SFU NatLangLab

CMPT 413/713: Natural Language Processing

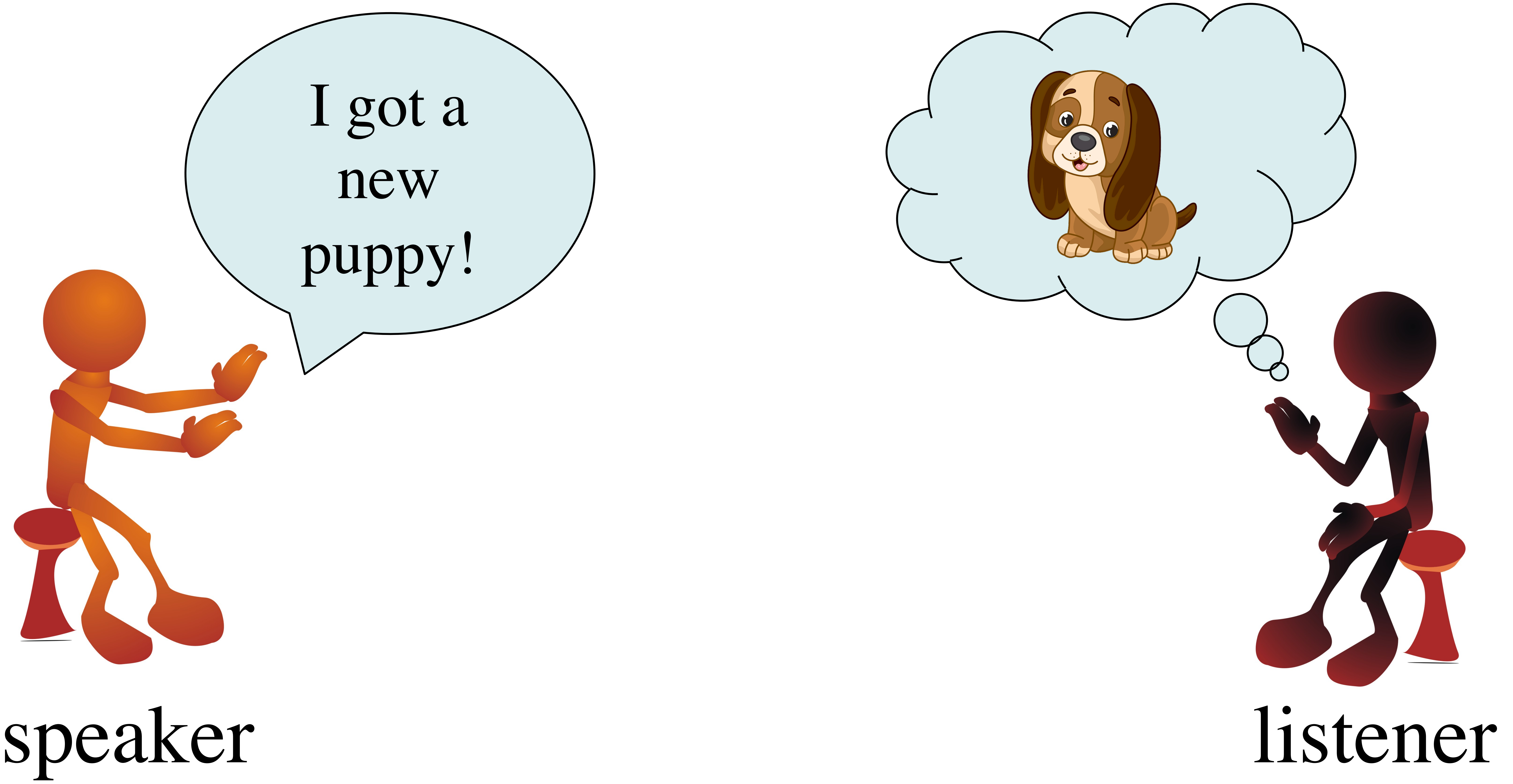
Grounded Natural Language

Spring 2024
2024-03-24

What is grounding?

Language is used to communicate about the world

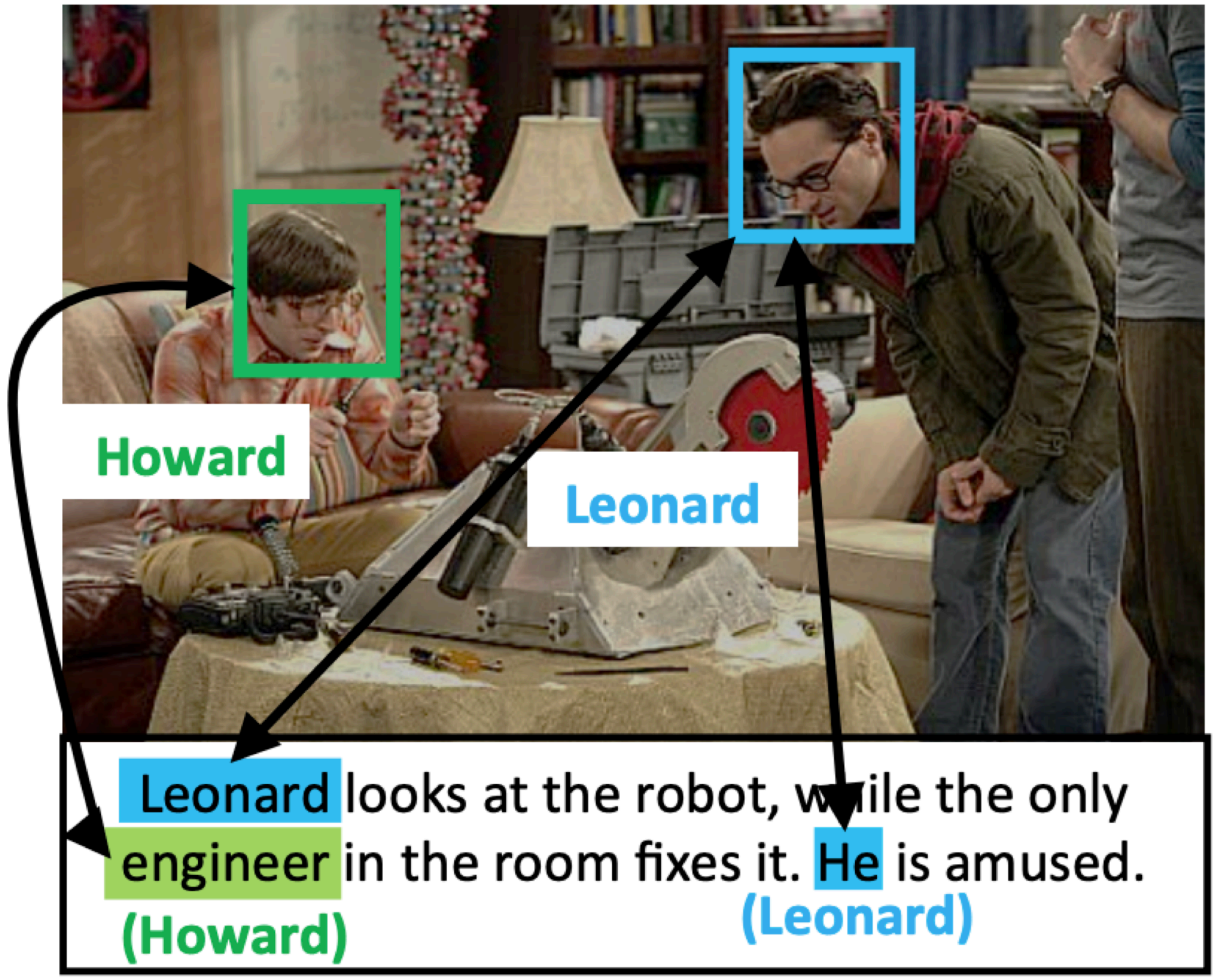
- Things, actions, abstract concepts



What is symbol grounding?

- Connecting **linguistic symbols** to their meaning

- Connecting words and sentences to what they represent

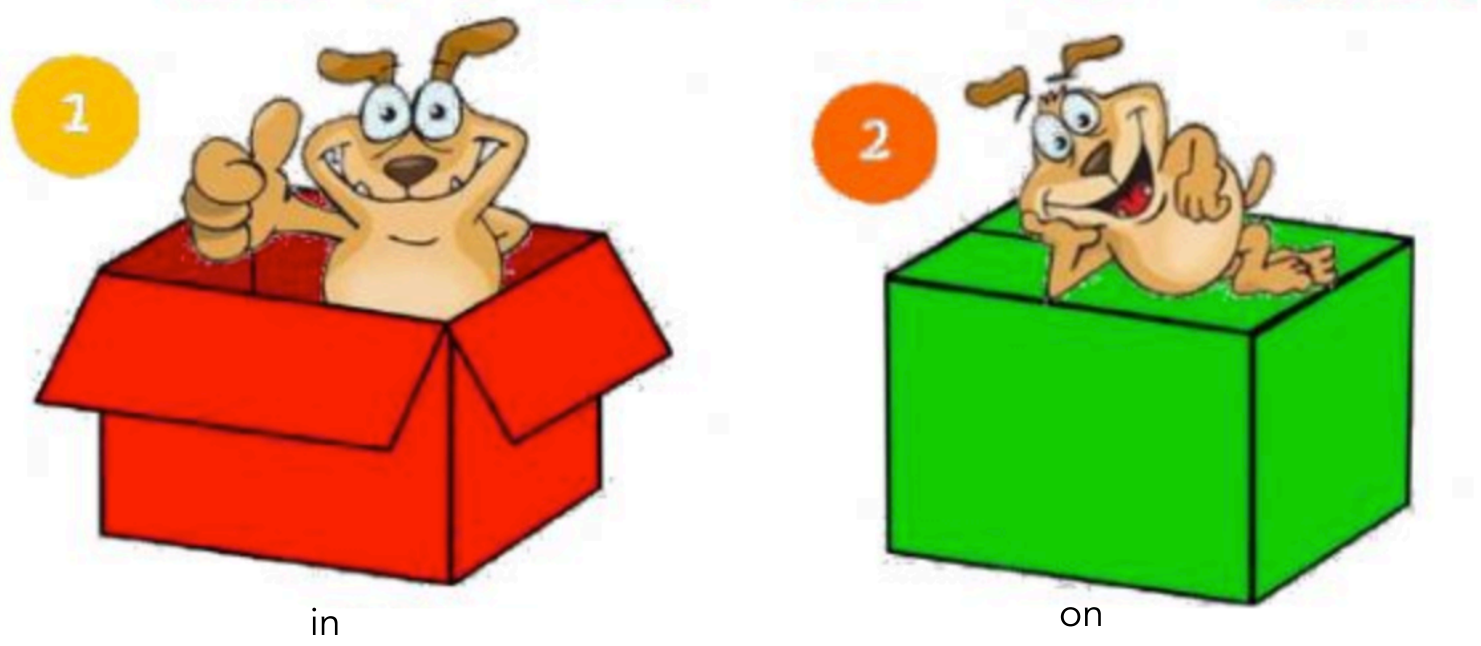


Linking people in videos with "their" names using coreference resolution [Ramanathan et al, 2014]

Actions

running

Spatial relations



Types of grounding

- **Perception**
  - Visual: *green* = [0,1,0] in RGB
  - Auditory: *loud* = >120 dB
  - Taste: sweet = >some threshold level of sensation on taste buds
  - High-level concepts:

cat

dog

Slide adapted from Danqi Chen and Karthik Narasimhan

Types of grounding

- **Temporal concepts**
  - *late evening* = after 6pm
  - *fast, slow* = describing rates of change
- **Actions**

running

eating

Slide adapted from Danqi Chen and Karthik Narasimhan

Some grounding tasks

- **Vision**
  - Captioning
  - Text to image generation and manipulation
  - Visual question answering (VQA)
  - Referring Expressions and Spatial reasoning
- **Interaction**
  - Instruction following
  - Text-based games

Image captioning

the girl is licking the spoon of batter

- Describe an image in a sentence

(MS COCO Captions, Chen et al., 2015)

Image captioning

the girl is licking the spoon of batter

- Describe an image in a sentence
- Requires recognizing objects, attributes, relations in image
- Caption must be fluent

(MS COCO Captions, Chen et al., 2015)

Captioning as multi-modal translation

French → Text encoder (e.g. RNN) → Text decoder (e.g. RNN) → *English*

 → Image encoder → Text decoder (e.g. RNN) → *English*

(Donahue et al., 2015, Vinyals et al., 2015)

Learning to connect linguistic symbols to the physical world

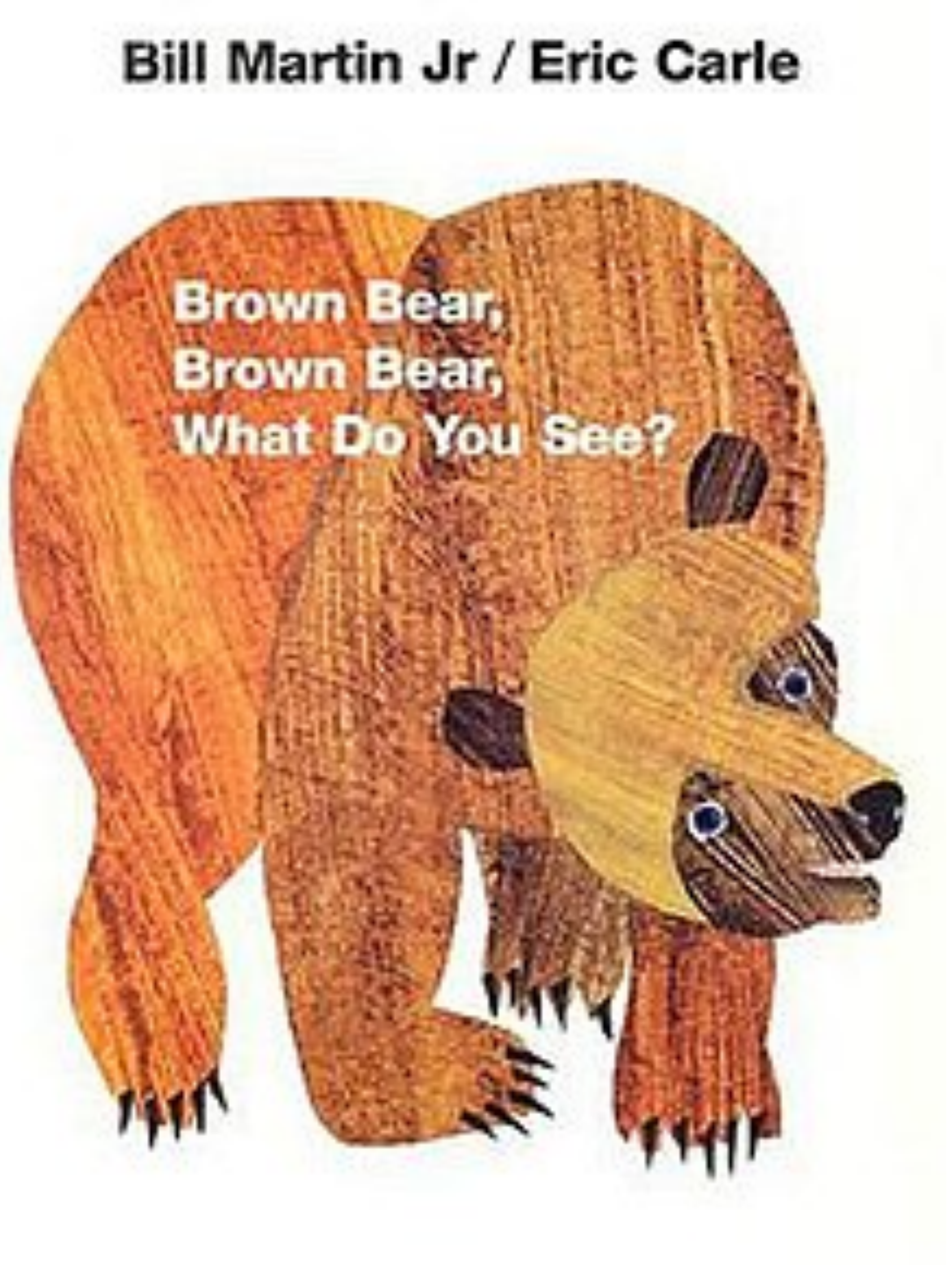
Children do not learn language from raw text
or passively watching TV

Natural way to learn language in the context of
its use in the physical and social world

This requires inferring the meaning of
utterances from their perceptual context

Children learn from multimodal sensory input and experience

Learning from multimodal information



Learn more about how children learn from Linda Smith: https://www.youtube.com/watch?v=dxli8qWJHLU

Choices in what to ground to

Connecting linguistic symbols to

- perceptual experiences and actions

- other symbols

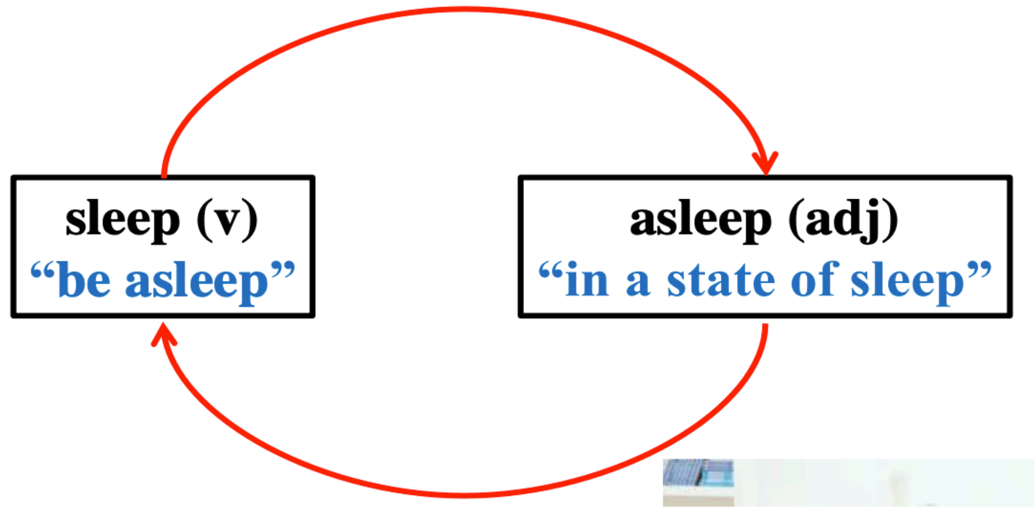
One hundred → 100

The Big Bang Theory → https://en.wikipedia.org/wiki/The_Big_Bang_Theory

Circular definitions

``Sleep” means ``be asleep”

sleep(n): ``a natural and periodic state of rest during which consciousness of the world is suspended”



- to executable programs

Create a key `key` if it does not exist in dict `dic` and append element `value` to value

```
dic.setdefault(key, []).append(value)
```

Meaning representations

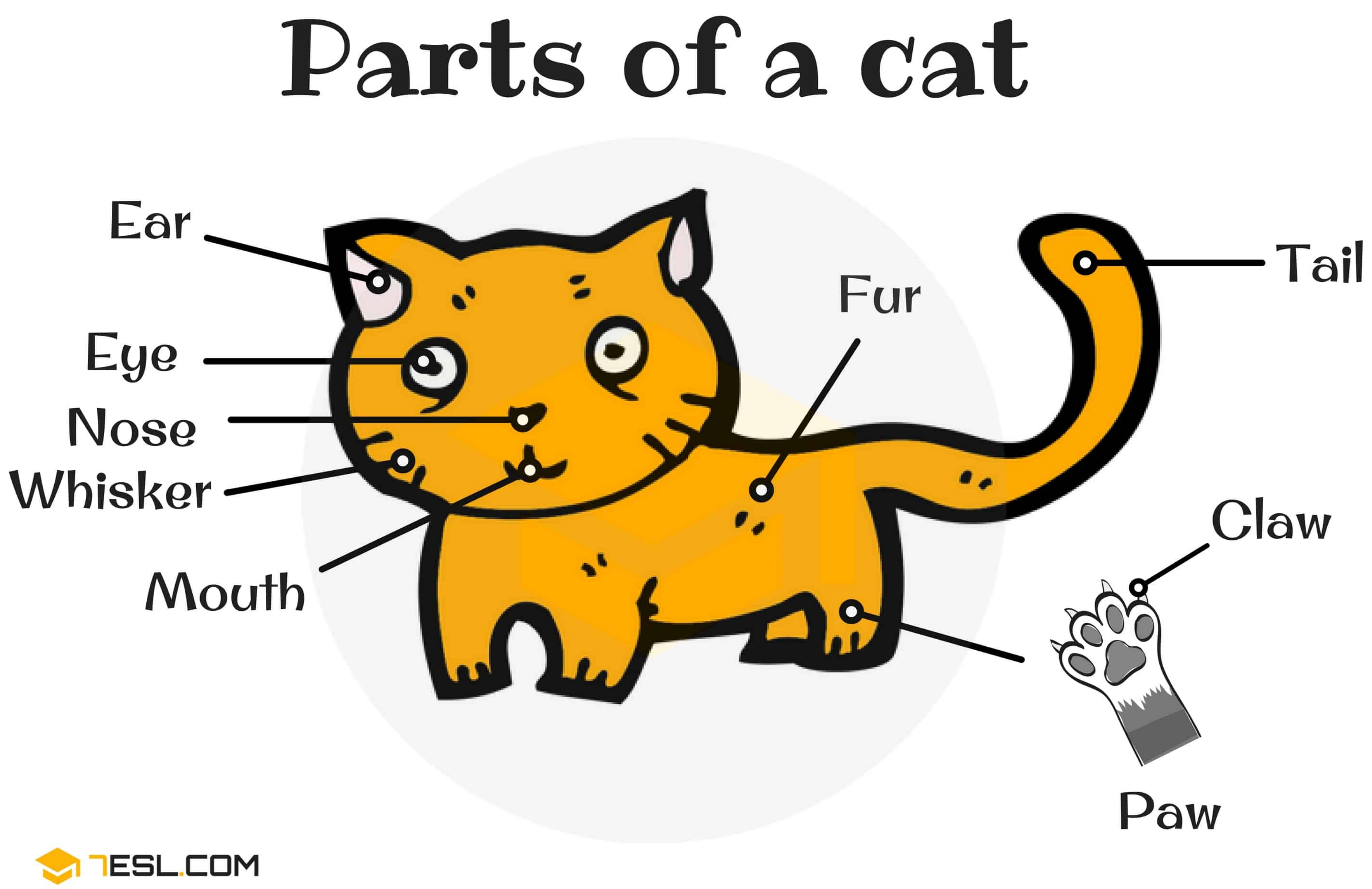
How do we represent the meaning of something?

"cat"

cat: a small domesticated carnivore, *Felis domestica* or *F. catus,* bred in a number of varieties.

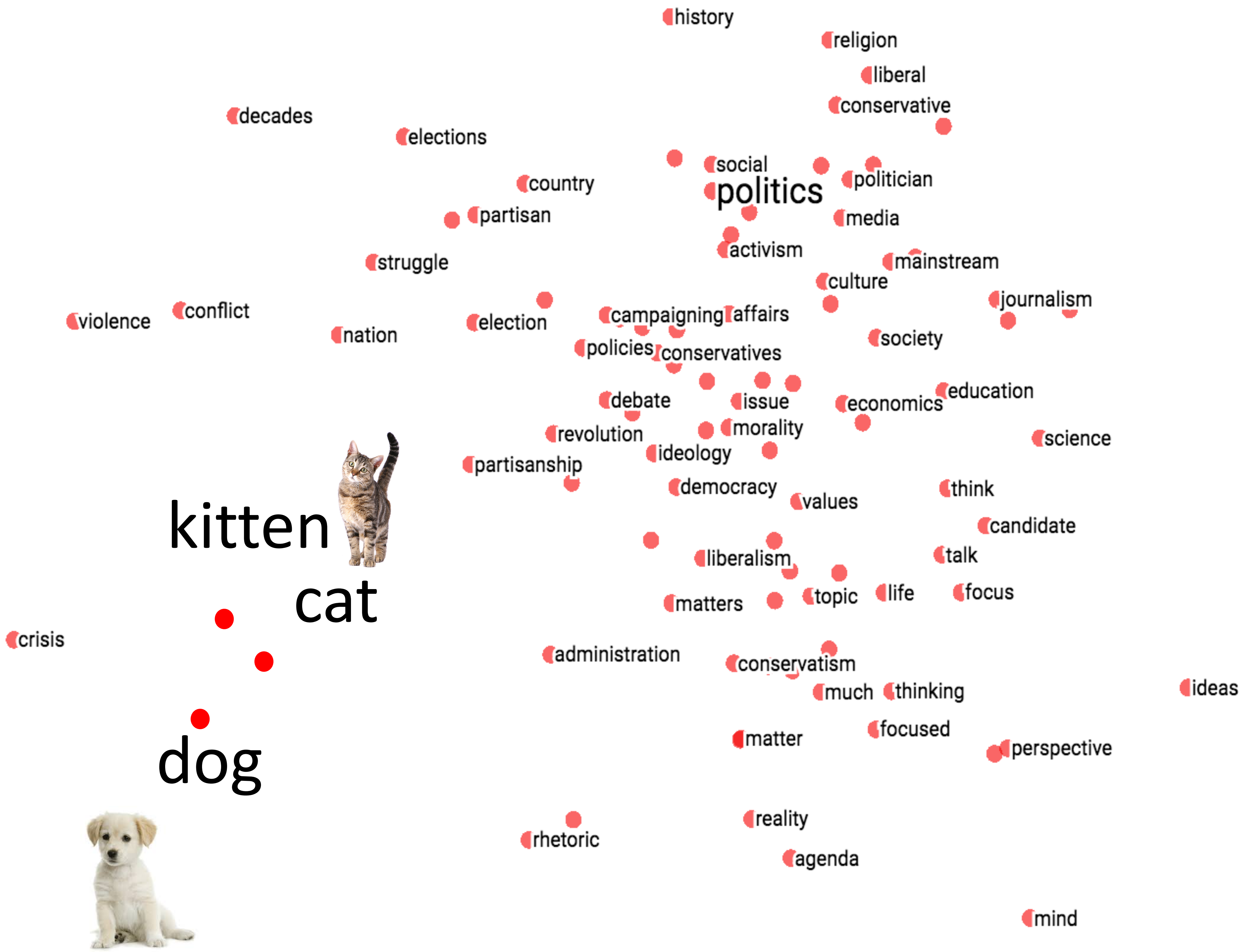
cat → {
  isMammal: true
  hasFur: true
  hasLegs: true
  meows: true
  barks: false
  height: 9.1 – 9.8 in
  weight: 7.9 – 9.9 lbs
  …
}

Attributed representation



Representations

Similar words closer to each other



Representing meaning as vectors

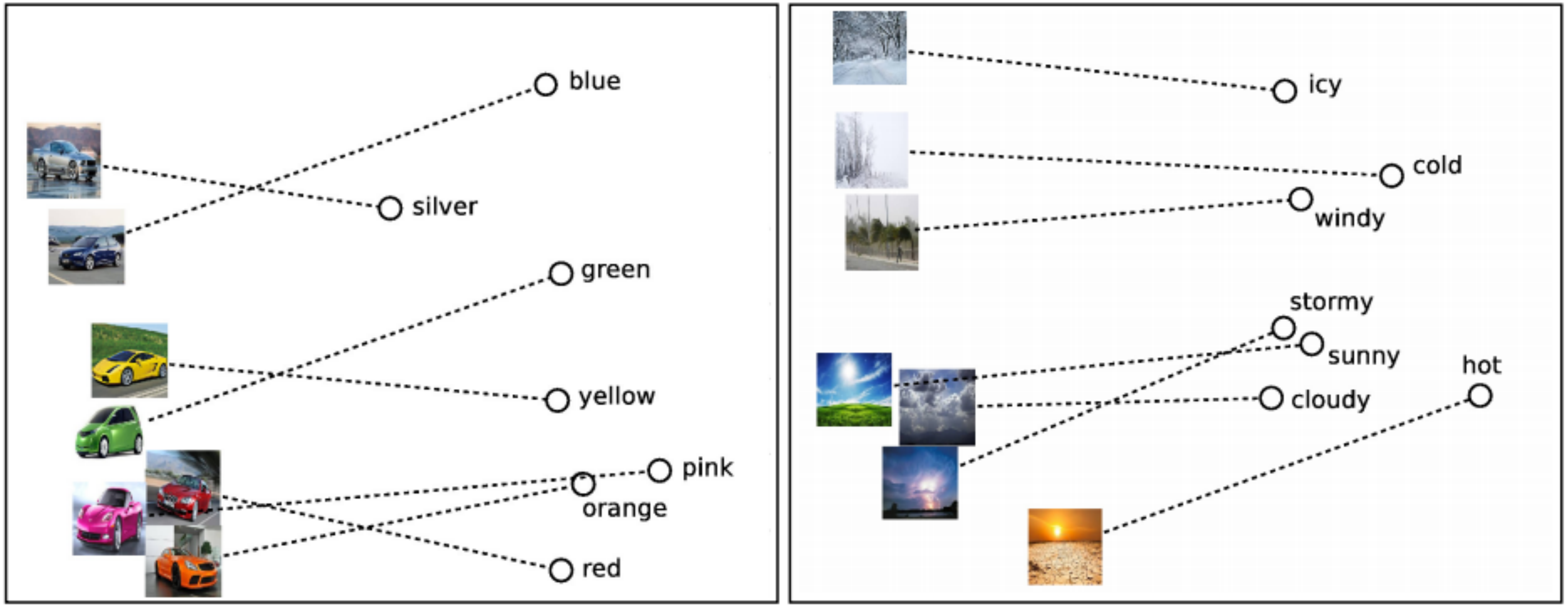
- common representation space
- enables information sharing
- can be learned from data

Embeddings in continuous vector space

cat = [0.04 1.79 -1.79 1.07 0.48]

dog = [0.61 1.84 -1.12 0.52 0.53]

Multimodal Embeddings



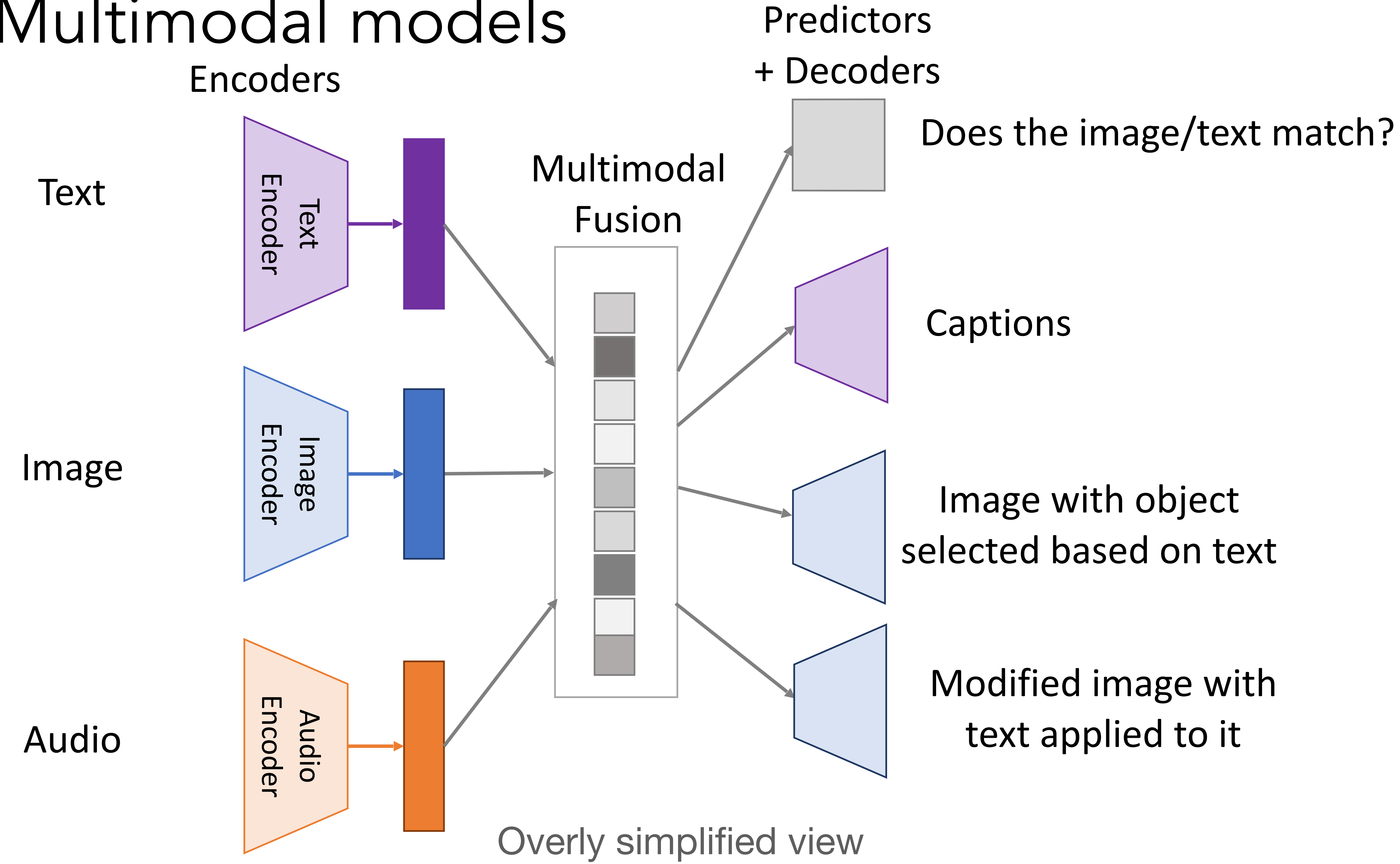
(a) Colors (b) Weather

Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

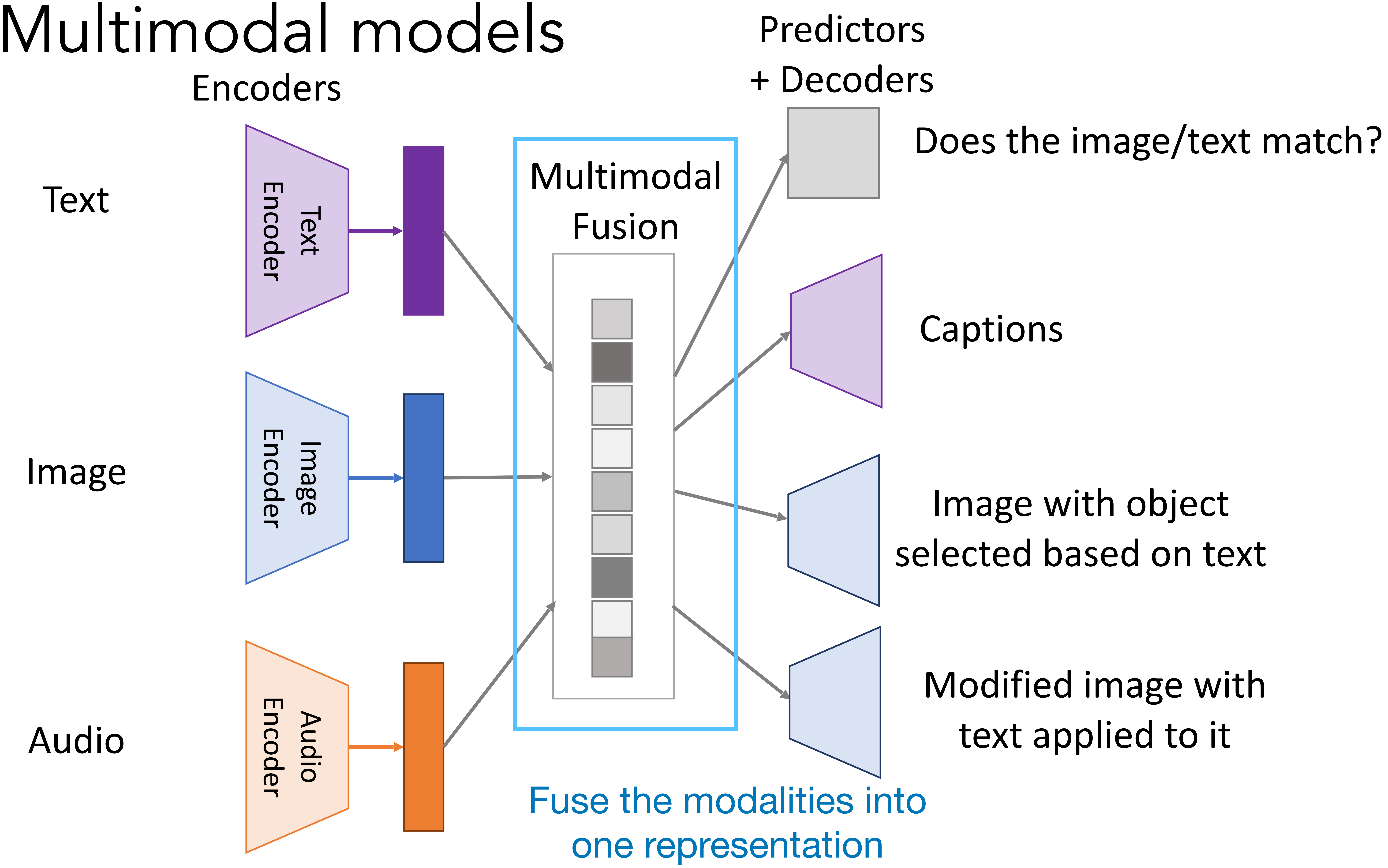
"Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models"
[Kiros, Salakhutdinov, Zemel TACL 2015]

Cross-modal models

Multimodal models

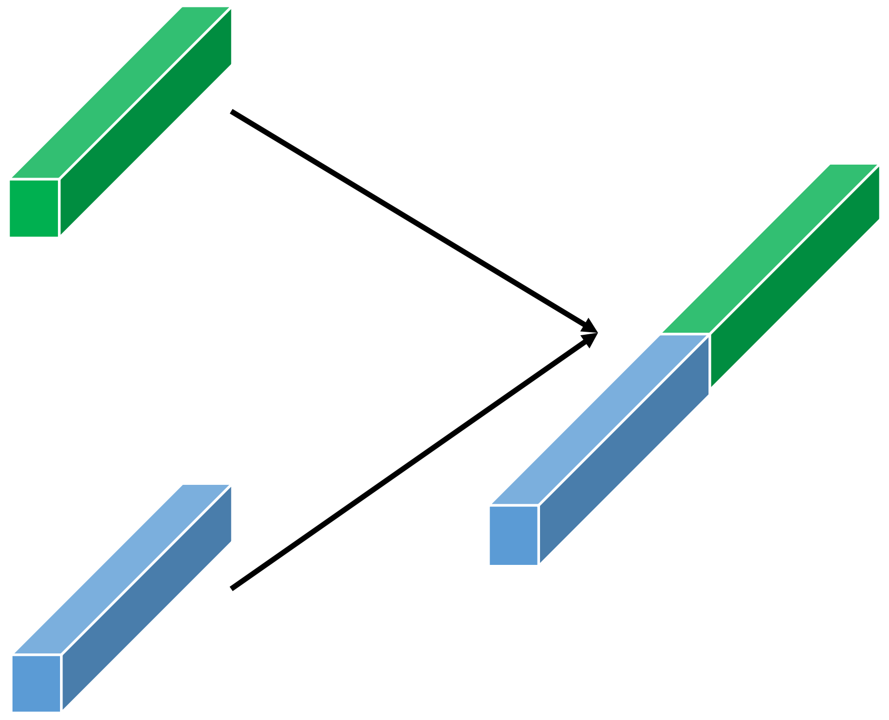


Multimodal models



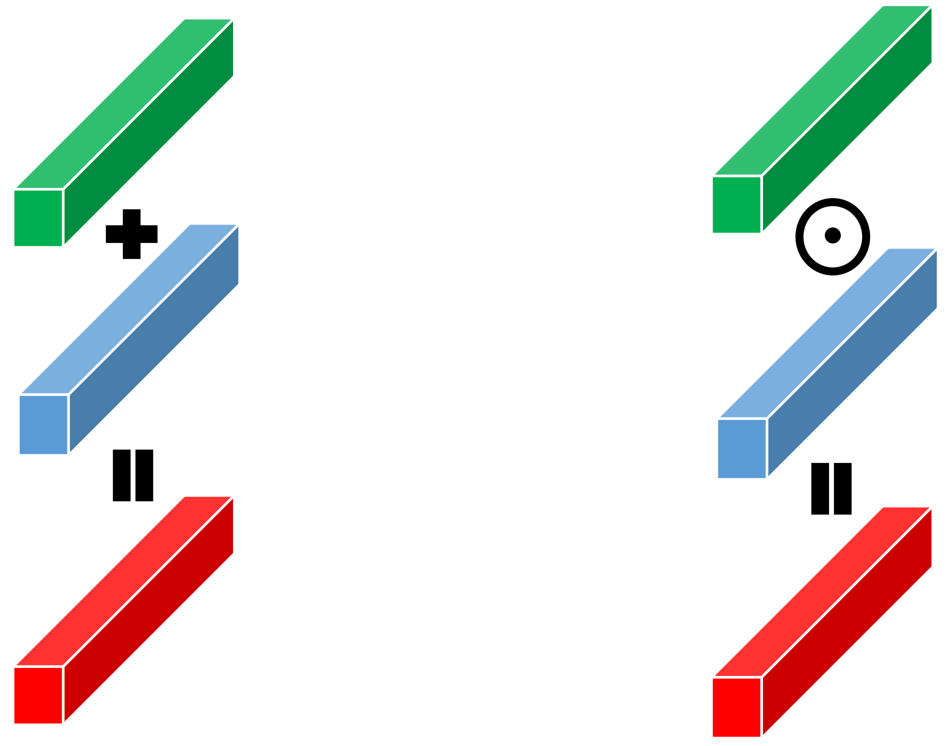
Multimodal Fusion

Concatenation

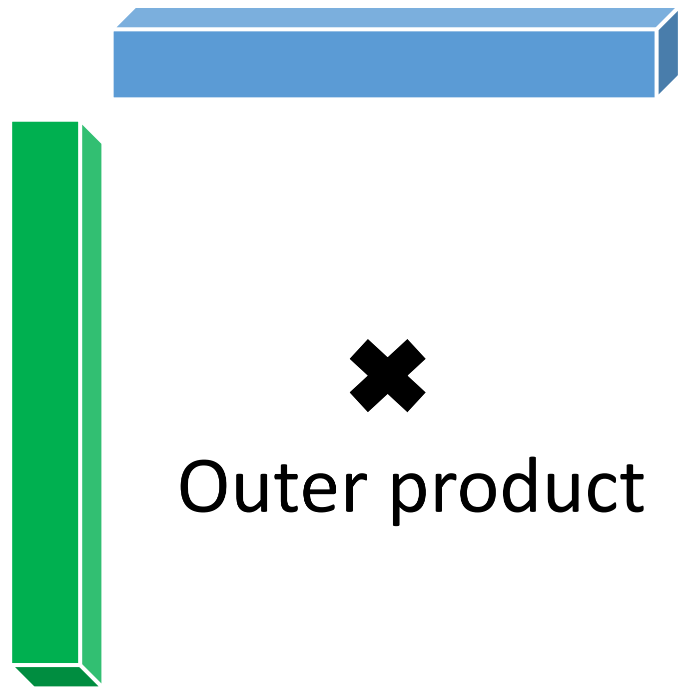


Element wise

Sum / Product



Bilinear Pooling



$$z = W\,[x \otimes q]$$

3000 2048 2048

12.5 billion !!!

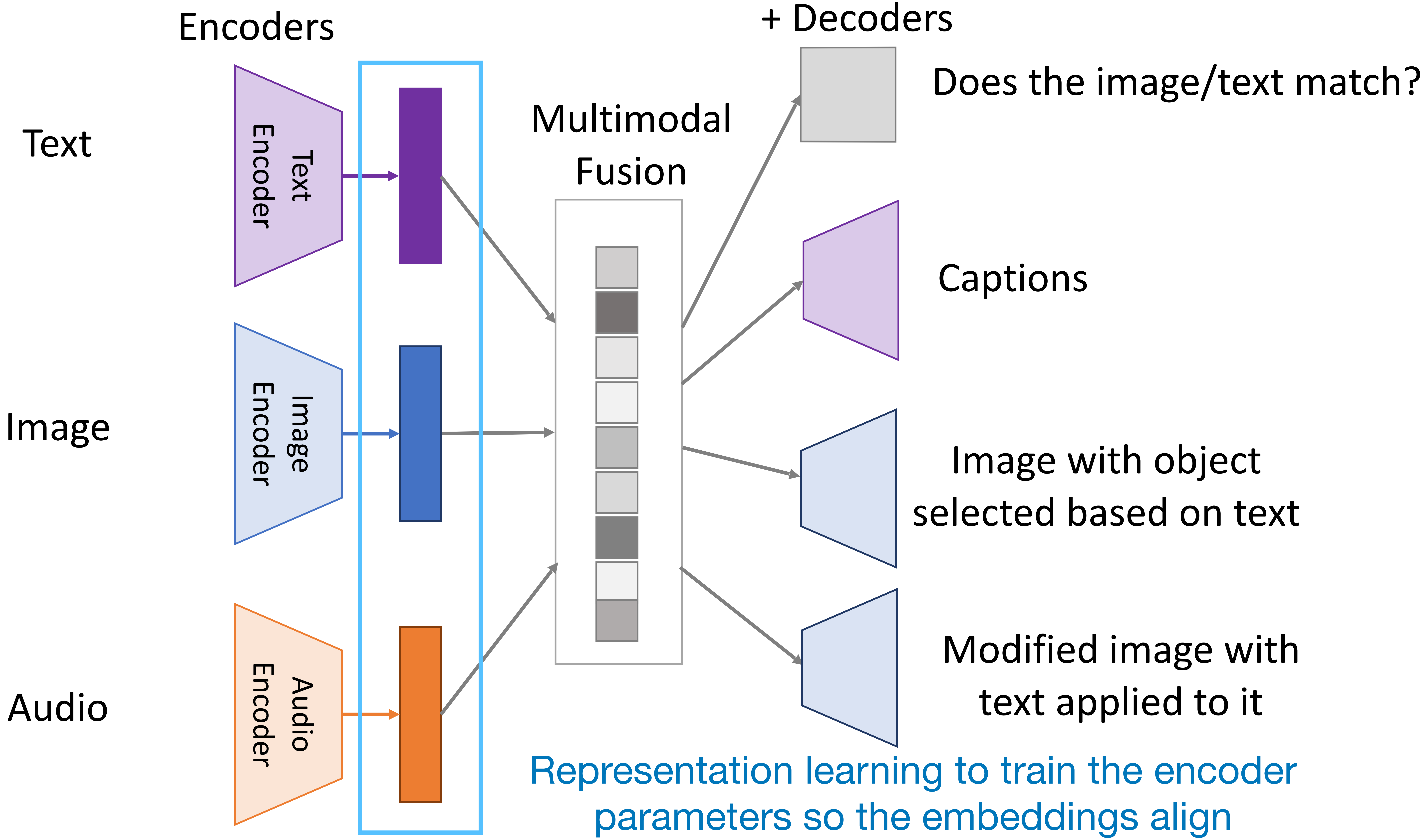
All elements can interact.
More flexible, but lots of weights!

Attention-based fusion

Use transformer to fuse modalities using attention

Image credits: Qi Wu

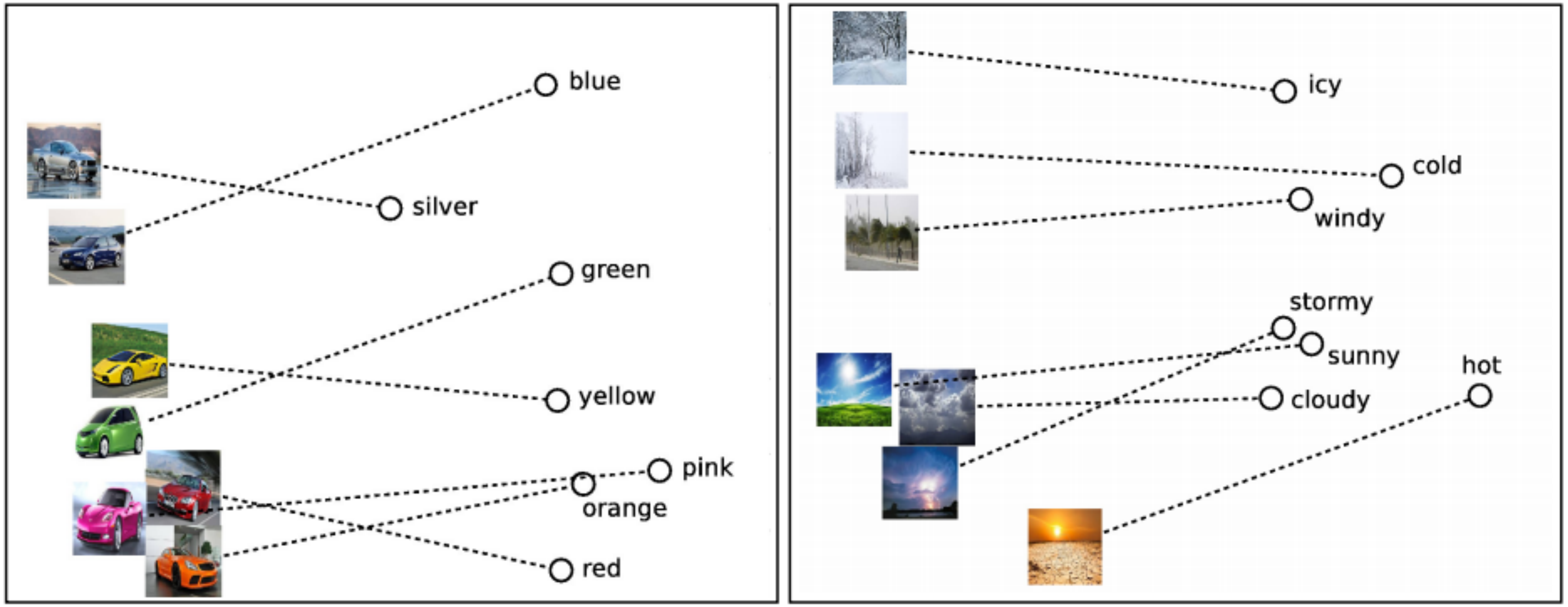
Multimodal models



Representation learning to train the encoder parameters so the embeddings align

Cross-modal Embeddings

Common representation for language and vision: vectors!



(a) Colors (b) Weather

Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models
[Kiros, Salakhutdinov, Zemel TACL 2015, https://arxiv.org/pdf/1411.2539.pdf]

Cross-modal Embeddings

Images and **class labels** are embedded into the same space

Image Embedding

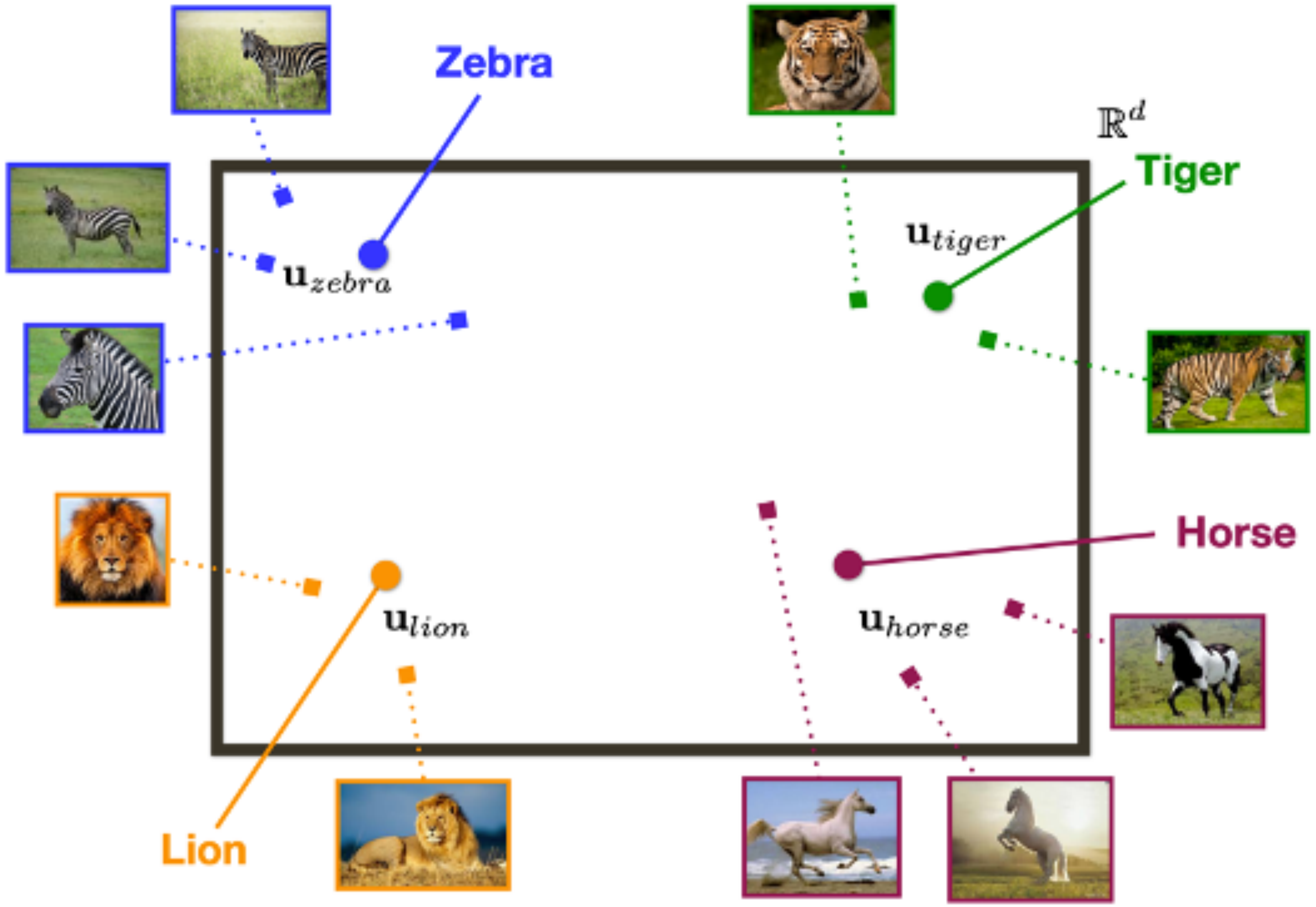
$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) : \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

Similarity in Embedding Space

$$S(\mathbf{u}, \mathbf{u}') = \frac{\mathbf{u}}{\|\mathbf{u}\|} \cdot \frac{\mathbf{u}'}{\|\mathbf{u}'\|}$$



Adapted from slide by Leonid Sigal

Cross-modal Embeddings

Correct label (more similar)

Other labels (less similar)

$$\mathcal{L}_C = \sum \max(0, \alpha - \boxed{S(\Psi(I_i), \mathbf{u}_{y_i})} + \boxed{S(\Psi(I_i), \mathbf{u}_{y_c})})$$

+ pair

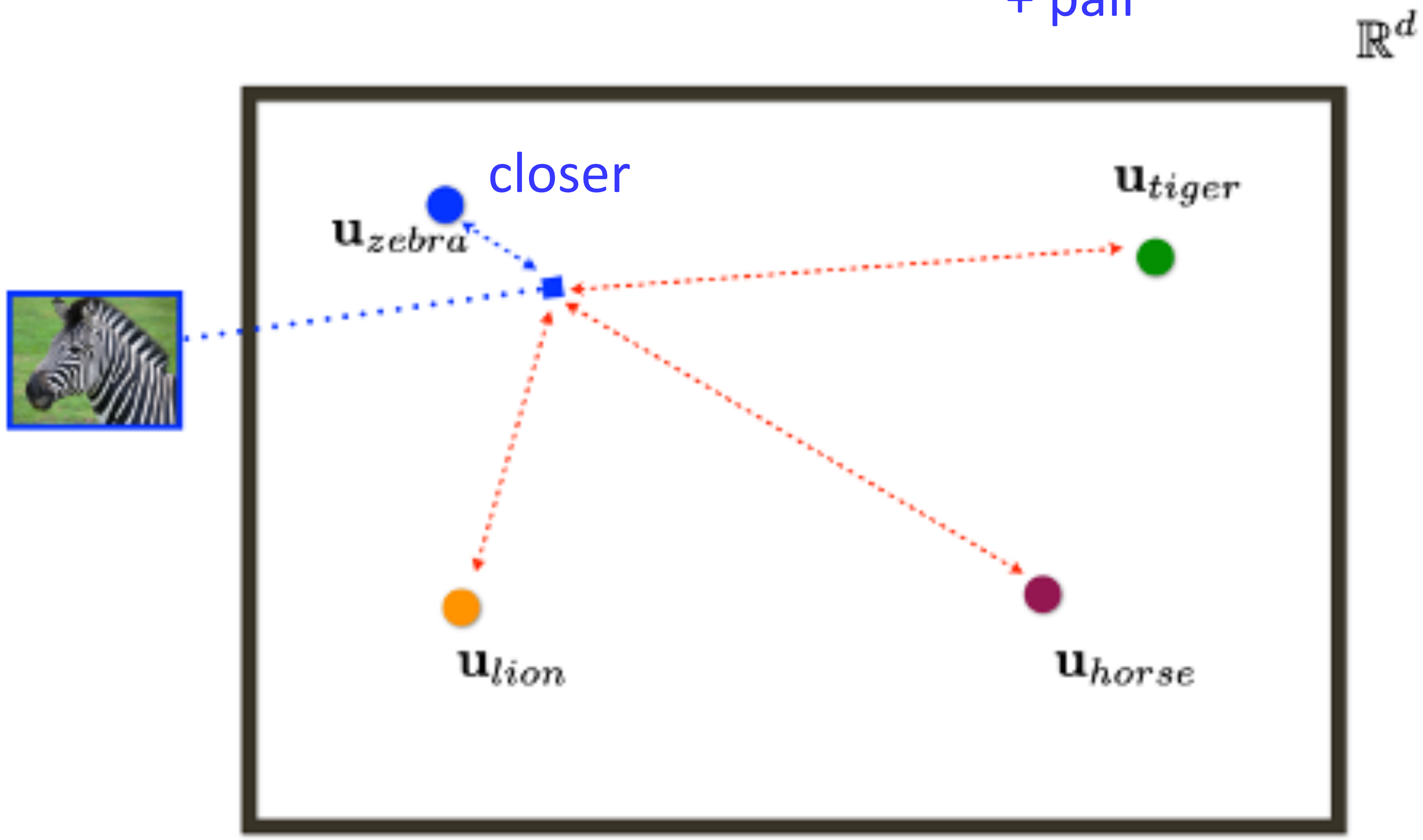
- pair

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) \colon \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$



Similarity in Embedding Space

$$S(\mathbf{u}, \mathbf{u}') = \frac{\mathbf{u}}{\|\mathbf{u}\|} \cdot \frac{\mathbf{u}'}{\|\mathbf{u}'\|}$$

Objective Function:

$$\min_{\mathbf{W}, \mathbf{U}} \sum_i^N \mathcal{L}_C(\mathbf{W}, \mathbf{U}, I_i, y_i) + \lambda_1 ||\mathbf{W}||_F^2 + \lambda_2 ||\mathbf{U}||_F^2$$

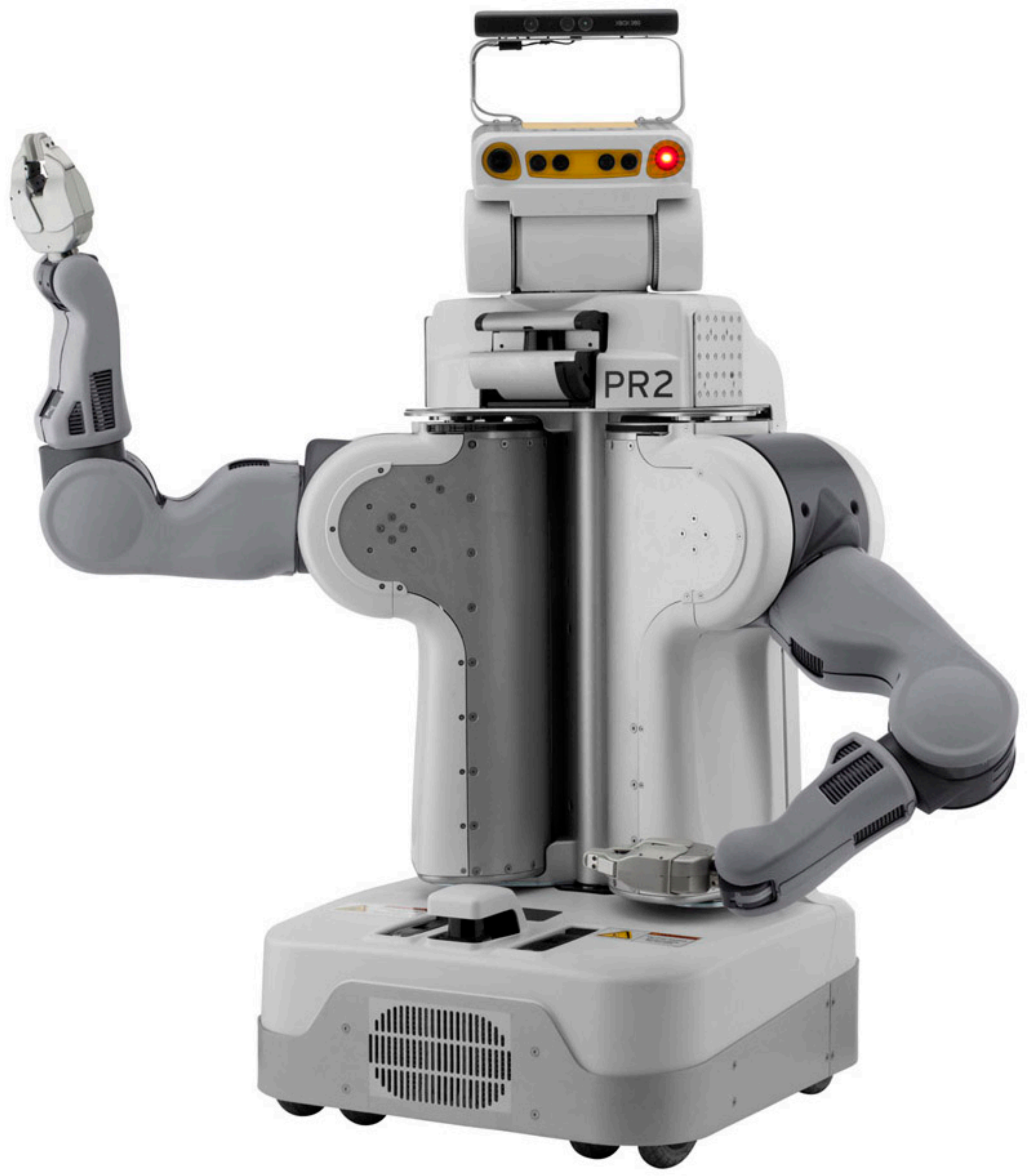
[Bengio *et al.*,, NIPS'10]

[Weinberger, Chapelle, NIPS'09]

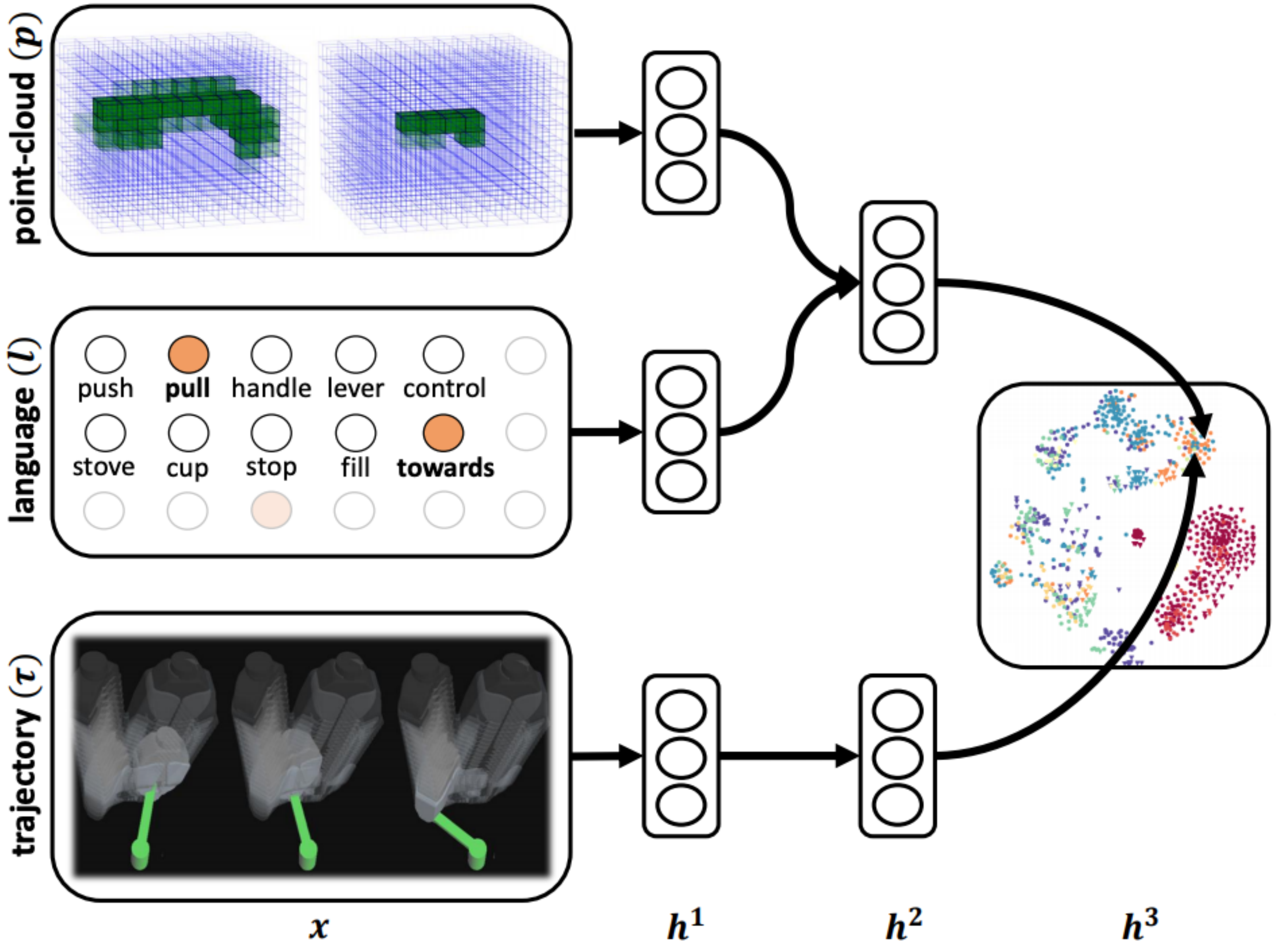
Adapted from slide by Leonid Sigal

Can embed anything!

PR2 robot



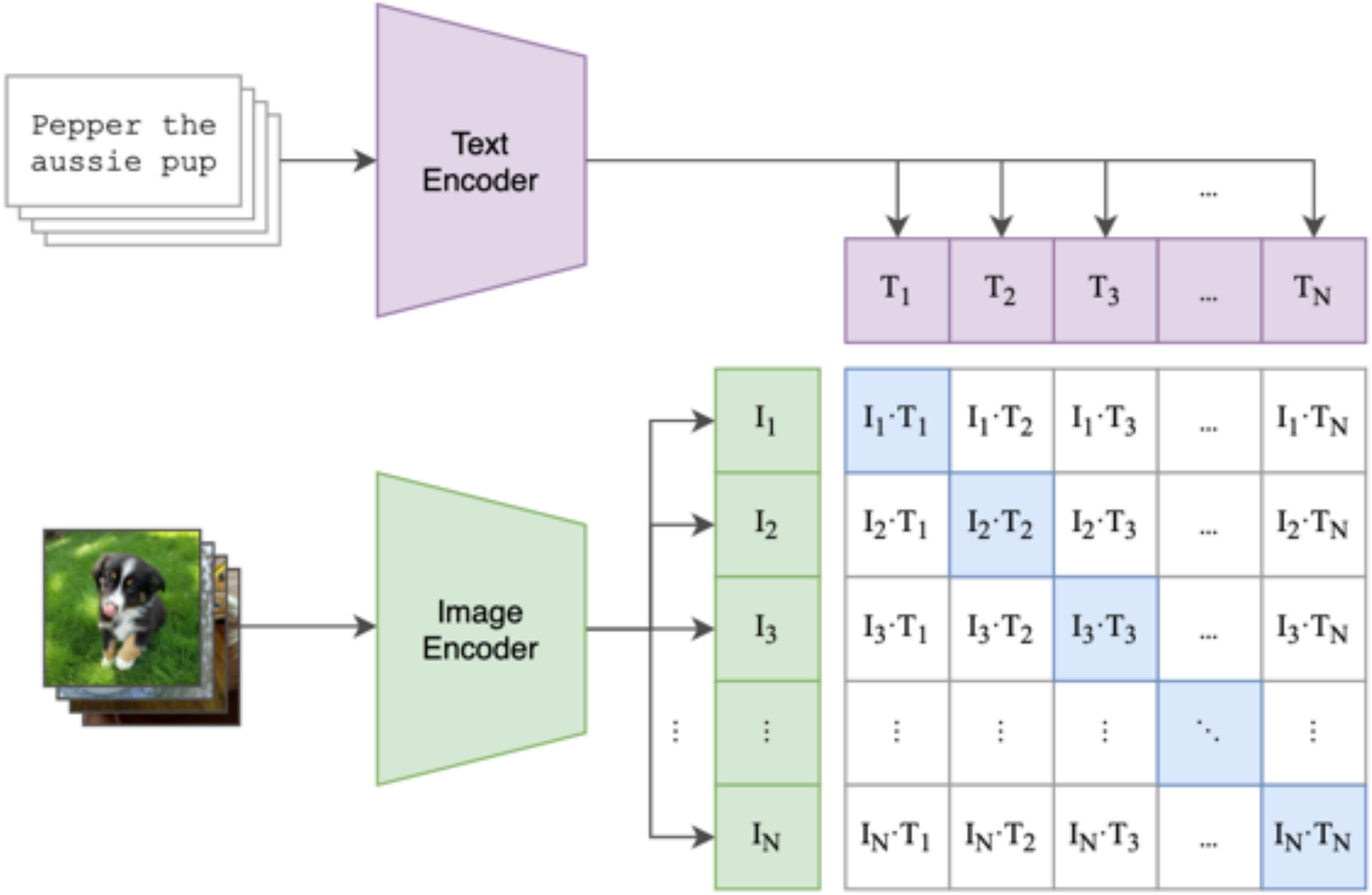
https://robots.ieee.org/robots/pr2



Deep Multimodal Embedding: Manipulating Novel Objects with Point-clouds, Language and Trajectories [Sung et al, 2015]

Contrastive pretraining

OpenAI CLIP

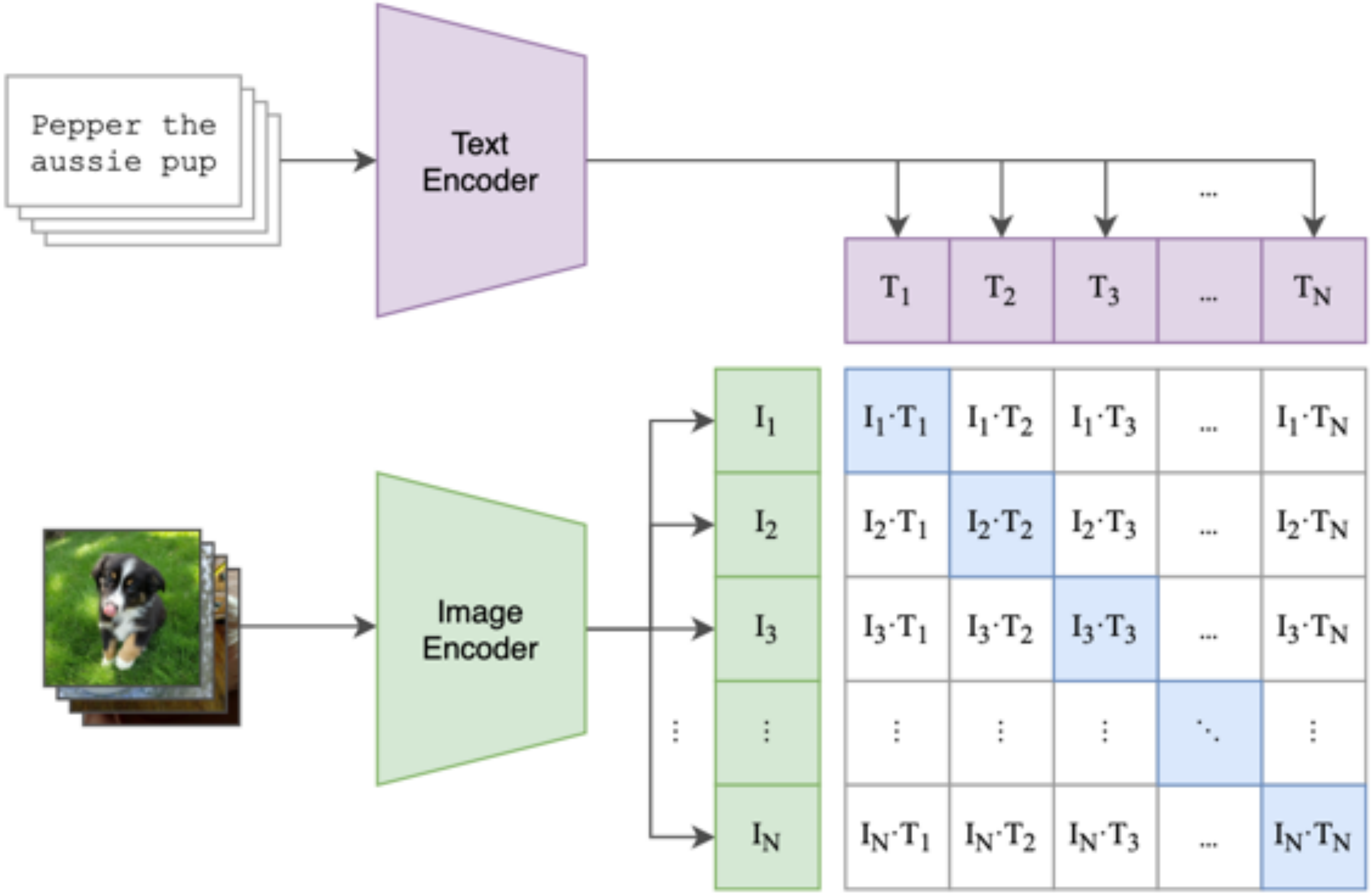


- Train on large amount of data
  - WebImageText: 400M text-image pairs
- Contrastive pretraining: does the text-image pair match?
  - Batch size N=32K
  - N positive pairs
  - $N^2 - N$ negative pairs
- Transformer based model for both vision and language

Learning transferable visual models from natural from natural language supervision [Radford et al, 2020]

Contrastive pretraining

OpenAI CLIP



- Contrastive Loss

NT-Xent loss (images and text)

$$l_j^{I \to T} = -\log \frac{\exp(\mathrm{sim}(I_j, T_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(I_j, T_k)/\tau)}$$

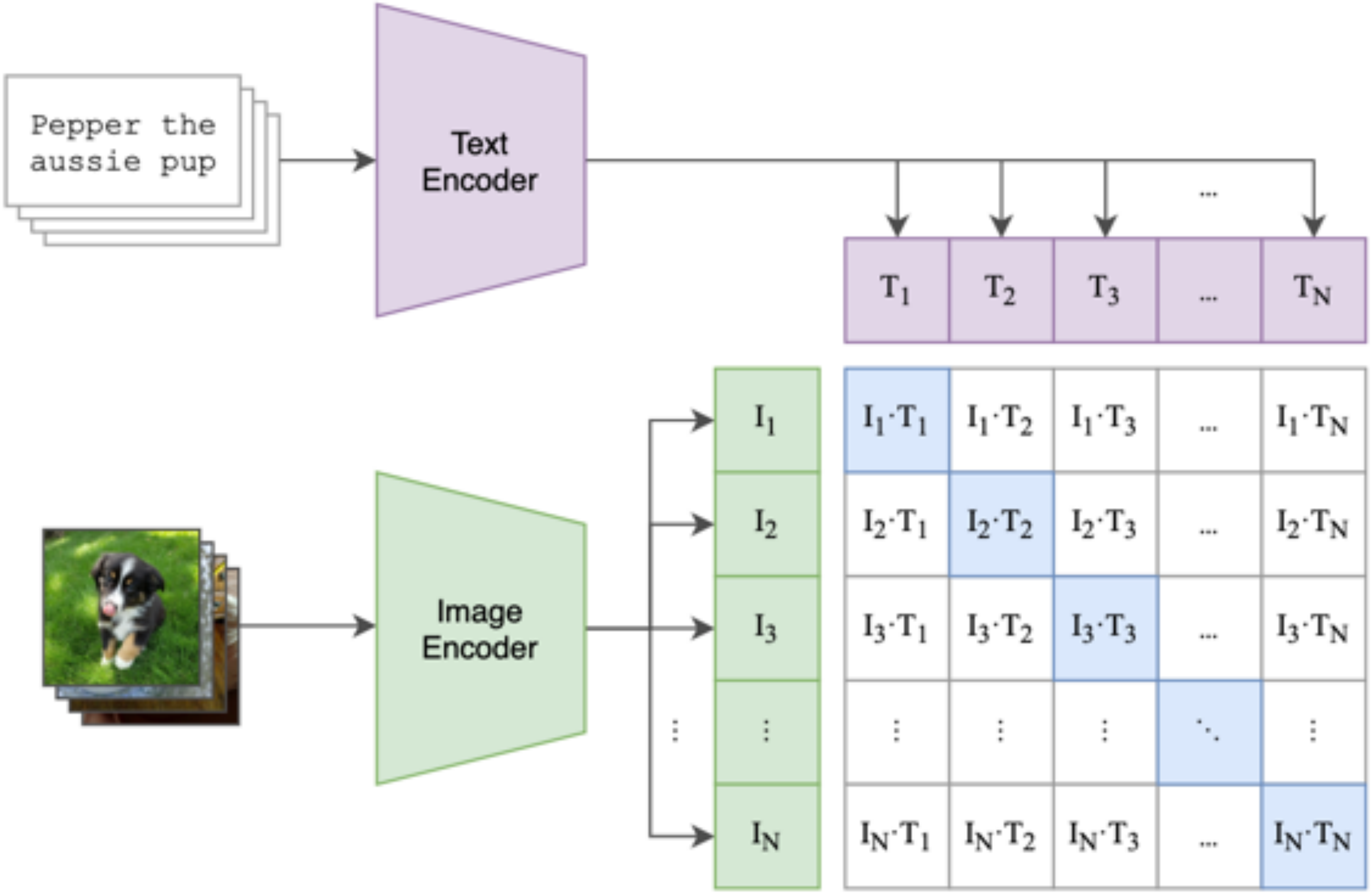
Symmetric Bimodal loss

$$L(I,T) = \frac{1}{N} \sum_{j=1}^{N} (\alpha l_j^{I \to T} + (1-\alpha) l_j^{T \to I})$$

Learning transferable visual models from natural from natural language supervision [Radford et al, 2020]

Contrastive pretraining

OpenAI CLIP



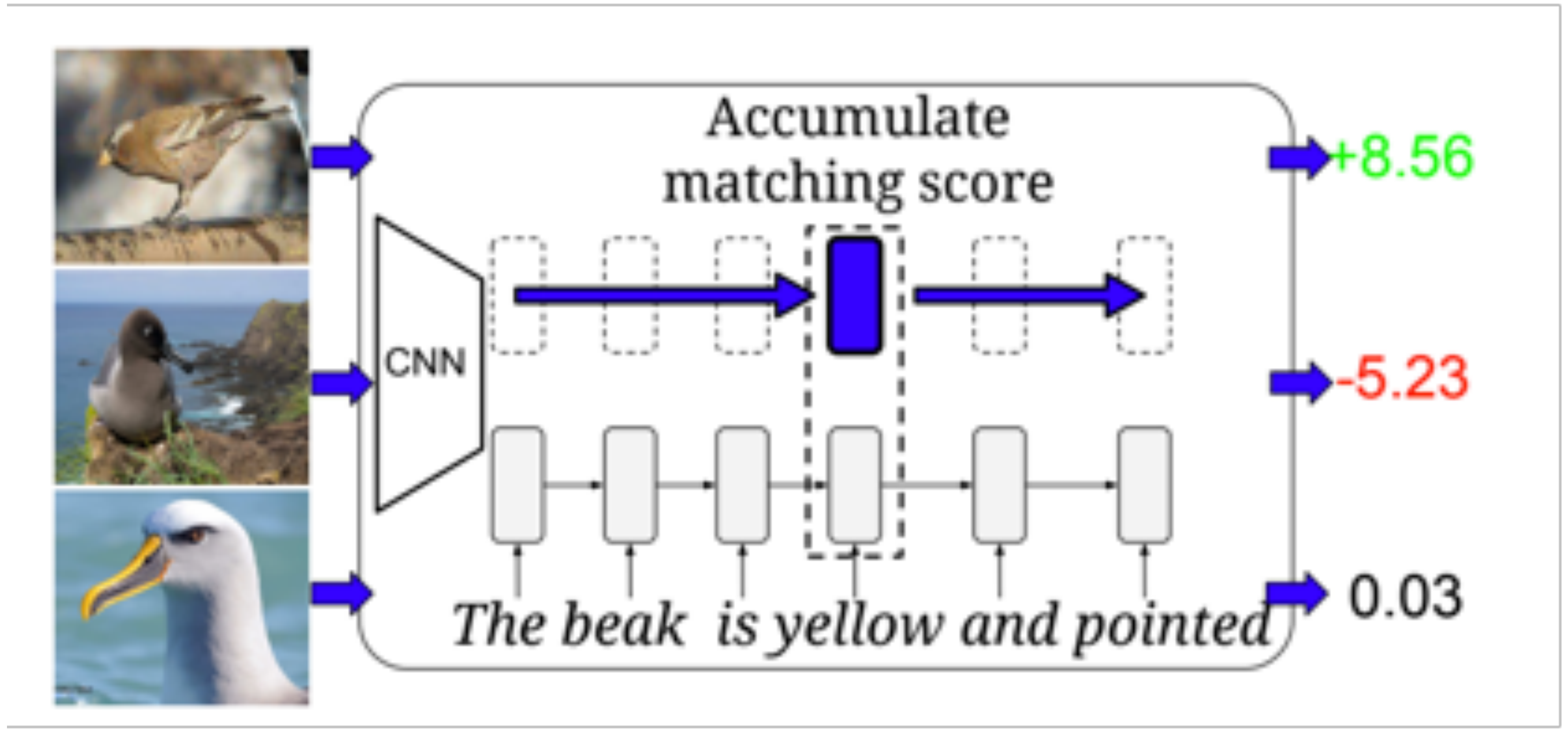

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Learning transferable visual models from natural from natural language supervision [Radford et al, 2020]

Aligned embeddings can be used for a variety of tasks

Retrieval

- Text to image/video retrieval
- Image/video to text retrieval



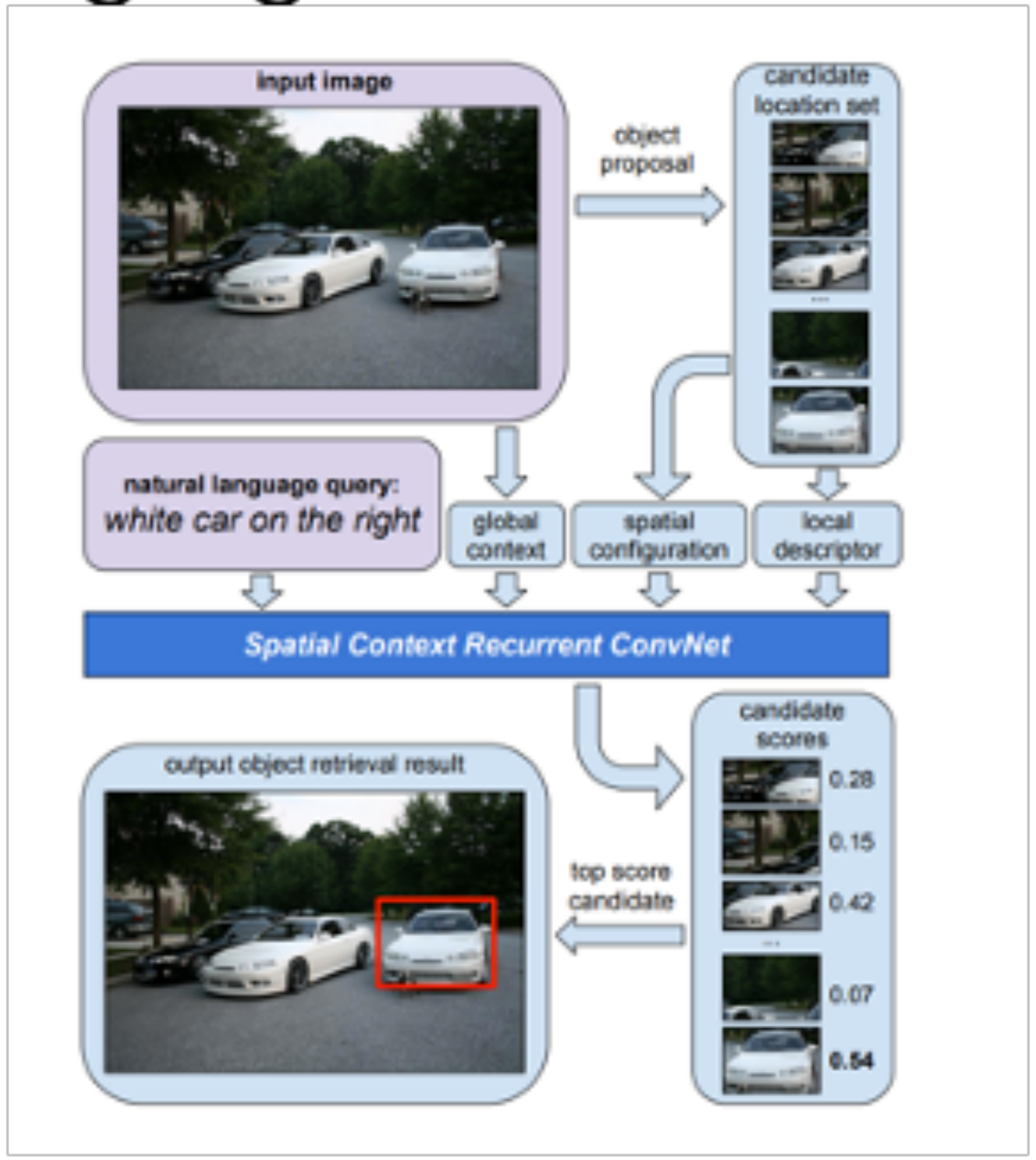
	Top-1 Acc (%)		AP@50 (%)	
Embedding	**DA-SJE**	**DS-SJE**	**DA-SJE**	**DS-SJE**
ATTRIBUTES	50.9	50.4	20.4	**50.0**
WORD2VEC	38.7	38.6	7.5	33.5
BAG-OF-WORDS	43.4	44.1	24.6	39.6
CHAR CNN	47.2	48.2	2.9	42.7
CHAR LSTM	22.6	21.6	11.6	22.3
CHAR CNN-RNN	54.0	54.0	6.9	45.6
WORD CNN	50.5	51.0	3.4	43.3
WORD LSTM	52.2	53.0	**36.8**	46.8
WORD CNN-RNN	**54.3**	**56.8**	4.8	48.7

CUB Birds

"Learning Deep Representations of Fine-Grained Visual Descriptions" (Reed et al, CVPR 2016)

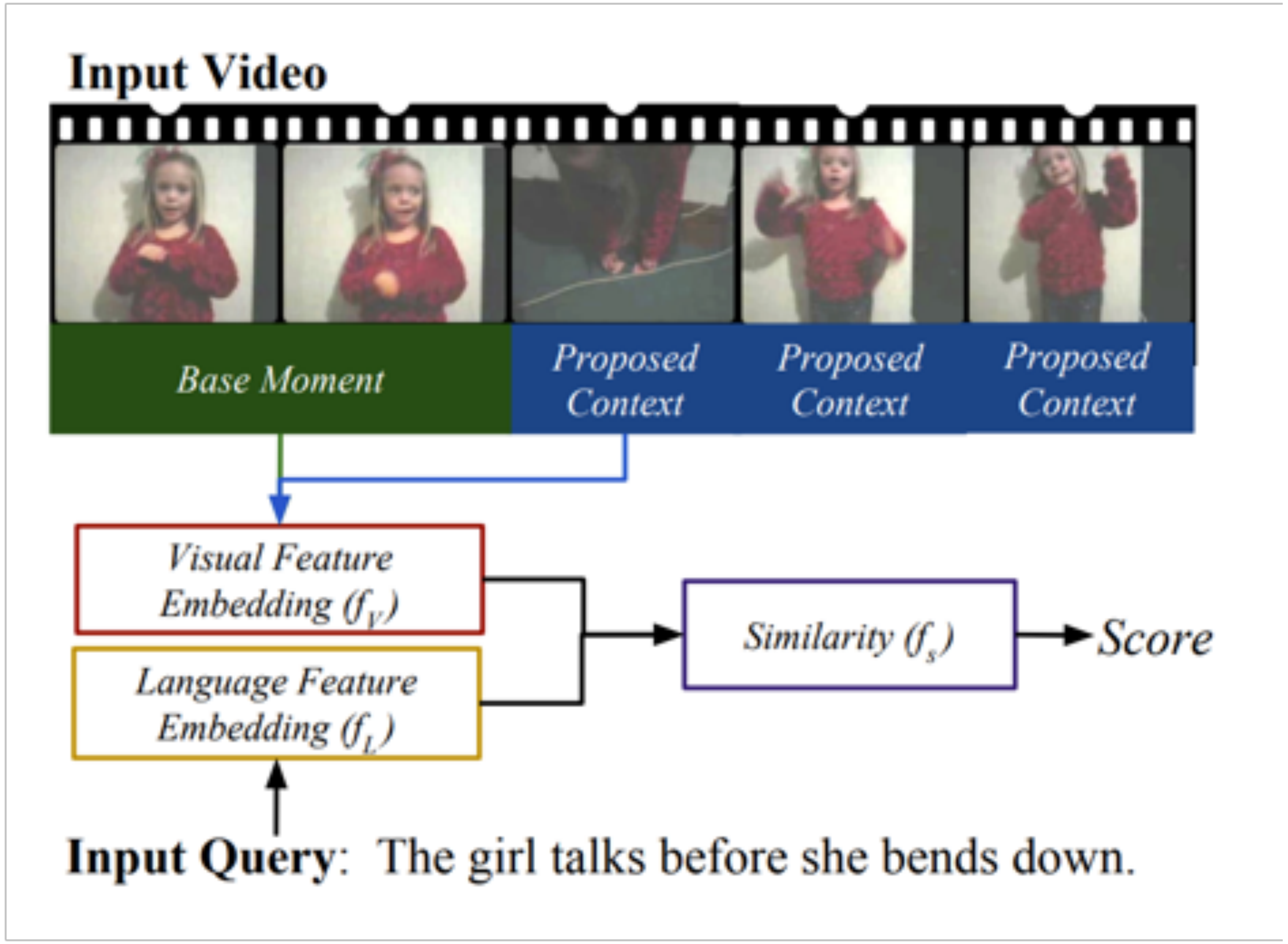
Grounding

Match image region to language



Natural Language Object Retrieval
(Hu et al, CVPR 2016)

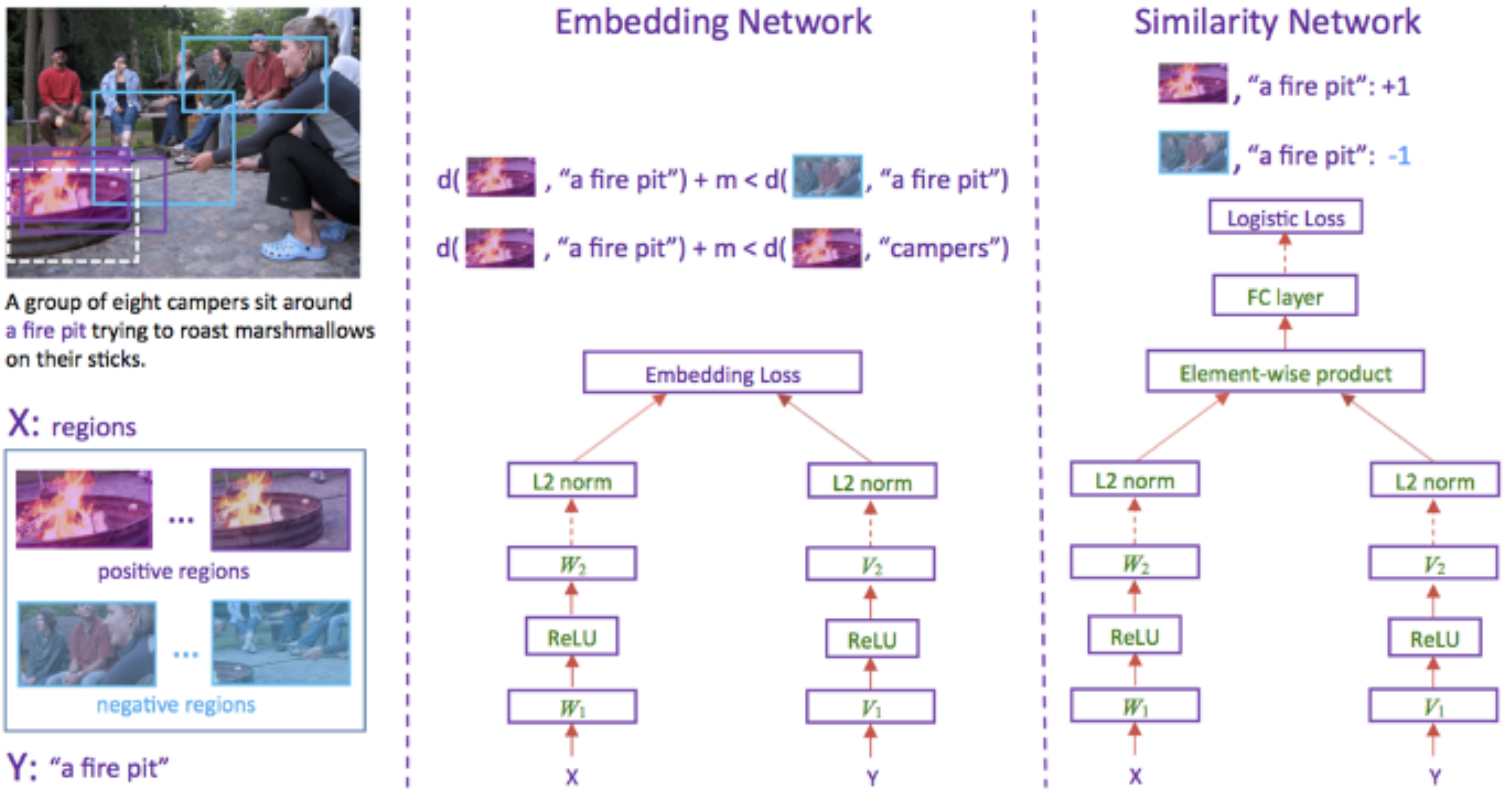
Match video frames to language



Localizing moments in video with temporal language
(Hendricks et al, EMNLP, 2018)

Grounding

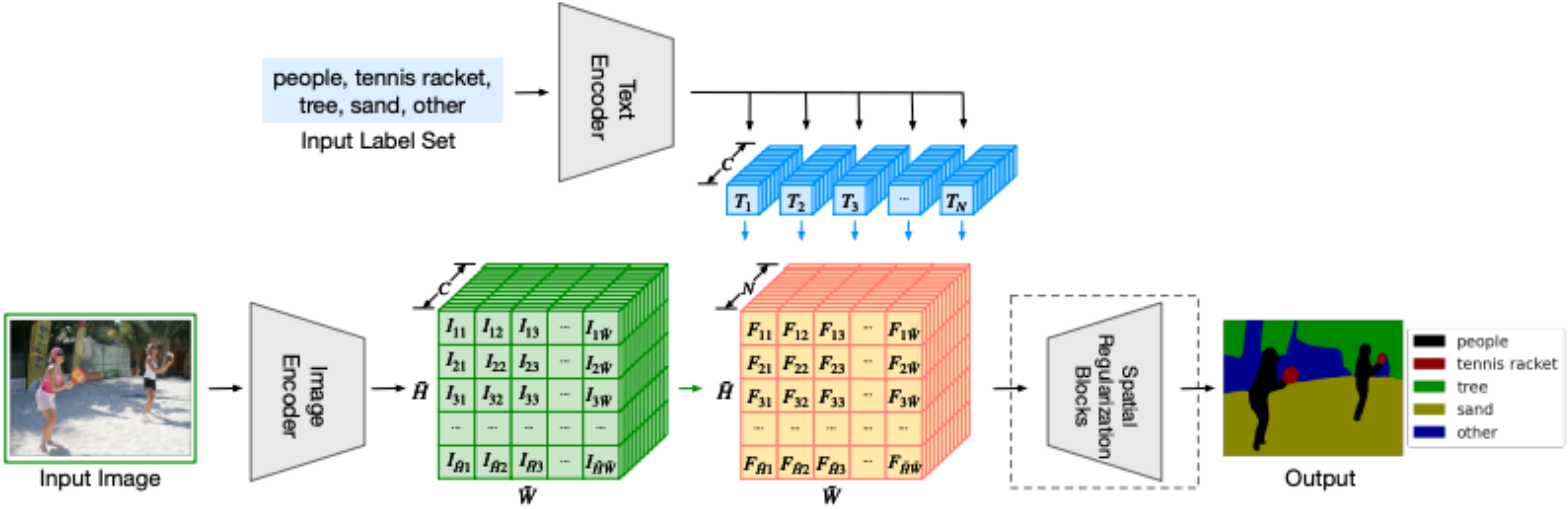
Phrase Localization



Learning Two-Branch Neural Networks for Image-Text Matching Tasks
(Wang et al, TPAMI 2018)

Language Driven Semantic Segmentation

Use CLIP as text encoder



Use dense prediction transformers architecture

Upsamples to obtain final output

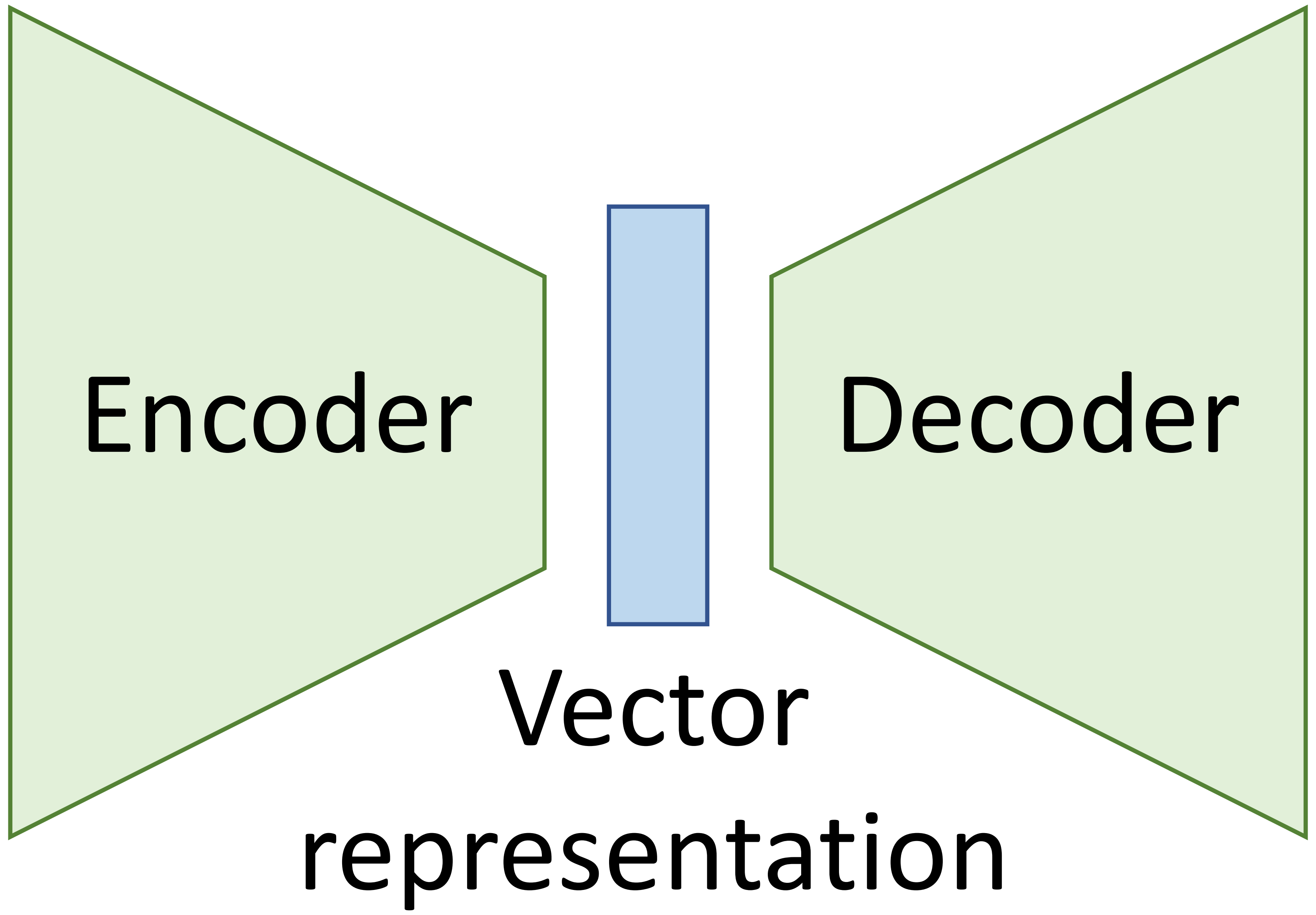
Language Driven Semantic Segmentation [Li et al, ICLR 2022]

Used in DALL-E

DALL-E (2021): 12B parameter version of GPT-3 trained to generate images from text descriptions

CLIP used to score and re-rank generated images

a teapot in the shape of a pikachu.
a teapot imitating a pikachu



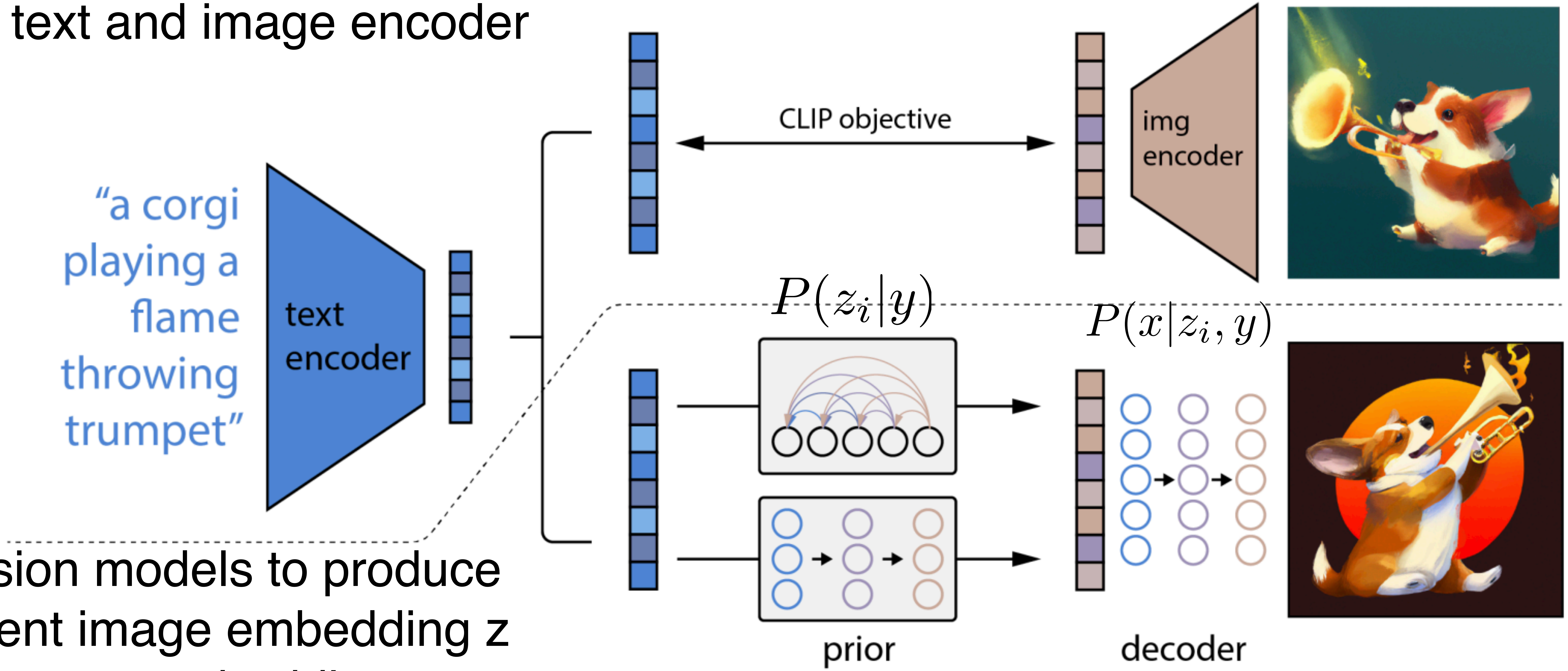
Image

Images are represented as sequence of tokens
(each image is encoded as 32x32 grid of tokens using discrete VAE to 8192 codewords)

[Zero-shot Text-to-Image Generation](https://openai.com/blog/dall-e/) [Ramesh et al, 2021]
(https://openai.com/blog/dall-e/)

DALLE-2: Text-to-Image generation with diffusion models

CLIP text and image encoder



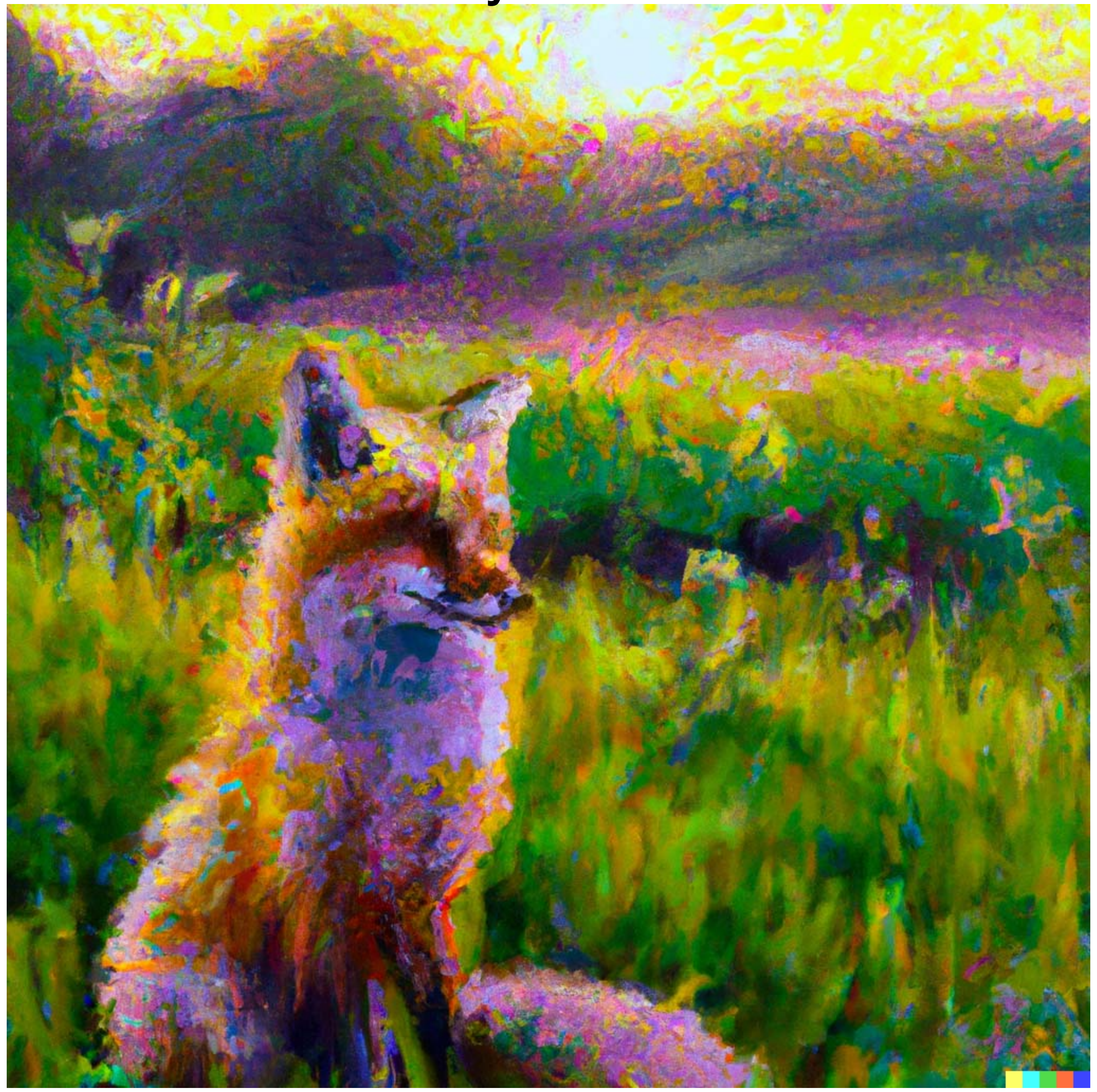
Diffusion models to produce
- latent image embedding z from text embedding y,
- image x from latent image embedding z

Hierarchical Text-Conditional Image Generation with CLIP Latents
https://arxiv.org/pdf/2204.06125.pdf [Ramesh et al, arXiv 2022]
https://openai.com/dall-e-2/

Text-to-Image Generation with Diffusion Models

OpenAI DALL-E 2

a painting of a fox sitting in a field at sunrise in the style of Claude Monet



https://openai.com/dall-e-2/

Google Imagen

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

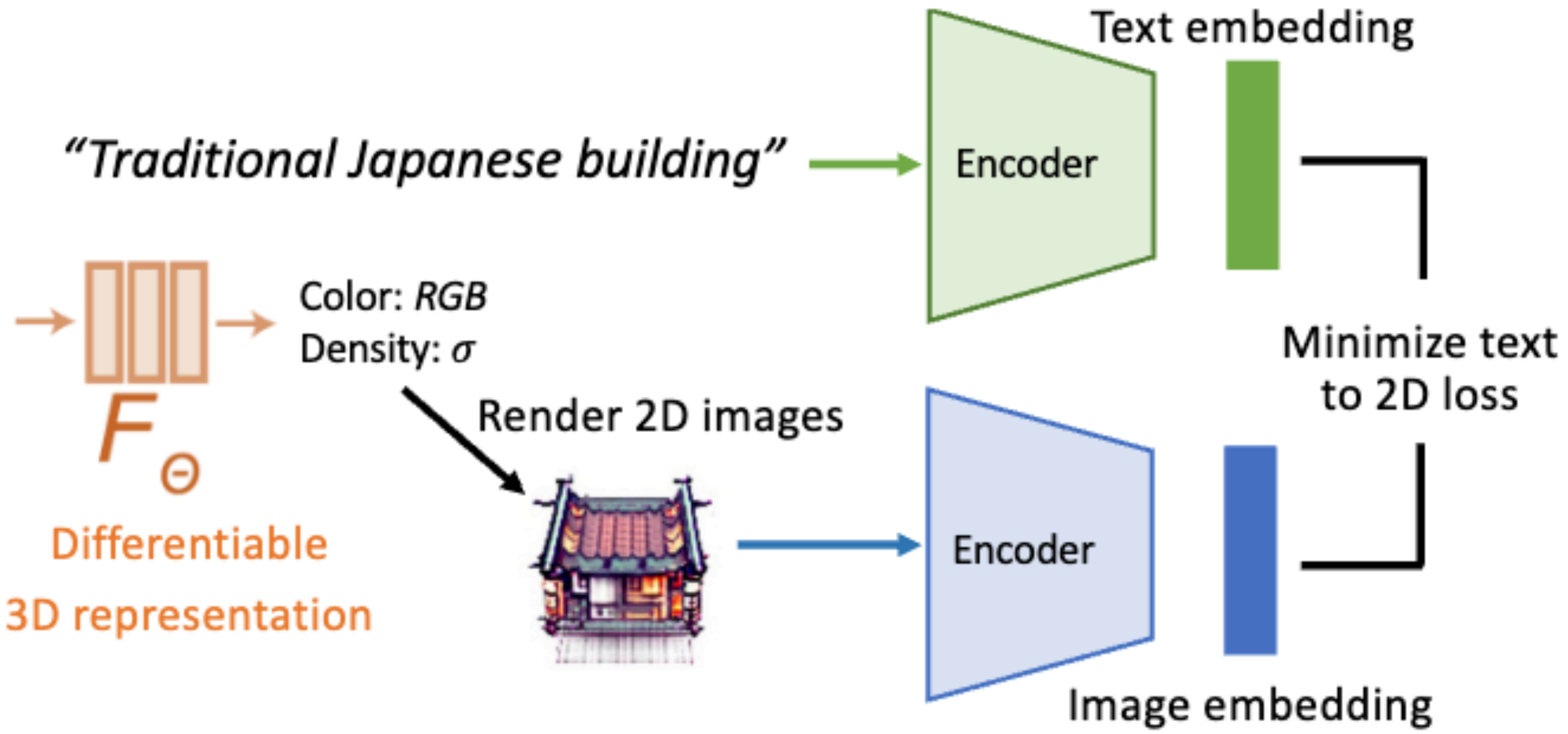
https://imagen.research.google/

Try it out yourself: https://huggingface.co/spaces/stabilityai/stable-diffusion, https://www.craiyon.com/, https://www.midjourney.com/home/

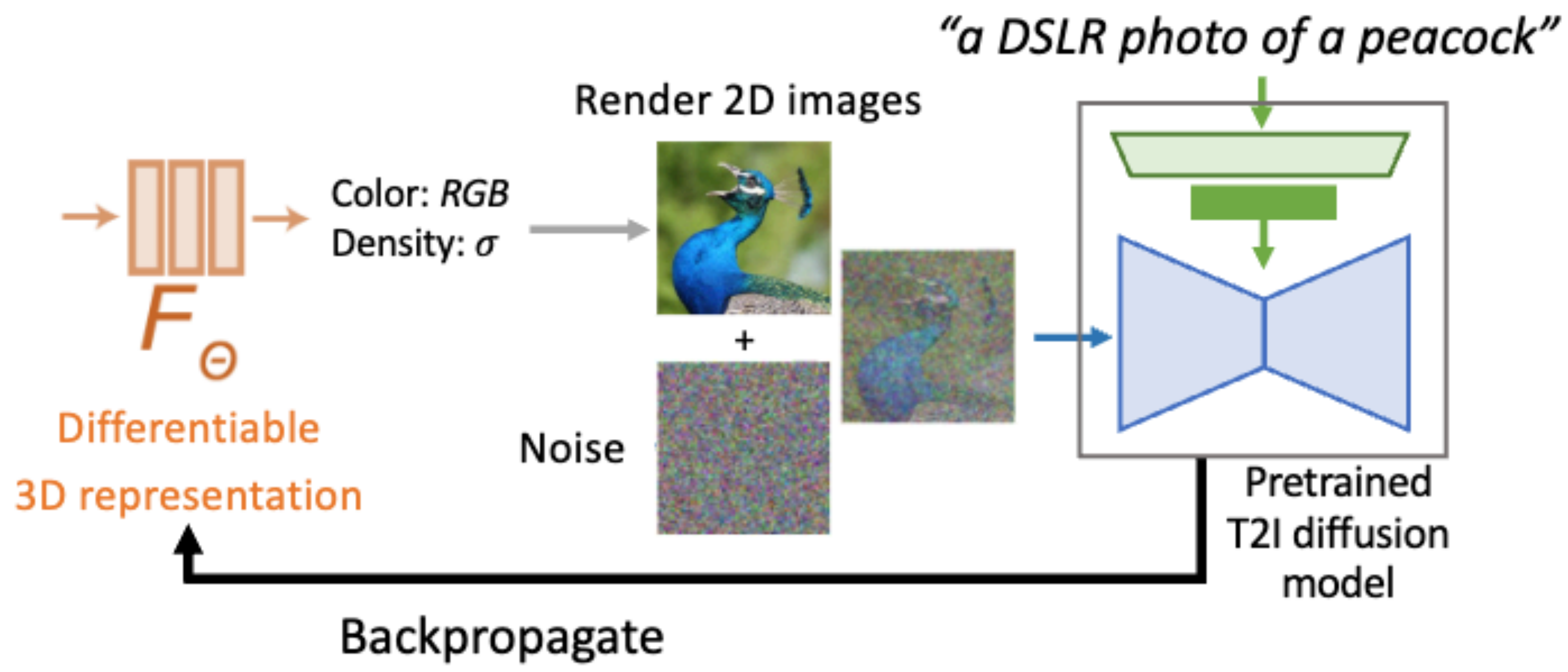
Text-to-3D generation

Optimize differentiable 3D representations with text-image models

Aligned text to image embeddings



Text-to-image diffusion model

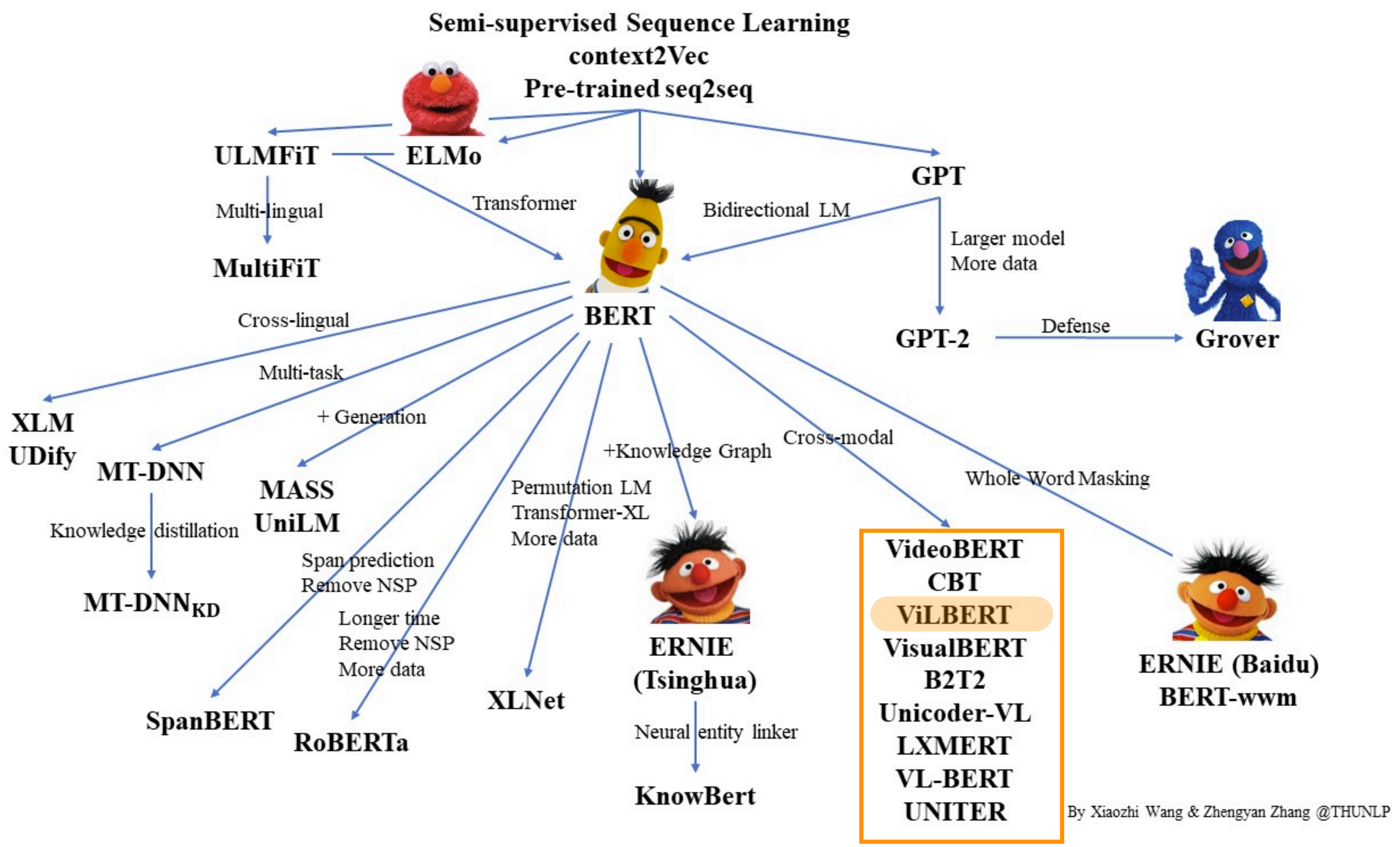


Text-to-3D with diffusion models

a DSLR photo of a squirrel wearing a purple hoodie

DreamFusion: Text-to-3D using 2D Diffusion
https://arxiv.org/abs/2209.14988 [Poole et al, 2022]
https://dreamfusion3d.github.io/

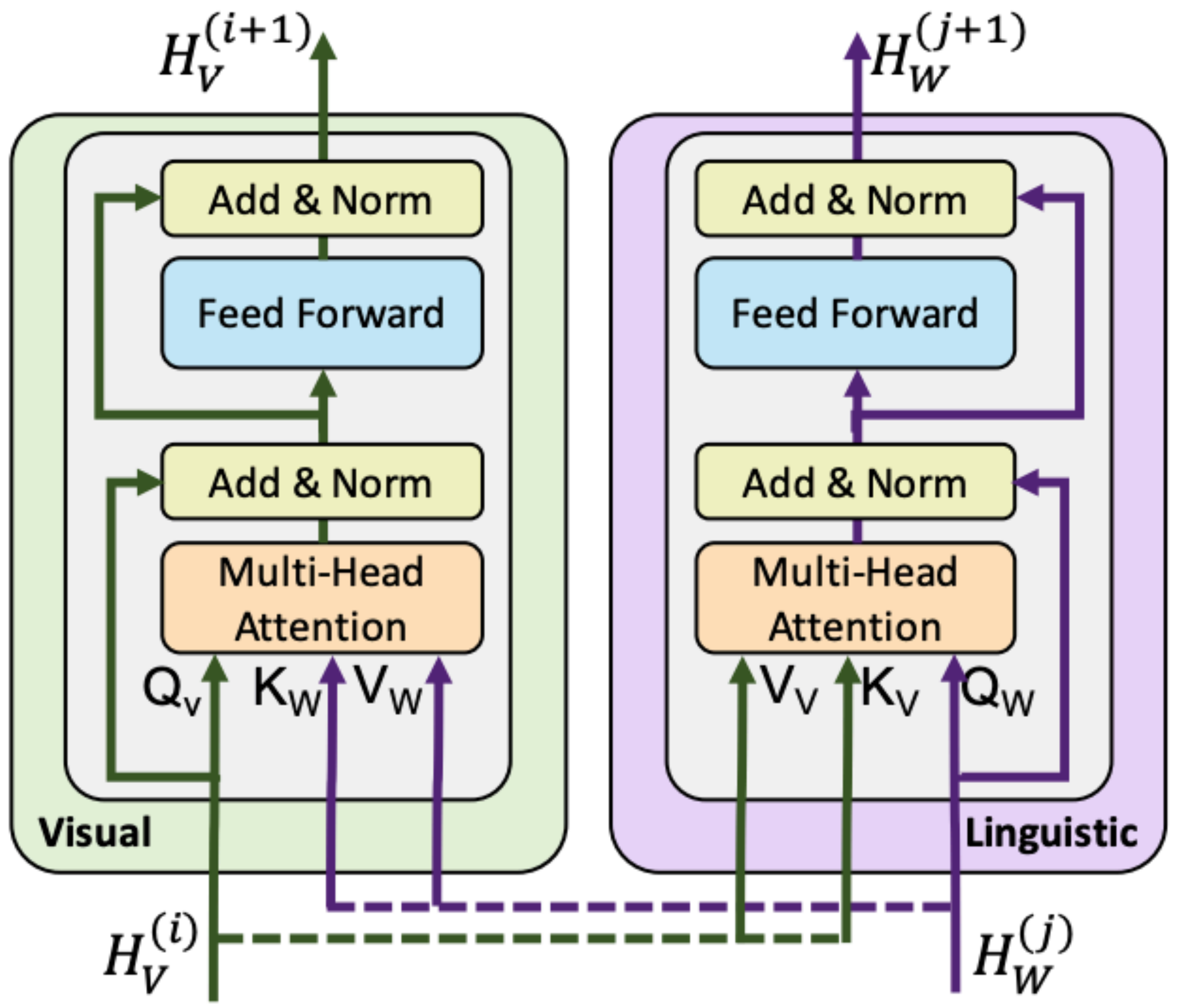
Beyond contrastive loss for multi-modal models

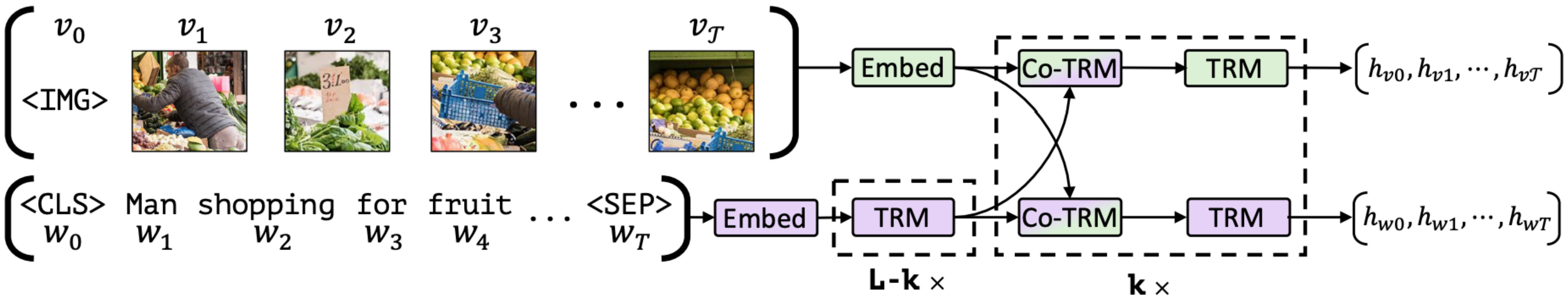

Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
Multi-lingual
MultiFiT
Transformer
BERT
Bidirectional LM
GPT
Larger model
More data
GPT-2
Defense
Grover
Cross-lingual
XLM
UDify
Multi-task
MT-DNN
Knowledge distillation
MT-DNN$_{KD}$
+ Generation
MASS
UniLM
Span prediction
Remove NSP
SpanBERT
Longer time
Remove NSP
More data
RoBERTa
Permutation LM
Transformer-XL
More data
XLNet
+Knowledge Graph
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
Cross-modal
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
Whole Word Masking
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

Pretrained representations for vision and language

Image represented as

- series of **image region features** (extracted from pre-trained object detection network)
- **Region position** encoded as $5d$ vector






ViLBERT

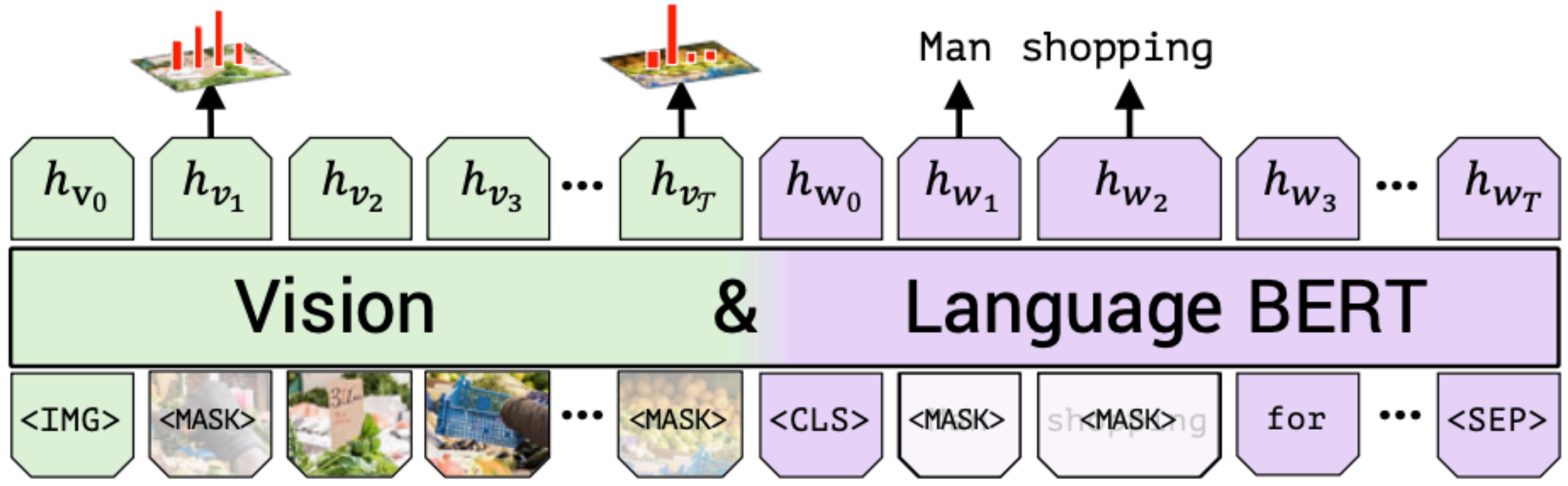
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

Pretrained representations for vision and language

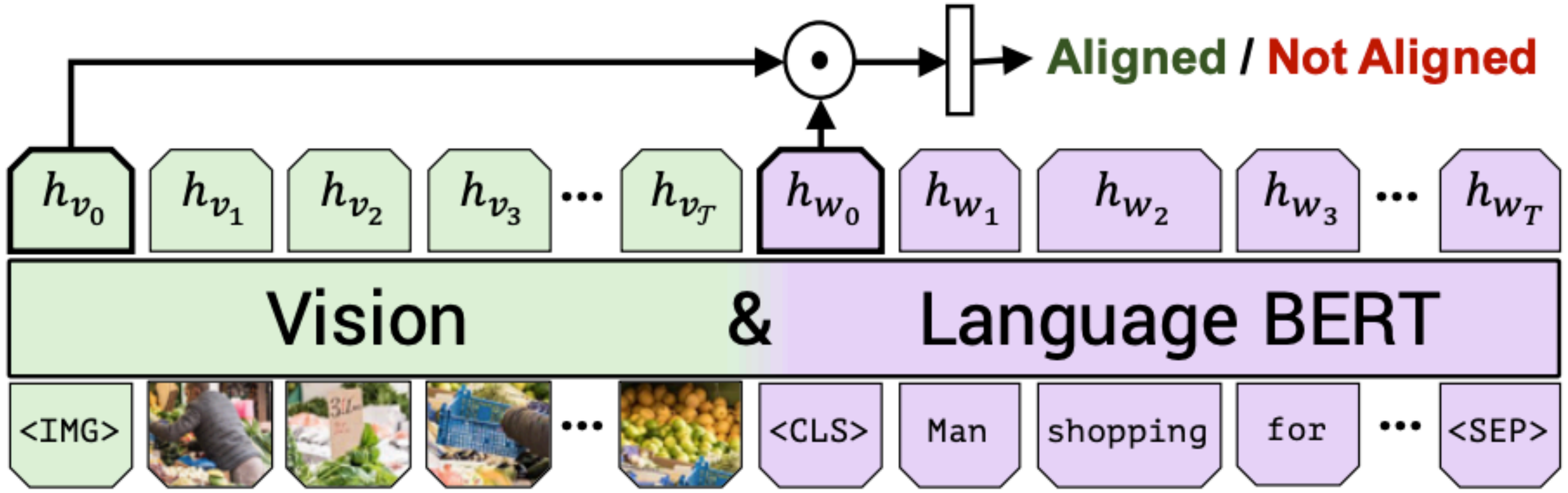
Predict semantic class distribution

Trained on

- Conceptual captions (~3.3M images with captions cleaned from alt-text labels)
- Two tasks to predict:
  - masked out words and semantic class distribution for masked out image regions
  - Is the image/description aligned?



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

ViLBERT

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

Pretrained representations for vision and language

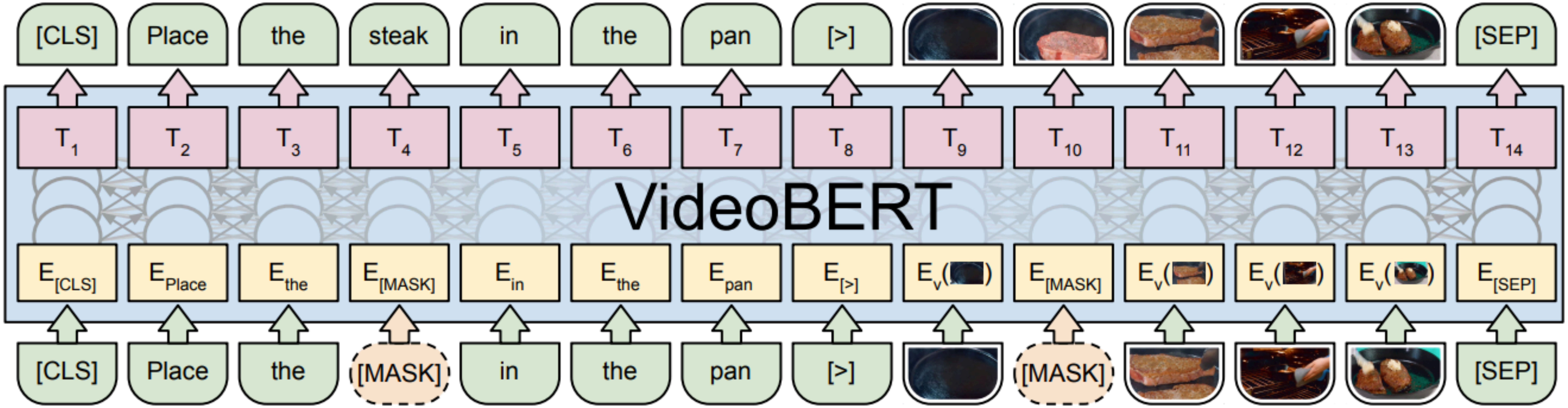
		VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	Method	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-	-
Ours	Single-Stream†	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
	ViLBERT†	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
	ViLBERT	**70.55 (70.92)**	**72.42 (73.3)**	**74.47 (74.6)**	**54.04 (54.8)**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Pretraining improves performance on variety of vision+language tasks!

ViLBERT

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, https://arxiv.org/pdf/1908.02265.pdf]

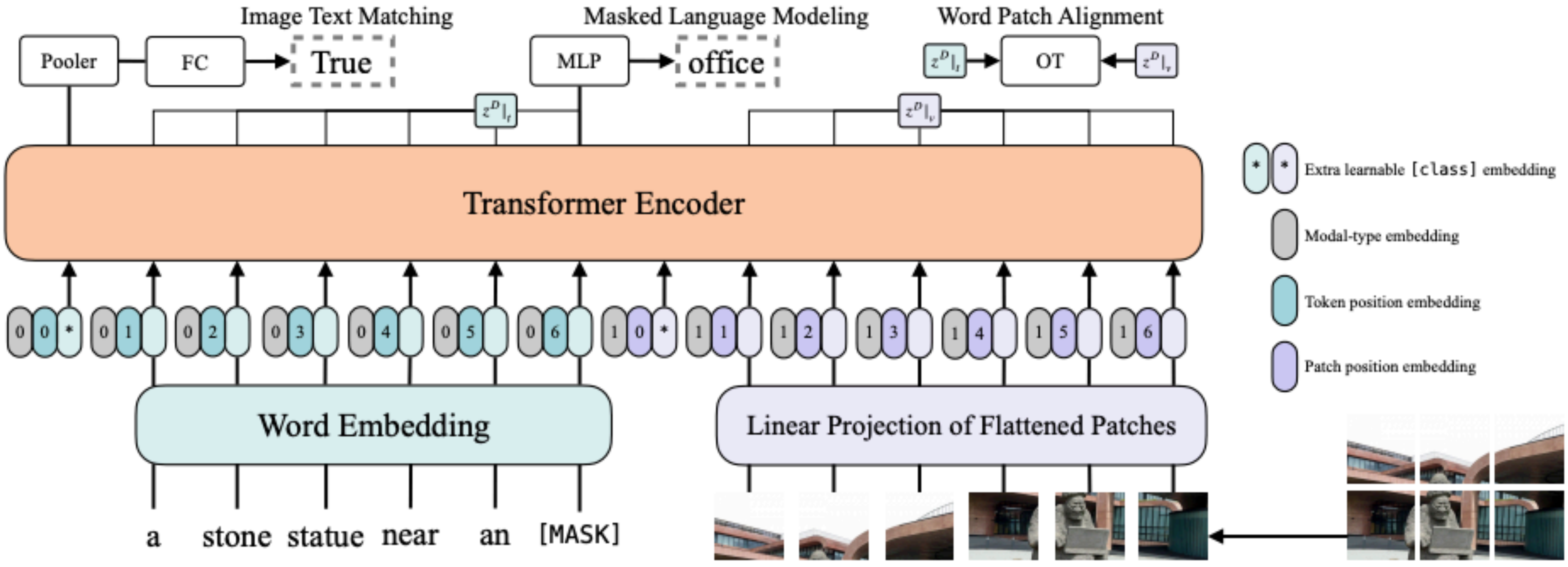
Masked modelling for video and language



VideoBERT: A Joint Model for Video and Language Representation Learning [Sun et al, ICCV 2019]

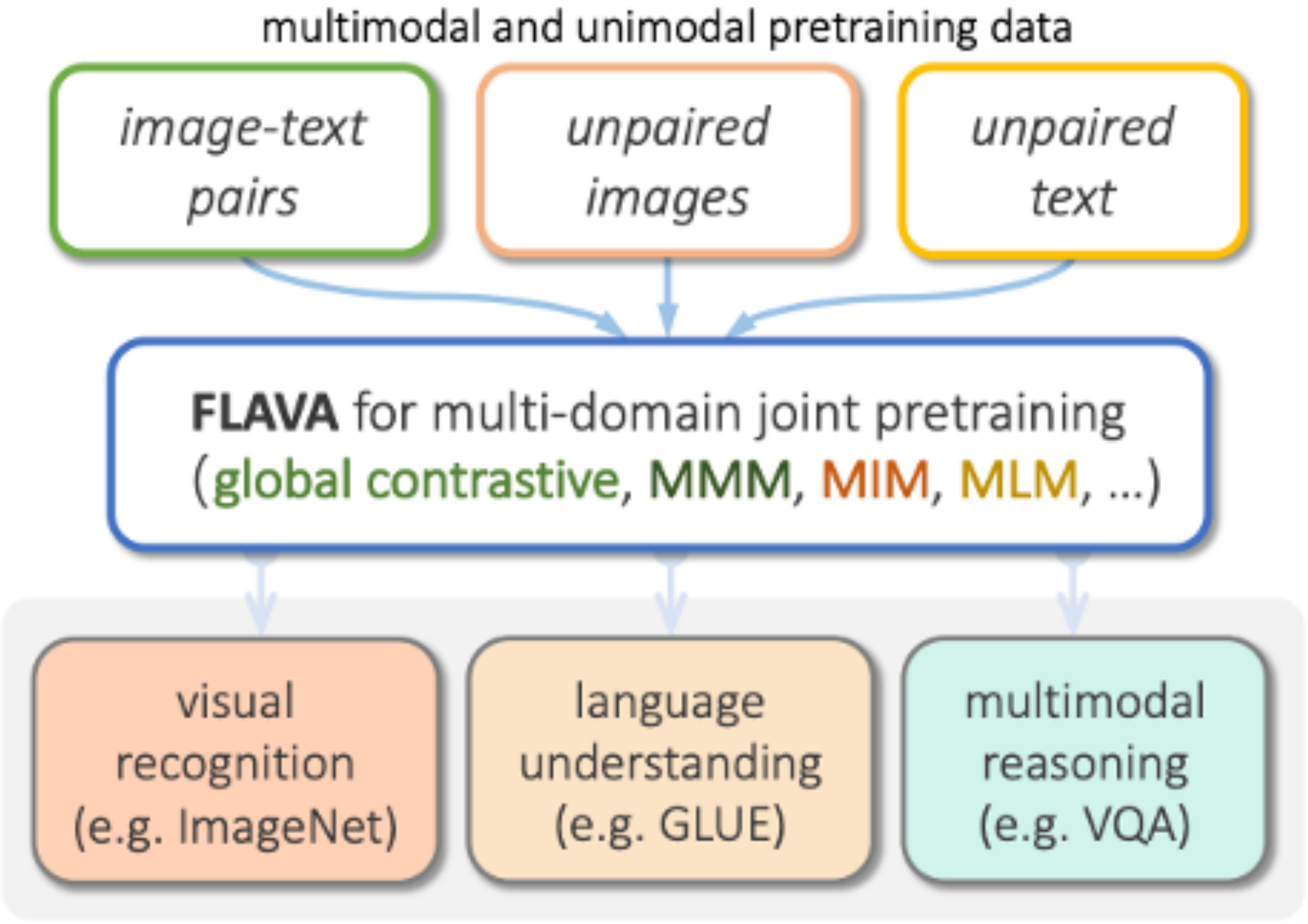
Combining masked modelling with contrastive learning

- Use image patches, no need for object detectors



[ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision](#) [Kim et al, ICML 2021]

Large multi-modal models: FLAVA

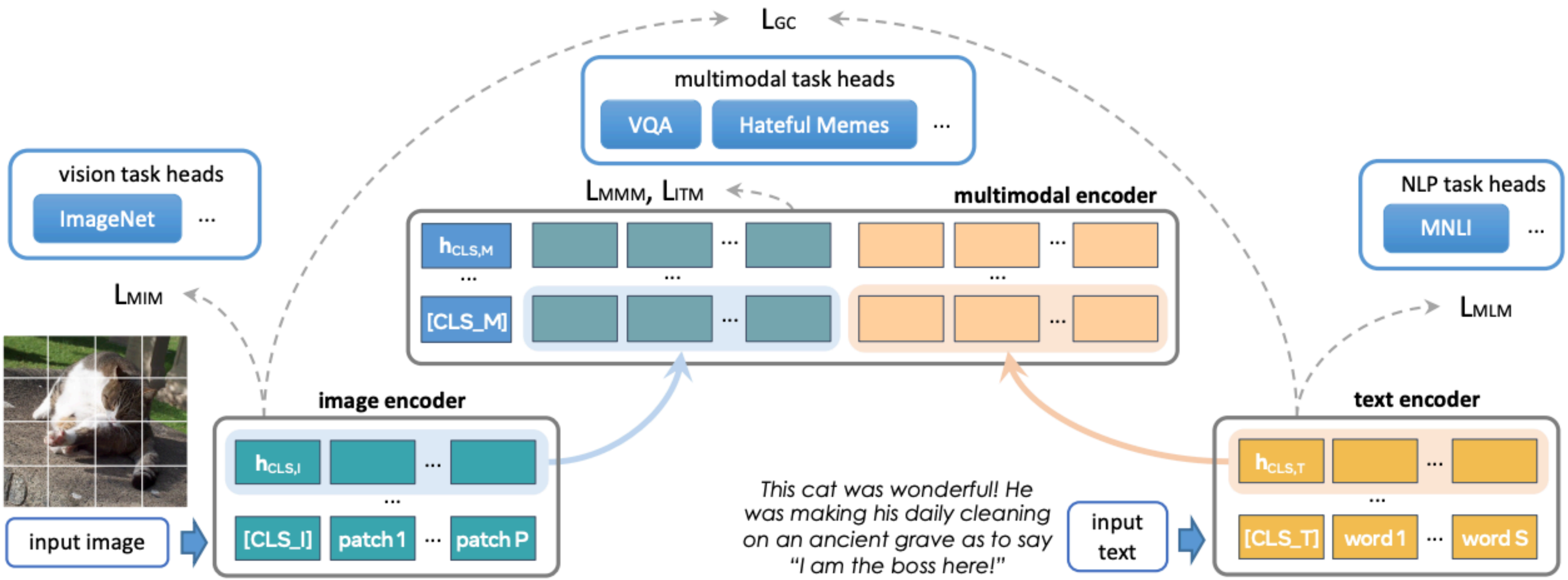


Meta AI

Masked Modeling
- Multimodal
- Image
- Language

FLAVA: A Foundational Language and Vision Alignment Model [Singh et al, CVPR 2022]

Large multi-modal models: FLAVA



FLAVA: A Foundational Language and Vision Alignment Model [Singh et al, CVPR 2022]

Large multi-modal models: FLAVA

- Pretrain unimodal encoders on unpaired image and text data
- Joint unimodal and multi-modal training
- Multi-modal training with paired image-text pairs

	#Image-Text Pairs	Avg. text length
COCO [66]	0.9M	12.4
SBU Captions [77]	1.0M	12.1
Localized Narratives [82]	1.9M	13.8
Conceptual Captions [92]	3.1M	10.3
Visual Genome [57]	5.4M	5.1
Wikipedia Image Text [99]	4.8M	12.8
Conceptual Captions 12M [14]	11.0M	17.3
Red Caps [27]	11.6M	9.5
YFCC100M [103], filtered	30.3M	12.7
Total	70M	12.1

COCO

A close up view of a pizza sitting on a table with a soda in the back.

CC12M

Jumping girl in a green summer dress stock illustration

FLAVA: A Foundational Language and Vision Alignment Model [Singh et al, CVPR 2022]

Large multi-modal models: FLAVA

- Pretrain unimodal encoders on unpaired image and text data
- Joint unimodal and multi-modal training
- Multi-modal training with paired image-text pairs
- Training details
  - Hyperparameters important for pretraining: Large batch size (8K), large weight decay (0.1) with learning rate (1e-3), long warm up (10K) with AdamW
  - Noted again the importance of having the layer-norm before the MHA

FLAVA: A Foundational Language and Vision Alignment Model [Singh et al, CVPR 2022]

Large multi-modal models: FLAVA

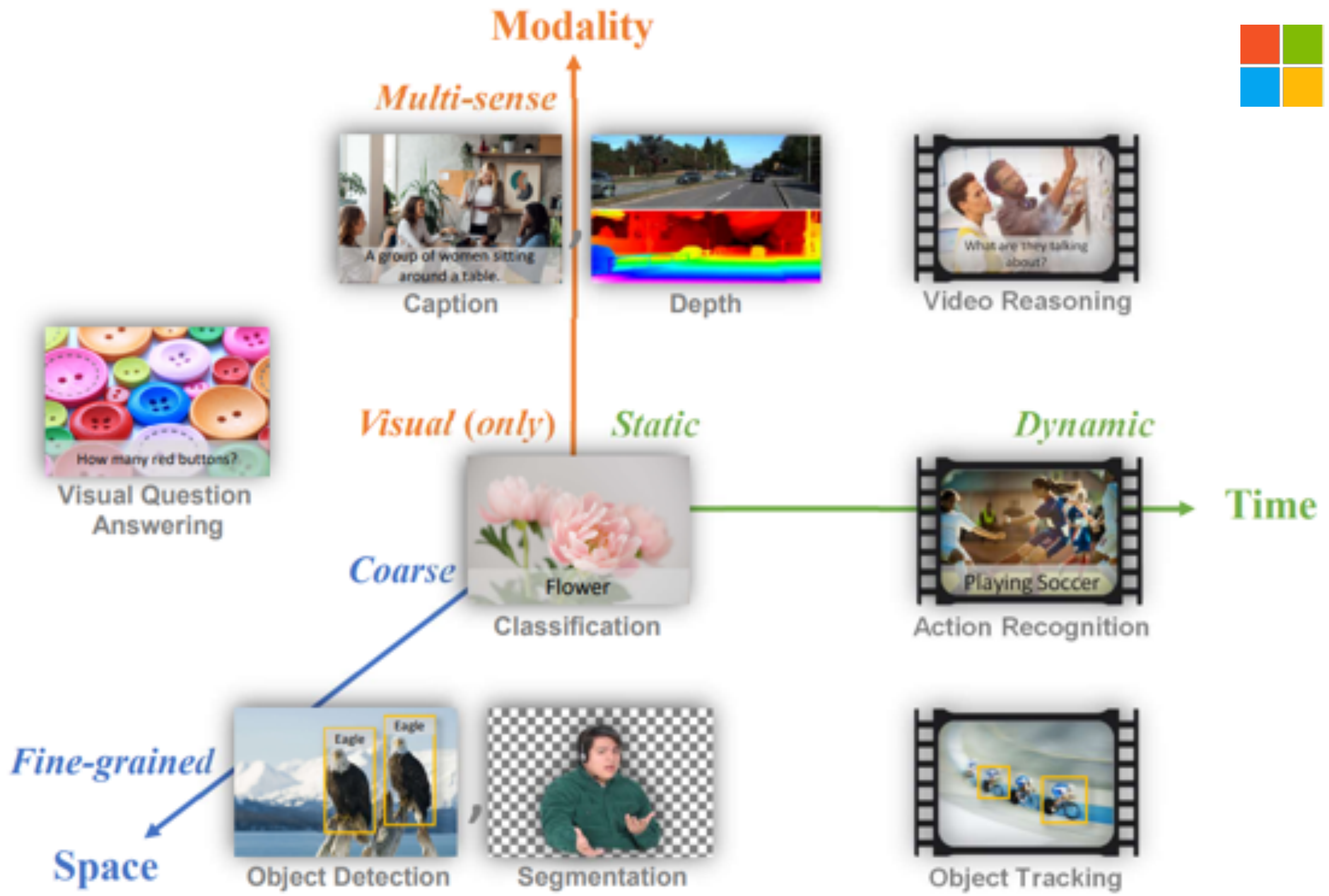
FLAVA model performance on variety of tasks

	public data		Multimodal Tasks			Language Tasks								ImageNet
			VQAv2	SNLI-VE	HM	CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	linear eval
1	✓	$\text{BERT}_{\text{base}}$ [28]	–	–	–	54.6	92.5	62.5	81.9/87.6	90.6/87.4	84.4	91.0	88.1	–
2	✗	CLIP-ViT-B/16 [83]	55.3	74.0	63.4	25.4	88.2	55.2	74.9/65.0	76.8/53.9	33.5	50.5	16.0	80.2
3	✗	$\text{SimVLM}_{\text{base}}$ [109]	77.9	84.2	–	46.7	90.9	63.9	75.2/84.4	90.4/87.2	83.4	88.6	–	80.6
4	✓	VisualBERT [63]	70.8	$77.3^{\dagger}$	$74.1^{\ddagger}$	38.6	89.4	56.6	71.9/82.1	89.4/86.0	**81.6**	87.0	81.8	–
5	✓	$\text{UNITER}_{\text{base}}$ [16]	72.7	78.3	–	37.4	89.7	55.6	69.3/80.3	89.2/85.7	80.9	86.0	75.3	–
6	✓	$\text{VL-BERT}_{\text{base}}$ [101]	71.2	–	–	38.7	89.8	55.7	70.6/81.8	89.0/85.4	81.2	86.3	82.9	–
7	✓	ViLBERT [70]	70.6	$75.7^{\dagger}$	$74.1^{\ddagger}$	36.1	90.4	53.7	69.0/79.4	88.6/85.0	79.9	83.8	77.9	–
8	✓	LXMERT [102]	72.4	–	–	39.0	90.2	57.2	69.7/80.4	75.3/75.3	80.4	84.2	75.3	–
9	✓	UniT [43]	67.0	73.1	–	–	89.3	–	–	90.6/ –	81.5	**88.0**	–	–
10	✓	CLIP-ViT-B/16 (PMD)	59.8	73.5	56.6	11.0	83.5	53.1	63.5/68.7	75.4/43.0	32.9	49.5	13.7	73.0
11	✓	FLAVA (ours)	**72.8**	**79.0**	**76.7**	**50.7**	**90.9**	**57.8**	**81.4/86.9**	**90.4/87.2**	80.3	87.3	**85.7**	**75.5**

FLAVA: A Foundational Language and Vision Alignment Model [Singh et al, CVPR 2022]

Large multi-modal, multi-lingual models: Florence

Microsoft



Florence: A New Foundation Model for Computer Vision [Yuan et al, CVPR 2022]

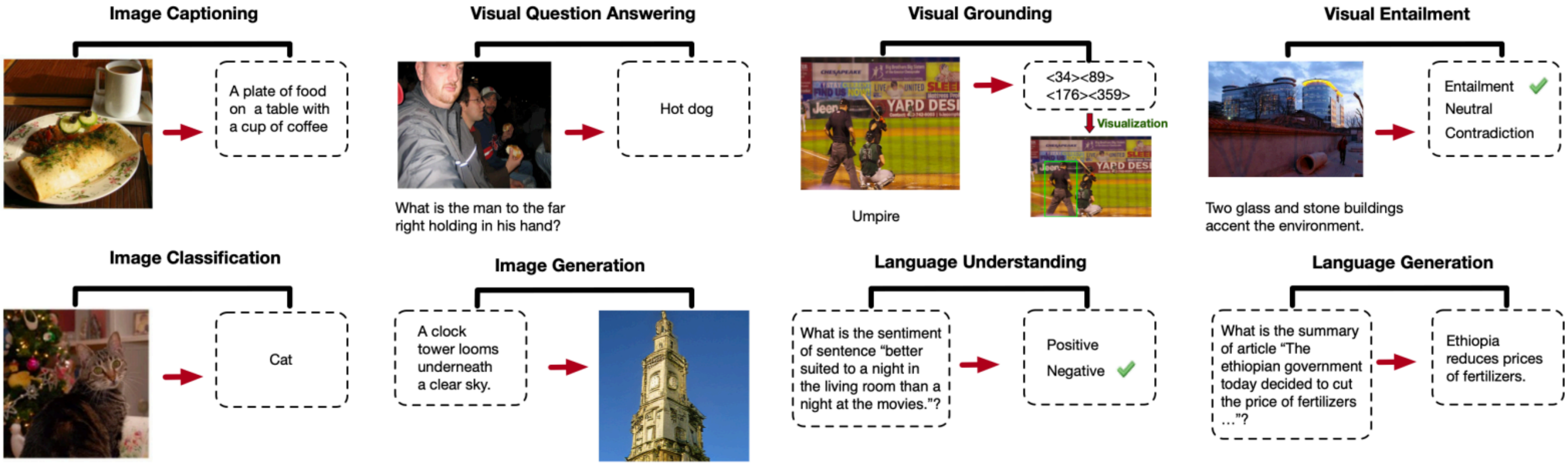
Large multi-modal models: OFA

Alibaba

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

OFA



OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

- Unified framework using transformers
  - Encoder-decoder architecture
- Treat all tasks as sequence-to-sequence
- Represent text, image patches, and objects as token sequences
  - Use BPE for text tokens
  - Use ResNet to obtain image patch features coded as tokens
  - Objects are represented as image region bounding box with label and encoded as location tokens (x1,y1,x2,y2) and BPE token (label)
- Pretrain on mix of vision, language, vision+language data

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

- Pretrain on mix of vision, language, vision+language data

Type	Pretraining Task	Source	#Image	#Label
Vision&Language	Image Captioning Image-Text Matching	CC12M, CC3M, SBU, COCO, VG-Cap	14.78M	15.25M
	Visual Question Answering	VQAv2, VG-QA, GQA	178K	2.92M
	Visual Grounding Grounded Captioning	RefCOCO, RefCOCO+, RefCOCOg, VG-Cap	131K	3.20M
Vision	Detection	OpenImages, Object365, VG, COCO	2.98M	3.00M
	Image Infilling	OpenImages, YFCC100M, ImageNet-21K	36.27M	-
Language	Masked Language Modeling	Pile (Filter)	-	140G*

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

- Instructions for task

Task	Dataset	Instruction	Target
Image Captioning	COCO	[**Image**] What does the image describe?	{**Caption**}
Visual Question Answering	VQA	[**Image**] {**Question**}	{**Answer**}
Visual Entailment	SNLI-VE	[**Image**] Can image and text1 "{**Text1**}" imply text2 "{**Text2**}"?	Yes/No/Maybe
Referring Expression Comprehension	RefCOCO, RefCOCO+, RefCOCOg	[**Image**] Which region does the text "{**Text**}" describe?	{**Location**}
Image Generation	COCO	What is the complete image? caption: {**Caption**}	{**Image**}
Image Classification	ImageNet-1K	[**Image**] What does the image describe?	{**Label**}
Single-Sentence Classification	SST-2	Is the sentiment of text "{**Text**}" positive or negative?	Positive/Negative
Sentence-Pair Classification	RTE	Can text1 "{**Text1**}" imply text2 "{**Text2**}"?	Yes/No
	MRPC	Does text1 "{**Text1**}" and text2 "{**Text2**}" have the same semantics?	Yes/No
	QQP	Is question "{**Question1**}" and question "{**Question2**}" equivalent?	Yes/No
	MNLI	Can text1 "{**Text1**}" imply text2 "{**Text2**}"?	Yes/No/Maybe
	QNLI	Does "{**Text**}" contain the answer to question "{**Question**}"?	Yes/No
Text Summarization	Gigaword	What is the summary of article "{**Article**}"?	{**Summary**}

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

- OFA model sizes

Model	#Param.	Backbone	Hidden size	Intermediate Size	#Head	#Enc. Layers	#Dec. Layers
OFA_{Tiny}	33M	ResNet50	256	1024	4	4	4
$\text{OFA}_{\text{Medium}}$	93M	ResNet101	512	2048	8	4	4
OFA_{Base}	182M	ResNet101	768	3072	12	6	6
$\text{OFA}_{\text{Large}}$	472M	ResNet152	1024	4096	16	12	12
OFA_{Huge}	930M	ResNet152	1280	5120	16	24	12

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework
[Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

OFA model performance on variety of tasks

Model	VQA		SNLI-VE	
	test-dev	test-std	dev	test
UNITER [14]	73.8	74.0	79.4	79.4
OSCAR [15]	73.6	73.8	-	-
VILLA [16]	74.7	74.9	80.2	80.0
VL-T5 [56]	-	70.3	-	-
VinVL [17]	76.5	76.6	-	-
UNIMO [46]	75.0	75.3	81.1	80.6
ALBEF [69]	75.8	76.0	80.8	80.9
METER [70]	77.7	77.6	80.9	81.2
VLMo [48]	79.9	80.0	-	-
SimVLM [22]	80.0	80.3	86.2	86.3
Florence [23]	80.2	80.4	-	-
OFA_{Tiny}	70.3	70.4	85.3	85.2
$\text{OFA}_{\text{Medium}}$	75.4	75.5	86.6	87.0
OFA_{Base}	78.0	78.1	89.3	89.2
$\text{OFA}_{\text{Large}}$	80.3	80.5	90.3	90.2
OFA	**82.0**	**82.0**	**91.0**	**91.2**

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val-u	test-u
VL-T5 [56]	-	-	-	-	-	-	-	71.3
UNITER [14]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA [16]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
MDETR [72]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
UNICORN [57]	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
OFA_{Tiny}	80.20	84.07	75.00	68.22	75.13	57.66	72.02	69.74
$\text{OFA}_{\text{Medium}}$	85.34	87.68	77.92	76.09	83.04	66.25	78.76	78.58
OFA_{Base}	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31
$\text{OFA}_{\text{Large}}$	90.05	92.93	85.26	85.80	89.87	79.22	85.89	86.55
OFA	**92.04**	**94.03**	**88.44**	**87.86**	**91.70**	**80.71**	**88.07**	**88.78**

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Large multi-modal models: OFA

Image captioning

Model	Cross-Entropy Optimization				CIDEr Optimization			
	BLEU@4	METEOR	CIDEr	SPICE	BLEU@4	METEOR	CIDEr	SPICE
VL-T5 [56]	34.5	28.7	116.5	21.9	-	-	-	-
OSCAR [15]	37.4	30.7	127.8	23.5	41.7	30.6	140.0	24.5
UNICORN [57]	35.8	28.4	119.1	21.5	-	-	-	-
VinVL [17]	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
UNIMO [46]	39.6	-	127.7	-	-	-	-	-
LEMON [71]	41.5	30.8	139.1	24.1	42.6	31.4	145.5	25.5
SimVLM [22]	40.6	**33.7**	143.3	**25.4**	-	-	-	-
OFA_{Tiny}	35.9	28.1	119.0	21.6	38.1	29.2	128.7	23.1
OFA_{Medium}	39.1	30.0	130.4	23.2	41.4	30.8	140.7	24.8
OFA_{Base}	41.0	30.9	138.2	24.2	42.8	31.7	146.7	25.8
OFA_{Large}	42.4	31.5	142.2	24.5	43.6	32.2	150.7	26.2
OFA	**43.9**	31.8	**145.3**	24.8	**44.9**	**32.5**	**154.9**	**26.6**

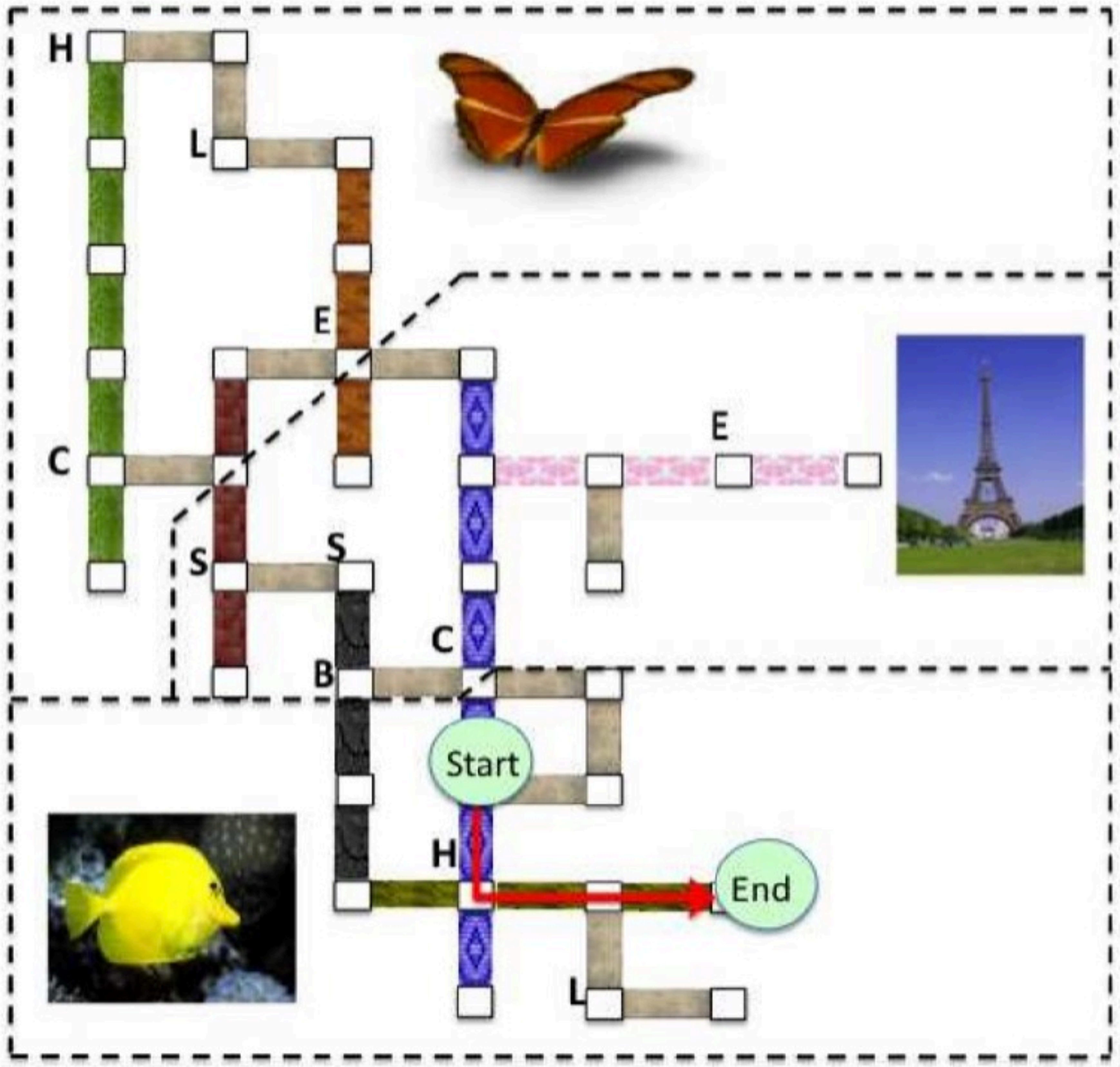
Text to Image Generation

Model	FID↓	CLIPSIM↑	IS↑
DALLE [50]	27.5	-	17.9
CogView [51]	27.1	33.3	18.2
GLIDE [77]	12.2	-	-
Unifying [78]	29.9	30.9	-
NÜWA [52]	12.9	34.3	27.2
OFA	**10.5**	**34.4**	**31.1**

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] https://arxiv.org/abs/2202.03052

Instruction following

Instruction Following

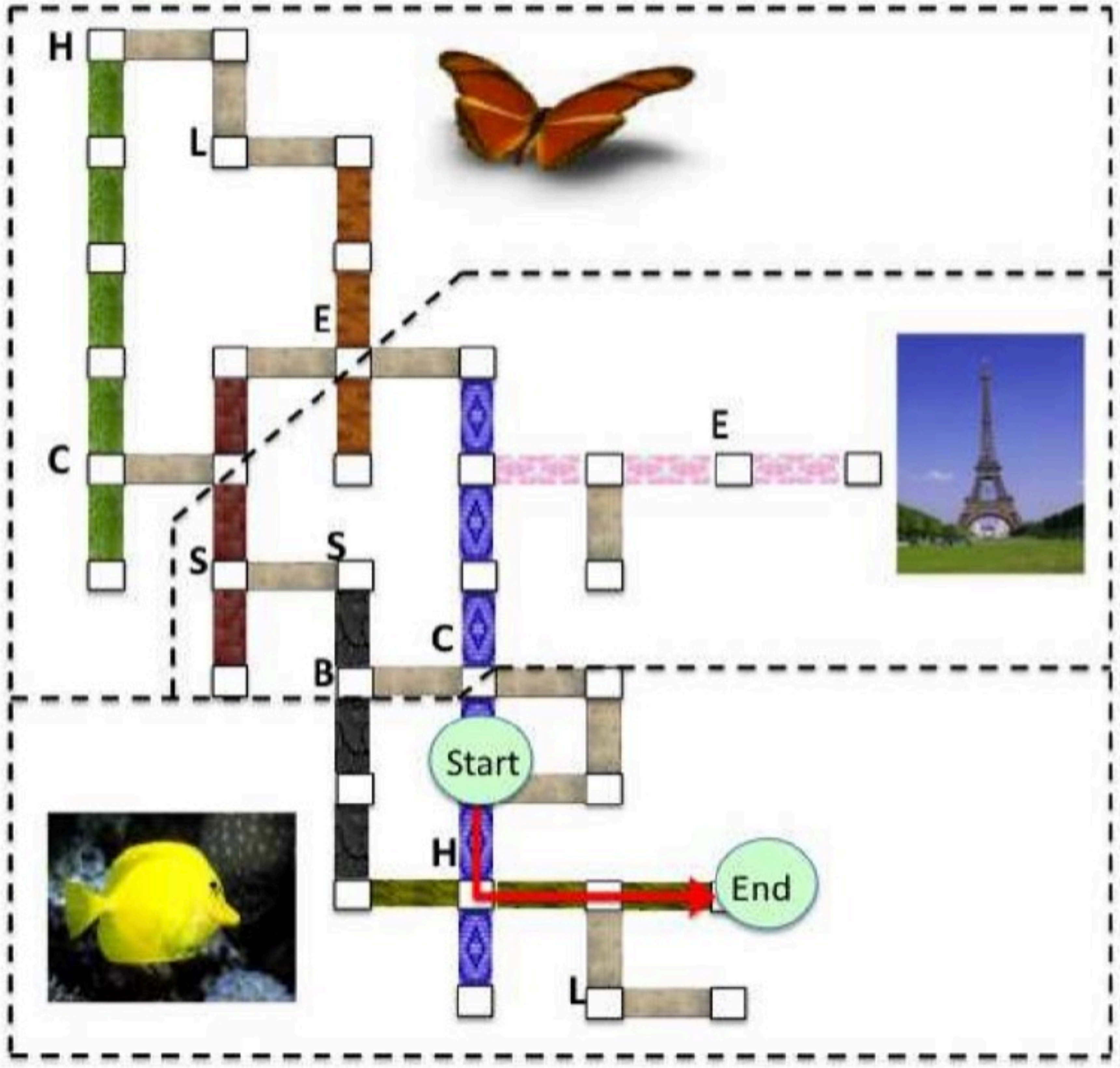


- Want to be able to follow instructions in a virtual environment
- "Go along the blue hall, then turn left away from the fish painting and walk to the end of the hallway"

(MacMahon et al., 2006)

(slide adapted from Greg Durrett)

Instruction Following



Instruction: "Go away from the lamp to the intersection of the red brick and wood"

Basic: Turn ( ) ,
Travel (steps: 1)

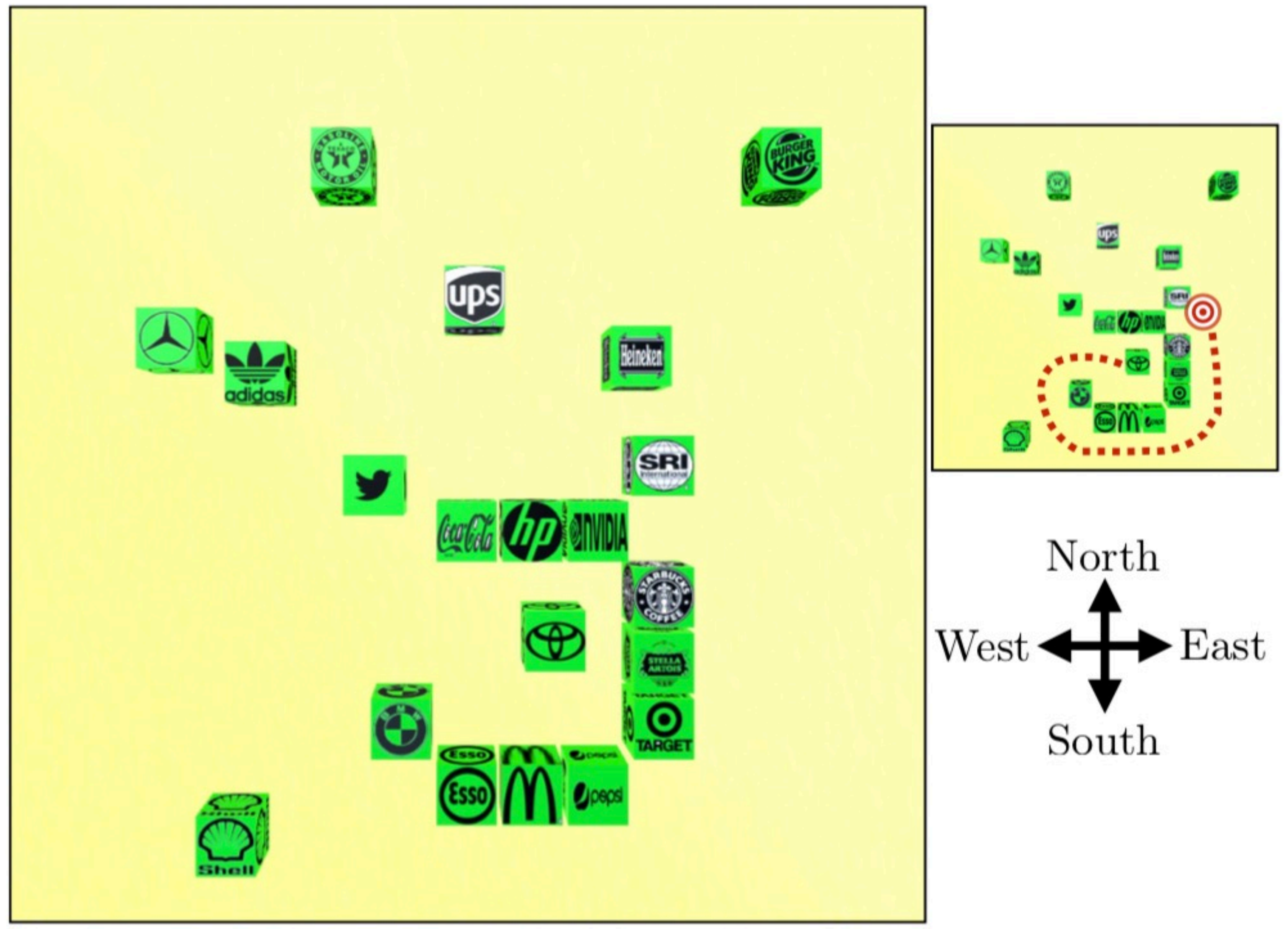
Landmarks: Turn ( ) ,
Verify (left: WALL , back: LAMP , back: HATRACK , front: BRICK HALL) ,
Travel (steps: 1) ,
Verify (side: WOOD HALL)

- Train semantic parser on (utterance, action) pairs
- Language is grounded in actions in the world

(Chen and Mooney, 2011)

(slide adapted from Greg Durrett)

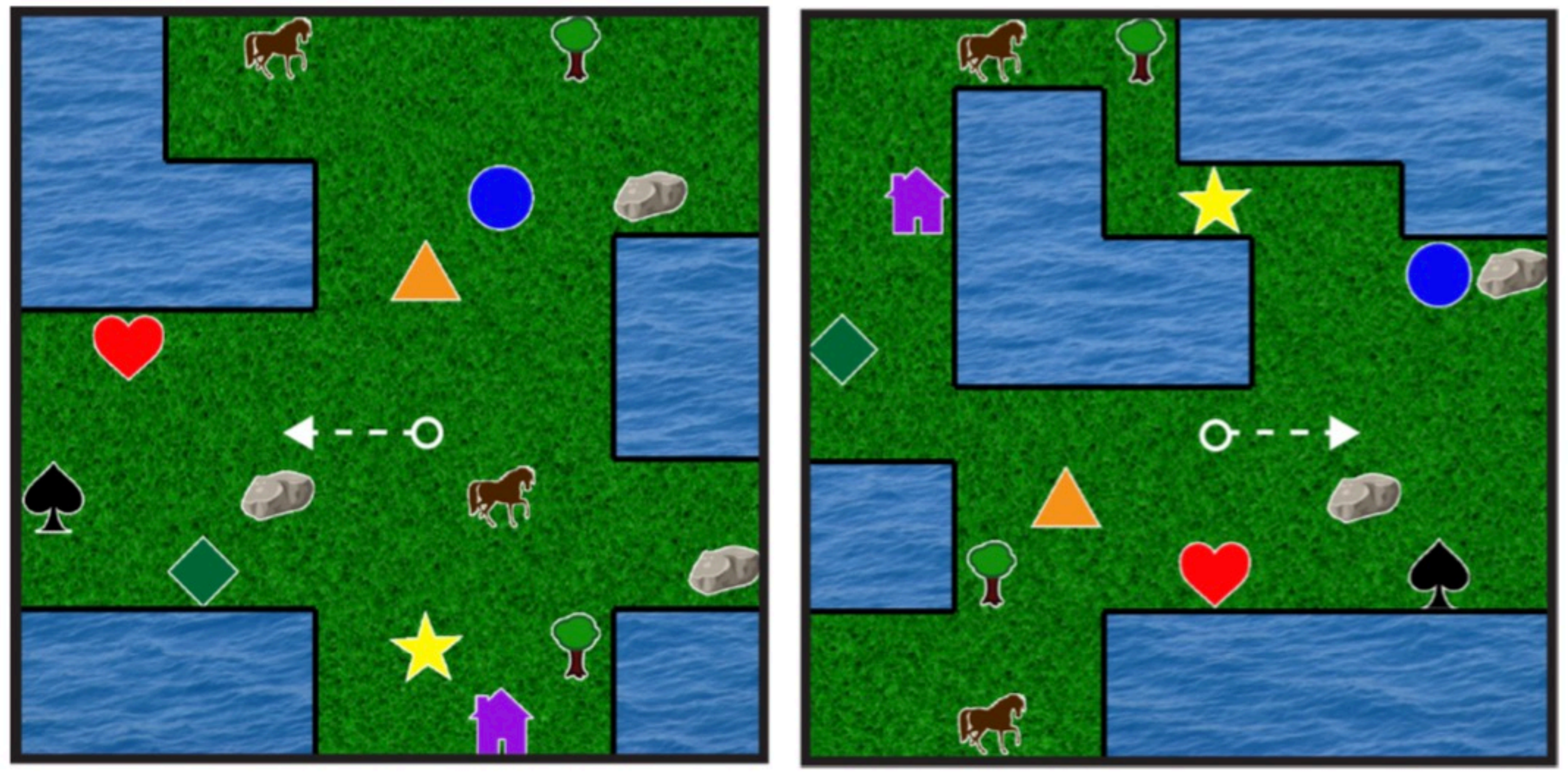
Spatial Reasoning



Put the Toyota block in the same row as the SRI block, in the first open space to the right of the SRI block
Move Toyota to the immediate right of SRI, evenly aligned and slightly separated
Move the Toyota block around the pile and place it just to the right of the SRI block
Place Toyota block just to the right of The SRI Block
Toyota, right side of SRI

Robotic Manipulation

(Bisk et al., 2016, Misra et al., 2017)



Reach the cell above the westernmost rock

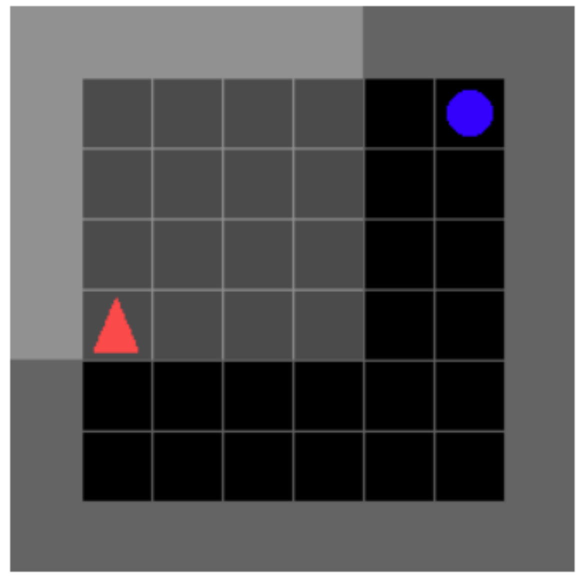
Autonomous navigation

(Janner et al., 2017)

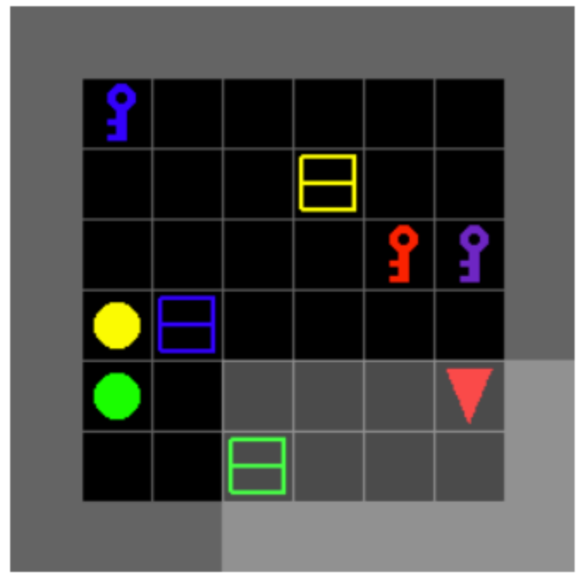
Frameworks for understanding grounded language (with perception and actions)

BabyAI

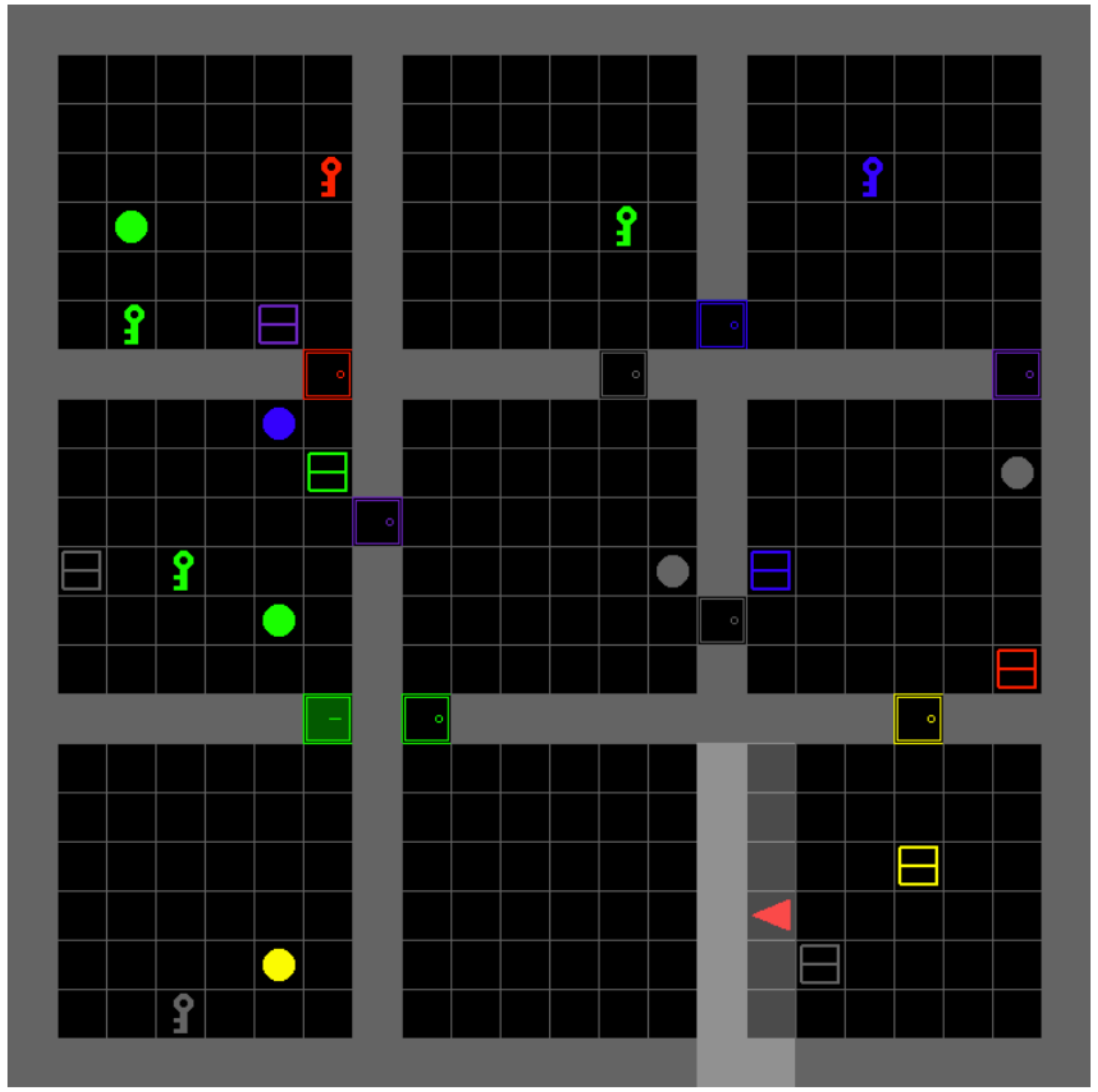
- Grid Environment
- Generated (synthetic language) using grammar
- Easy to hard levels
- Studies grounding and compositionality



(a) GoToObj: "go to the blue ball"



(b) PutNextLocal: "put the blue key next to the green ball"

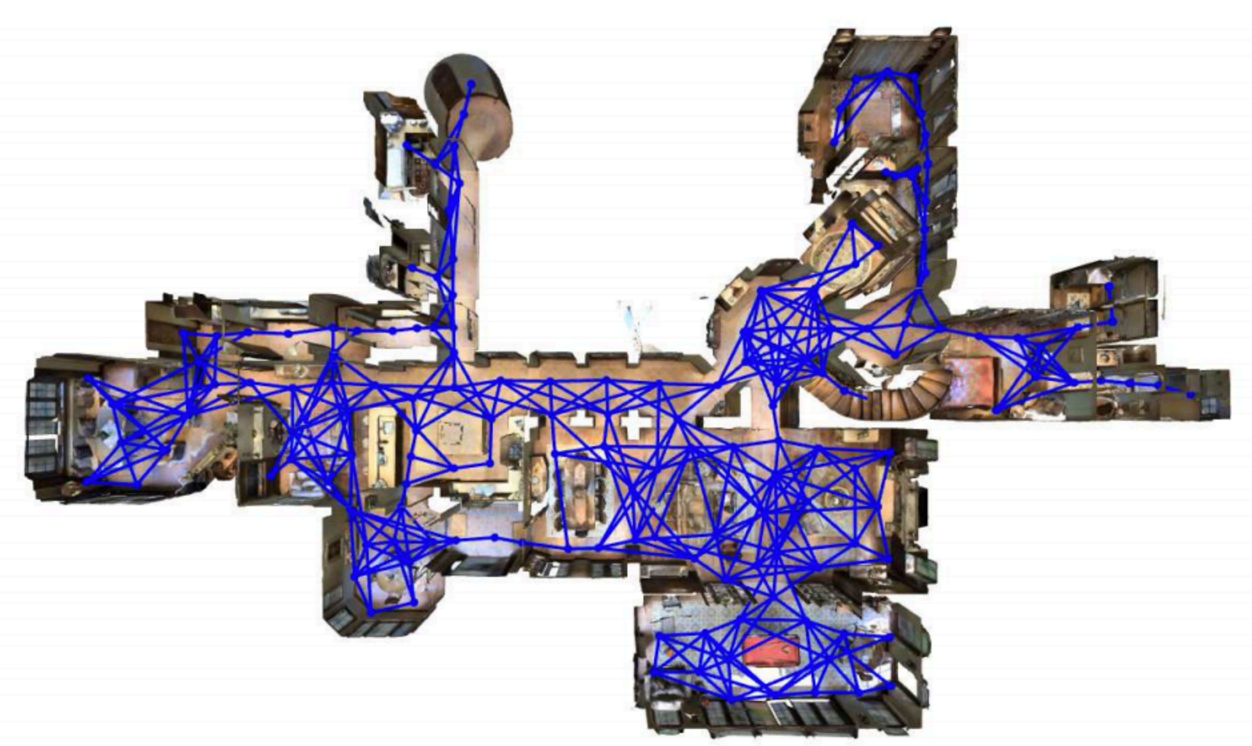


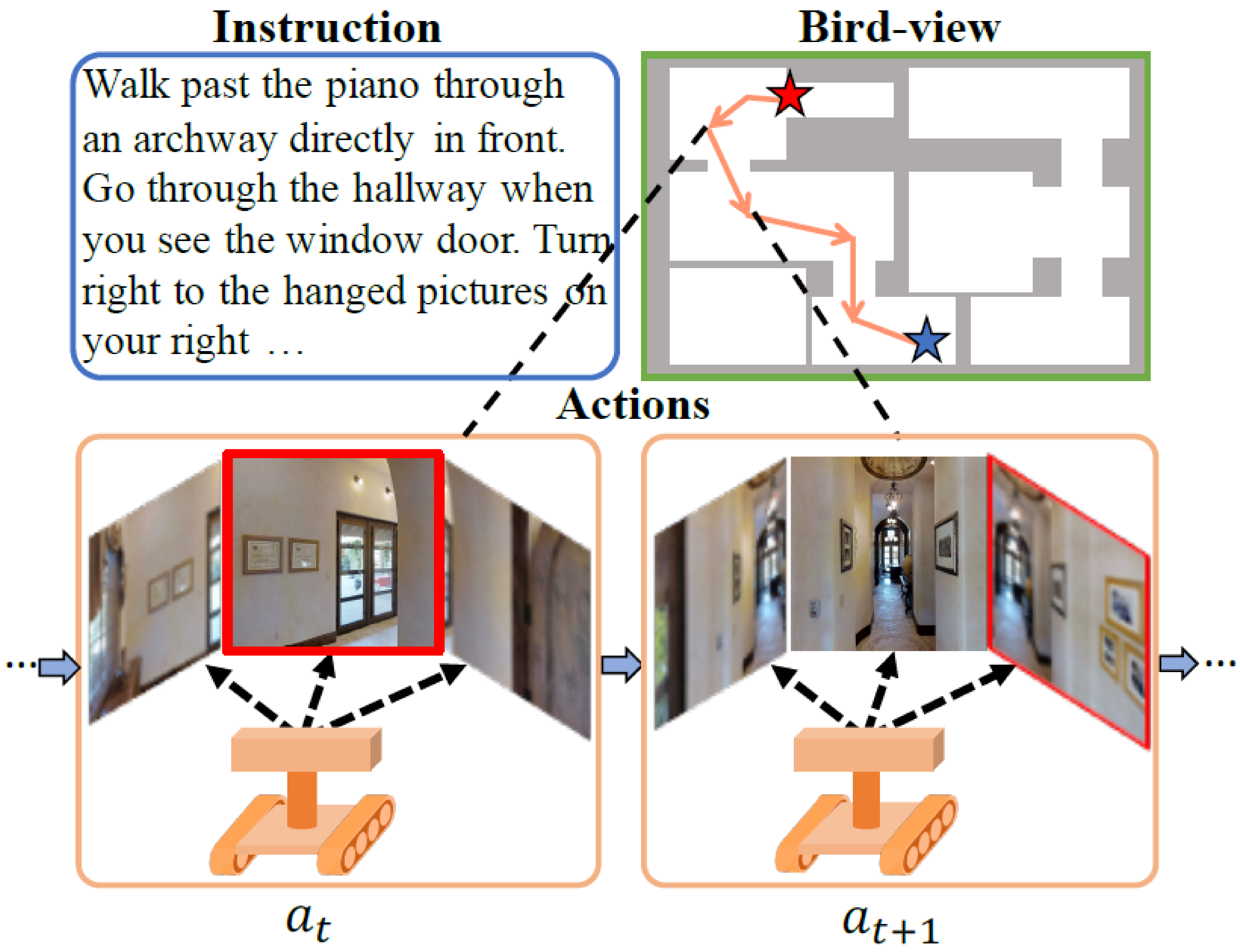
(c) BossLevel: "pick up the grey box behind you, then go to the grey key and open a door". Note that the green door near the bottom left needs to be unlocked with a green key, but this is not explicitly stated in the instruction.

BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning
[Chevalier-Boisvert et al 2018, https://arxiv.org/pdf/1810.08272.pdf]

Vision-and-language Navigation

- More realistic houses
- Human instructions navigation
- Discrete action space
- Navigation graph

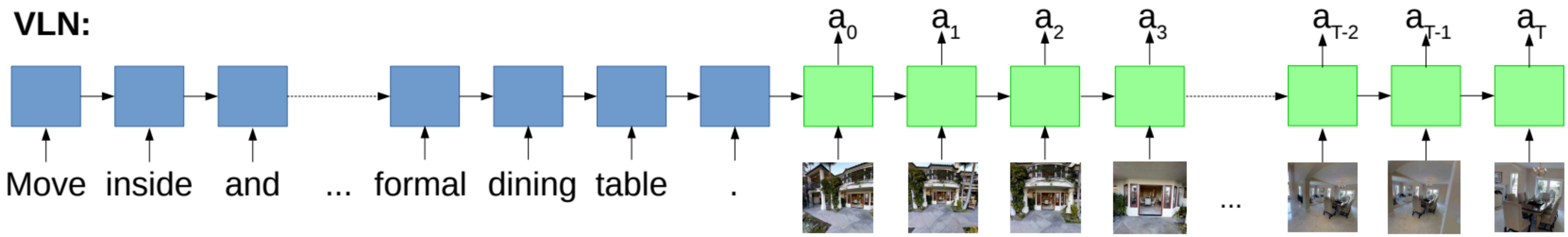




Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments [Anderson et al 2018, https://bringmeaspoon.org/]

Vision-and-language Navigation

- Sequence of words to sequence of actions!



Input Images at each time step

Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments
[Anderson et al 2018, https://bringmeaspoon.org/]

Vision-and-language Navigation

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

ALFRED

- More realistic houses
- Sequence of human instructions for common household tasks
- Study embodied language understanding

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks
[Shridhar et al 2019, https://askforalfred.com/]

ALFRED

A Benchmark for Interpreting
Grounded Instructions for Everyday Tasks

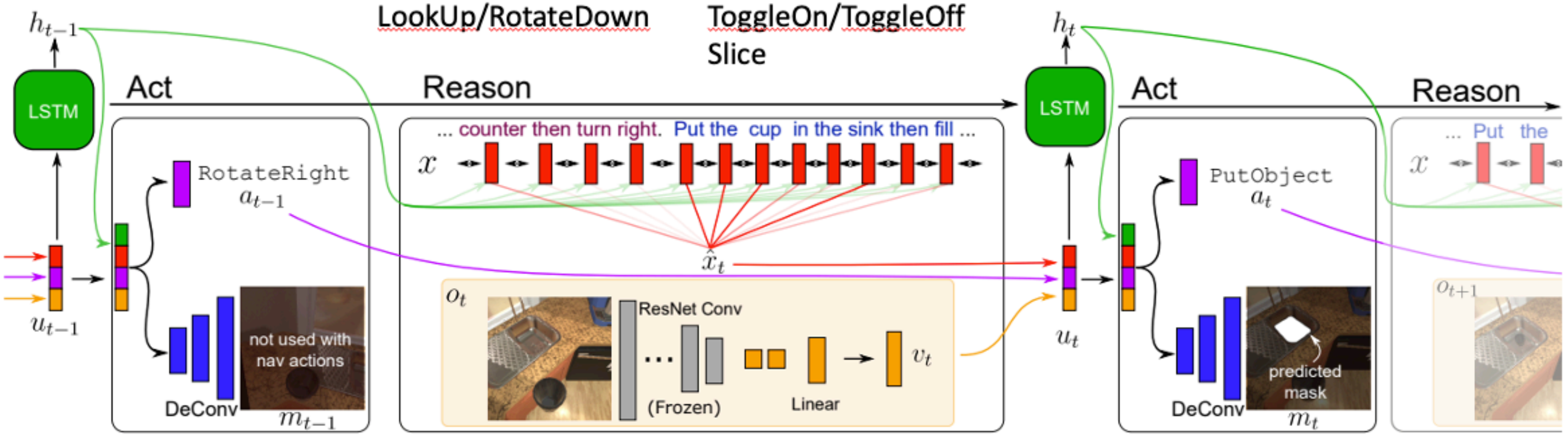
ALFRED agent model

- Seq2seq model (CNN vision, LSTM language)
- Predicts action + binary mask of object from concatenated input
  - 13 actions (5 navigation + 7 interaction + stop)

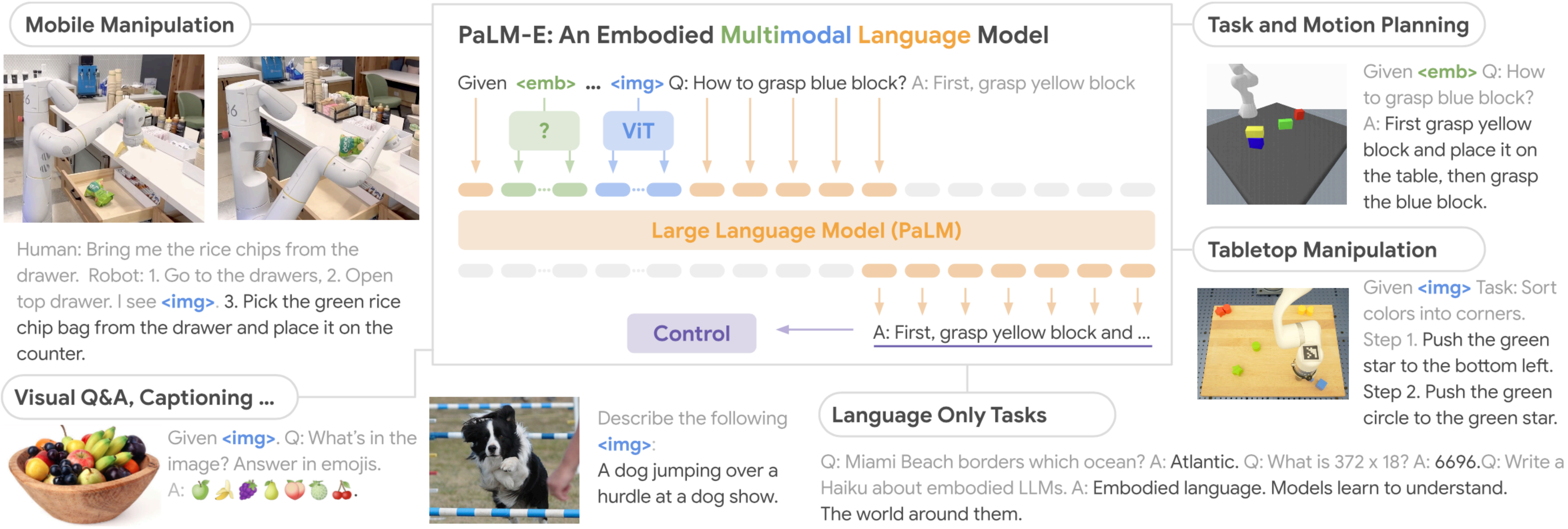
Concatenated input
Vision
Language
Last action

Navigation
MoveAhead
RotateLeft/RotateRight
LookUp/RotateDown

Interaction
Pickup/Put
Open/Close
ToggleOn/ToggleOff
Slice



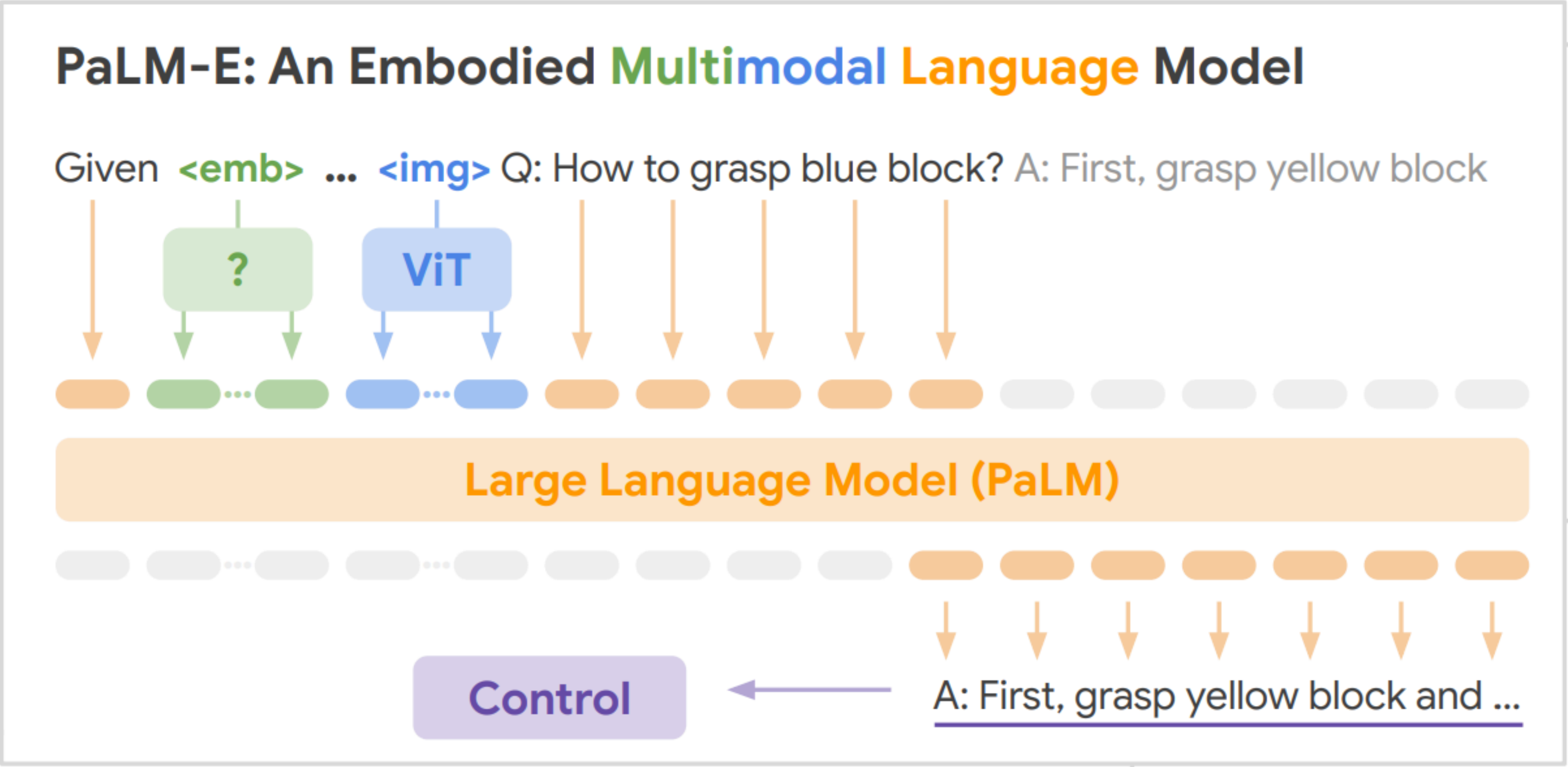
Toward multimodal agents



PaLM-E: An Embodied Multimodal Language Model [Dreiss et al, Google, 2023]
https://palm-e.github.io/

PaLM-E

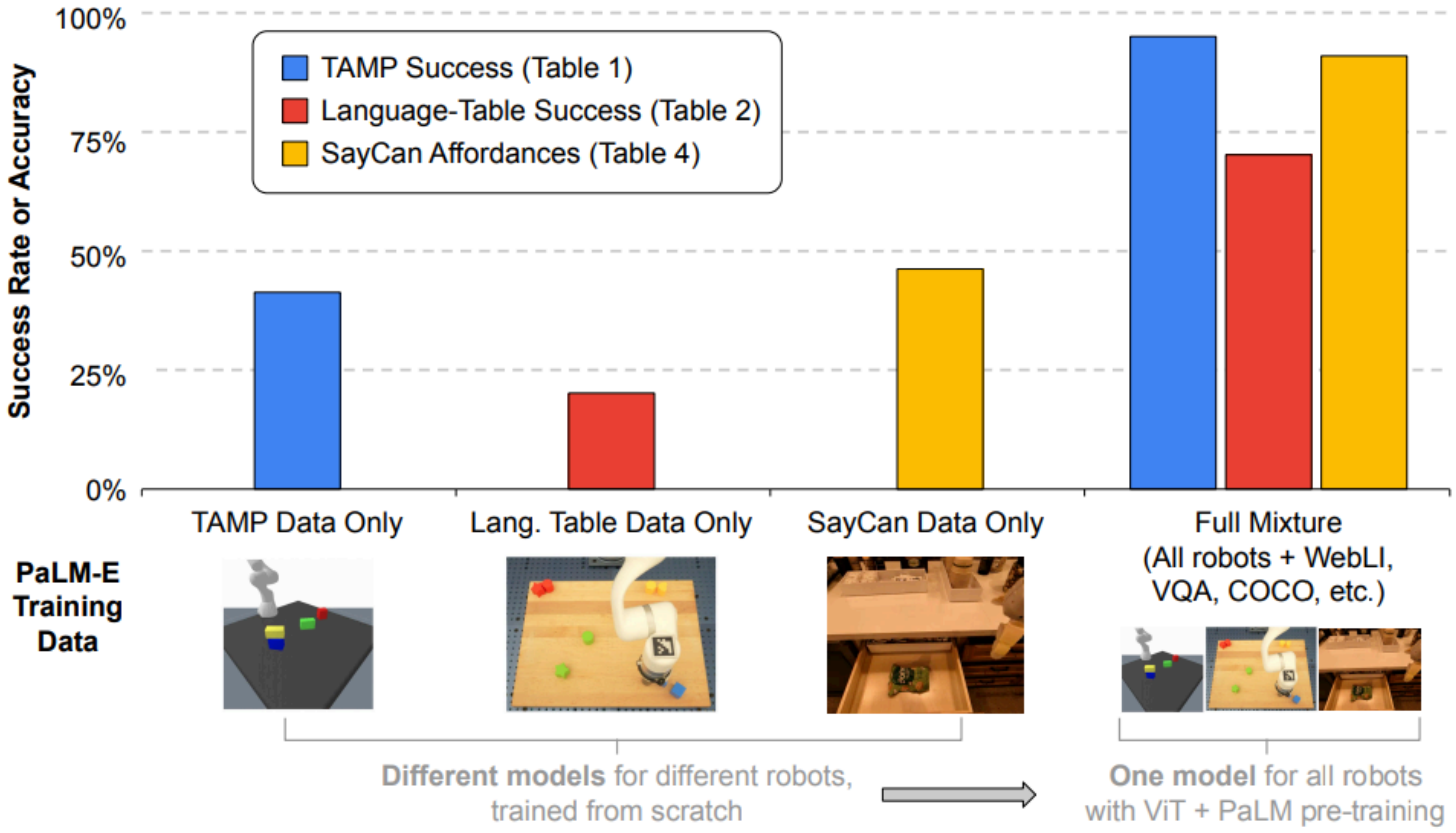
- ViT (22B parameters) + PaLM (562B parameters)
- Decoder only LLM
- Multimodal information injected as continuous vectors into PaLM



PaLM-E: An Embodied Multimodal Language Model [Dreiss et al, Google, 2023]
https://palm-e.github.io/

PaLM-E

- Train on mixture of data



PaLM-E: An Embodied Multimodal Language Model [Dreiss et al, Google, 2023]
https://palm-e.github.io/

PaLM-E

CMPT 839 / CMPT 983

Advanced NLP / Grounded Natural Language Understanding