	**88.88**
	R@10	-	72.80	-	81.50	-	-	91.48	**93.52**
IR (Flickr)	R@1	48.60	58.20	-	68.30	-	-	71.50	**73.66**
	R@5	77.70	84.90	-	90.30	-	-	91.16	**93.06**
	R@10	85.20	91.52	-	94.60	-	-	95.20	**95.98**
IR (COCO)	R@1	38.60	-	-	44.50	-	-	48.42	**51.72**
	R@5	69.30	-	-	74.40	-	-	76.68	**78.41**
	R@10	80.40	-	-	84.00	-	-	85.90	**86.93**
ZS TR (Flickr)	R@1	-	-	-	61.60	-	-	75.10	**77.50**
	R@5	-	-	-	84.80	-	-	93.70	**96.30**
	R@10	-	-	-	90.10	-	-	95.50	**98.50**
TR (Flickr)	R@1	67.90	-	-	82.30	-	-	84.70	**88.20**
	R@5	90.30	-	-	95.10	-	-	97.10	**98.40**
	R@10	95.80	-	-	97.80	-	-	99.00	**99.00**
TR (COCO)	R@1	50.40	-	-	59.60	-	-	63.28	**66.60**
	R@5	82.20	-	-	85.10	-	-	87.04	**89.42**
	R@10	90.00	-	-	91.80	-	-	93.08	**94.26**
Ref-COCO	val	87.51		-	-	-	-	91.64	**91.84**
	testA	89.02						92.26	**92.65**
	testB	87.05	-	-	-	-	-	90.46	**91.19**
	val^d	77.48	-	-	-	-	-	81.24	**81.41**
	testA^d	83.37						86.48	**87.04**
	testB^d	70.32	-	-	-	-	-	73.94	**74.17**
Ref-COCO+	val	75.38	-	78.44	-	-	-	82.84	**84.04**
	testA	80.04	-	81.30	-	-	-	85.70	**85.87**
	testB	69.30	-	71.18	-	-	-	78.11	**78.89**
	val^d	68.19	72.34	71.84	-	-	-	74.72	**74.94**
	testA^d	75.97	78.52	77.59	-	-	-	80.65	**81.37**
	testB^d	57.52	62.61	60.57	-	-	-	65.15	**65.35**
Ref-COCOg	val	81.76	-	-	-	-	-	86.52	**87.85**
	test	81.75	-	-	-	-	-	86.52	**87.73**
	val^d	68.22	-	-	-	-	-	74.31	**74.86**
	test^d	69.46	-	-	-	-	-	74.51	**75.77**

Modalities are supporting each other

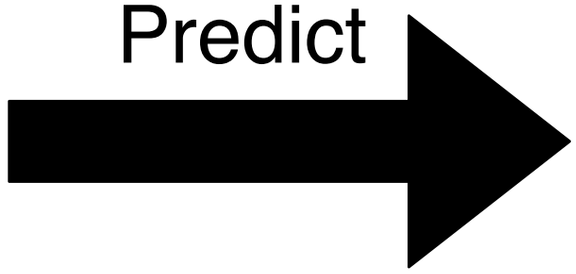


Modalities are supporting each other

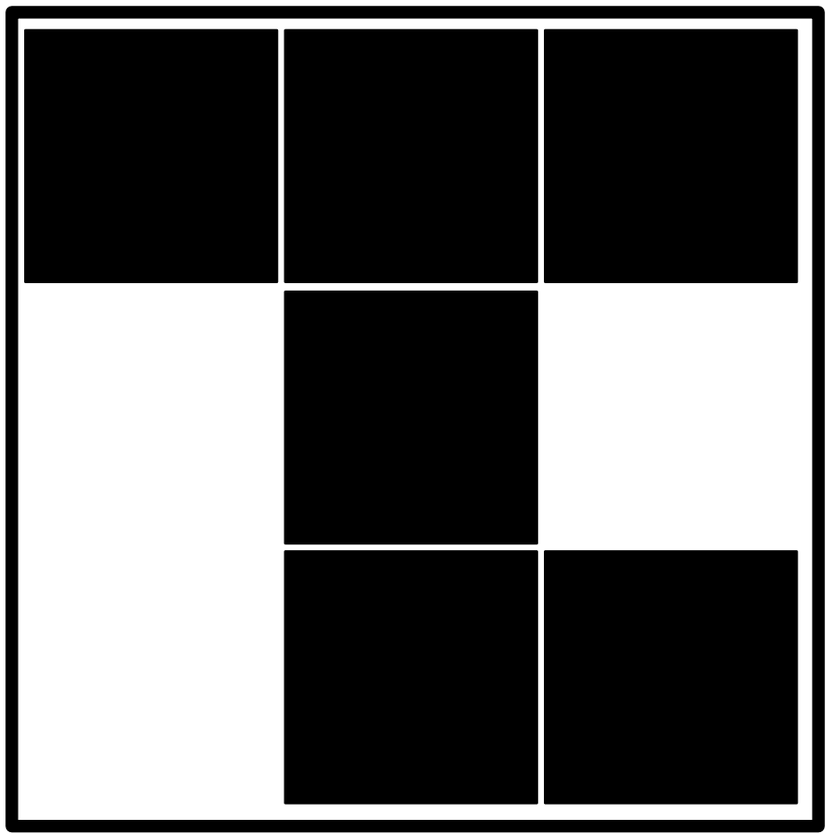
Modality 3



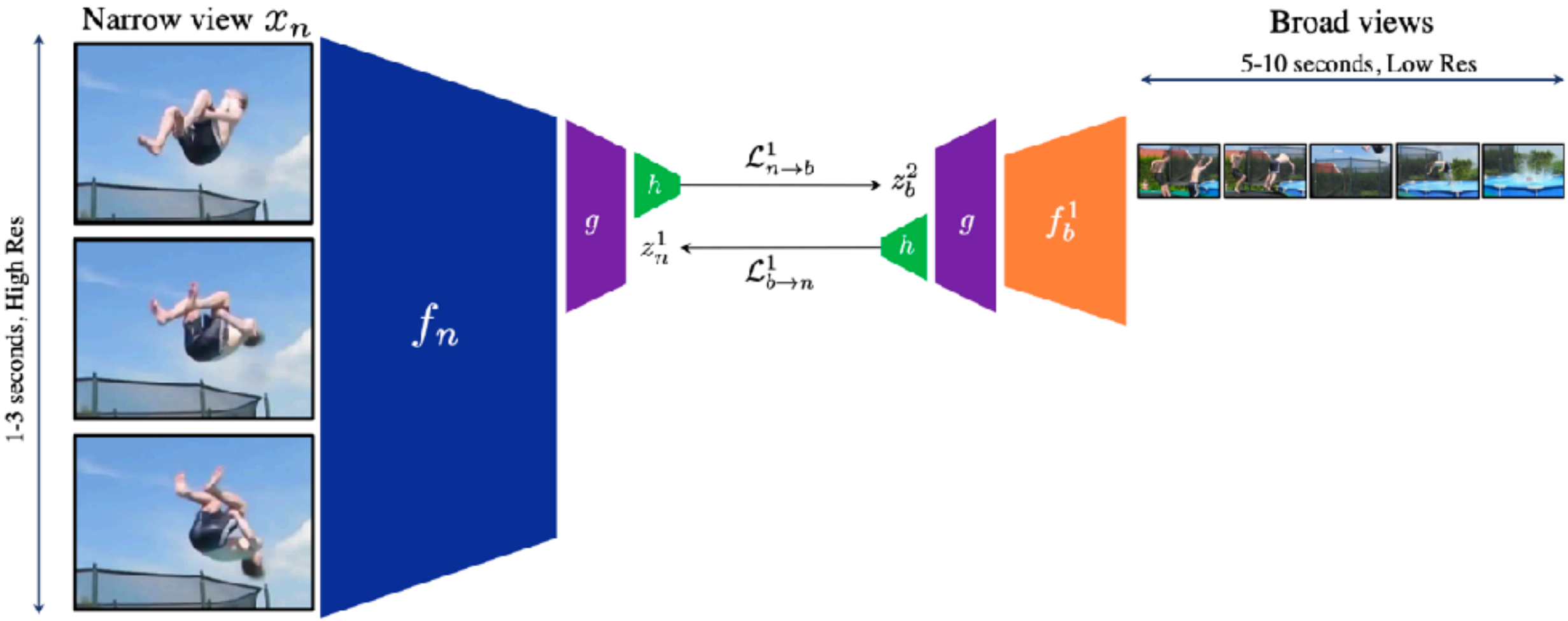
Predict



Modality 4

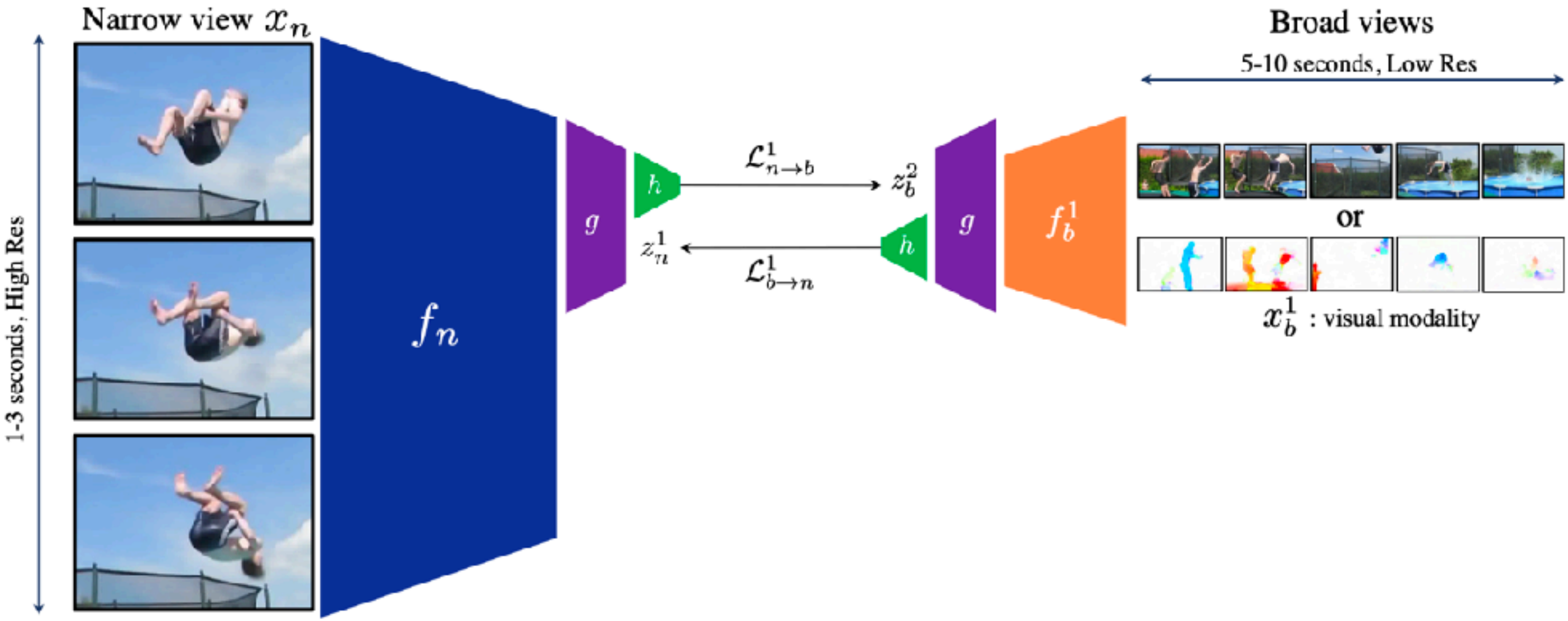


Using more modalities



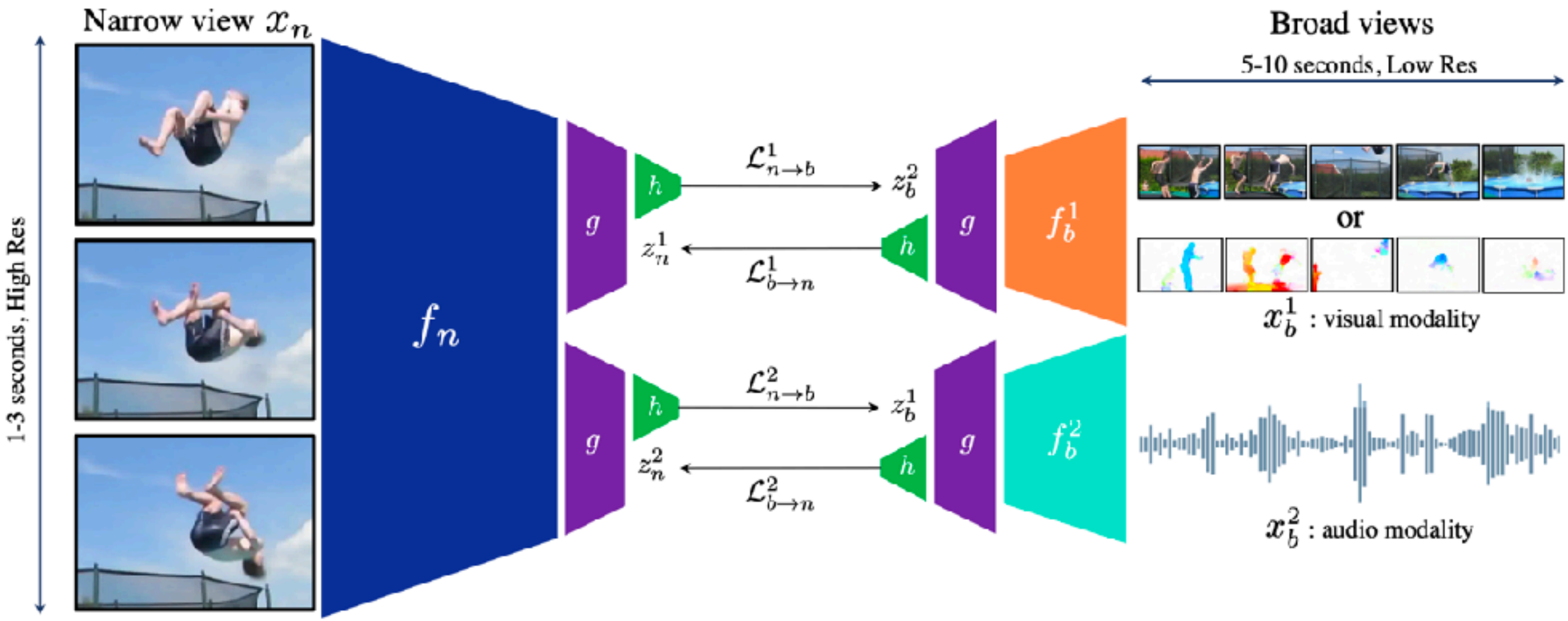
A. Recasens et al. “Broaden Your Views for Self-Supervised Video Learning”. ICCV’21

Using more modalities



A. Recasens et al. “Broaden Your Views for Self-Supervised Video Learning”. ICCV’21

Using more modalities



A. Recasens et al. “Broaden Your Views for Self-Supervised Video Learning”. ICCV’21

BRAVE: Results

	UCF101
SMART [1]	98.6
RGB [ours, 2]	90.5
Flow [ours, 2]	92.1
Flow + Audio [ours, 2]	93.2 (96.9)*

[1] S. Gowda et al. “SMART Frame Selection for Action Recognition”. AAAI-21
[2] A. Recasens et al. “Broaden Your Views for Self-Supervised Video Learning”. ICCV’21

* denotes fine-tuning

Learning multiple-languages via vision

G. Sigurdsson et al. “Visual Grounding in Video for Unsupervised Word Translation” CVPR’20

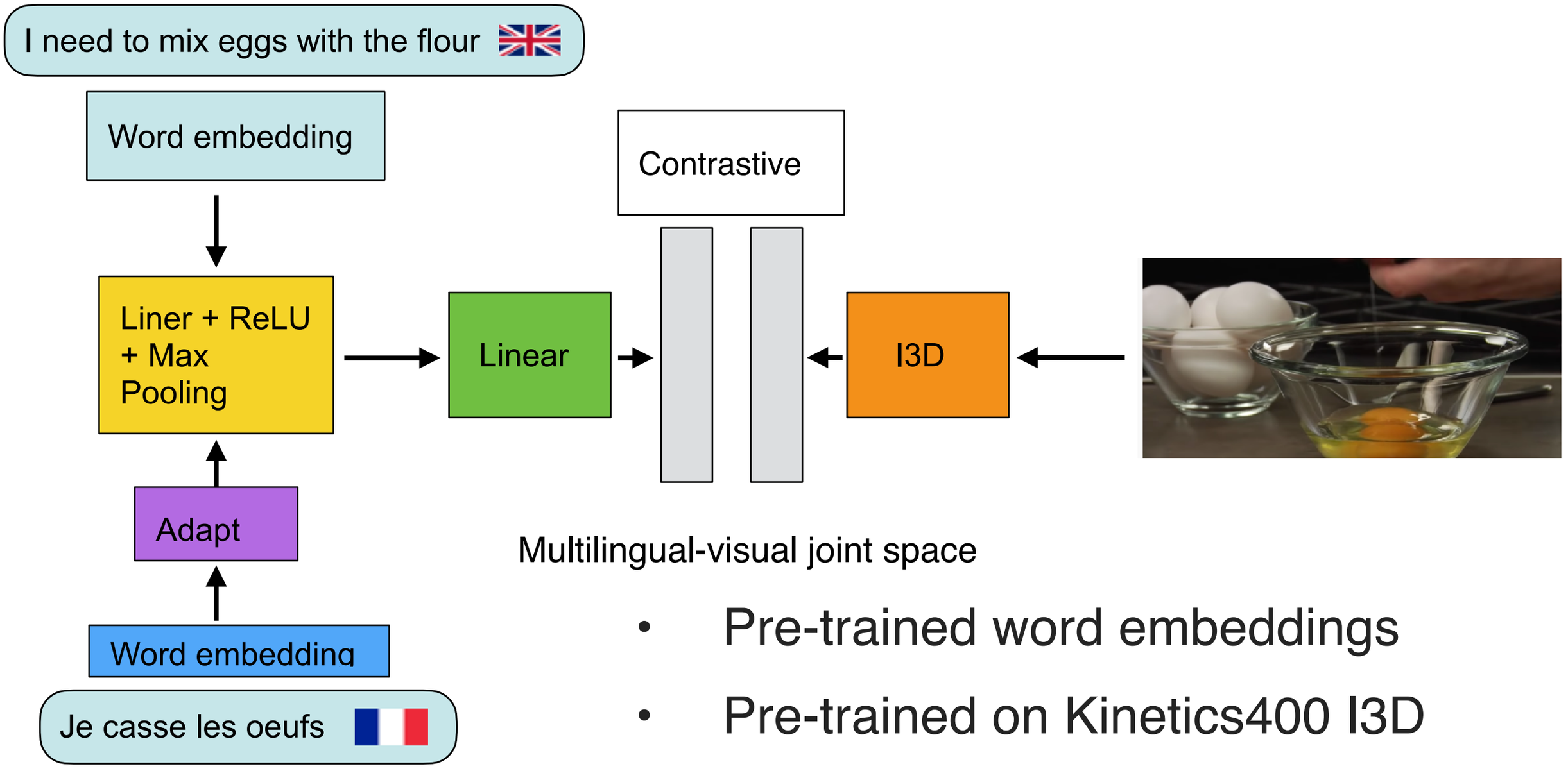
Different languages
Unpaired videos
Not exactly the same situation
(but similar)

Je casse les oeufs. = I break eggs

I need to mix eggs with the flour

- Word mapping between two languages
  - Embeddings in monolingual languages but no mapping between languages
- Can vision help in language?
  - Monolingual languages have their own videos
  - Videos are not paired
- Analogy to a small child born in a bilingual family

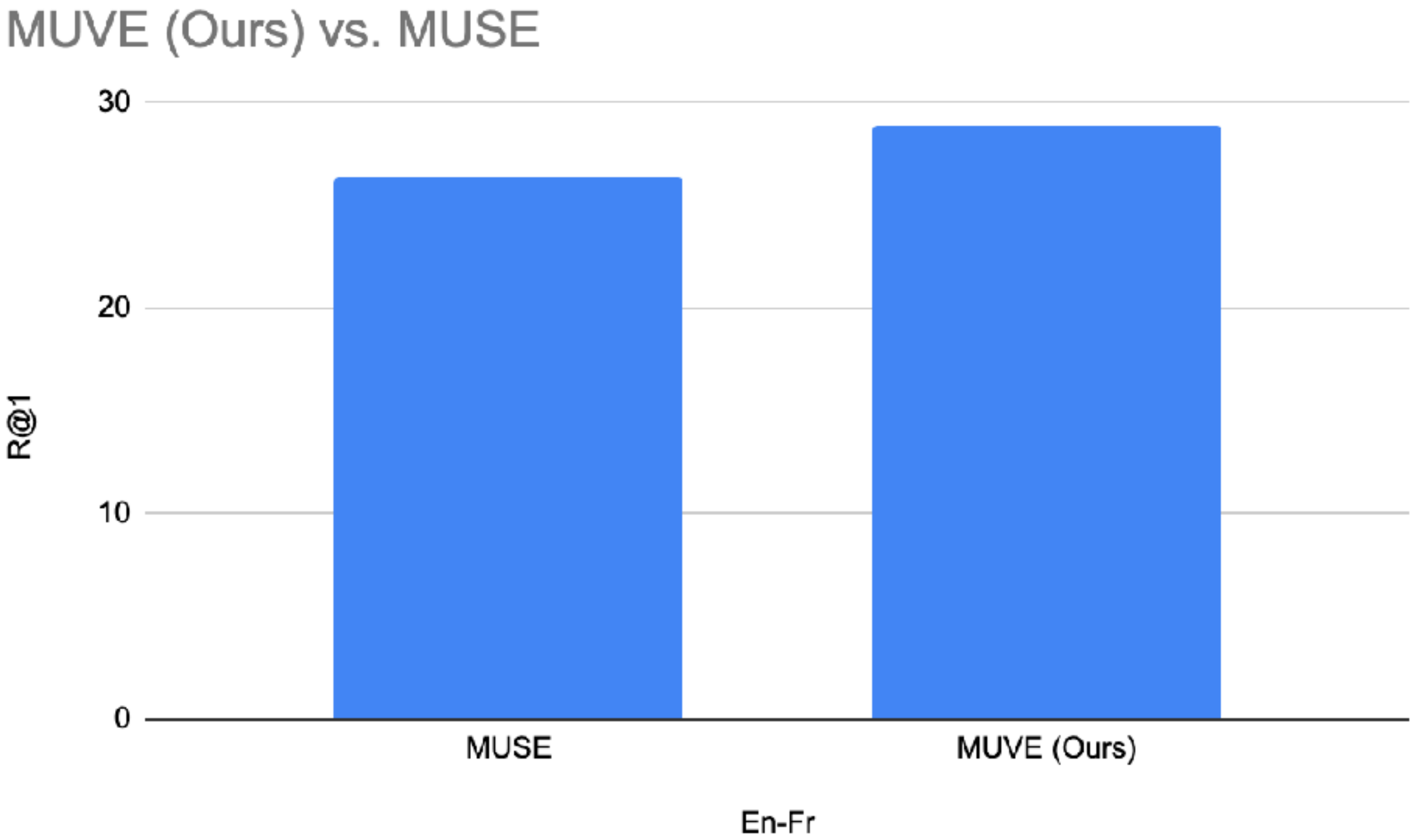
Learning multiple languages via vision



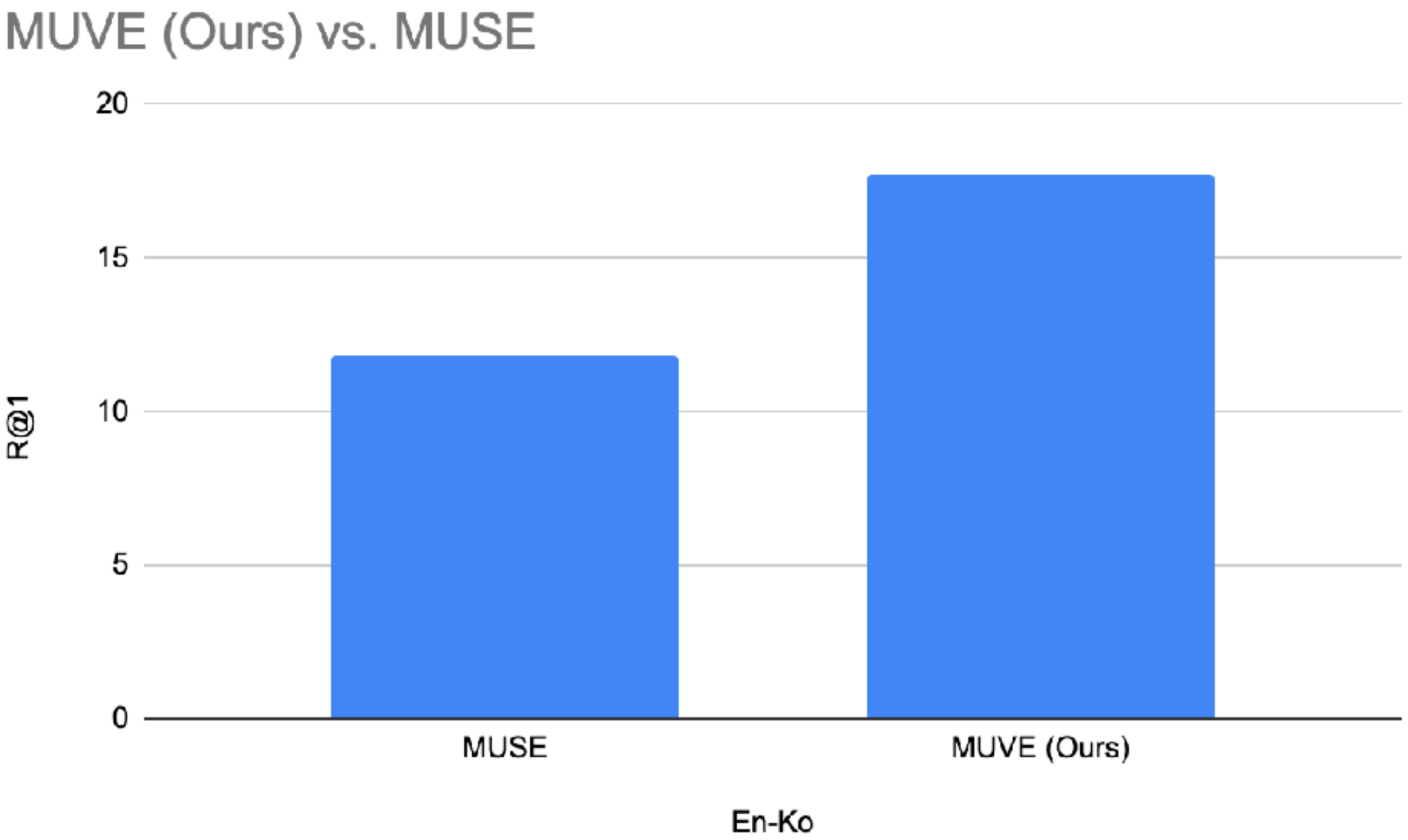
- Pre-trained word embeddings
- Pre-trained on Kinetics400 I3D
- Change language and videos
- Use AdaptLayer as the initial seed for the unsupervised word mapping

G. Sigurdsson et al. “Visual Grounding in Video for Unsupervised Word Translation” CVPR’20

Results on similar languages



Results on different languages



Self-supervised learning

- Self-supervised learning is a scalable approach to train neural networks

Self-supervised learning

- Self-supervised learning is a scalable approach to train neural networks
- The gap between self-supervised and supervised learning becomes lower

Self-supervised learning

- Self-supervised learning is a scalable approach to train neural networks
- The gap between self-supervised and supervised learning becomes lower
- We will still need annotations, but that could be more important for benchmarking networks

Self-supervised learning

- Self-supervised learning is a scalable approach to train neural networks
- The gap between self-supervised and supervised learning becomes lower
- We will still need annotations, but that could be more important for benchmarking networks
- Is self-supervised learning still supervised (choice of the pre-text task or augmentations)?

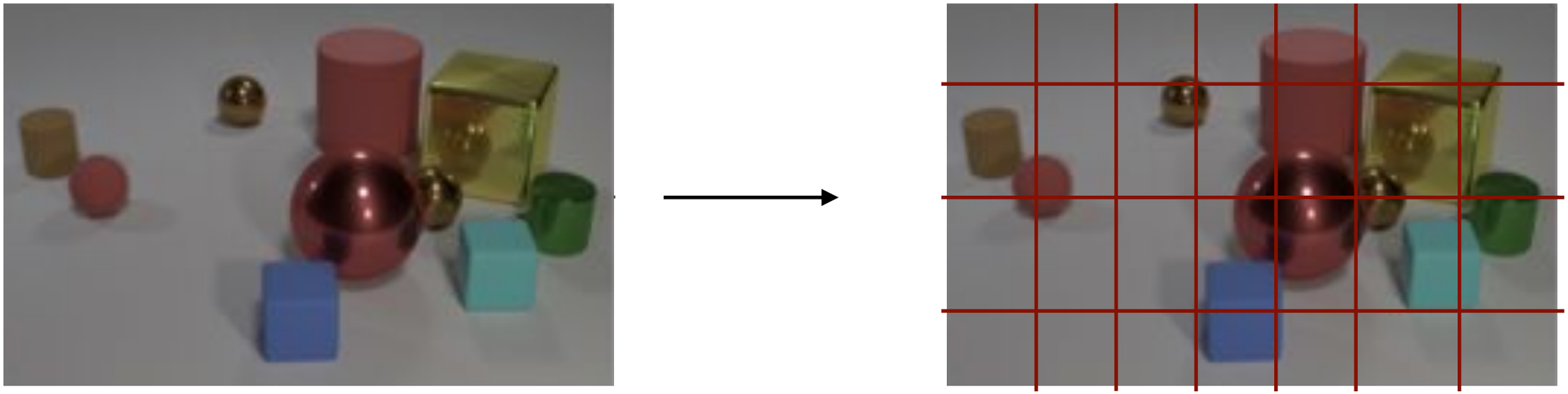
Plan

- Vision + Language (2)
  - Self-supervised learning
  - **“Vision as a Language”**
  - Generation and AI Art
  - Scalability

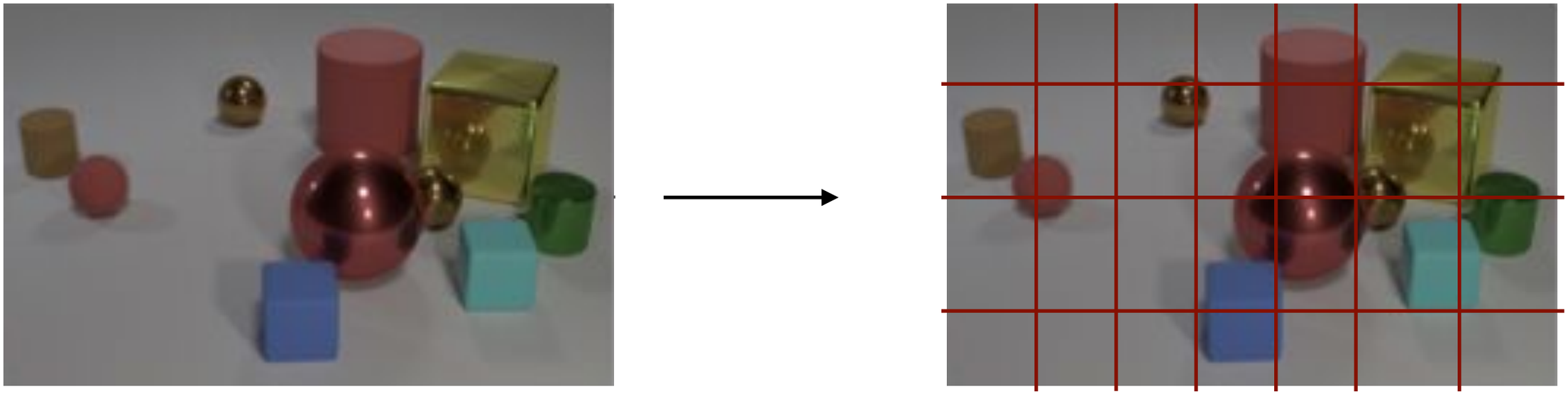
Tokenisation

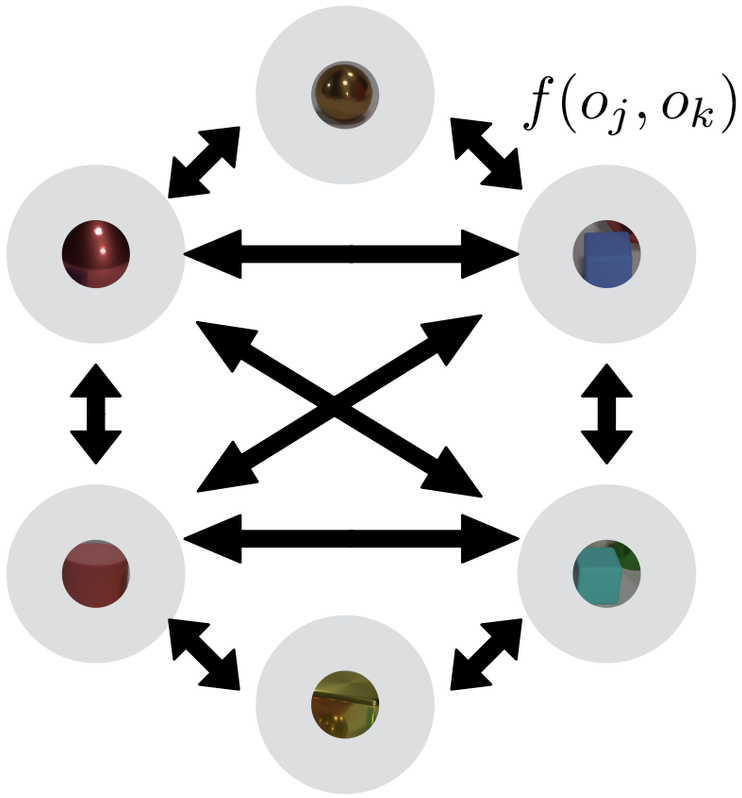
Masked Language Modeling (MLM)

Tokenisation

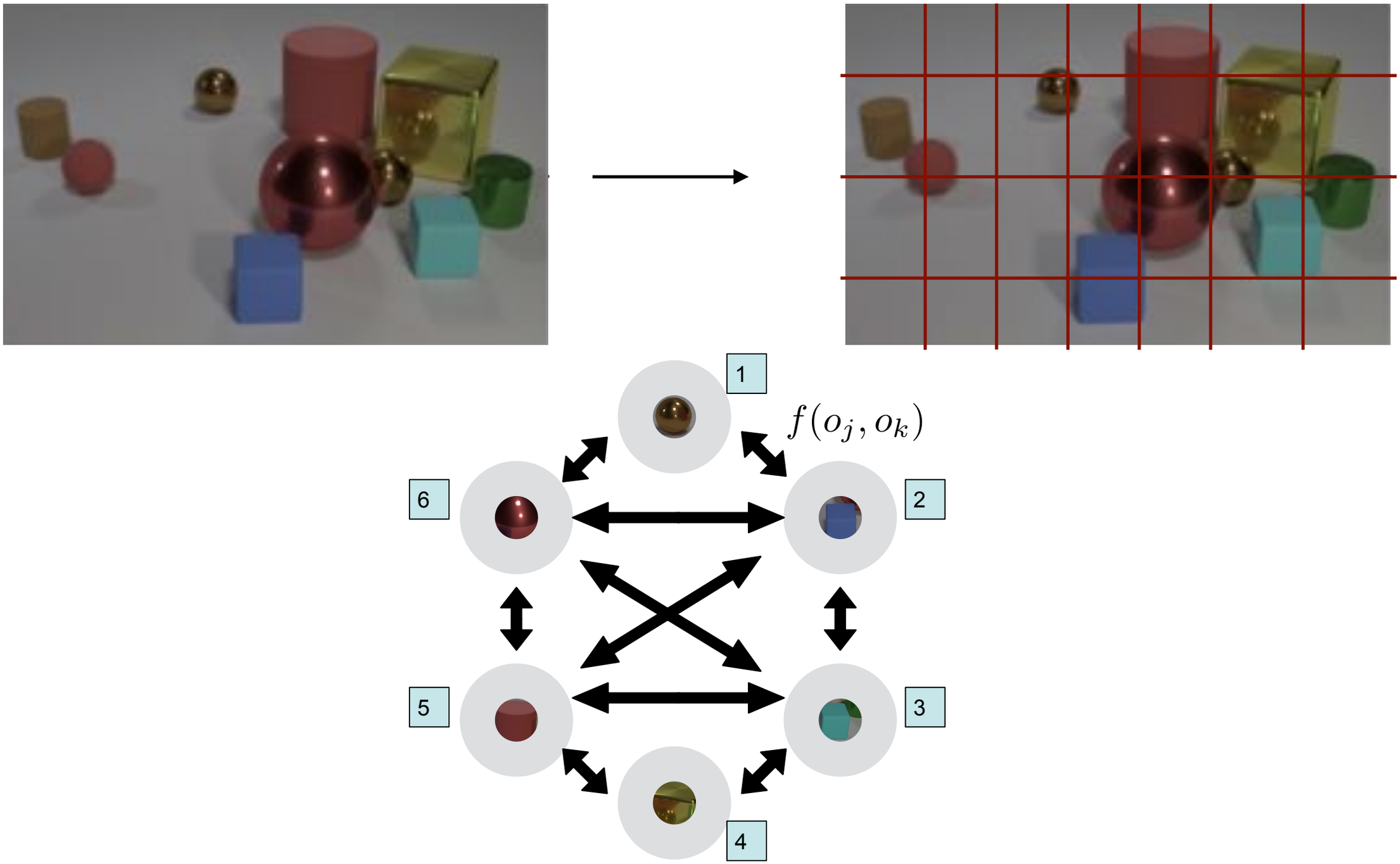


Tokenisation & Transformer





Tokenisation & Transformer



“Vision as Language”

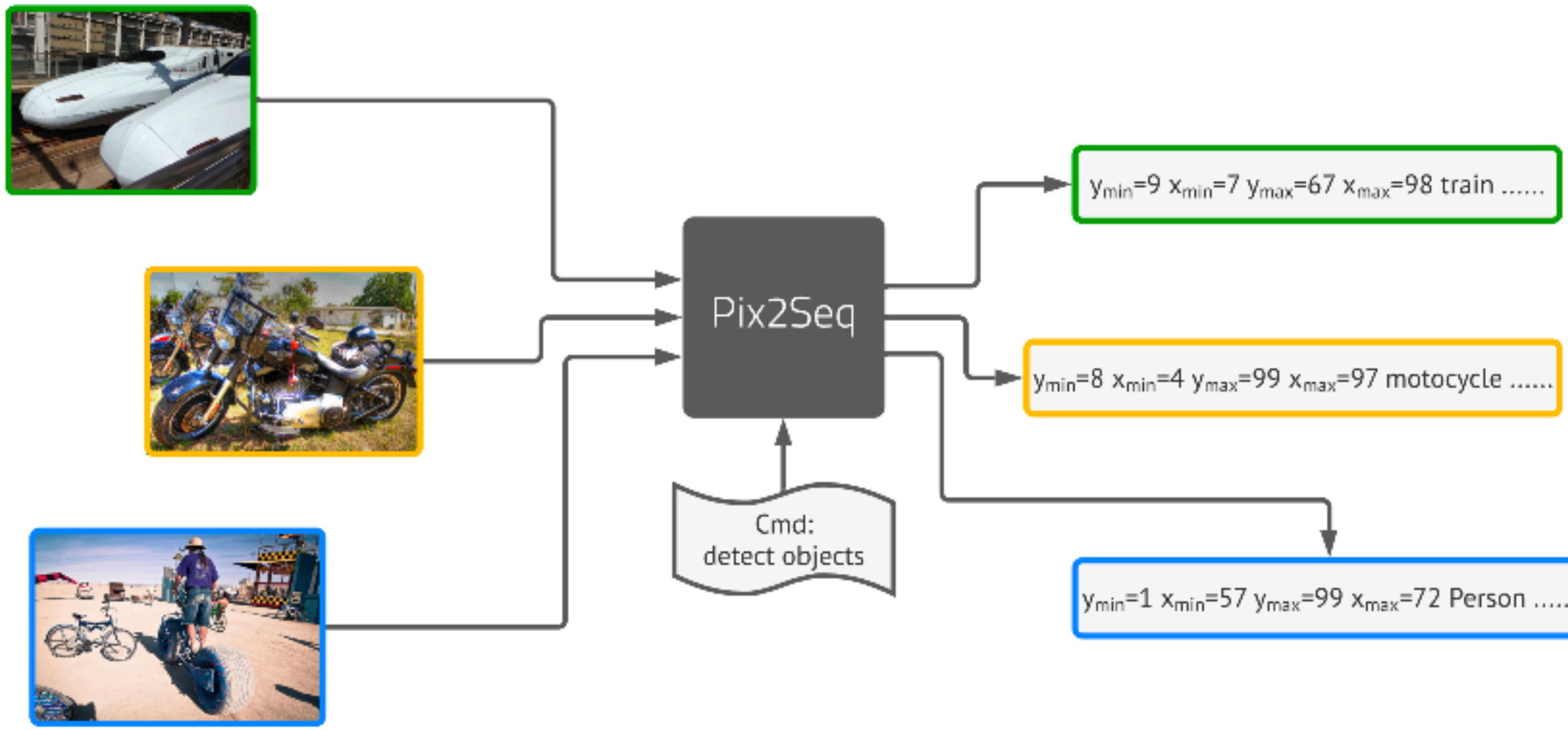


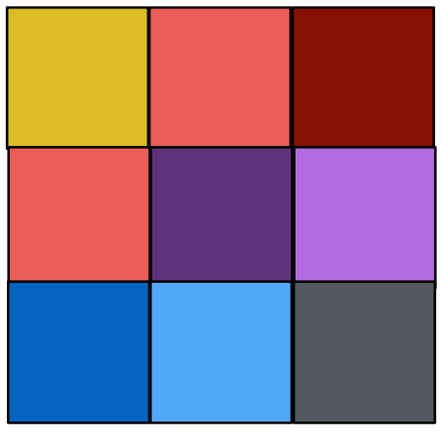
Figure 1: Illustration of Pix2Seq framework for object detection. The neural net perceives an image and generates a sequence of tokens that correspond to bounding boxes and class labels.

T. Chen et al. “Pix2seq: A Language Modeling Framework for Object Detection”

Plan

- Vision + Language (2)
  - Self-supervised learning
  - “Vision as a Language”
  - **Generation and AI Art**
  - Scalability

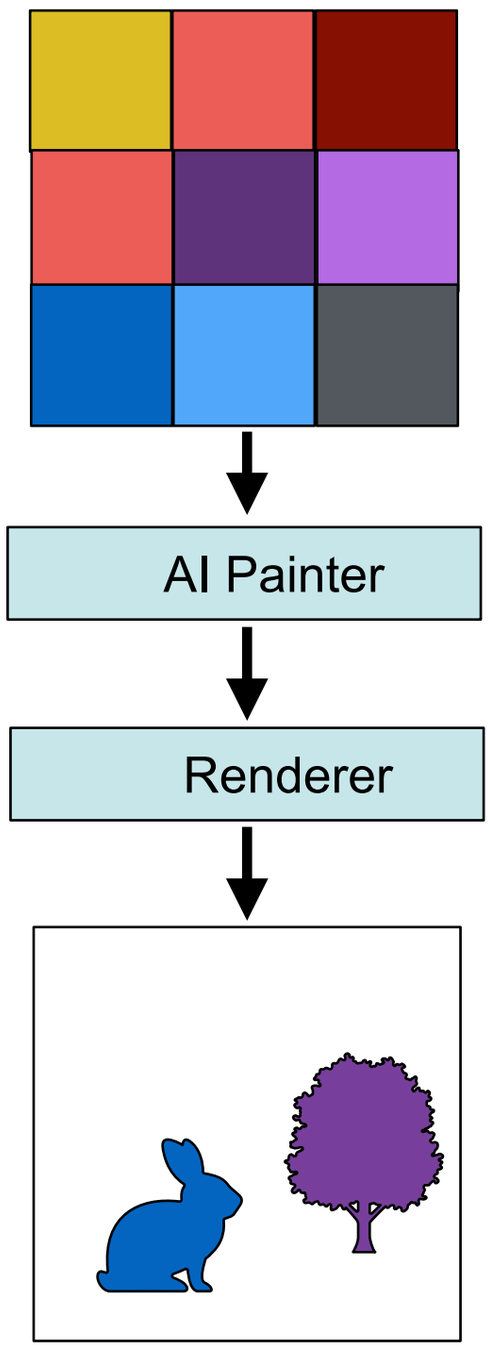
AI Art



“A blue rabbit and a purple tree”

Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

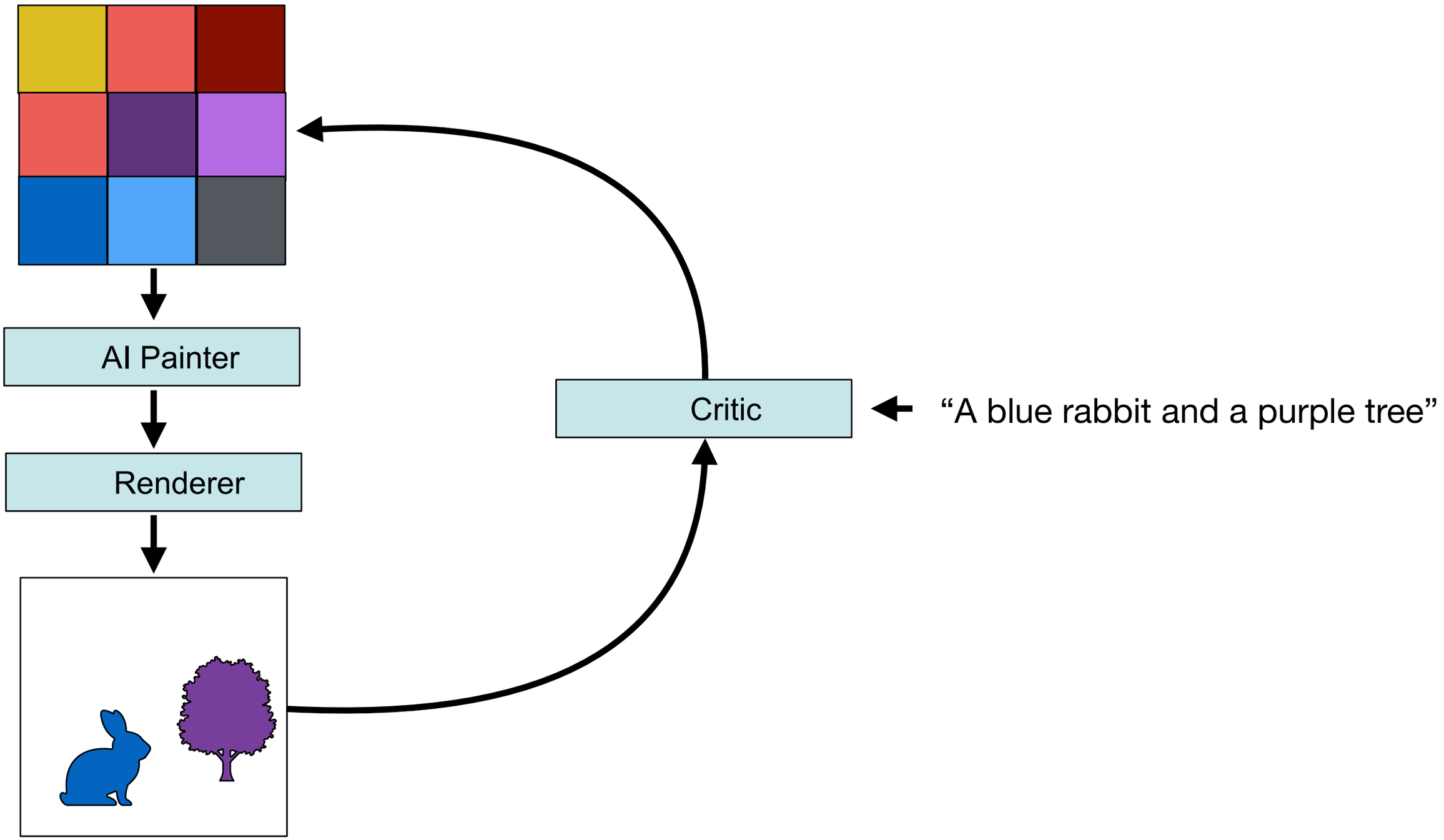
AI Art



“A blue rabbit and a purple tree”

Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

AI Art



Open AI “CLIP: Connecting Text and Images”
C. Fernando et al. “Arnheim 2.0”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Amplifying social norms

J. Pałucha “Polish Casino”

Role of Art in our society

Documenting events

C. Stanfield “Battle of Trafalgar”

Amplifying social norms

J. Pałucha “Polish Casino”

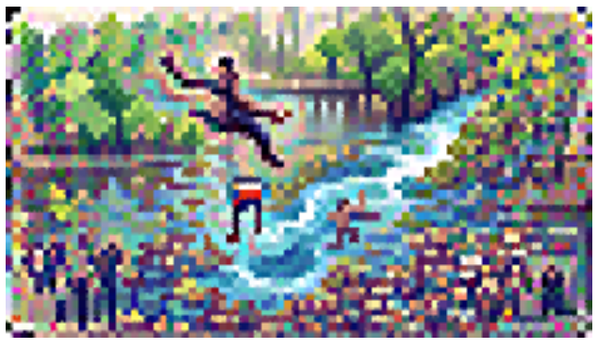
Investigating space and time

G. Balla “Dynamism of a dog on a leash”

Verb understanding

A man jumping into a river

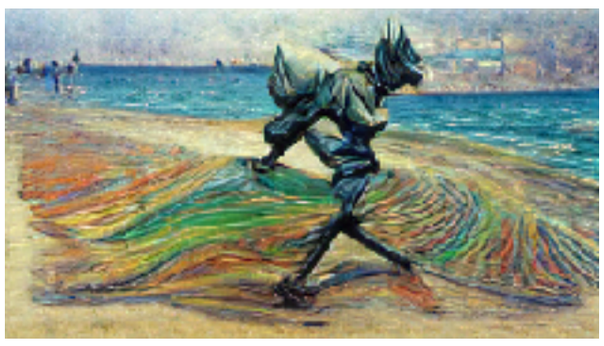


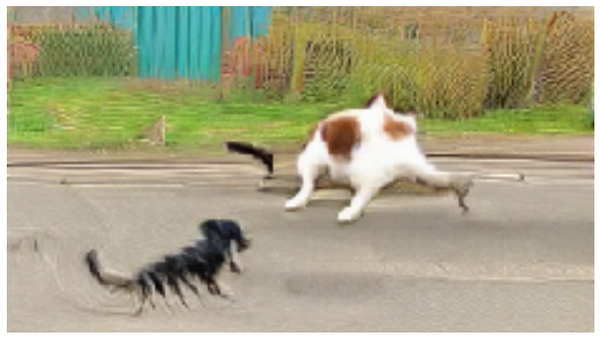
A person jogs on the beach



Umberto Boccioni



A cat chasing a dog



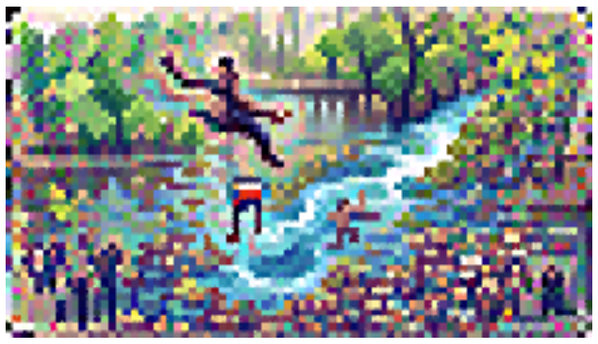

A dog chasing a cat


Based on “Pixray PixelDraw”

Verb understanding

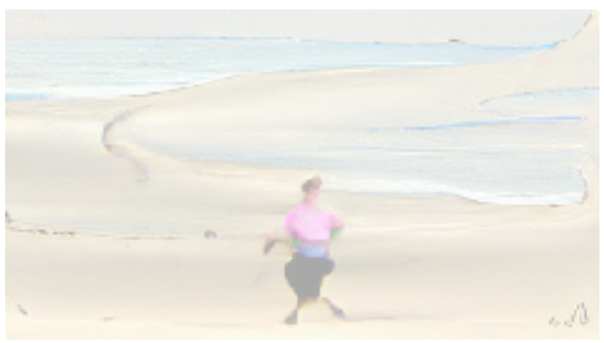
A man jumping into a river



A person jogs on the beach





Umberto Boccioni

A cat chasing a dog


A dog chasing a cat


Based on “Pixray PixelDraw”

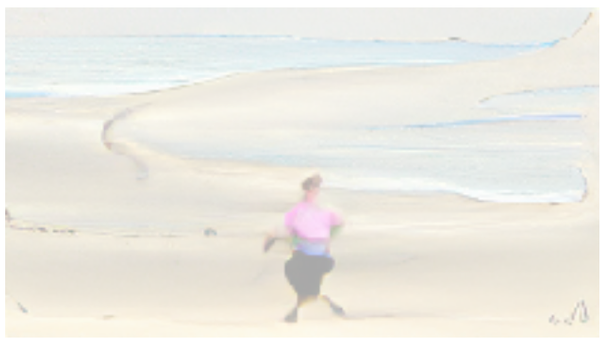
Verb understanding

A man jumping into a river




A person jogs on the beach

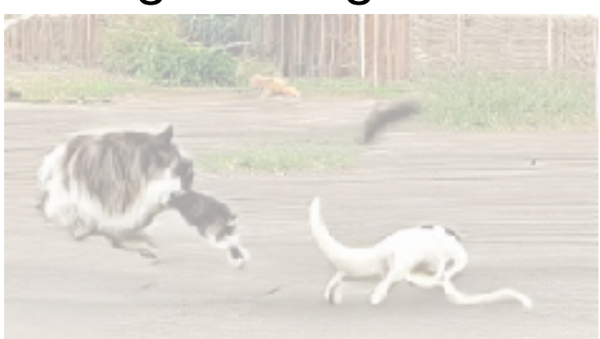




Umberto Boccioni

A cat chasing a dog


A dog chasing a cat




Based on “Pixray PixelDraw”

Multiplicity

CLIP-VQGAN “A cat chasing a dog”
AI-creativity

Based on “Pixray PixelDraw”

G. Balla “Dynamism of a dog on a leash”. Futurism.

The real thing.

Preposition understanding

A cup under a table

#Lucio Fontana, Paul Gauguin

Based on “Pixray PixelDraw”

Preposition understanding

A cup under a table

#Lucio Fontana, Paul Gauguin

Based on “Pixray PixelDraw”

Intuitive physics

A ball falling off the table

Based on “Pixray PixelDraw”

Intuitive physics, combinatorial creativity

A ball falling off the table

Half-car half-bike

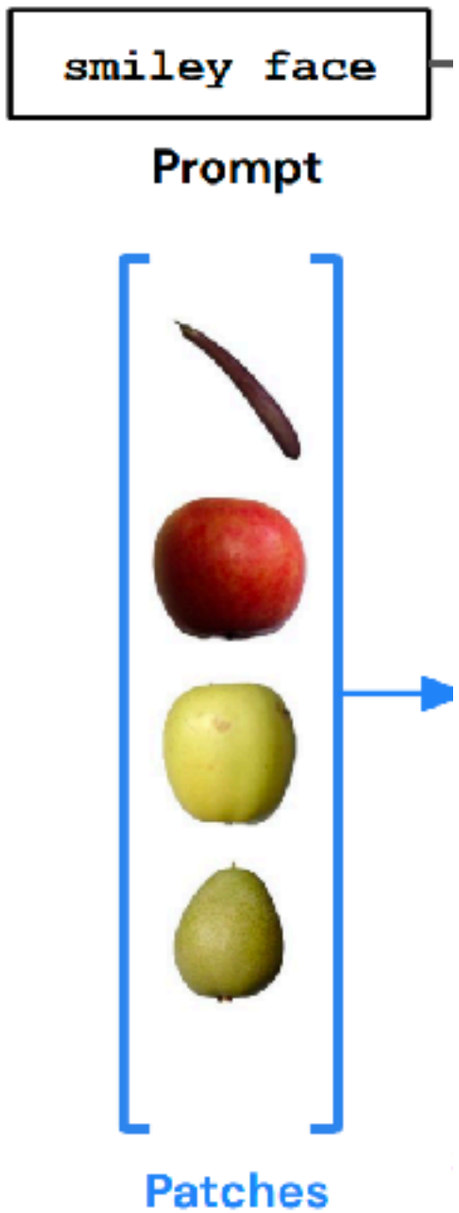
Cubical ball

Based on “Pixray PixelDraw”

Colorless green ideas sleep furiously. Grounded!

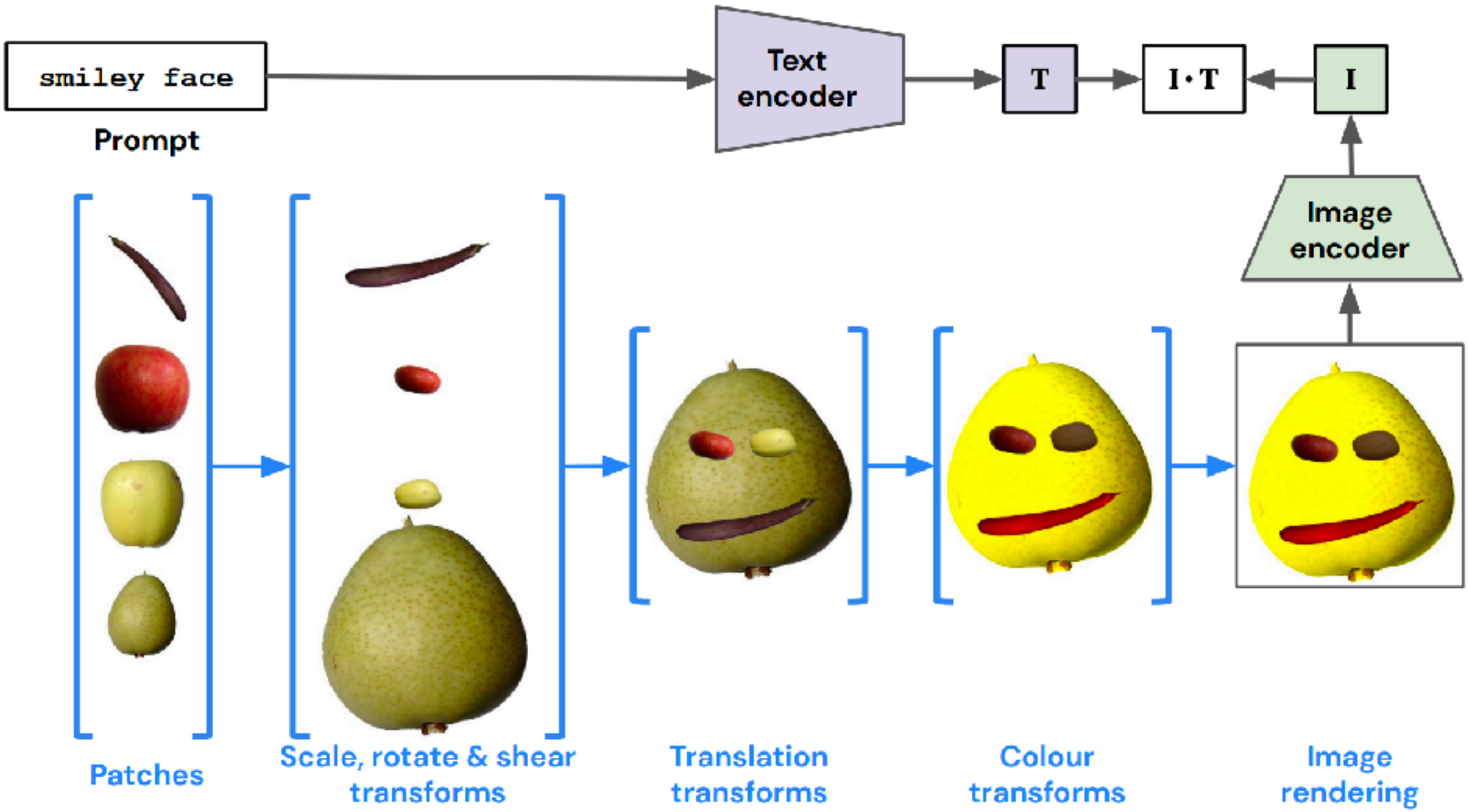
Based on “Pixray PixelDraw”

“Collage”



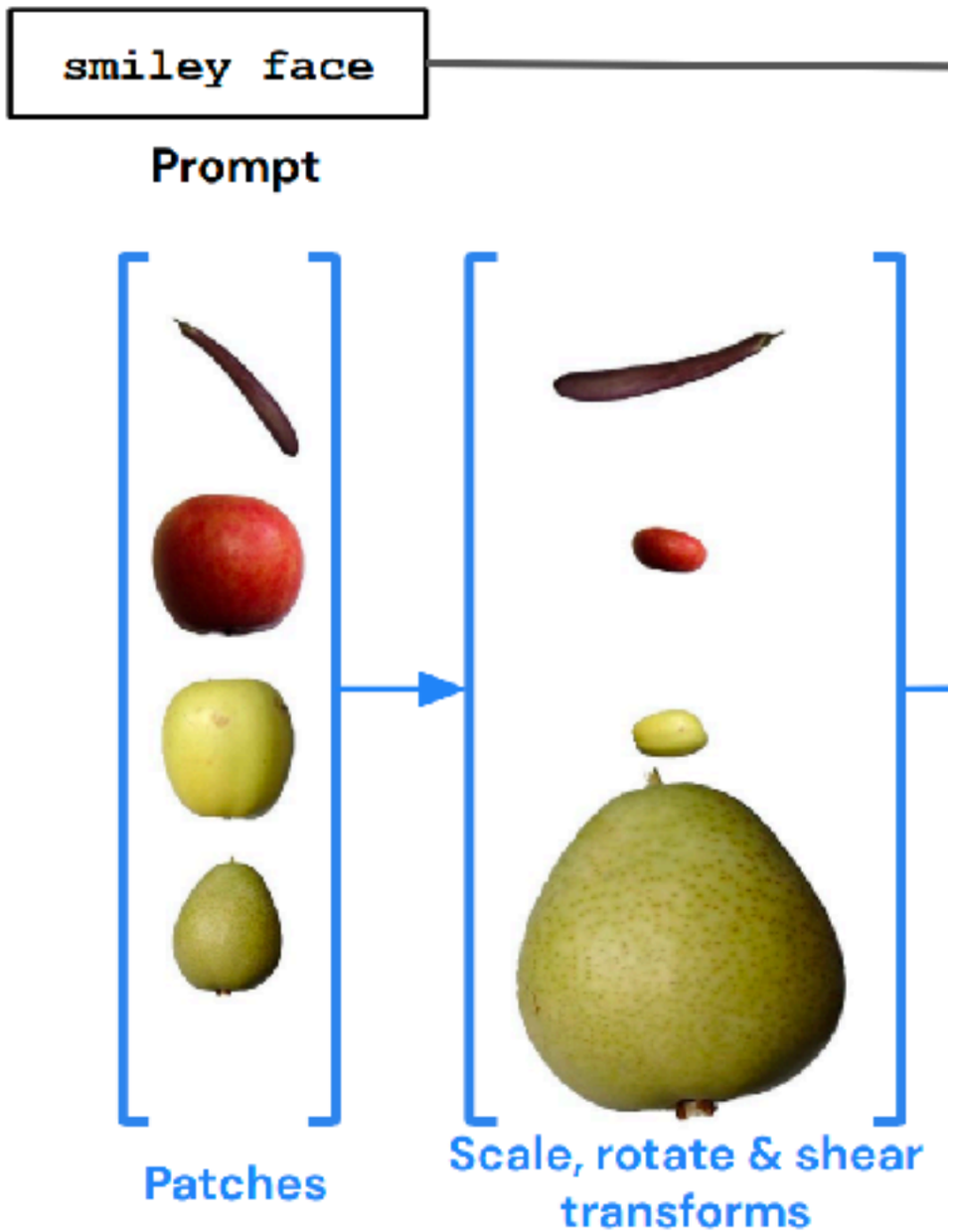
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



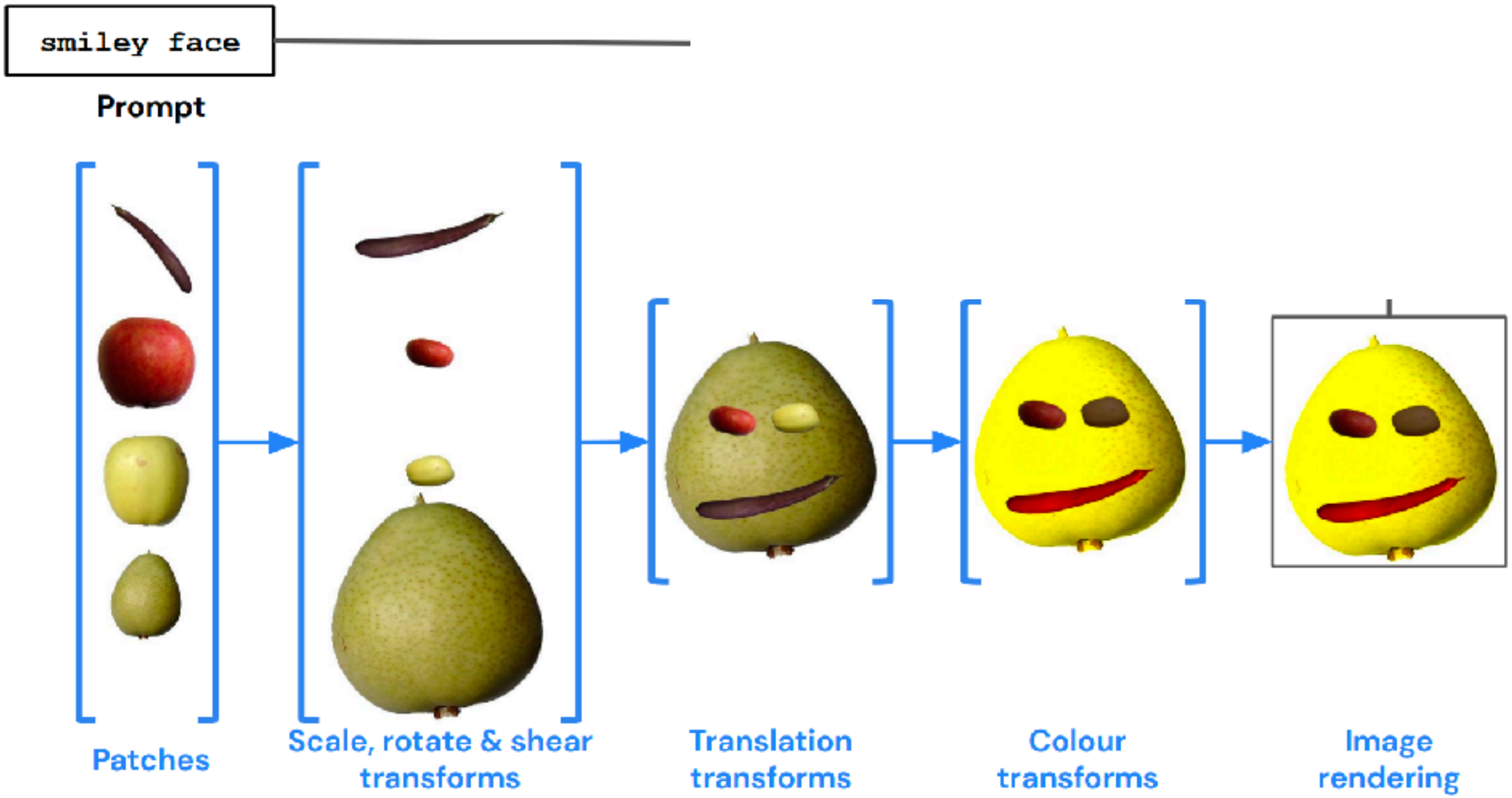
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



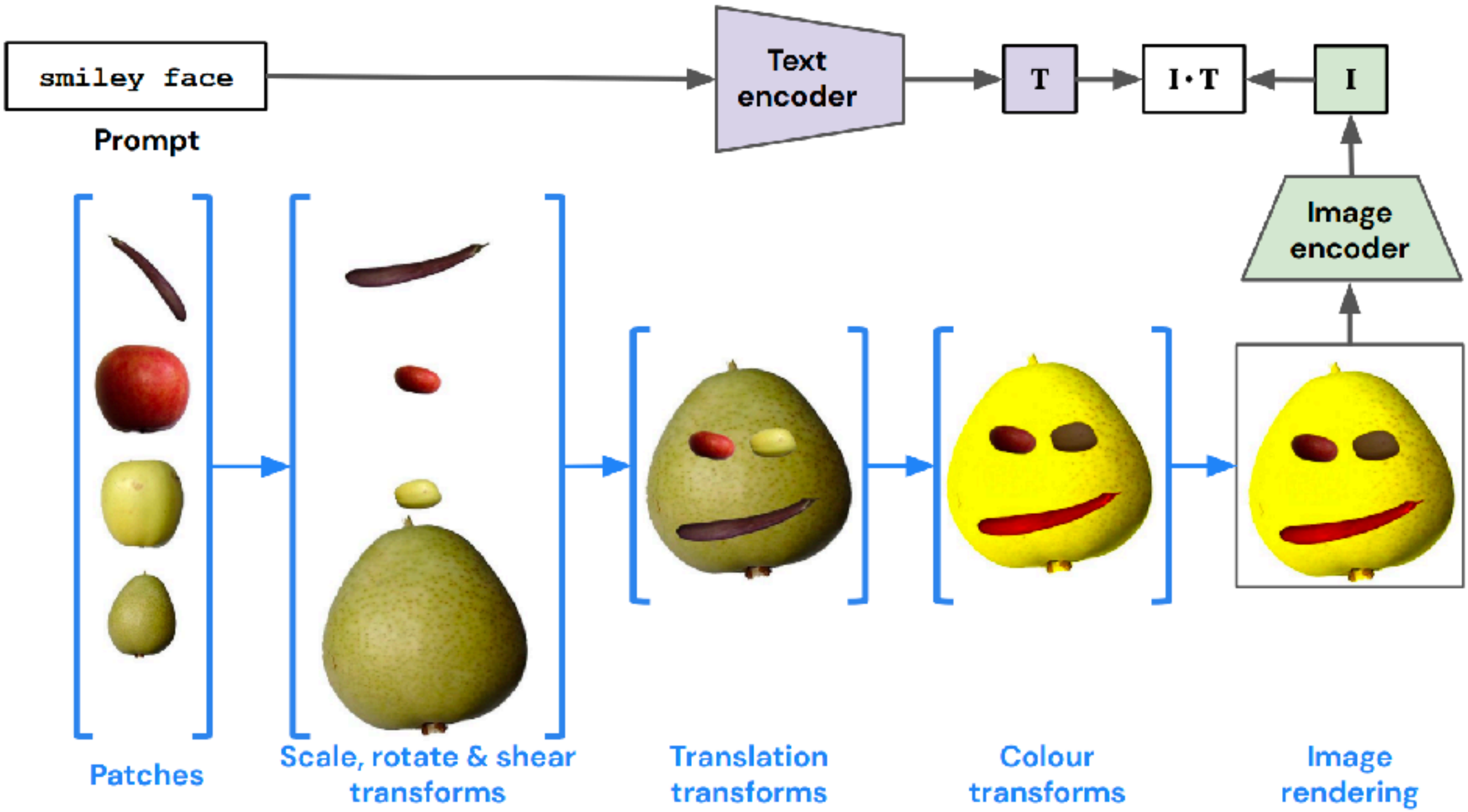
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



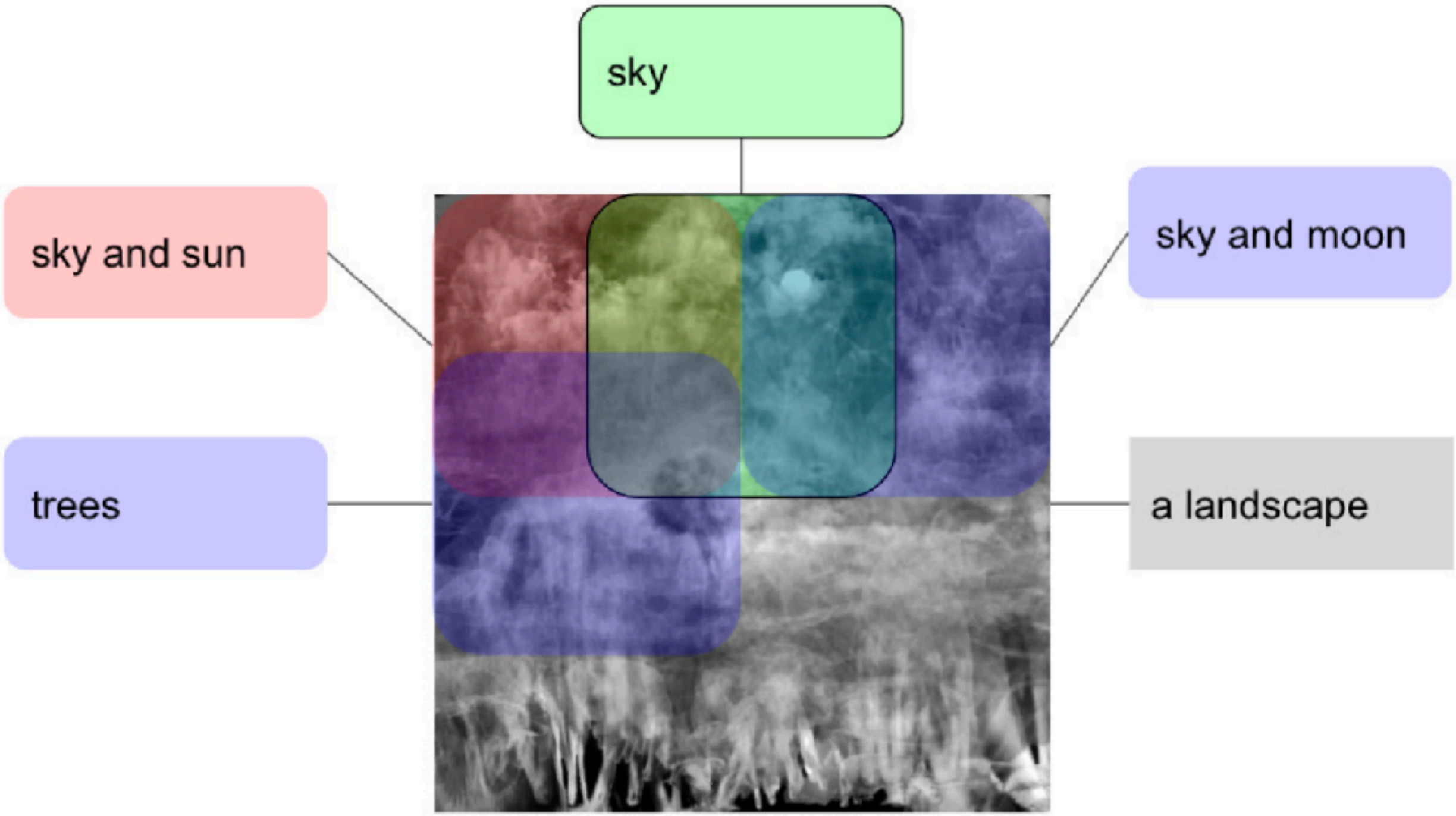
P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

“Collage”



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Scaling up to larger images



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

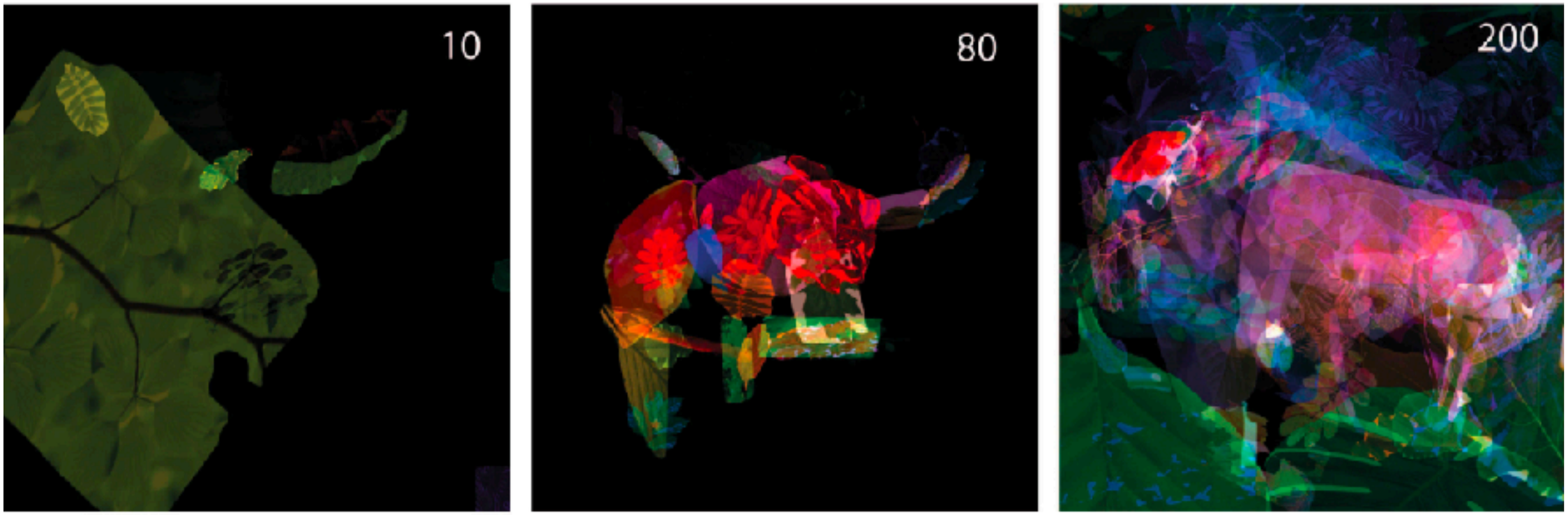
The Fall of the Damned after Rubens and Eaton

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Underwater coral

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Level of abstraction (Bull)



P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Different assets (animals, fruits, animals)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Different assets (animals, fruits, animals)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Different assets (animals, fruits, animals)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

Turing (animals, broken plates)

P. Mirowski et al. “CLIP-CLOP: CLIP-Guided Collage and Photomontage”

DALLE-2 (Open AI)

An astronaut riding a horse in a photorealistic style

Conclusions

- Fun
- Understanding creativity
- Helping artists (tools)
- Understanding of what networks understand

Plan

- Vision + Language (2)
  - Self-supervised learning
  - “Vision as a Language”
  - Generation and AI Art
  - **Scalability**

Why?

- Learning at scale is among the most successful paradigms

Why?

- Learning at scale is among the most successful paradigms
- Scale in terms of data and learning
  - Self-supervised or unsupervised learning

Why?

- Learning at scale is among the most successful paradigms
- Scale in terms of data and learning
  - Self-supervised or unsupervised learning
- Scale in terms of sequence lengths
  - Integrating more context, more “hours of life”

Why?

- Learning at scale is among the most successful paradigms
- Scale in terms of data and learning
  - Self-supervised or unsupervised learning
- Scale in terms of sequence lengths
  - Integrating more context, more “hours of life”
- Efficiency
  - Time
  - Memory consumption
  - Parallelism

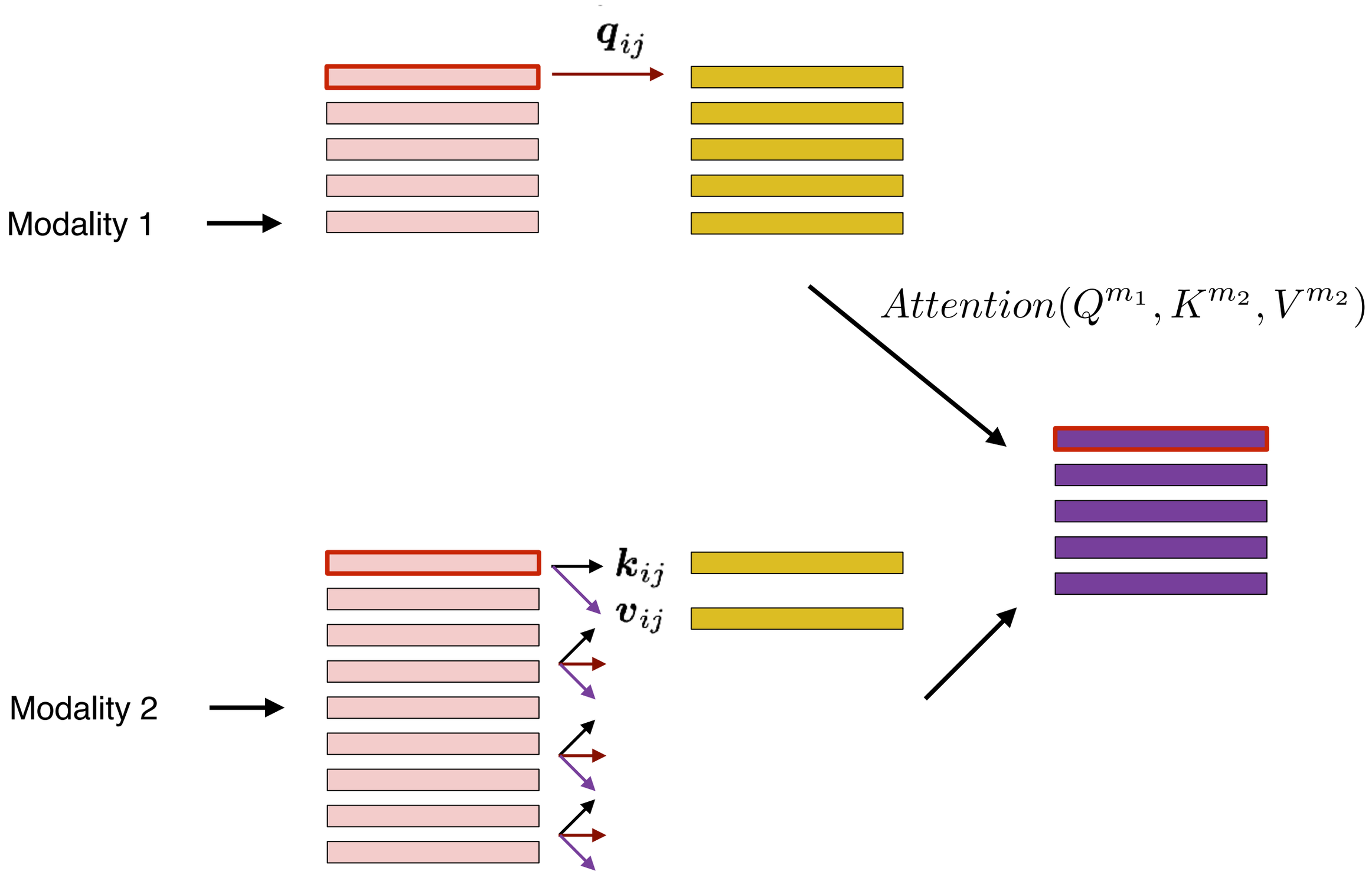
Why?

- Learning at scale is among the most successful paradigms
- Scale in terms of data and learning
  - Self-supervised or unsupervised learning
- Scale in terms of sequence lengths
  - Integrating more context, more “hours of life”
- Efficiency
  - Time
  - Memory consumption
  - Parallelism
    - Data parallelism -> Easy
    - Model parallelism -> Hard

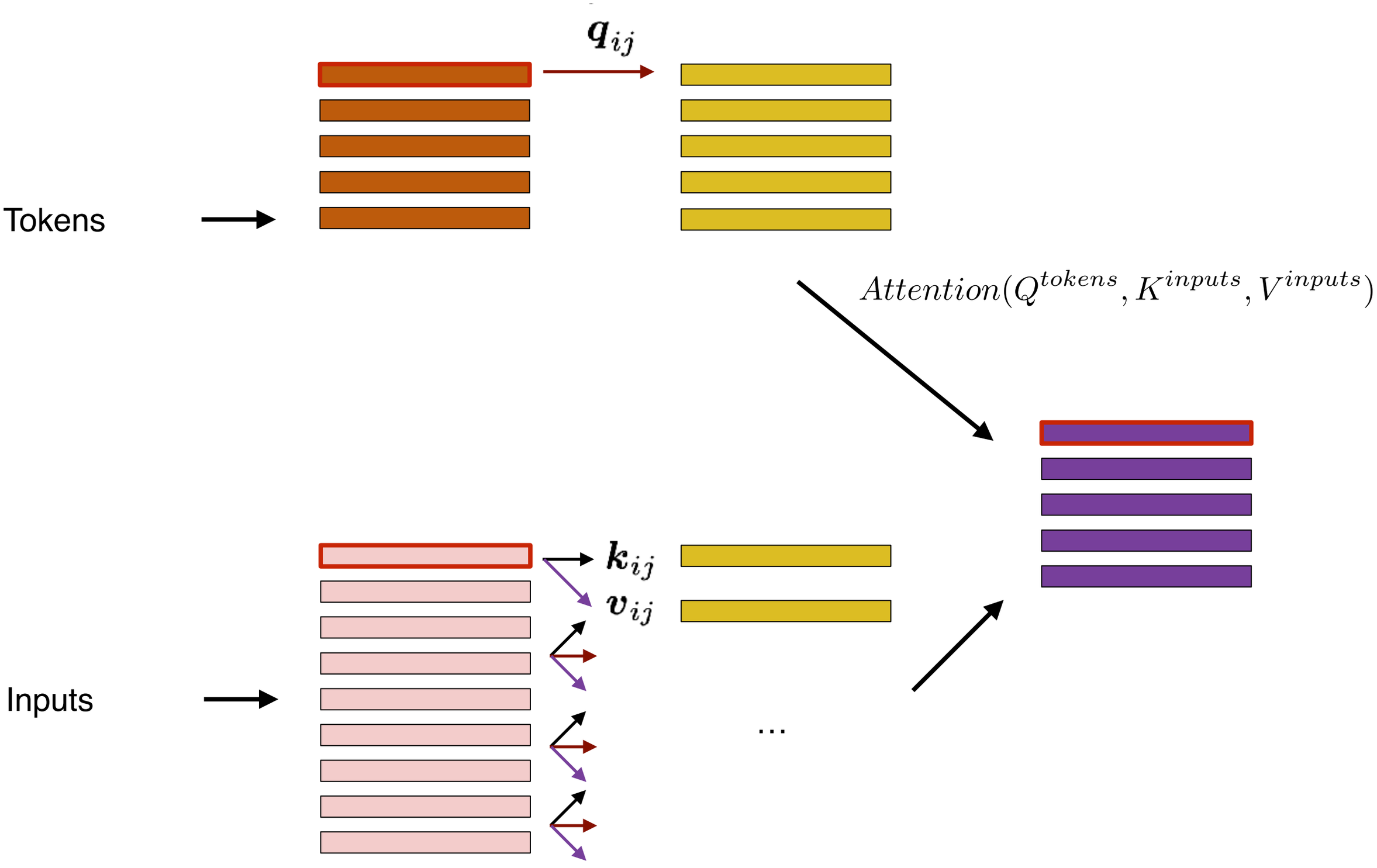
Why?

- Learning at scale is among the most successful paradigms
- Scale in terms of data and learning
  - Self-supervised or unsupervised learning
- Scale in terms of sequence lengths
  - Integrating more context, more “hours of life”
- Efficiency
  - Time
  - Memory consumption
  - Parallelism
    - Data parallelism -> Easy
    - Model parallelism -> Hard
- All is becoming more prominent when dealing with longer documents or videos

Taming quadratic complexity



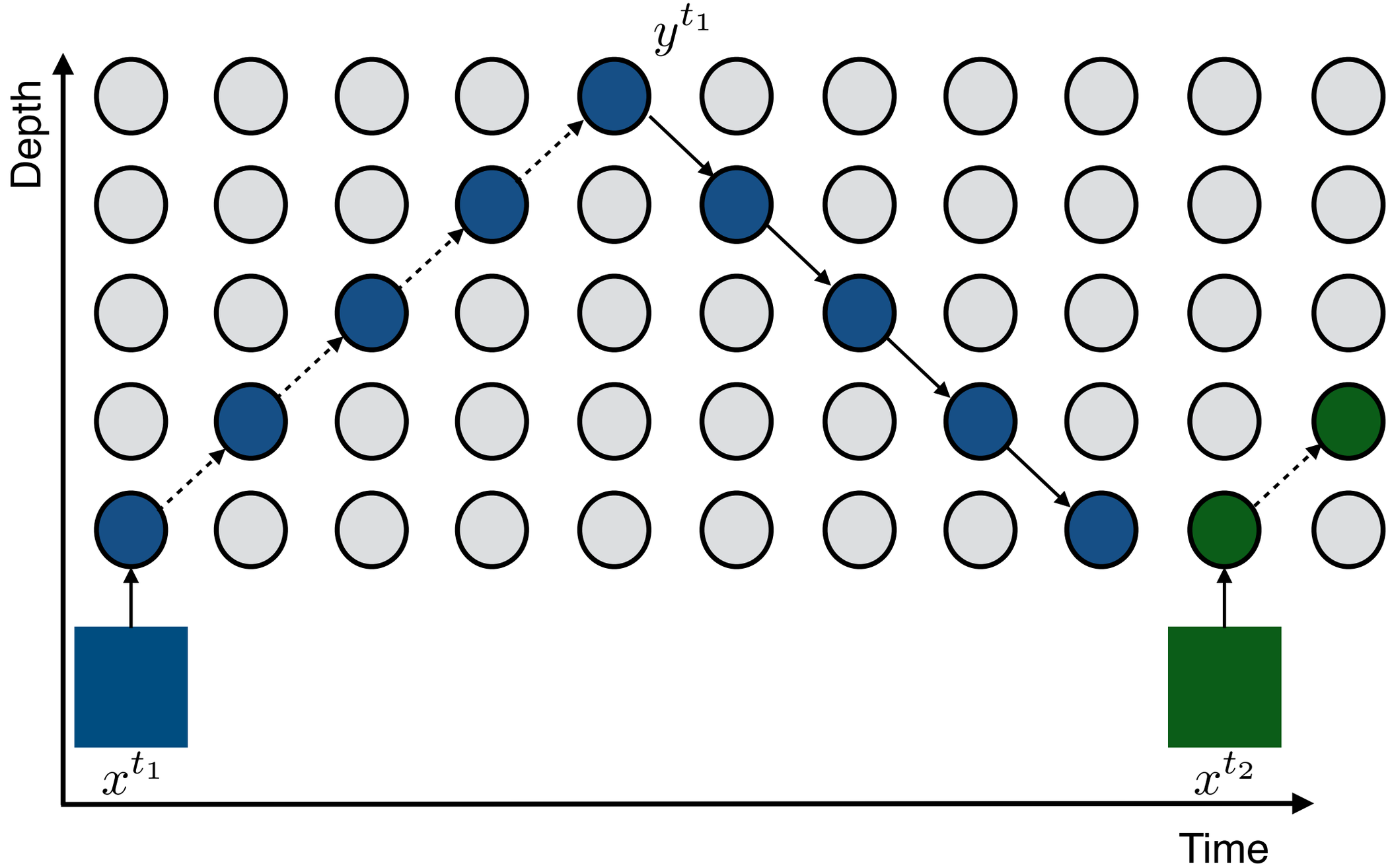
Taming quadratic complexity



D. Jaegle et al. “Perceiver: General perception with iterative attention”

Issues with backprop

- "Arrow-of-time" view on backprop



M. Malinowski et al. "Sideways"

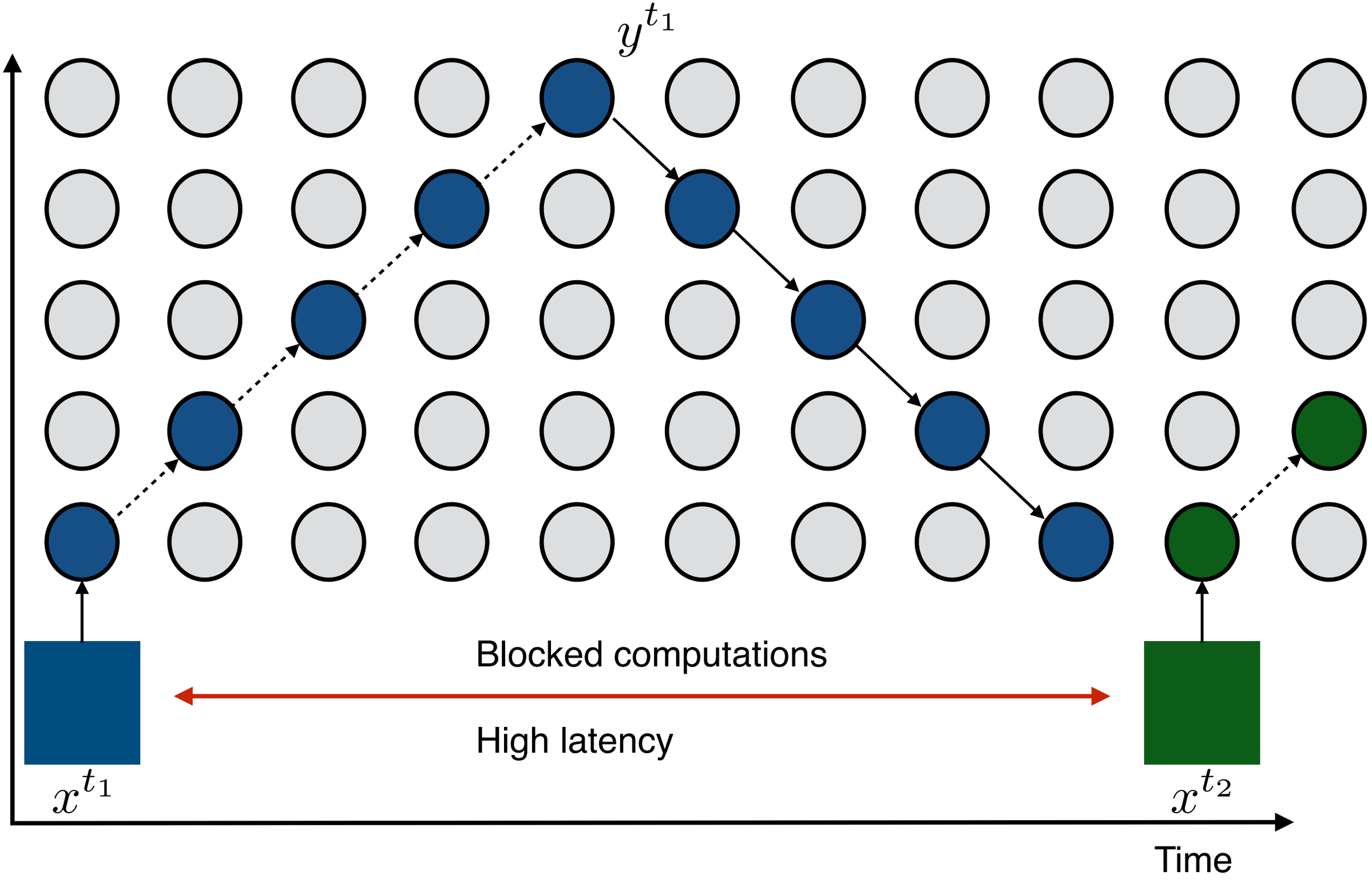
Issues with backprop

- “Arrow-of-time” view on backprop

1

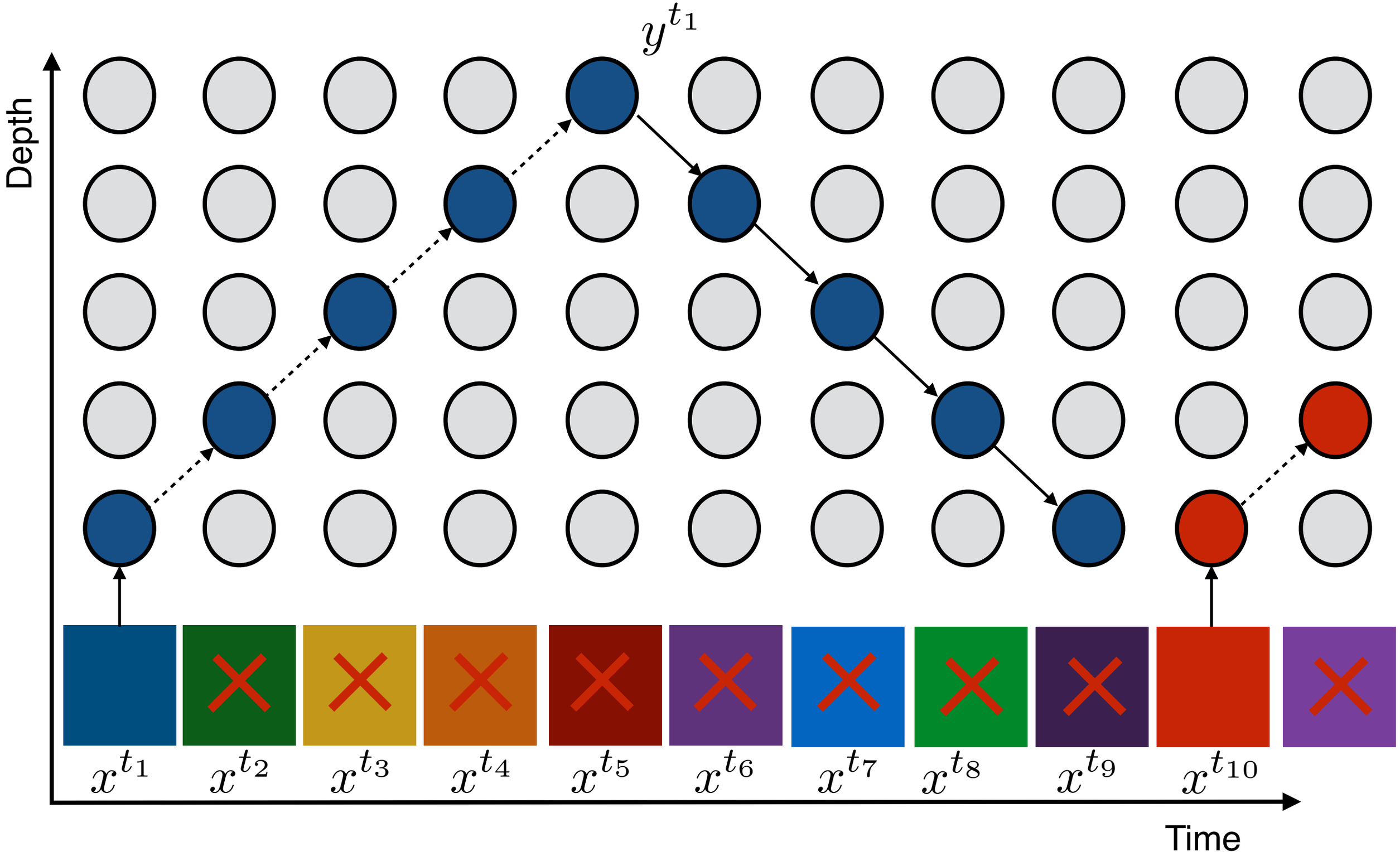
M. Malinowski et al. “Sideways”

Issues with backprop



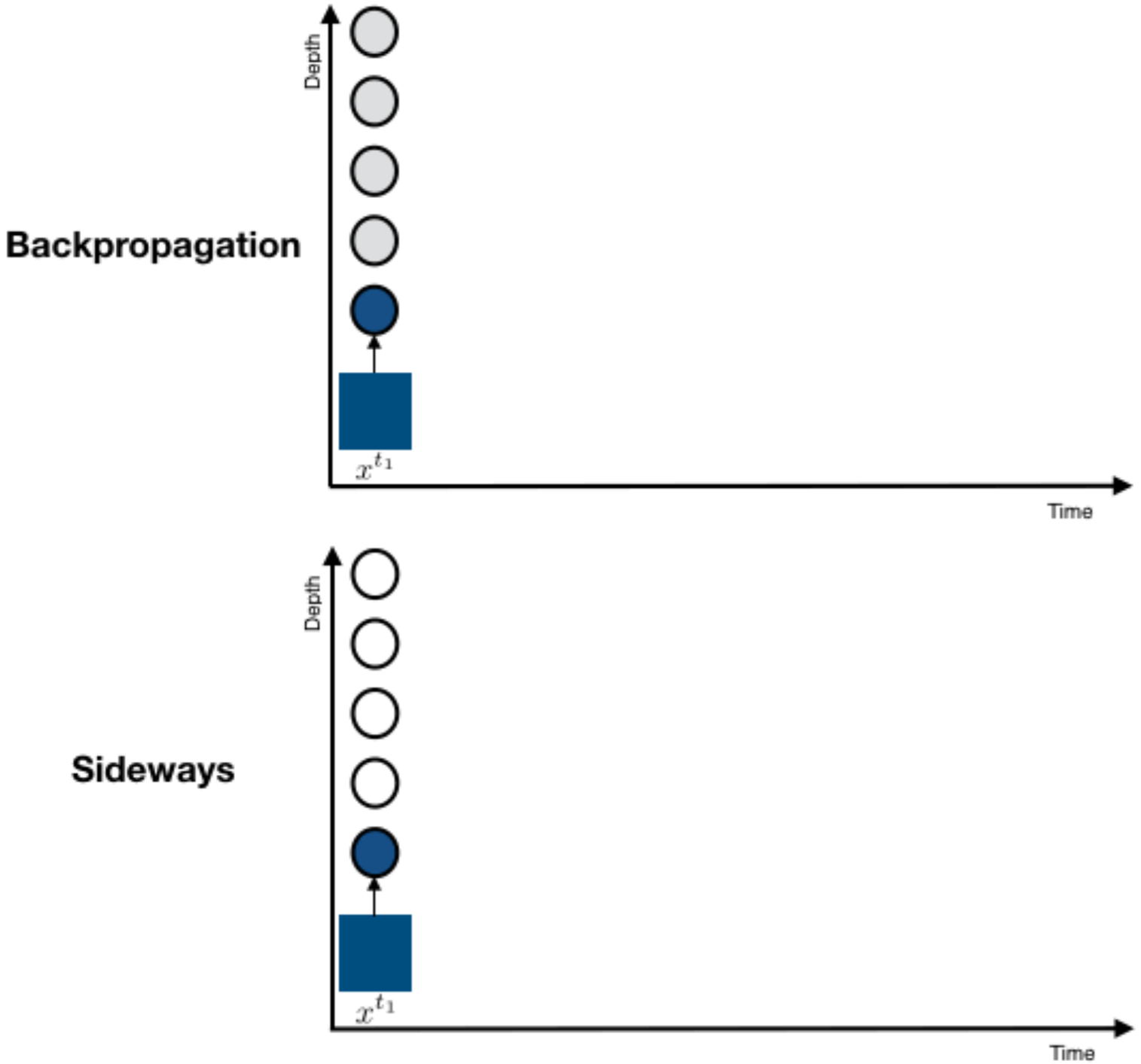
M. Malinowski et al. “Sideways”

Issues with backprop



M. Malinowski et al. “Sideways”

Sideways vs backprop



M. Malinowski et al. “Sideways”

Thank you for your attention

- Going beyond a single modality is exciting
- We are multimodal
- Questions about scale
- Training might become “easier"