



CMPT 413/713: Natural Language Processing

Grounded Natural Language and Multimodal Models

Spring 2026
2026-03-30

What is grounding?

Language is used to communicate about **the world**

- Things, actions, abstract concepts



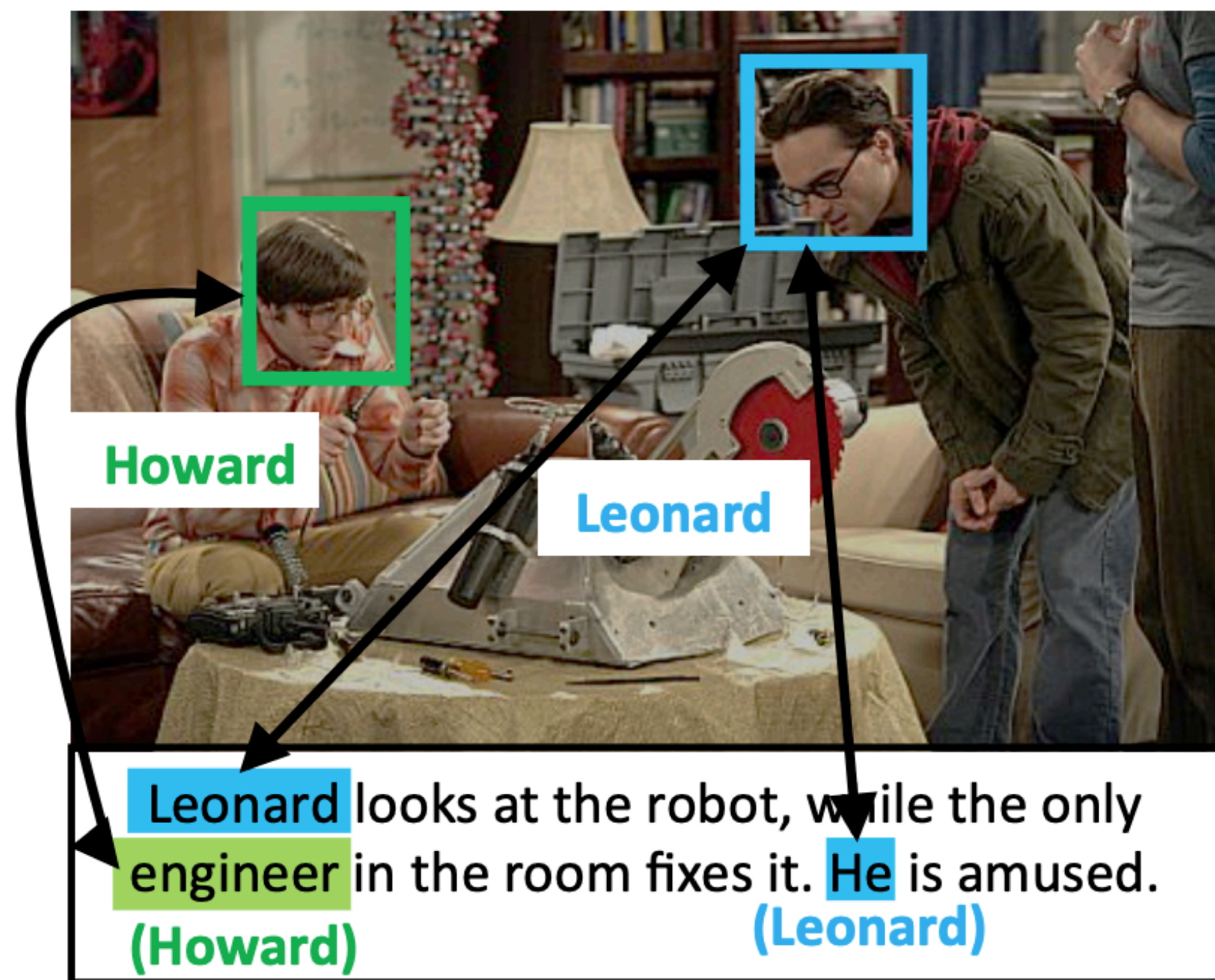
speaker



listener

What is symbol grounding?

- Connecting **linguistic symbols** to their **meaning**
- Connecting words and sentences to what they represent



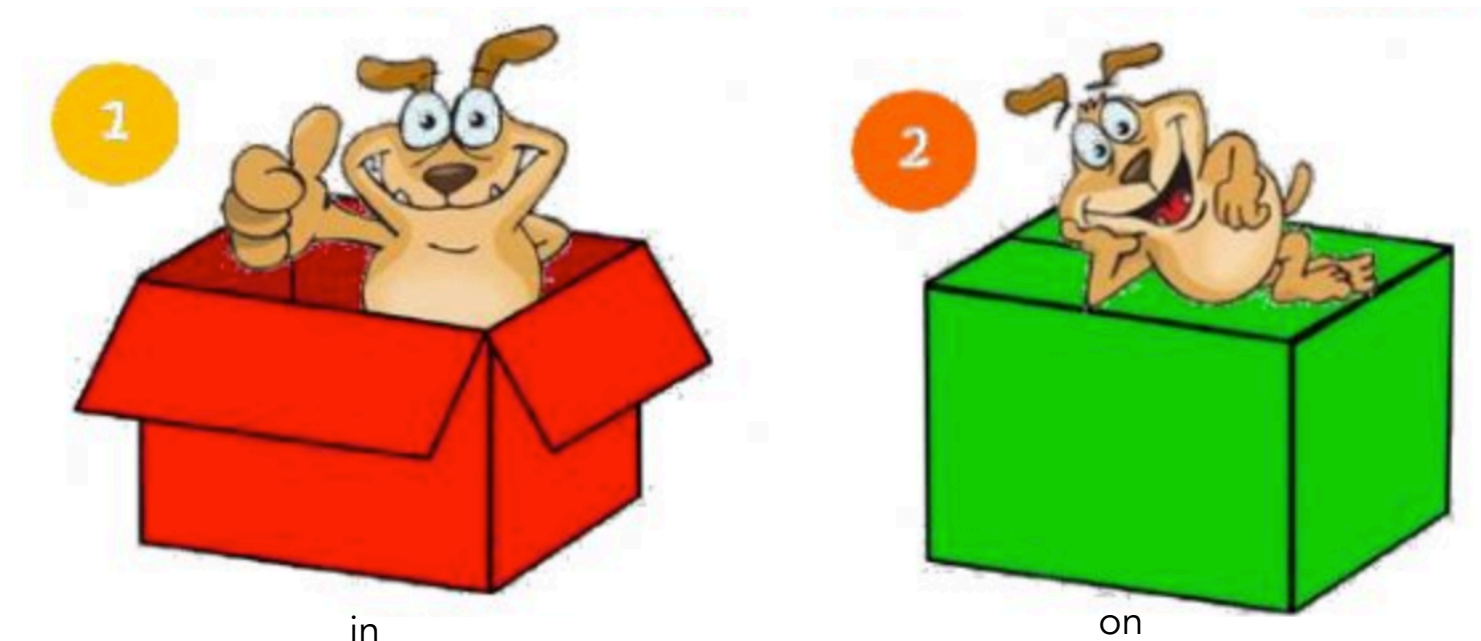
Linking people in videos with “their” names using coreference resolution [Ramanathan et al, 2014]

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

Types of grounding

- ▶ **Temporal concepts**

- ▶ *late evening* = after 6pm
- ▶ *fast, slow* = describing rates of change

- ▶ **Actions**



running



eating

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



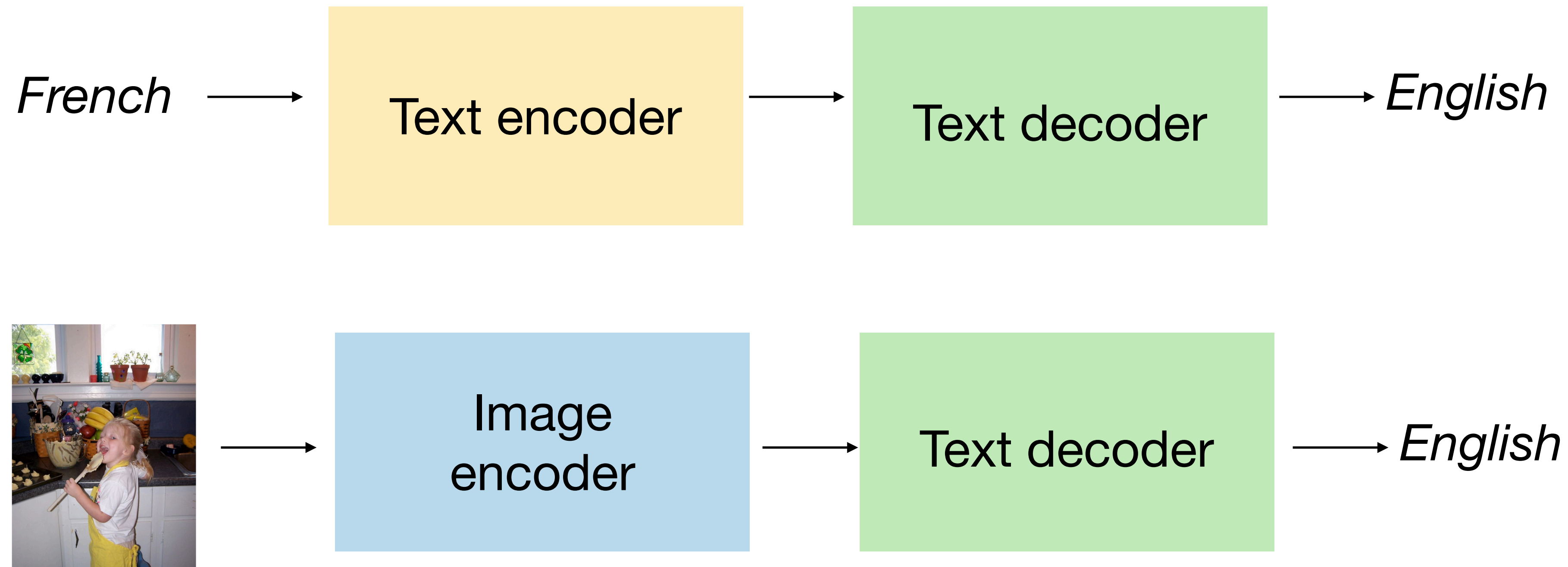
Image captioning

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

Captioning as multi-modal translation



(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

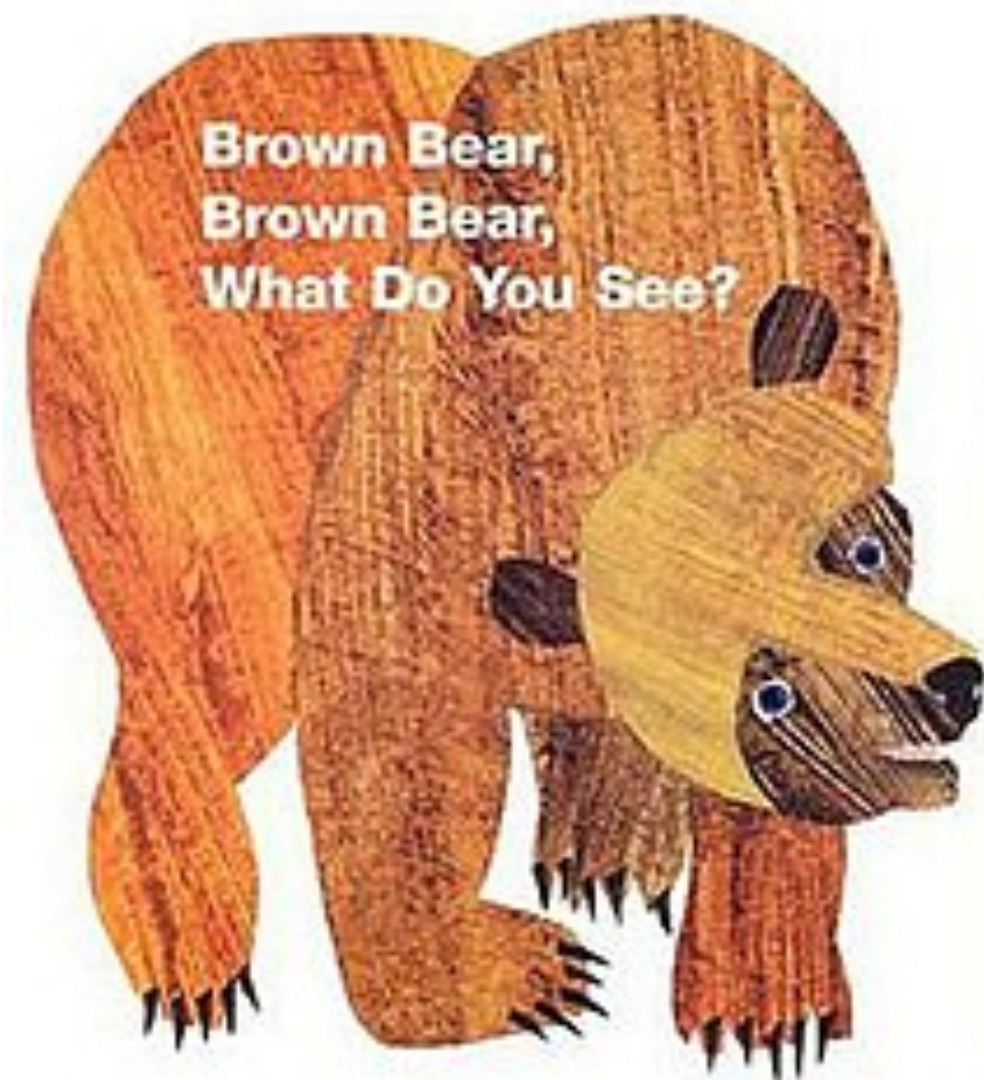
Natural way to learn language in the context of
its use in the **physical** and **social** world

This requires inferring the meaning of
utterances from their perceptual context

Children learn from **multimodal** sensory input and **experience**

Learning from multimodal information

Bill Martin Jr / Eric Carle



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

One hundred → 100

- other symbols

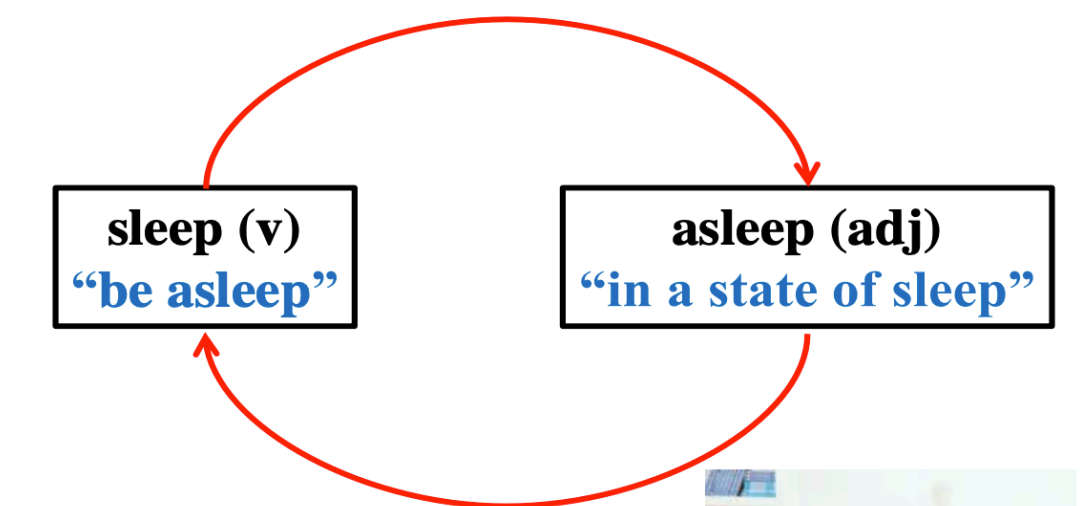
The Big Bang Theory →
[https://en.wikipedia.org/wiki/
The_Big_Bang_Theory](https://en.wikipedia.org/wiki/The_Big_Bang_Theory)

- to executable programs

Circular definitions

“Sleep” means “be asleep”

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



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?



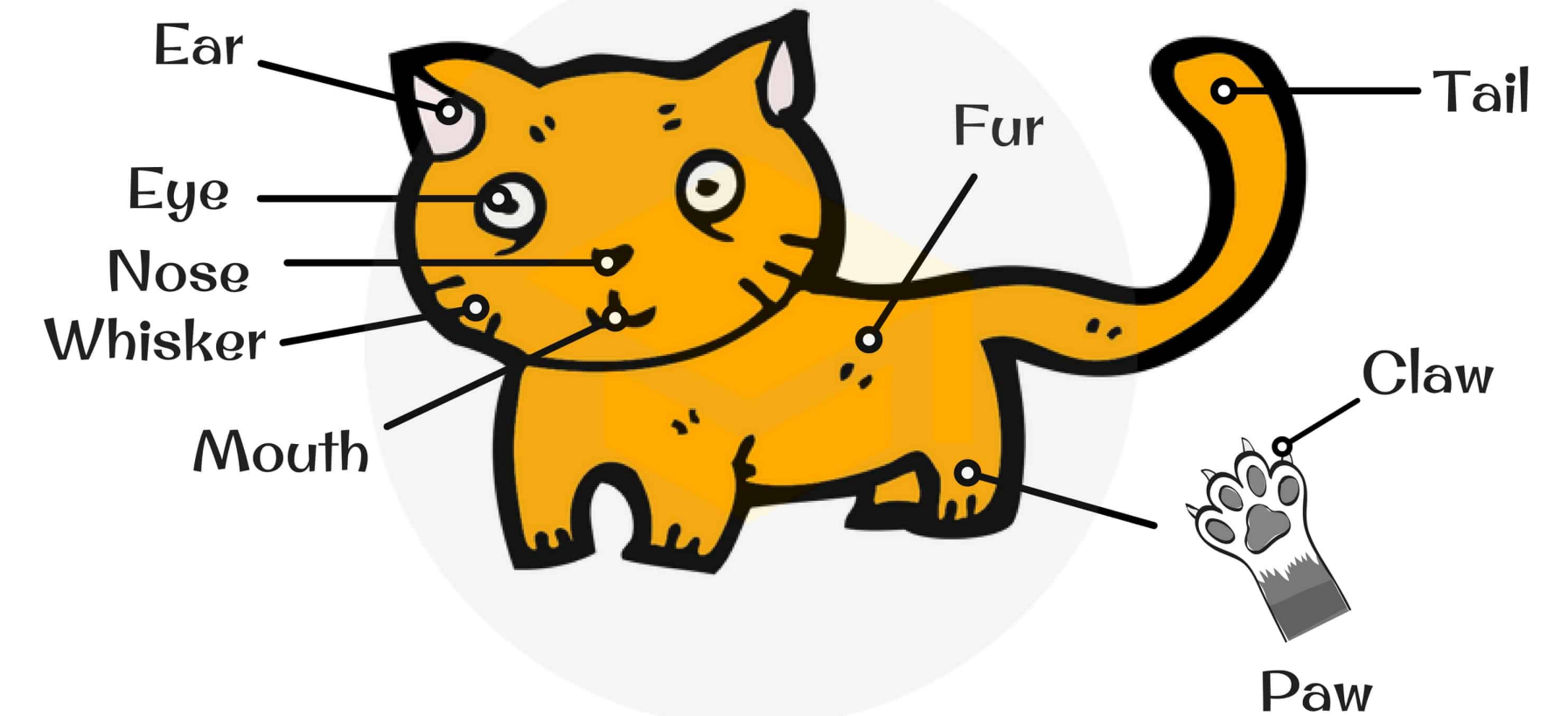
“cat”

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

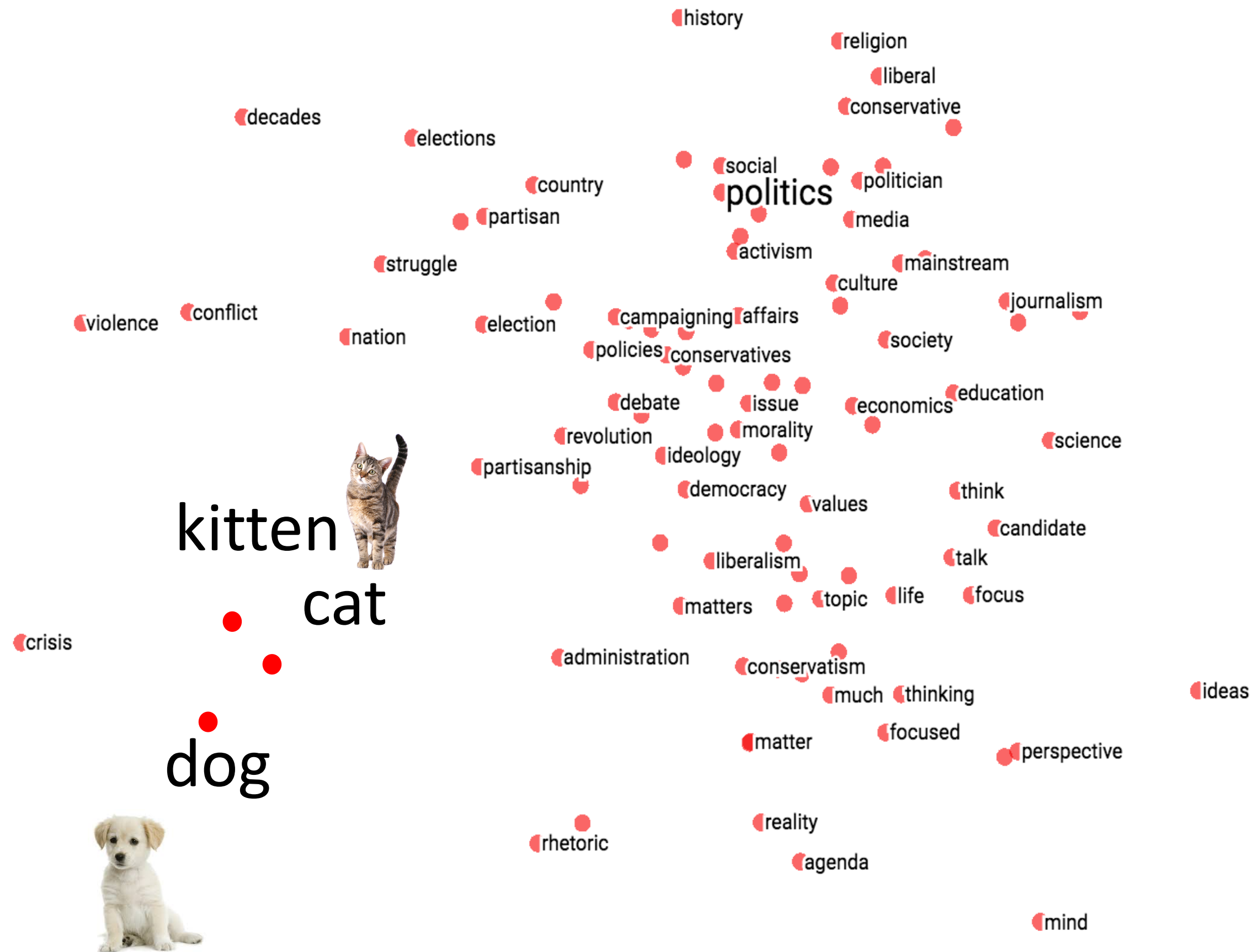
Parts of a cat



TESL.COM

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

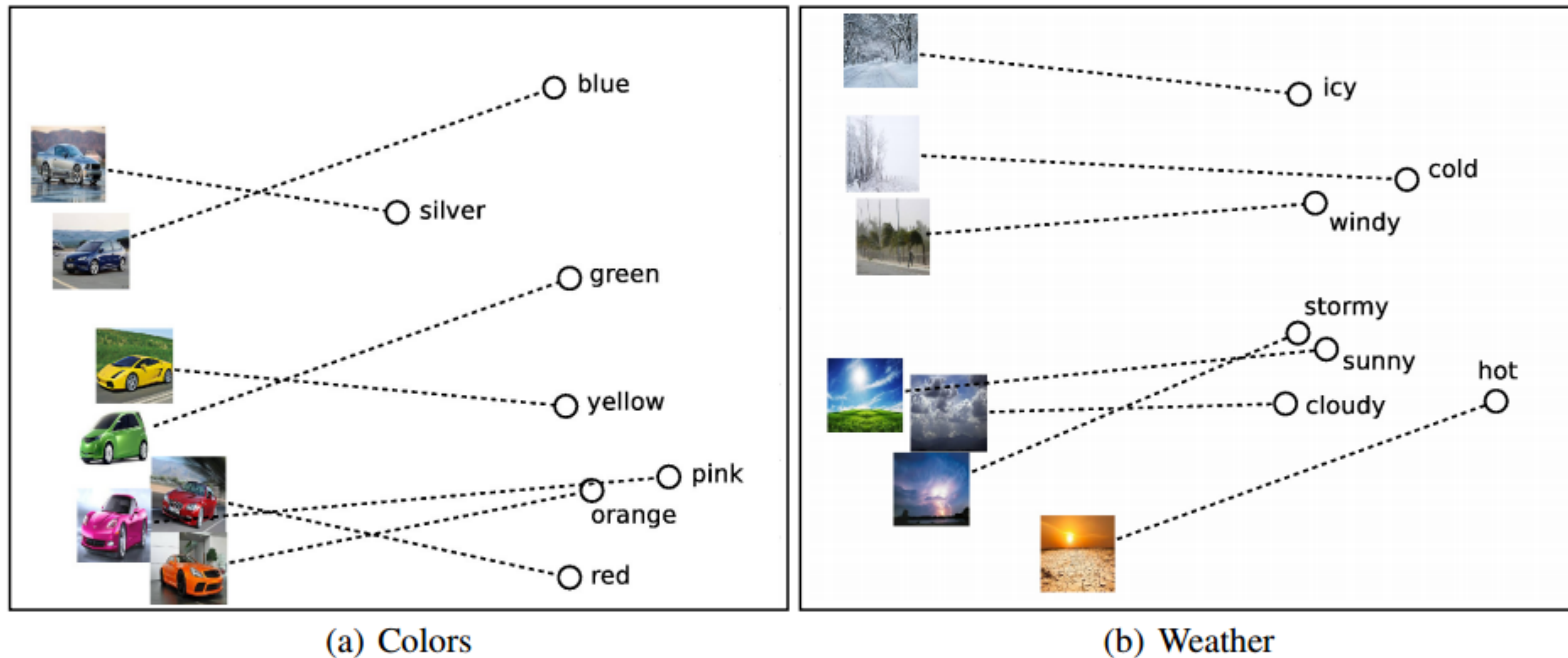


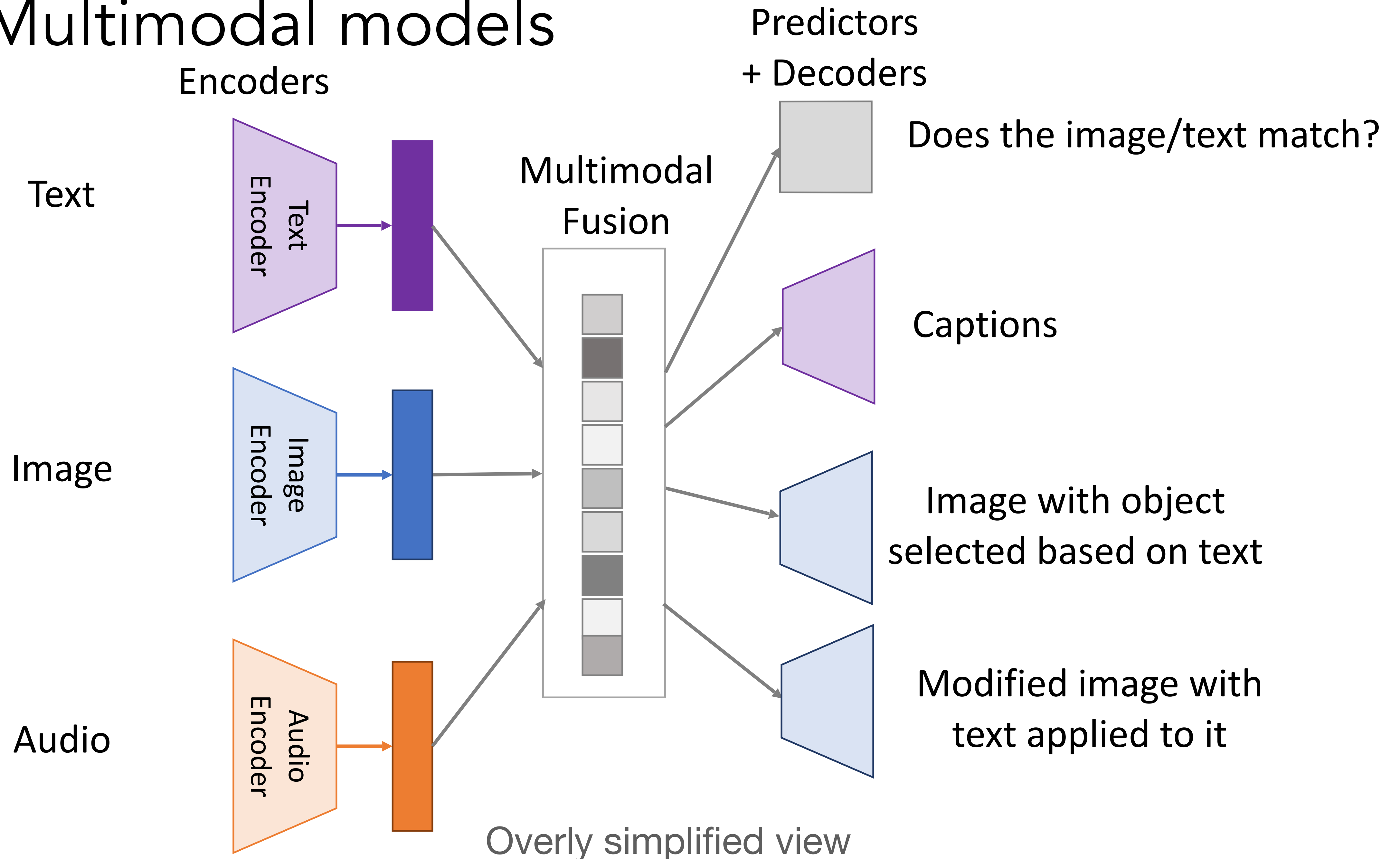
Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

“Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”

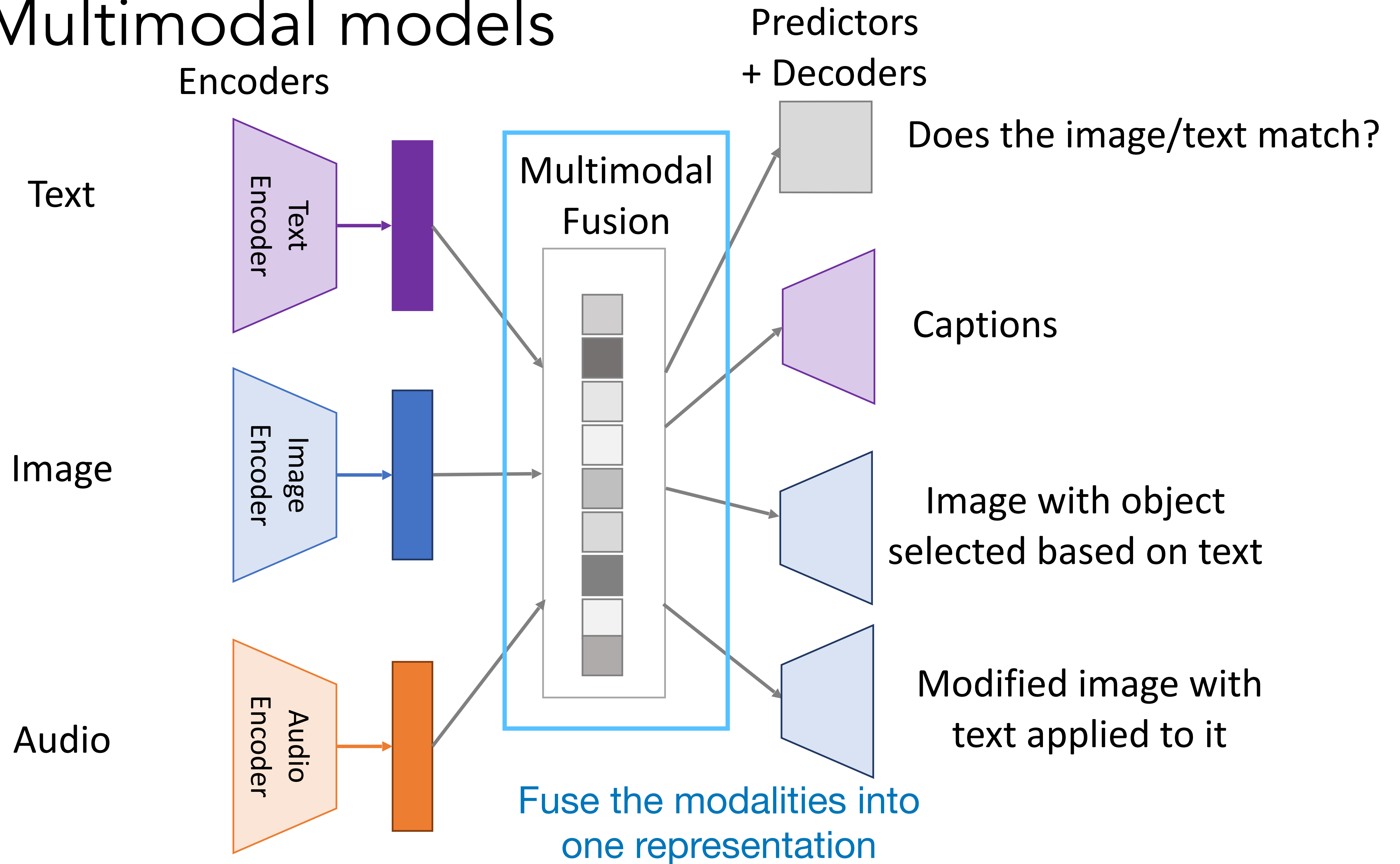
[Kiros, Salakhutdinov, Zemel TACL 2015]

Cross-modal models

Multimodal models

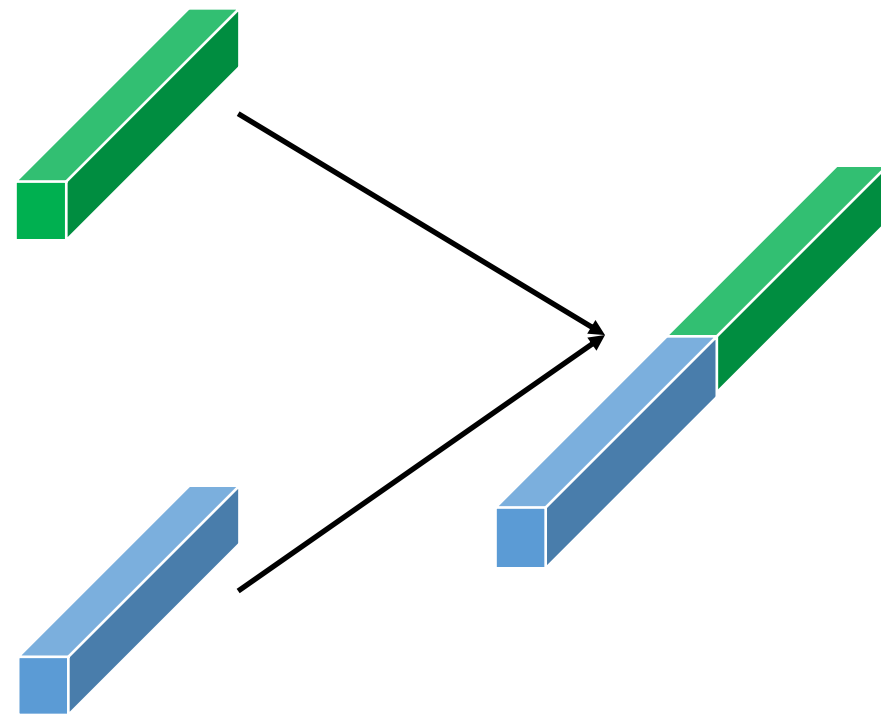


Multimodal models



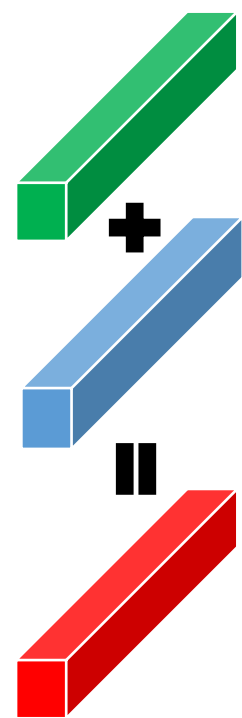
Multimodal Fusion

Concatenation

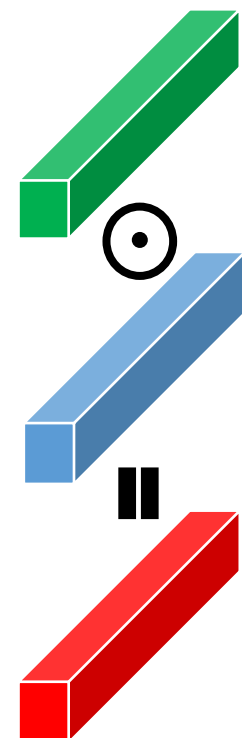


Element wise

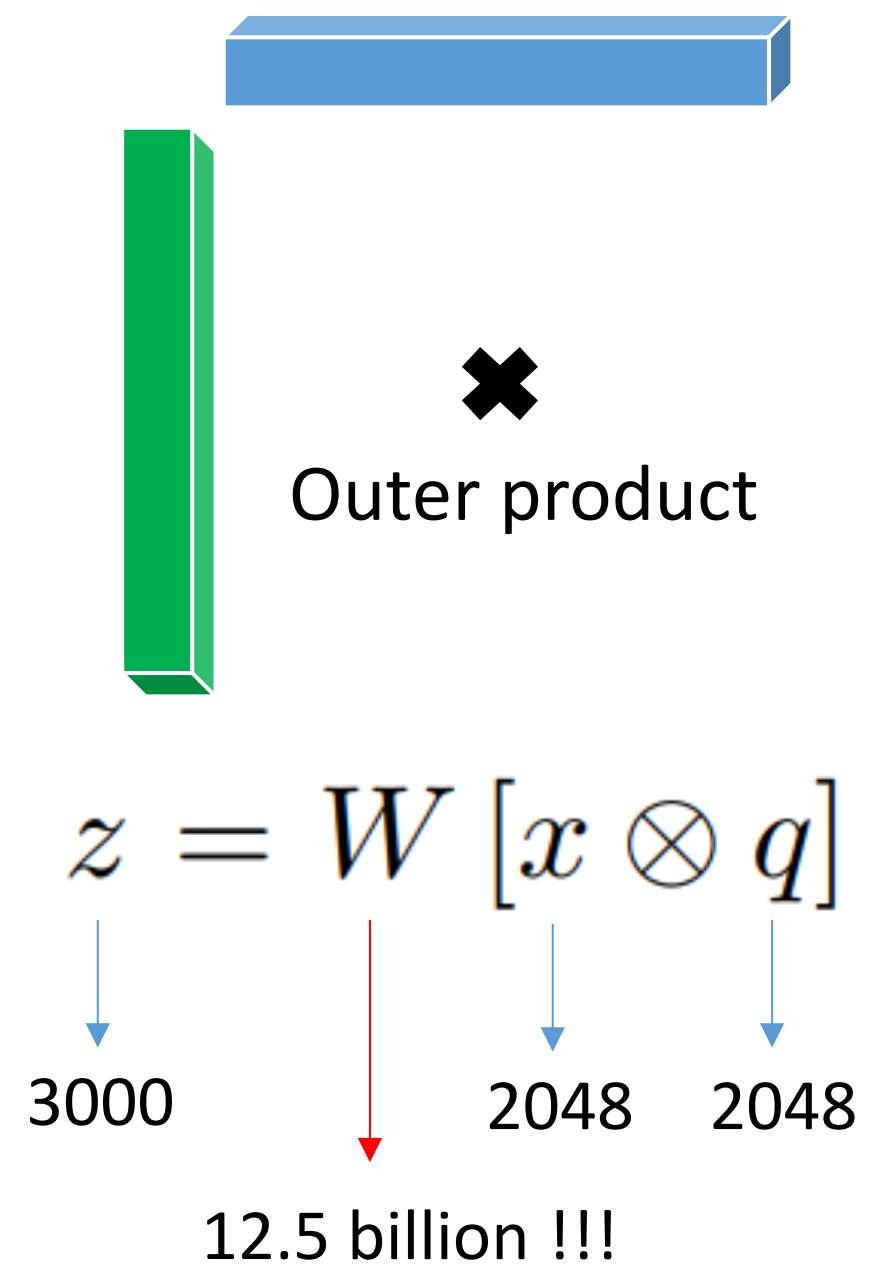
Sum



Product



Bilinear Pooling



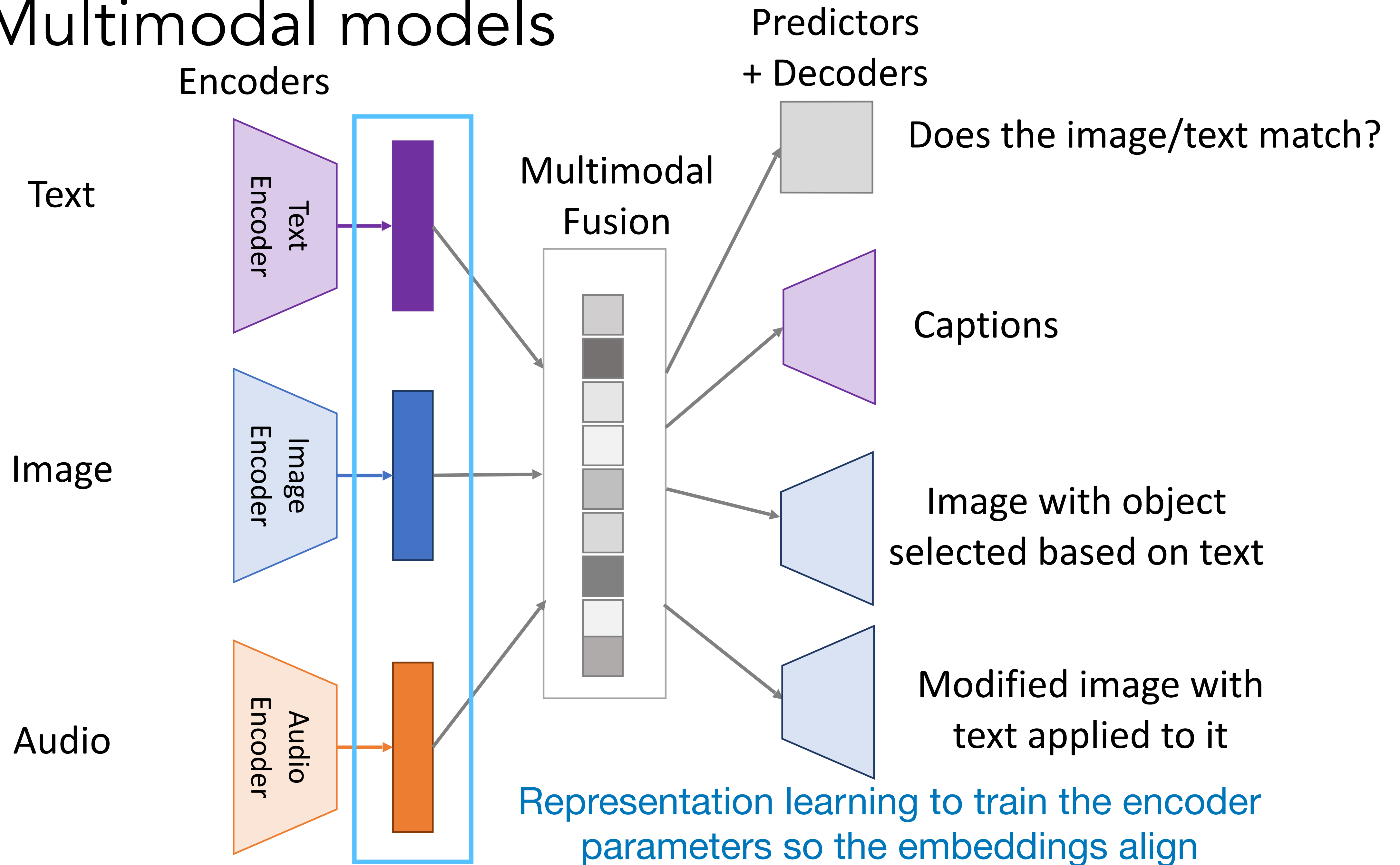
All elements can interact.
More flexible, but lots of weights!

Attention-based fusion



Use transformer to fuse modalities using attention

Multimodal models



Multimodal Embeddings

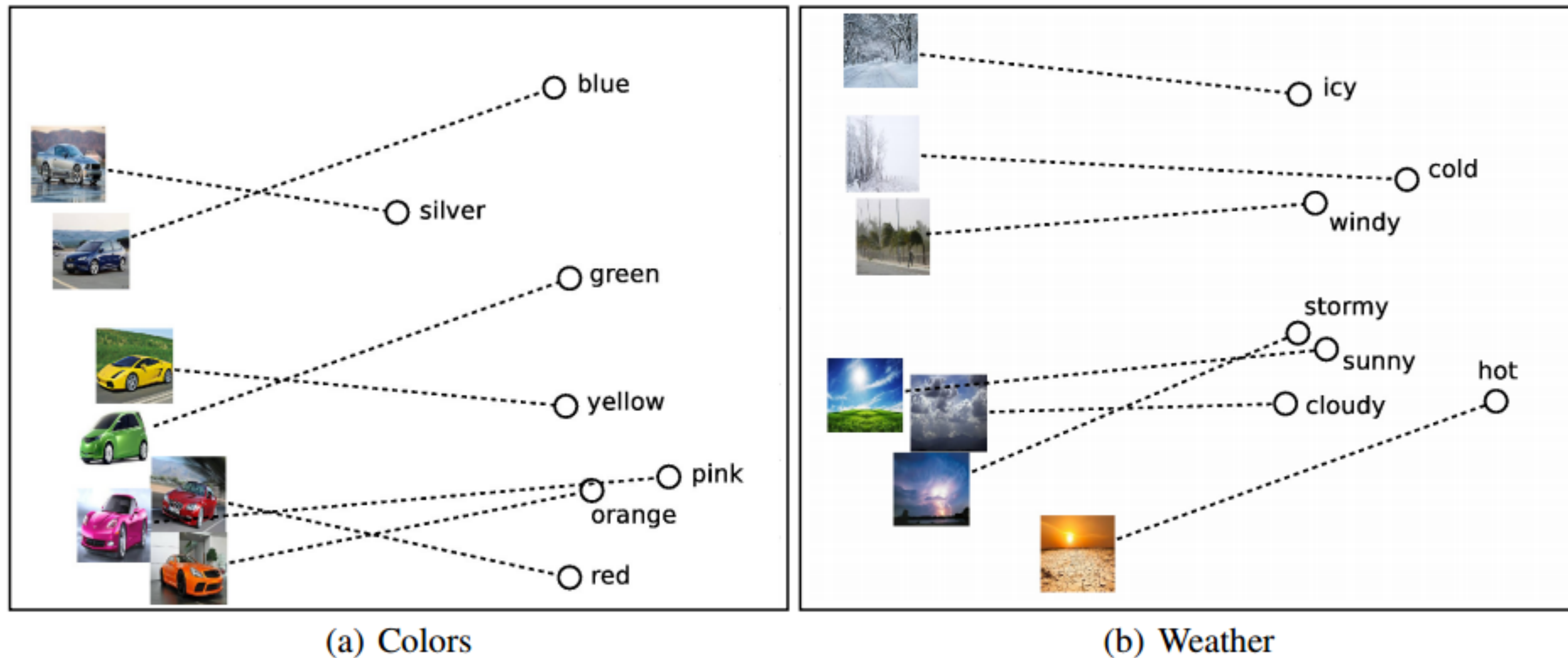


Figure 5: PCA projection of the 300-dimensional word and image representations for (a) cars and colors and (b) weather and temperature.

“Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”

[Kiros, Salakhutdinov, Zemel TACL 2015]

Cross-modal Embeddings

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

Image Embedding 

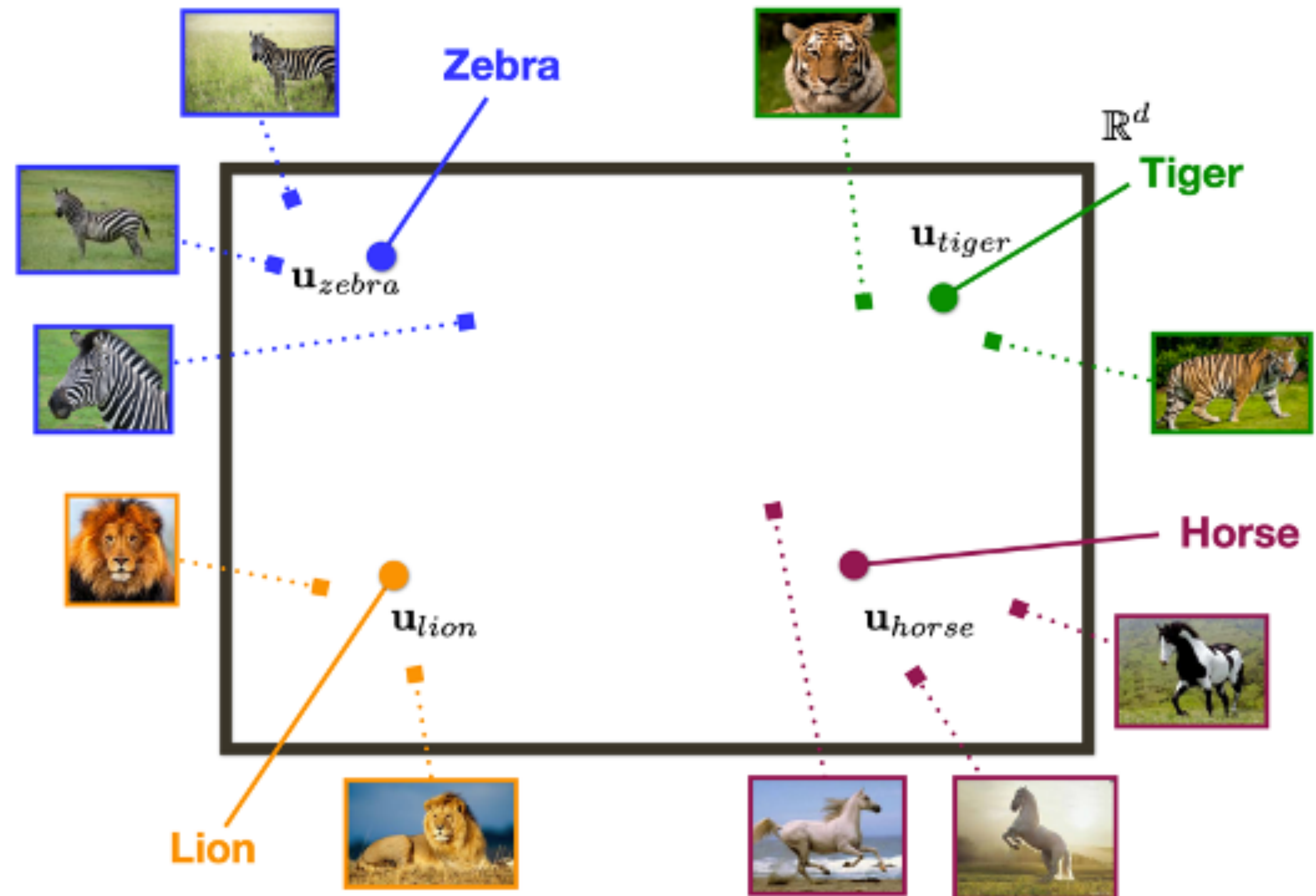
$$\Psi(I_i) = \mathbf{W} \cdot \text{CNN}(I_i; \Theta): \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding 

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

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

Image Embedding



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

Label Embedding



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

Similarity in Embedding Space

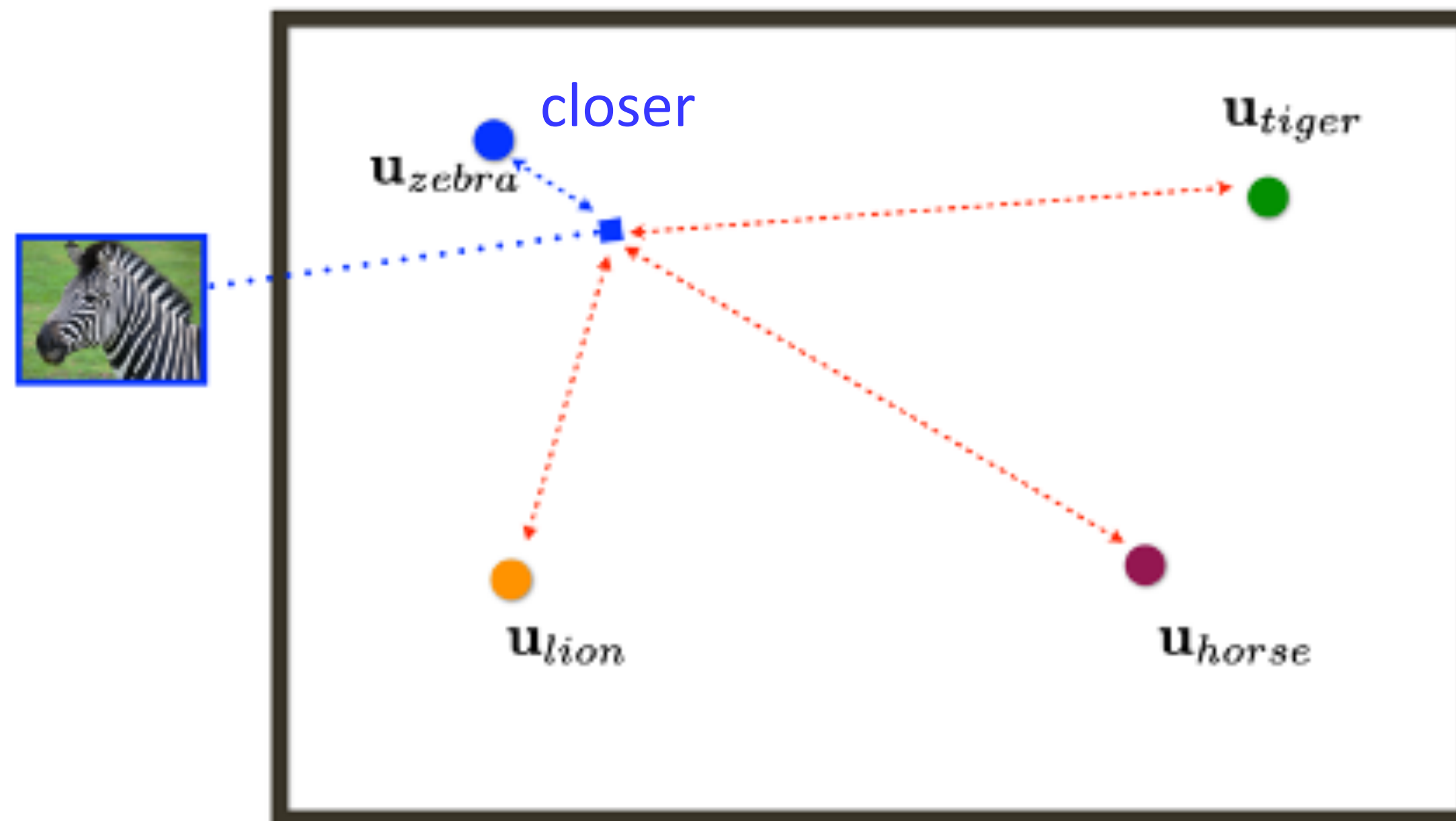
$$S(\mathbf{u}, \mathbf{u}') = \frac{\mathbf{u}}{\|\mathbf{u}\|} \cdot \frac{\mathbf{u}'}{\|\mathbf{u}'\|}$$

Objective Function:

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

$$\mathcal{L}_C = \sum \max(0, \alpha - \underbrace{S(\Psi(I_i), \mathbf{u}_{y_i})}_{\text{Correct label (more similar)}} + \underbrace{S(\Psi(I_i), \mathbf{u}_{y_c})}_{\text{Other labels (less similar)}})$$

+ pair - pair

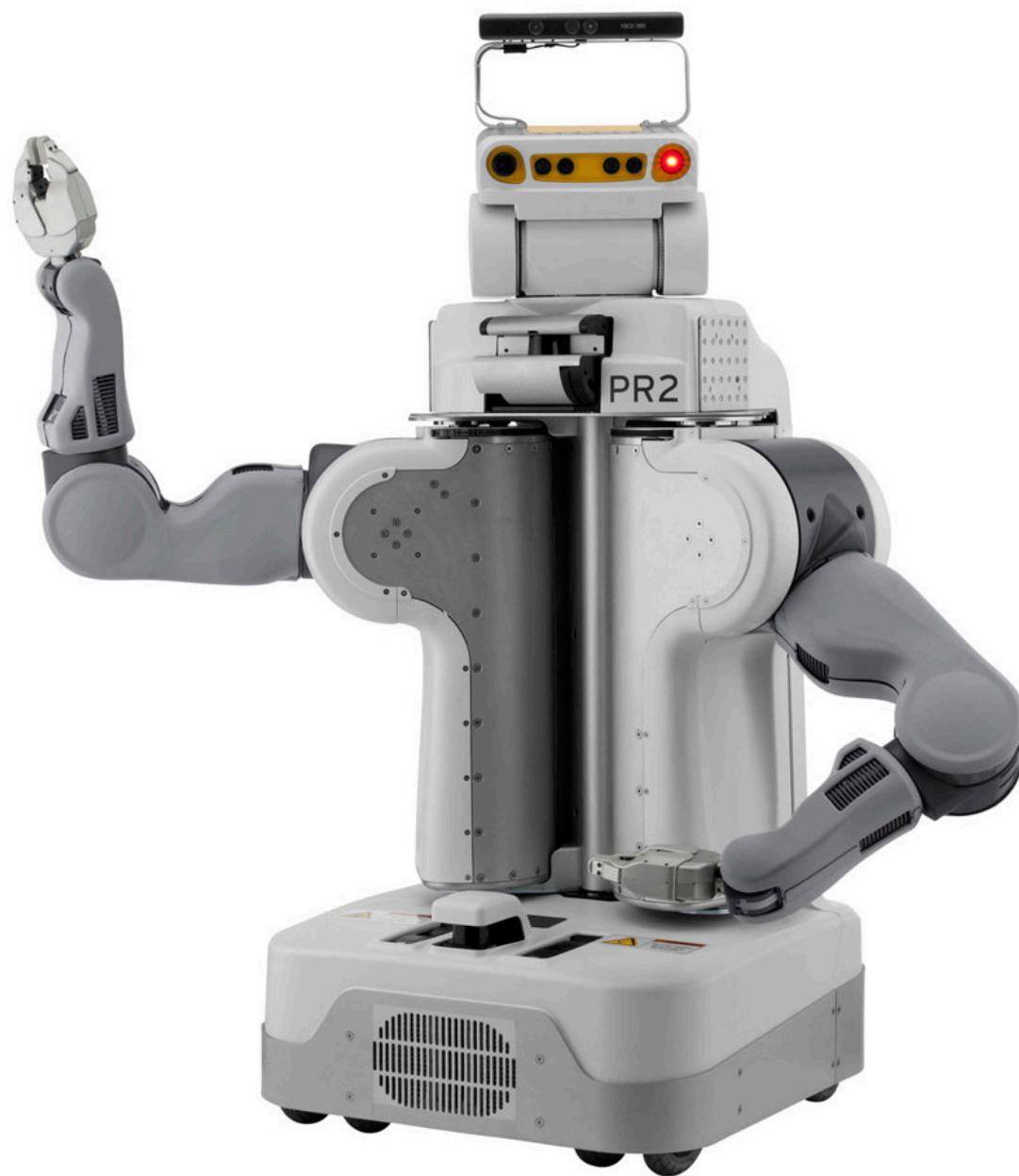


[Bengio et al., NIPS'10]

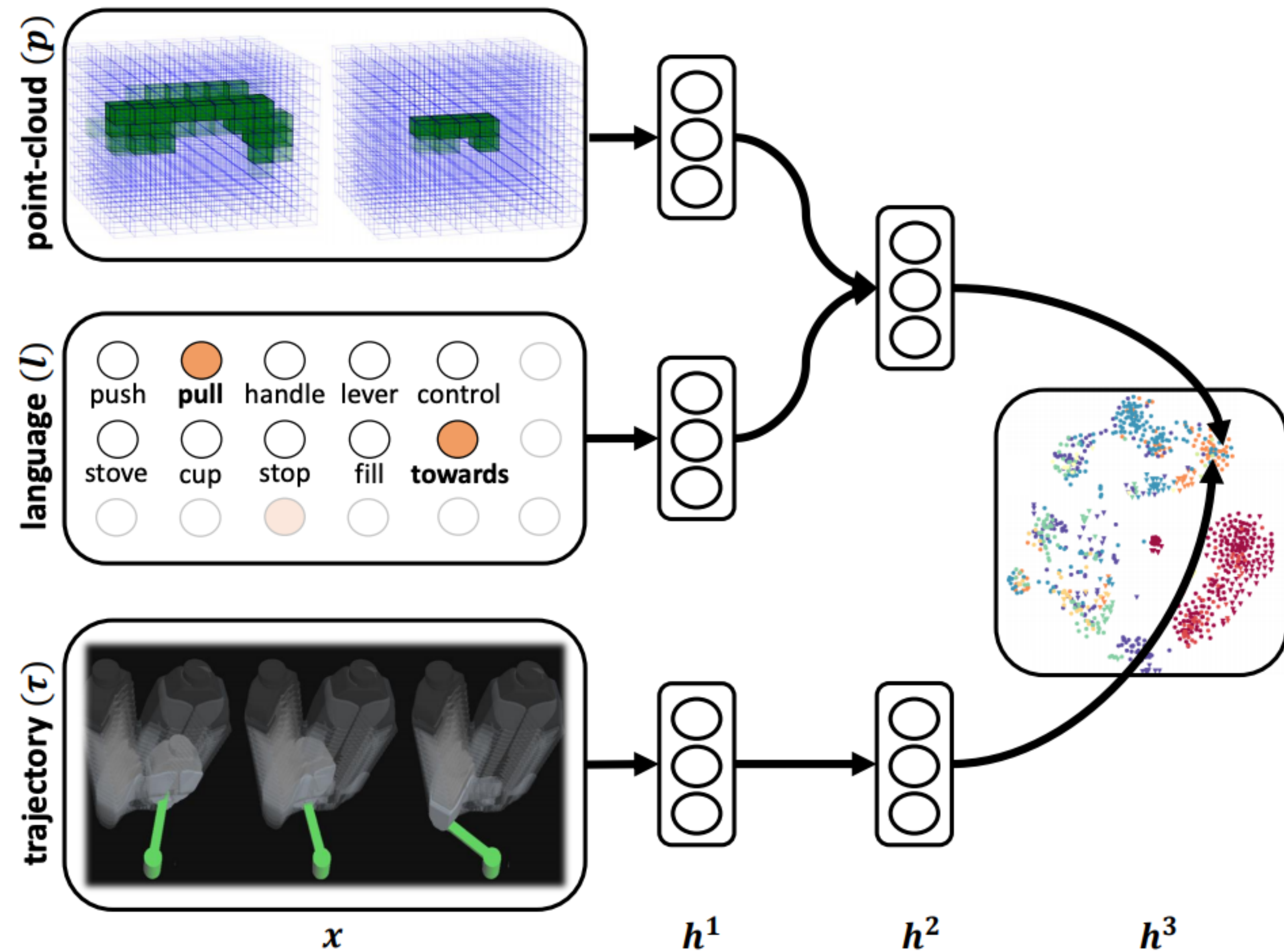
[Weinberger, Chapelle, NIPS'09]

Can embed anything!

PR2 robot

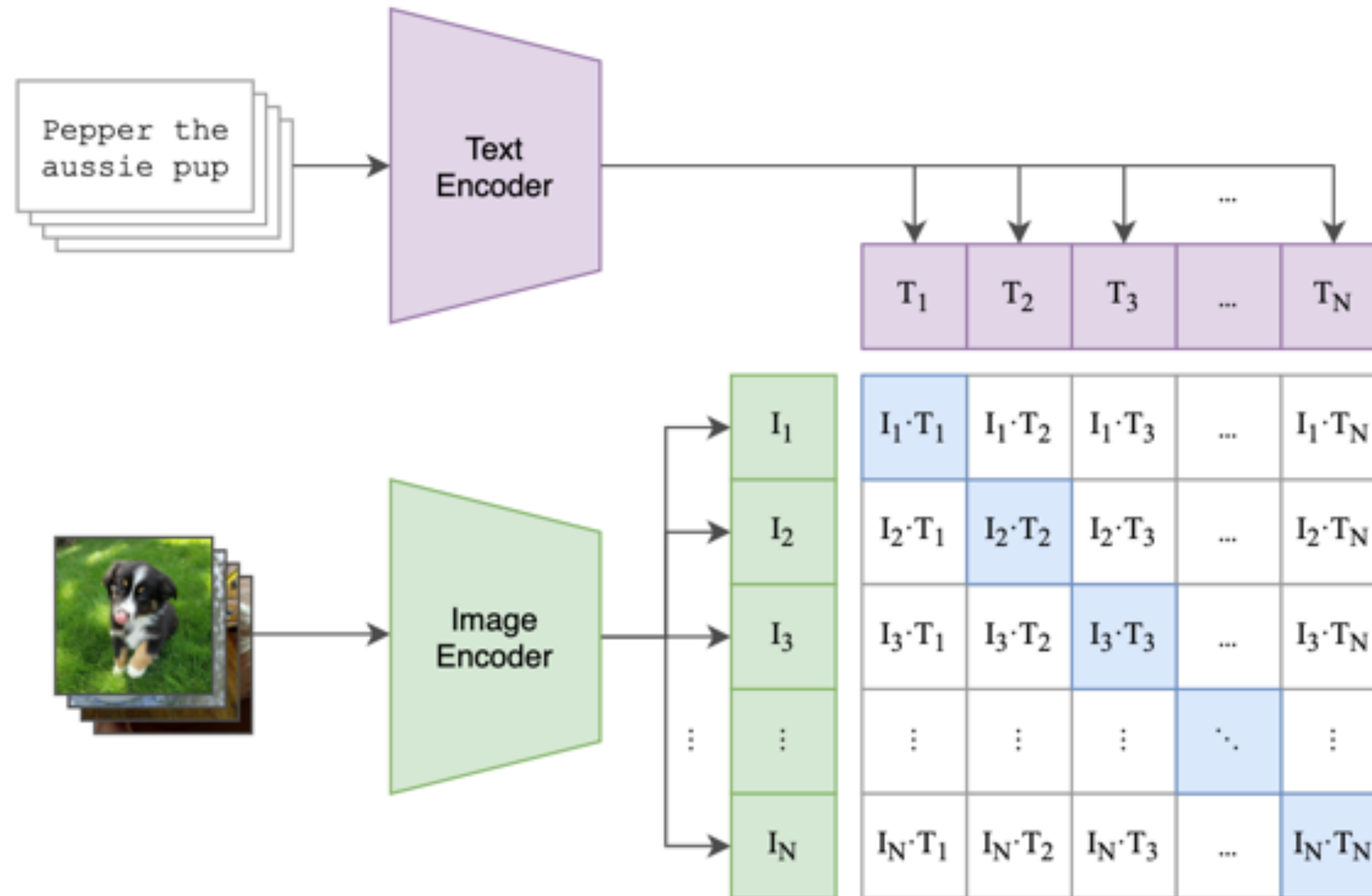


<https://robots.ieee.org/robots/pr2>



Contrastive pretraining

OpenAI CLIP

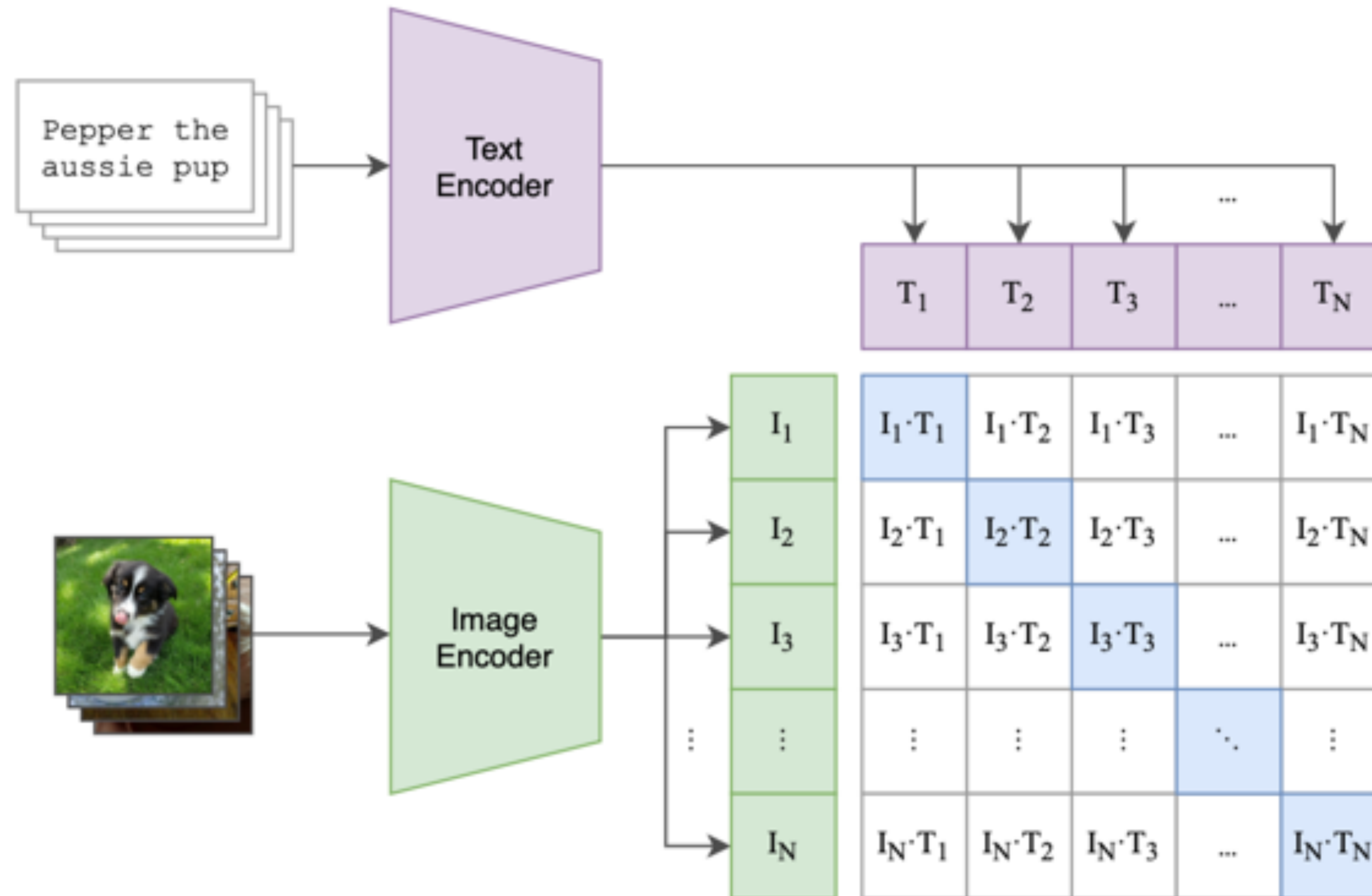


- Train on large amount of data
- WebImageText: 400M text-image pairs
- Contrastive pretraining: does the text-image pair match?
 - Batch size $N=32K$
 - N positive pairs
 - $N^2 - N$ negative pairs
- Transformer based model for both vision and language

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

Contrastive pretraining

OpenAI CLIP



- Contrastive Loss

NT-Xent loss (images and text)

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

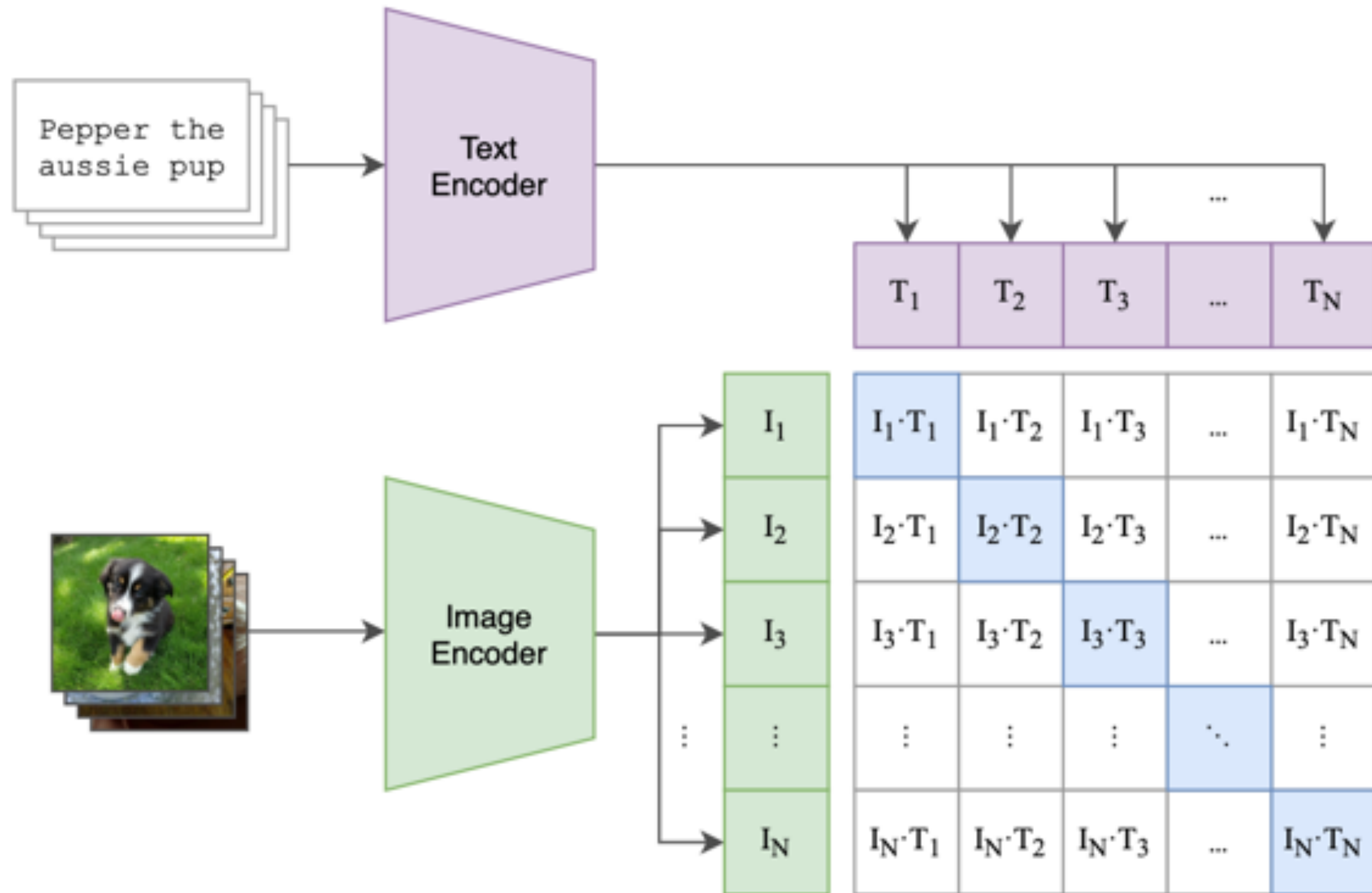
Symmetric Bimodal loss

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

Learning transferable visual models 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)

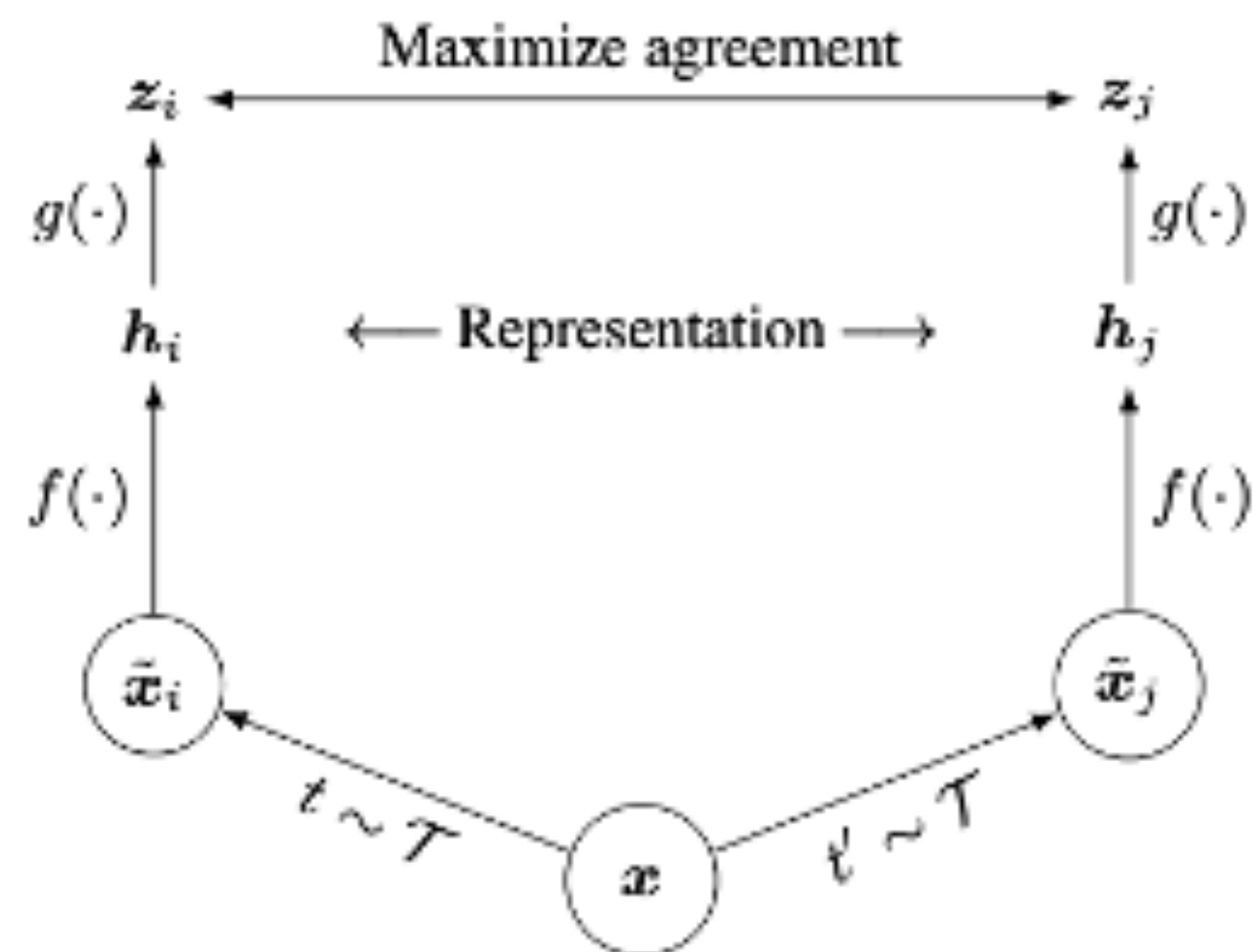
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# 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
```

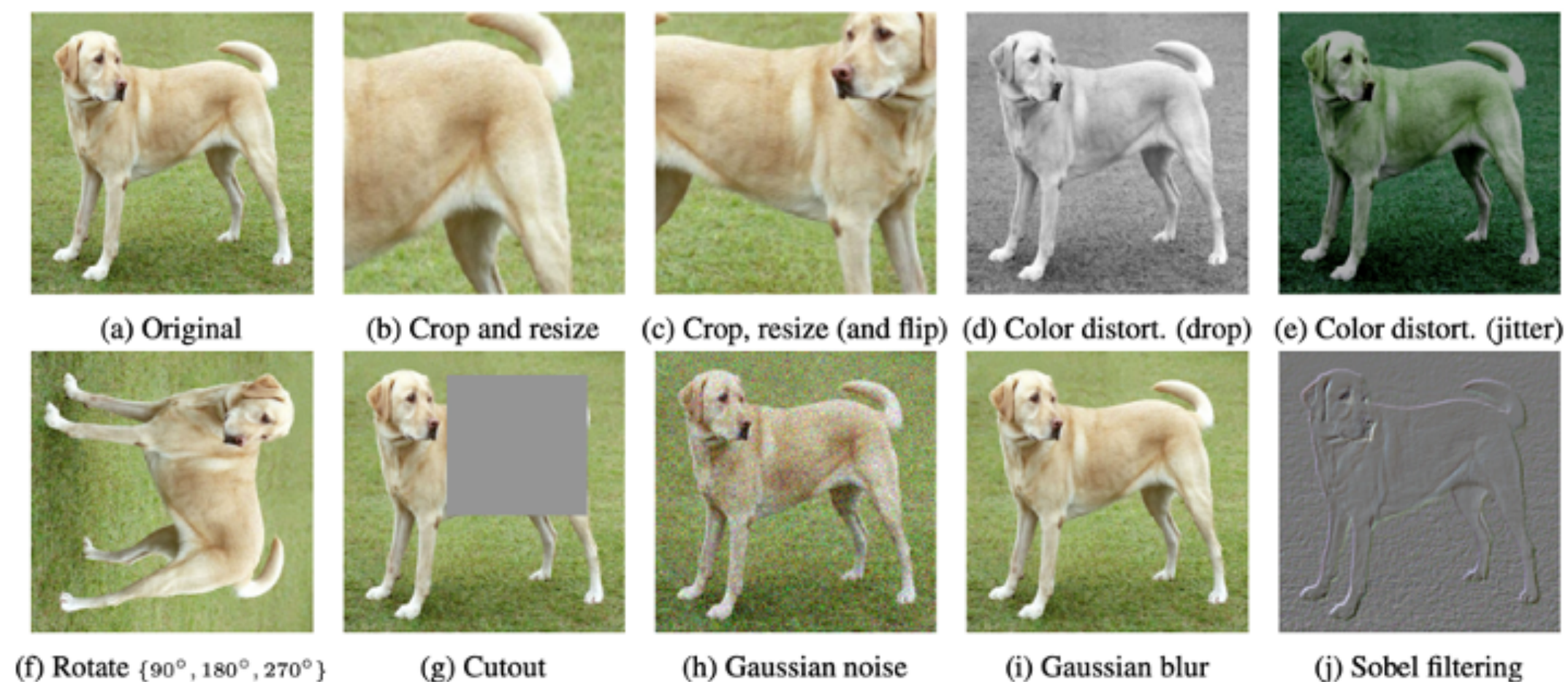
Learning transferable visual models from natural language supervision [Radford et al, 2020]

Contrastive loss is very common

Unimodal self-supervised learning



Images

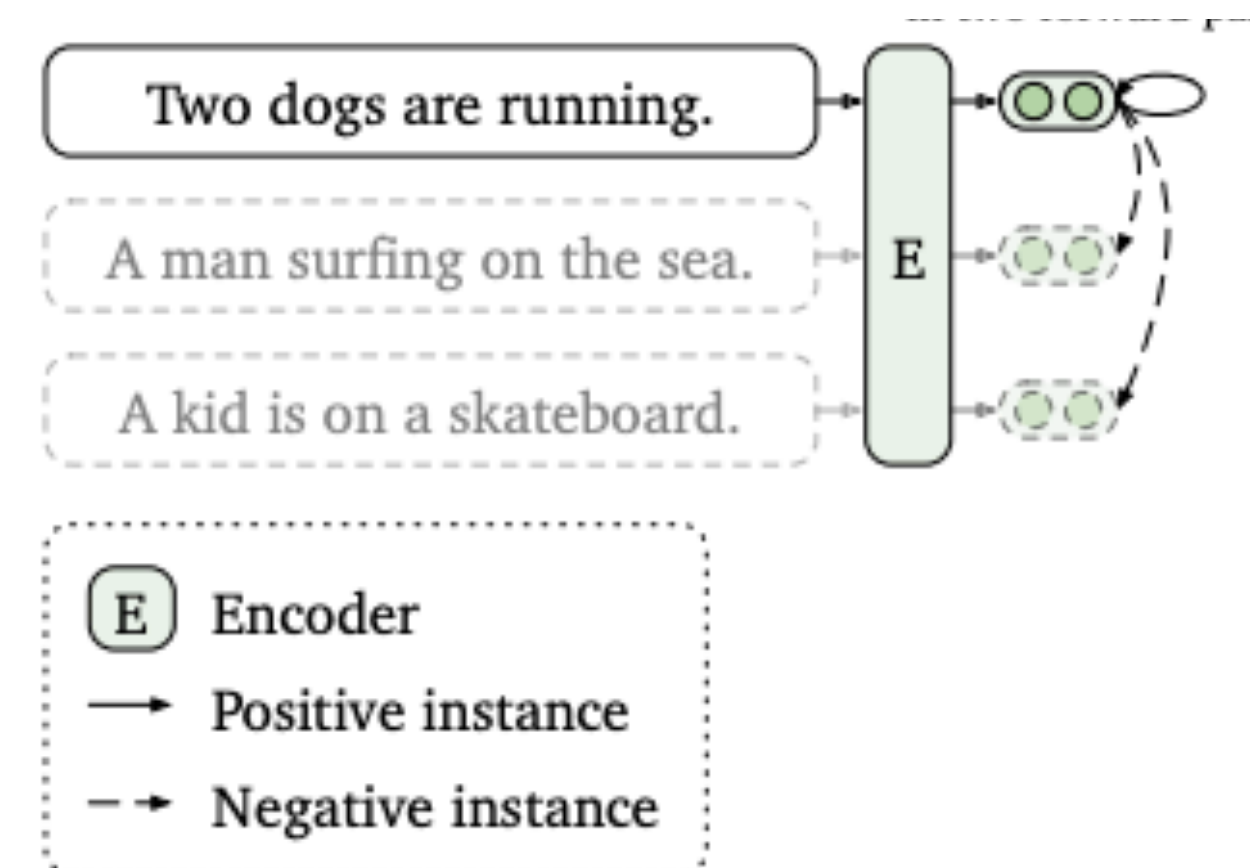


A Simple Framework for Contrastive Learning of Visual Representations [Chen et al., ICML 2020]

Two parallel encoders

Positive pair is two views
of the same input

Text



Encoder: BERT [CLS]
with MLP

Data augmentation:

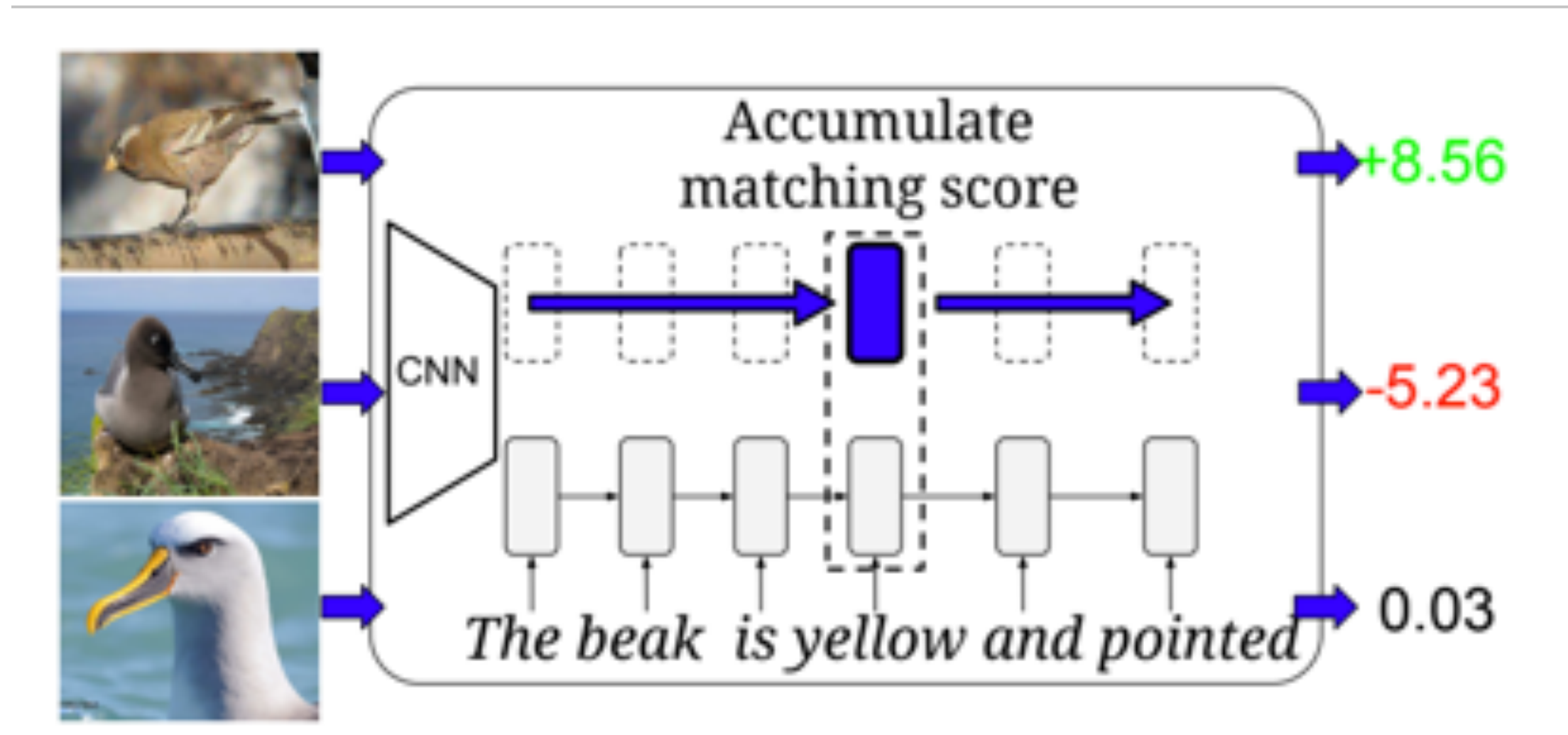
- Drop tokens
- Add dropout to network

SimCSE: Simple Contrastive Learning of Sentence Embeddings [Gao et al, 2022]

Aligned embeddings can be used for a
variety of tasks

Retrieval

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



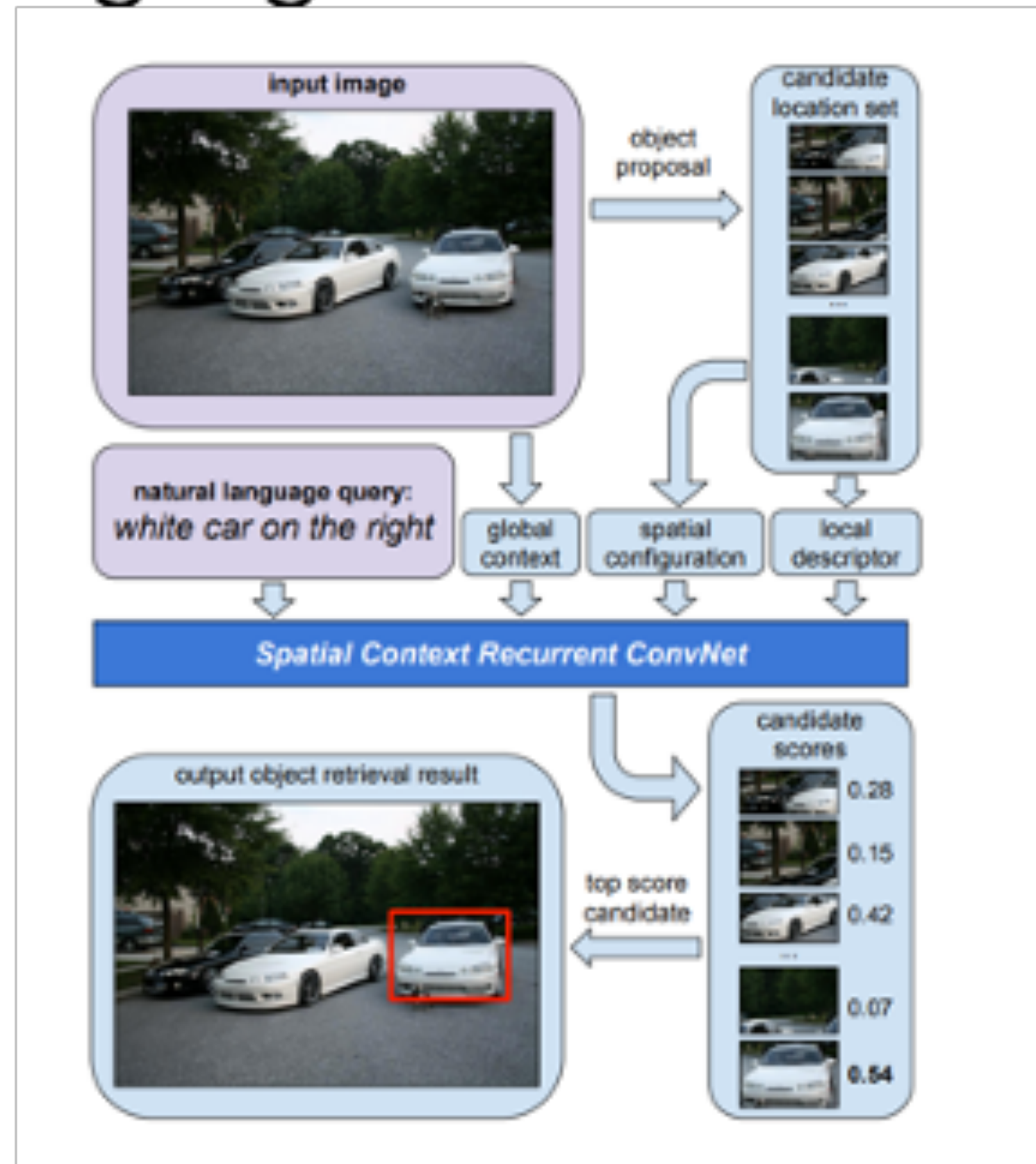
Embedding	Top-1 Acc (%)		AP@50 (%)	
	DA-SJE	DS-SJE	DA-SJE	DS-SJE
ATTRIBUTES	50.9	50.4	20.4	50.0
WORD2VEC	38.7	38.6	7.5	33.5
BAG-OF-WORDS	43.4	44.1	24.6	39.6
CHAR CNN	47.2	48.2	2.9	42.7
CHAR LSTM	22.6	21.6	11.6	22.3
CHAR CNN-RNN	54.0	54.0	6.9	45.6
WORD CNN	50.5	51.0	3.4	43.3
WORD LSTM	52.2	53.0	36.8	46.8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

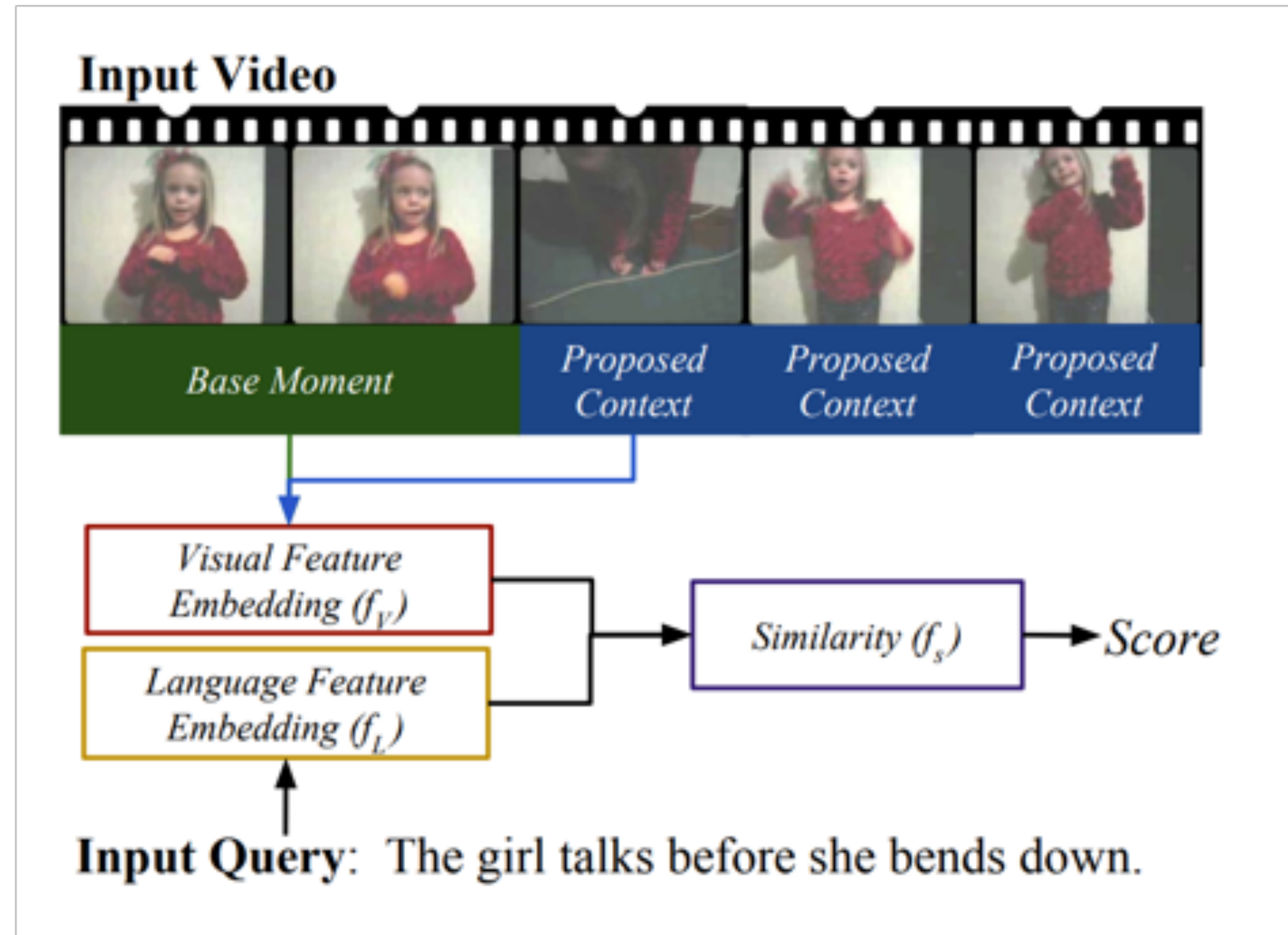
Grounding

Match image region to language



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

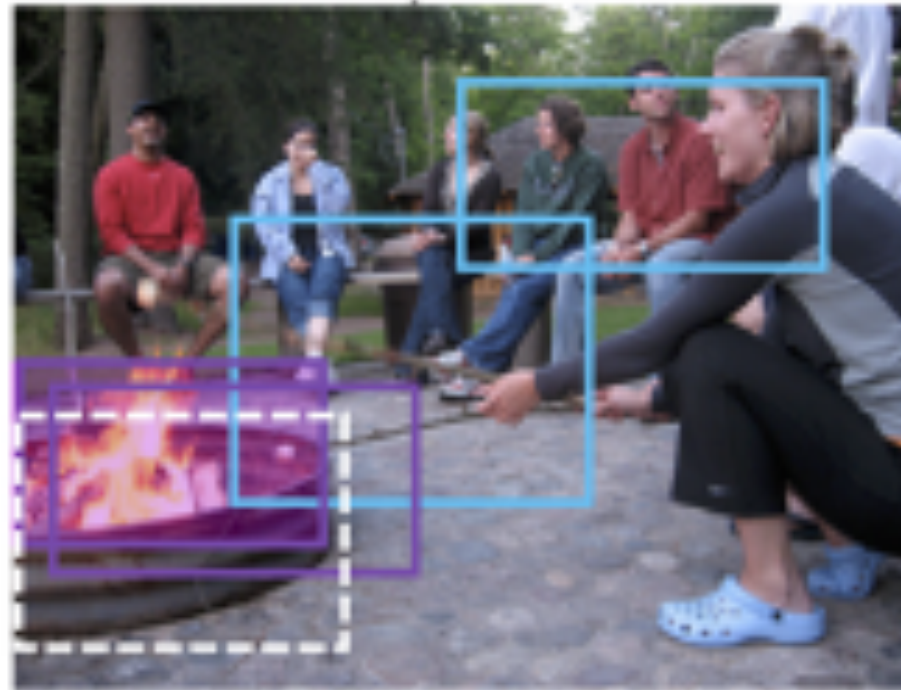
Match video frames to language



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

Grounding

Phrase Localization



A group of eight campers sit around a fire pit trying to roast marshmallows on their sticks.

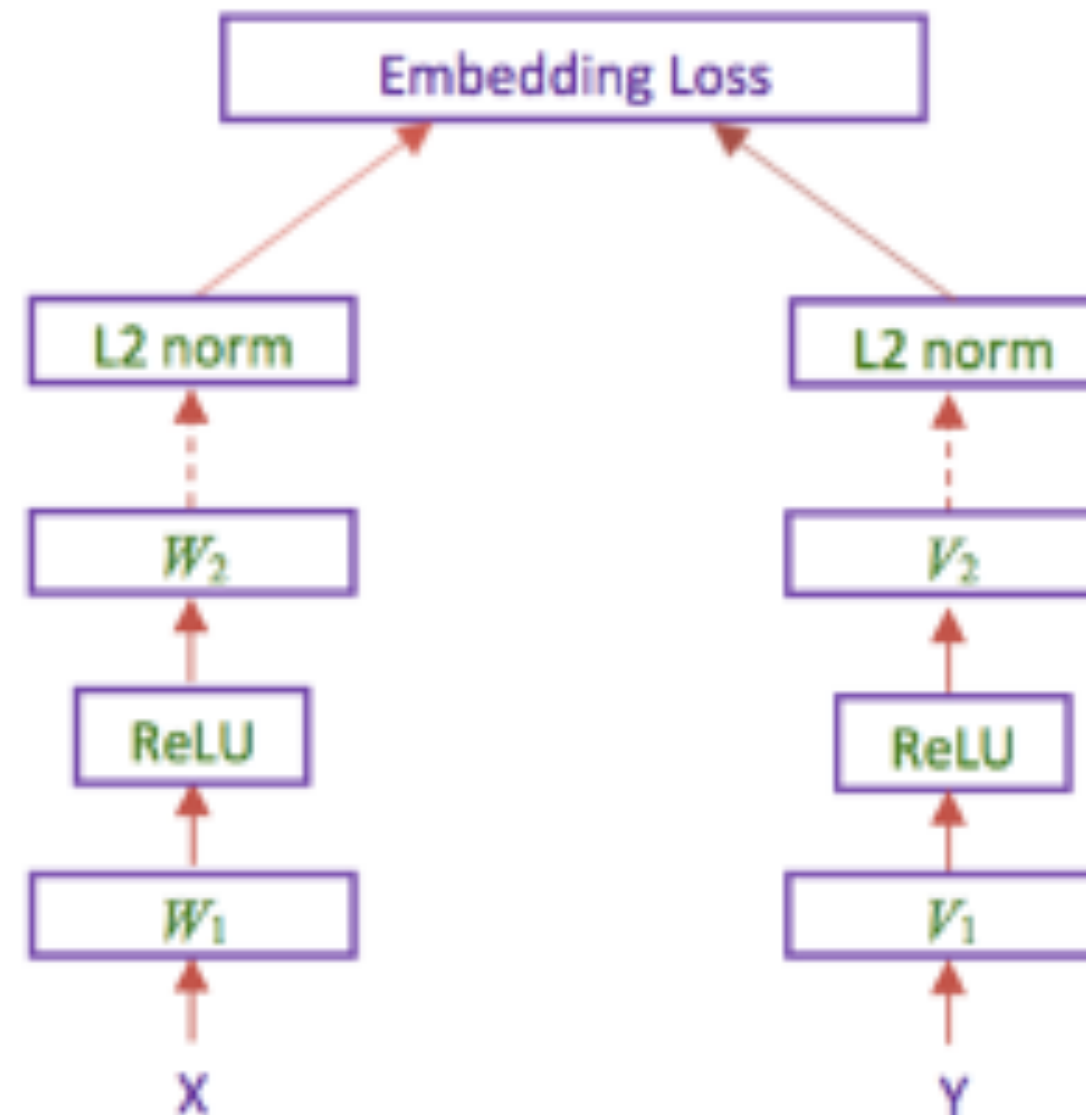
X: regions



Y: "a fire pit"

Embedding Network

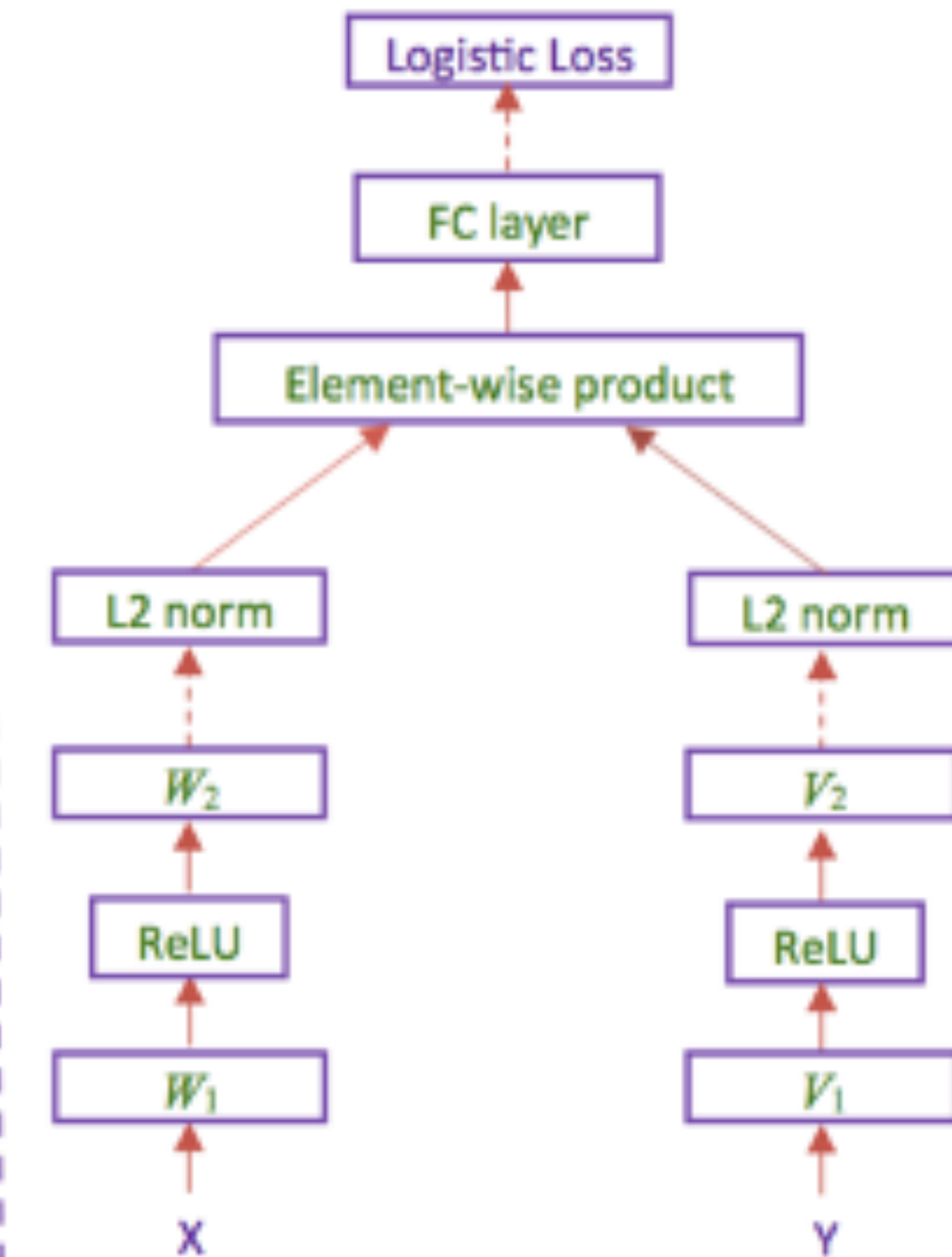
$$d(\text{fire pit image}, \text{"a fire pit"}) + m < d(\text{campers image}, \text{"a fire pit"})$$
$$d(\text{fire pit image}, \text{"a fire pit"}) + m < d(\text{fire pit image}, \text{"campers"})$$



Similarity Network

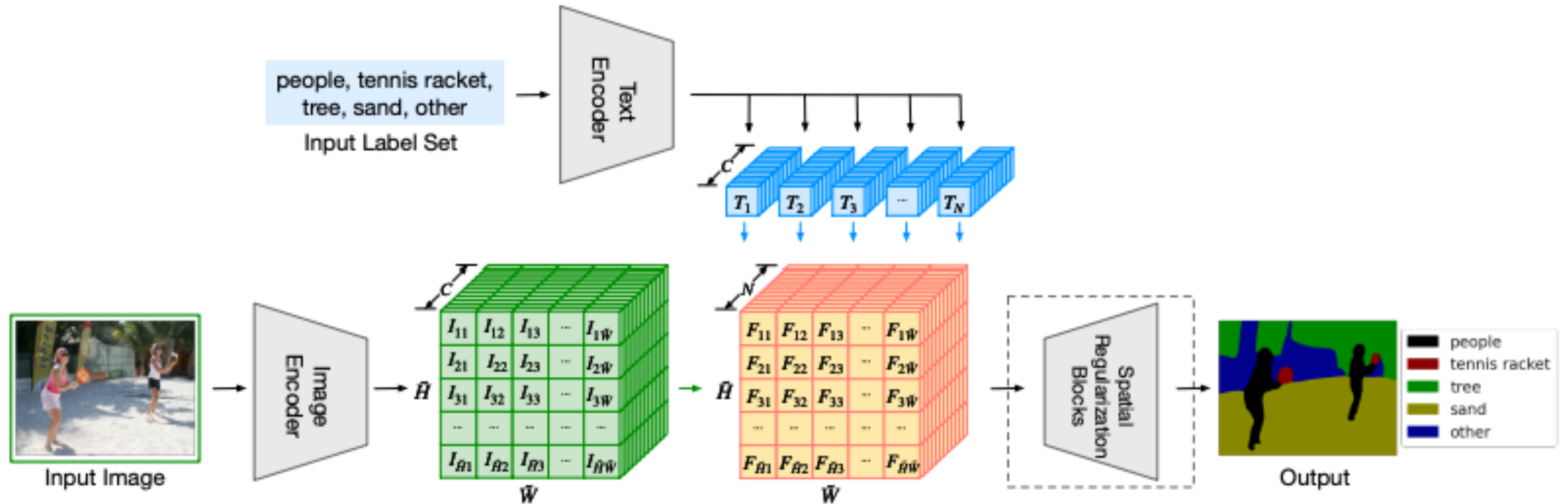
, "a fire pit": +1

, "a fire pit": -1



Language Driven Semantic Segmentation

Use CLIP as text encoder



Use dense prediction
transformers architecture

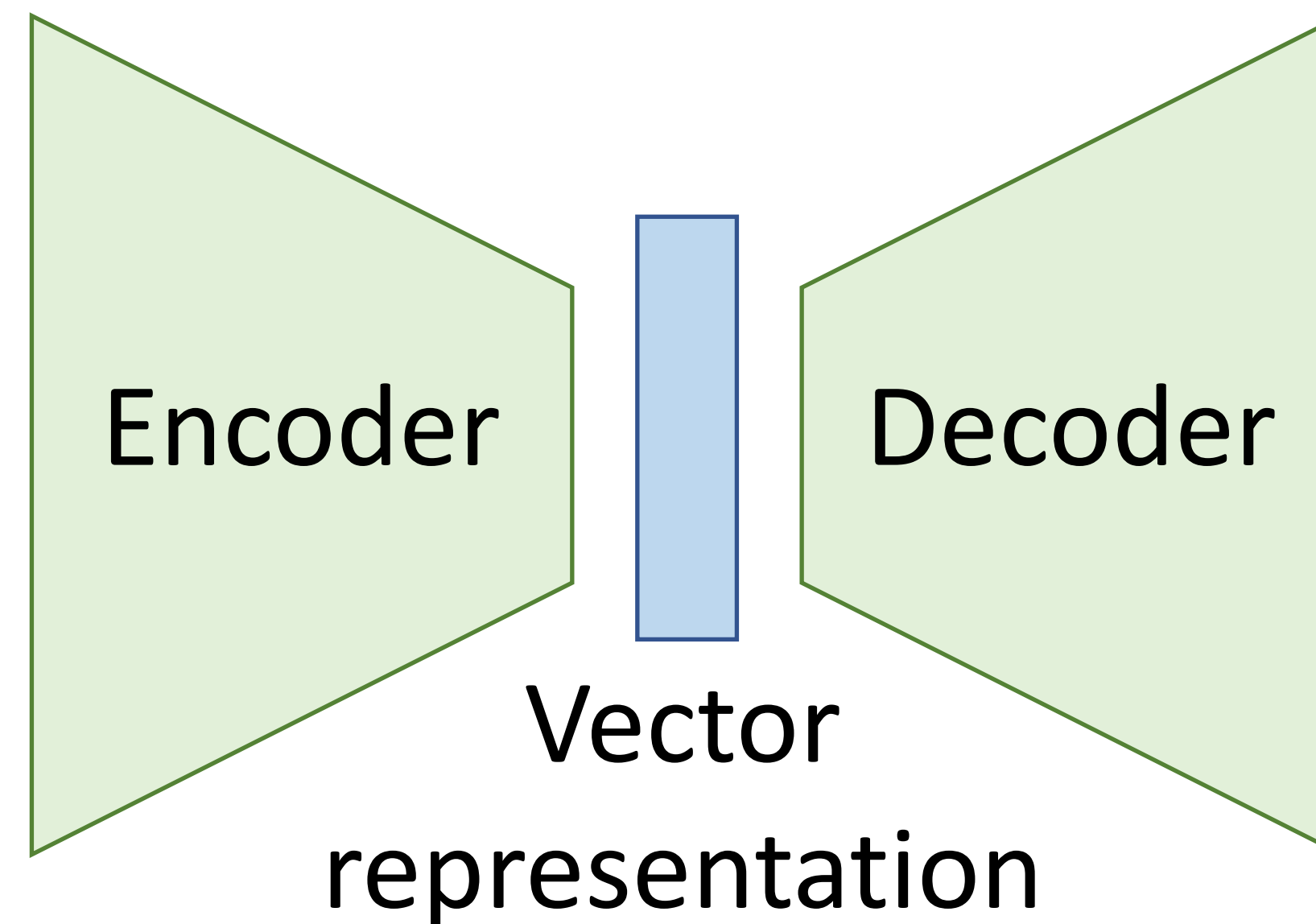
Upsamples to
obtain final output

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



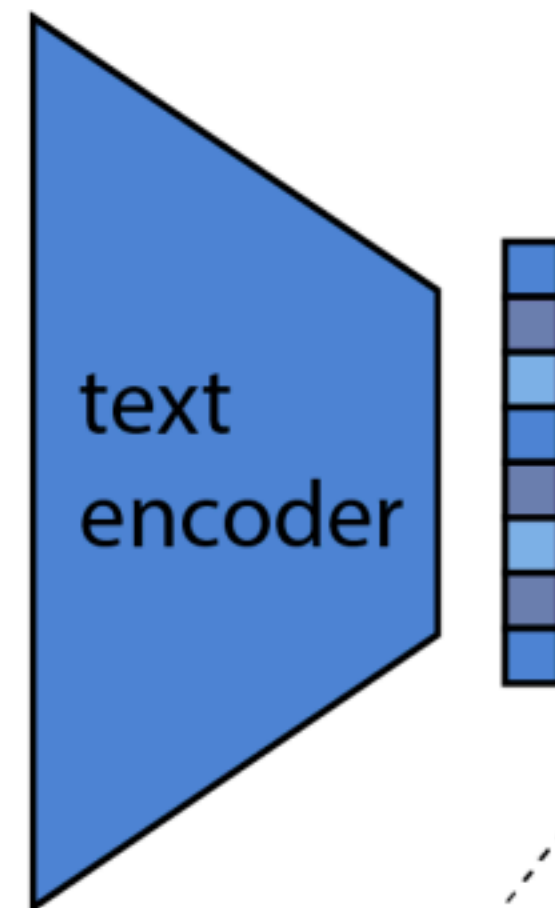
Images are represented as
sequence of tokens
(each image is encoded as 32x32
grid of tokens using discrete VAE to
8192 codewords)

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

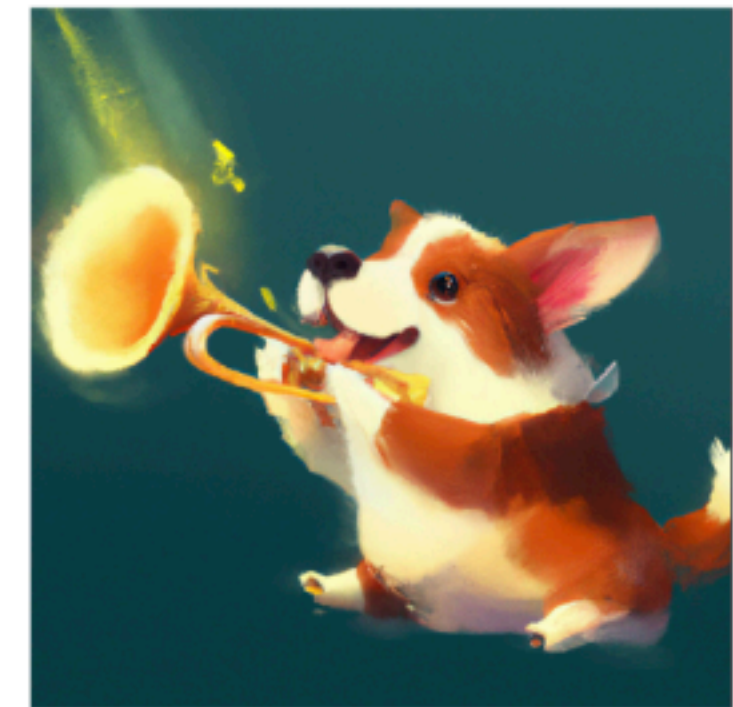
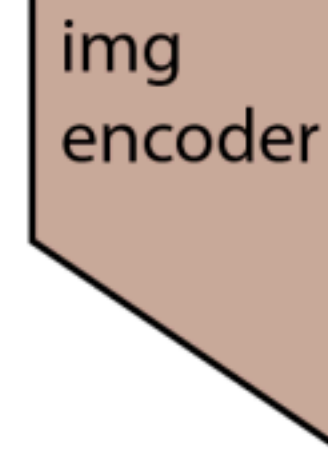
DALLE-2: Text-to-Image generation with diffusion models

CLIP text and image encoder

"a corgi
playing a
flame
throwing
trumpet"

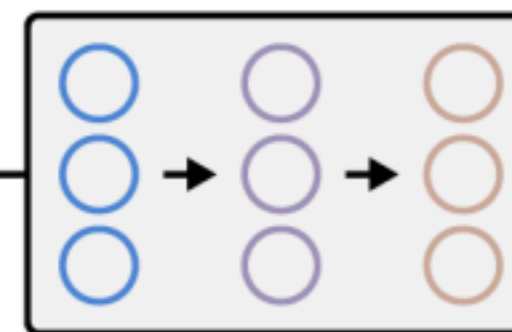
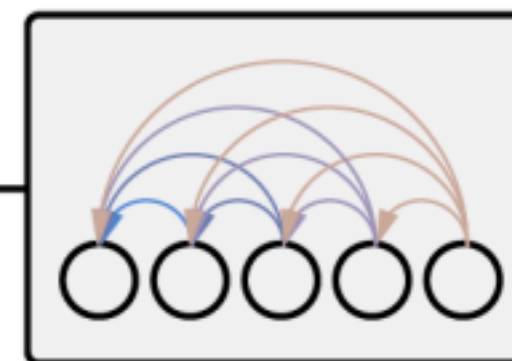


CLIP objective

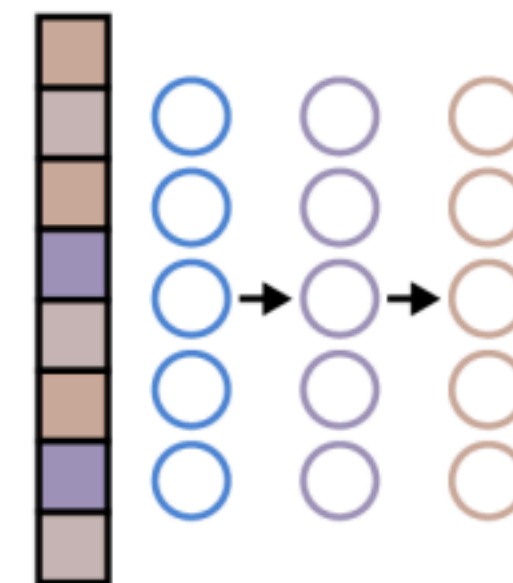


$$P(z_i|y)$$

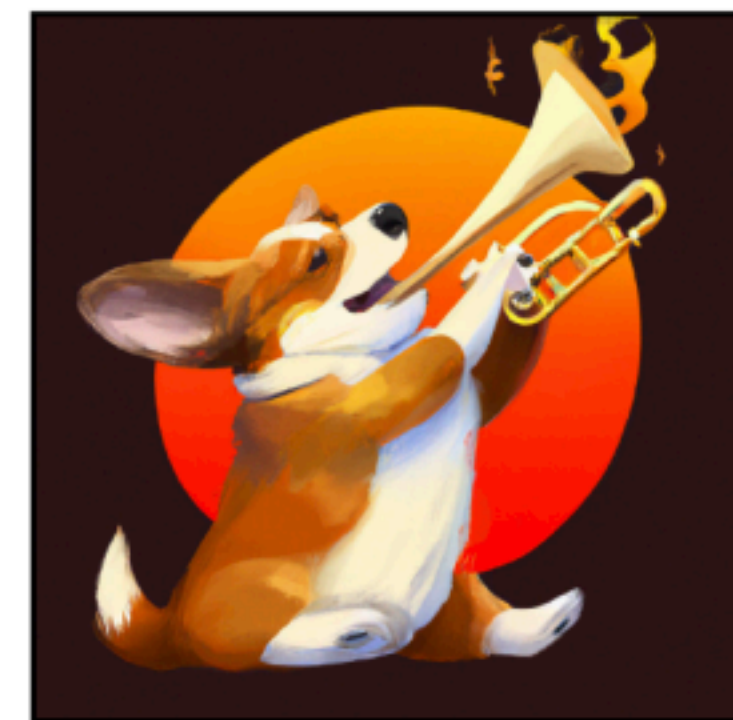
$$P(x|z_i, y)$$



prior



decoder



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

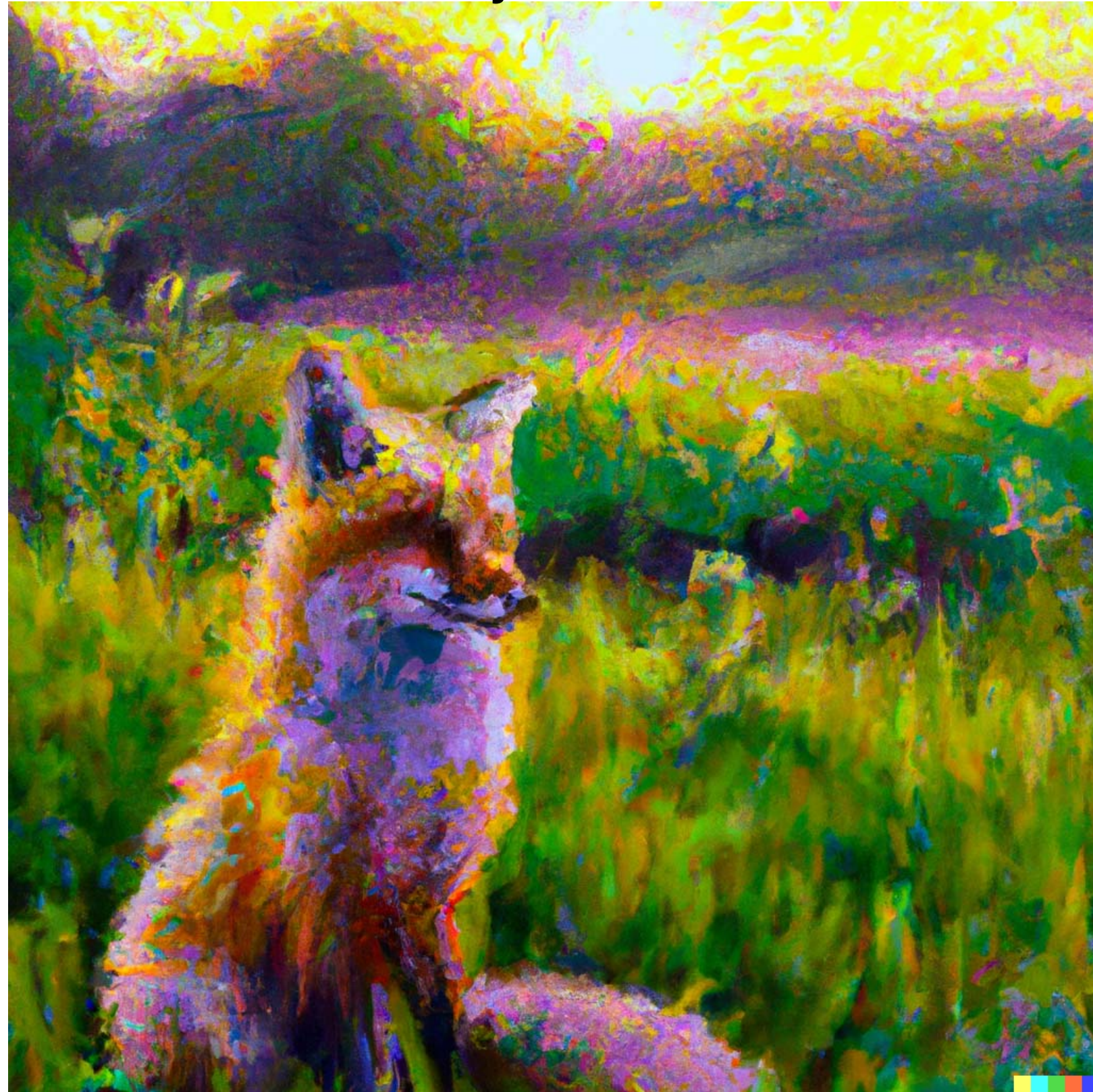
Hierarchical Text-Conditional Image Generation with CLIP Latents
<https://arxiv.org/pdf/2204.06125.pdf> [Ramesh et al, arXiv 2022]

<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/>

 **Imagen**

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



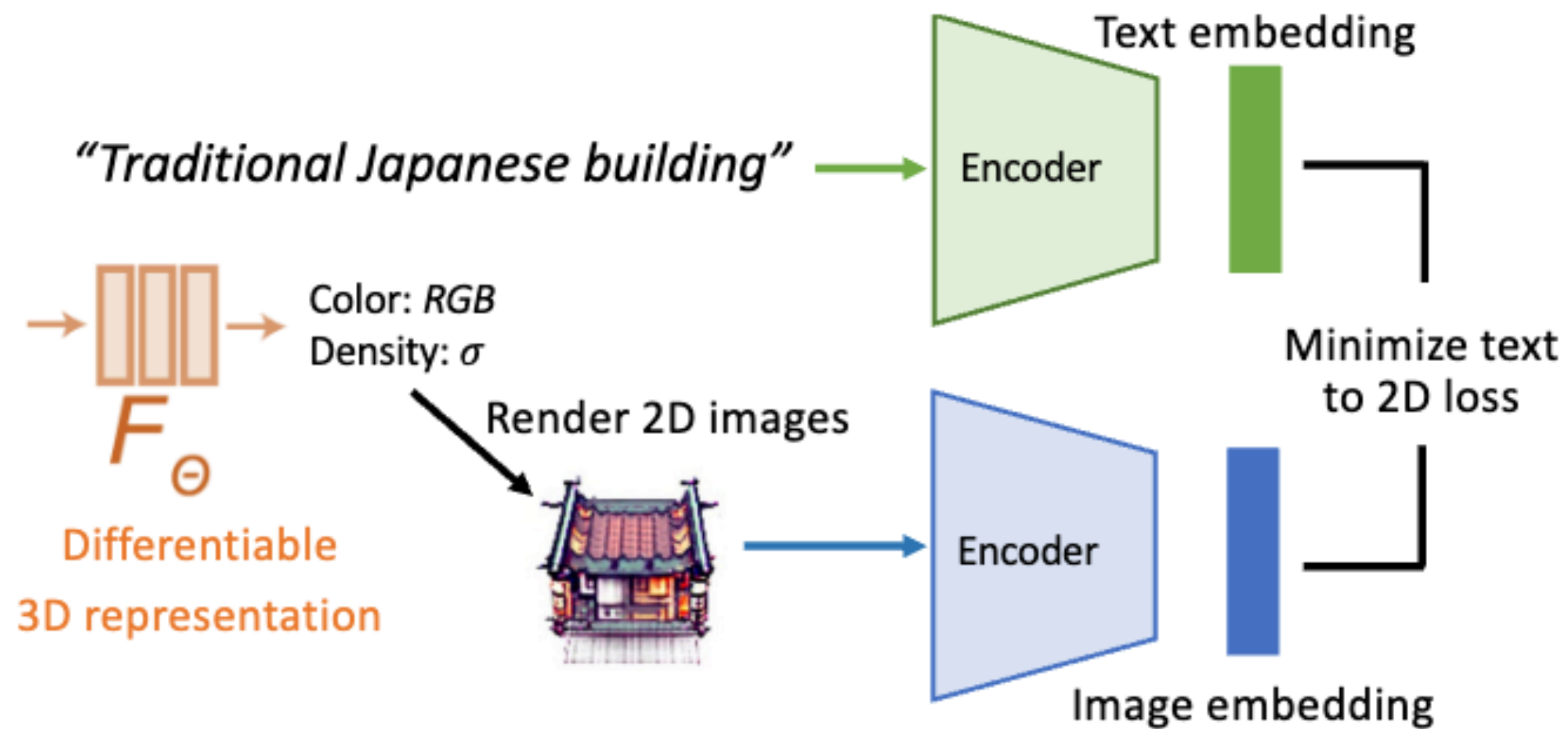
<https://imagen.research.google/>

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

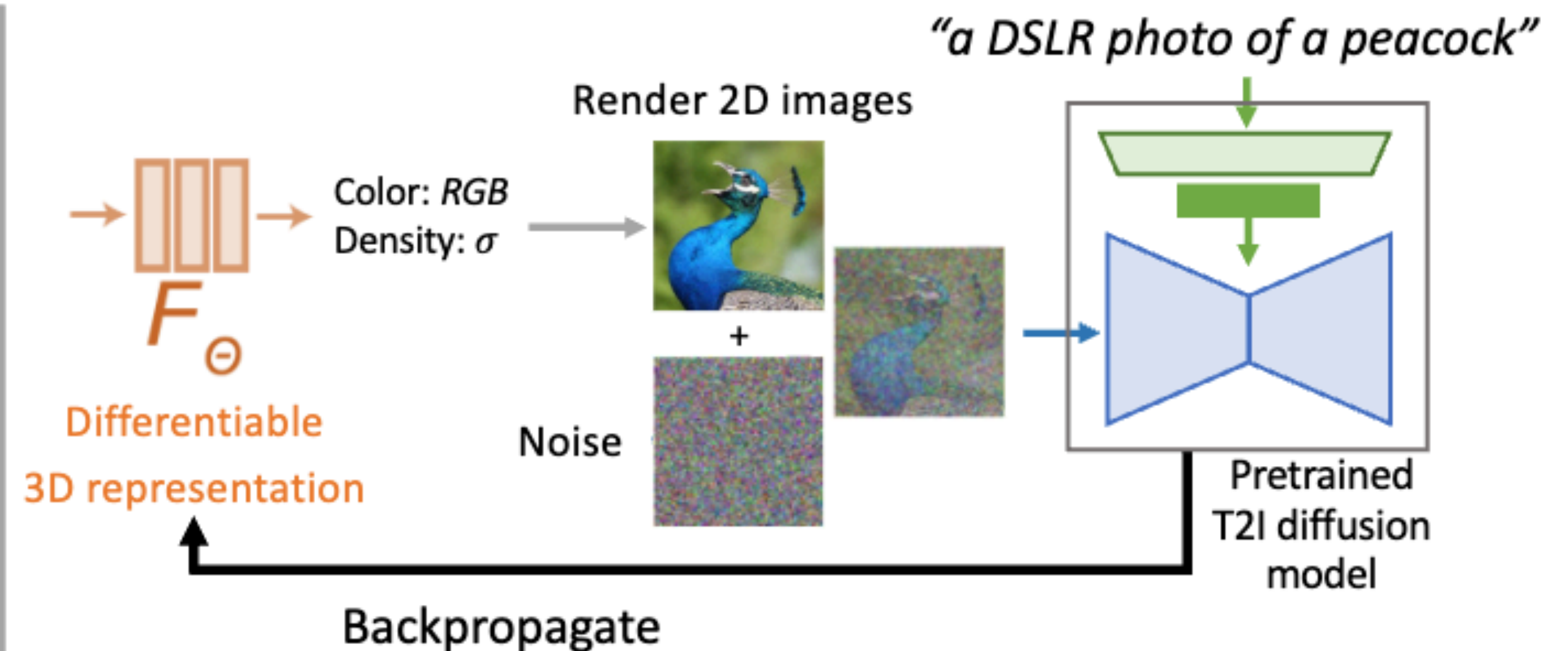
Text-to-3D generation

Optimize differentiable 3D representations with text-image models

Aligned text to image embeddings

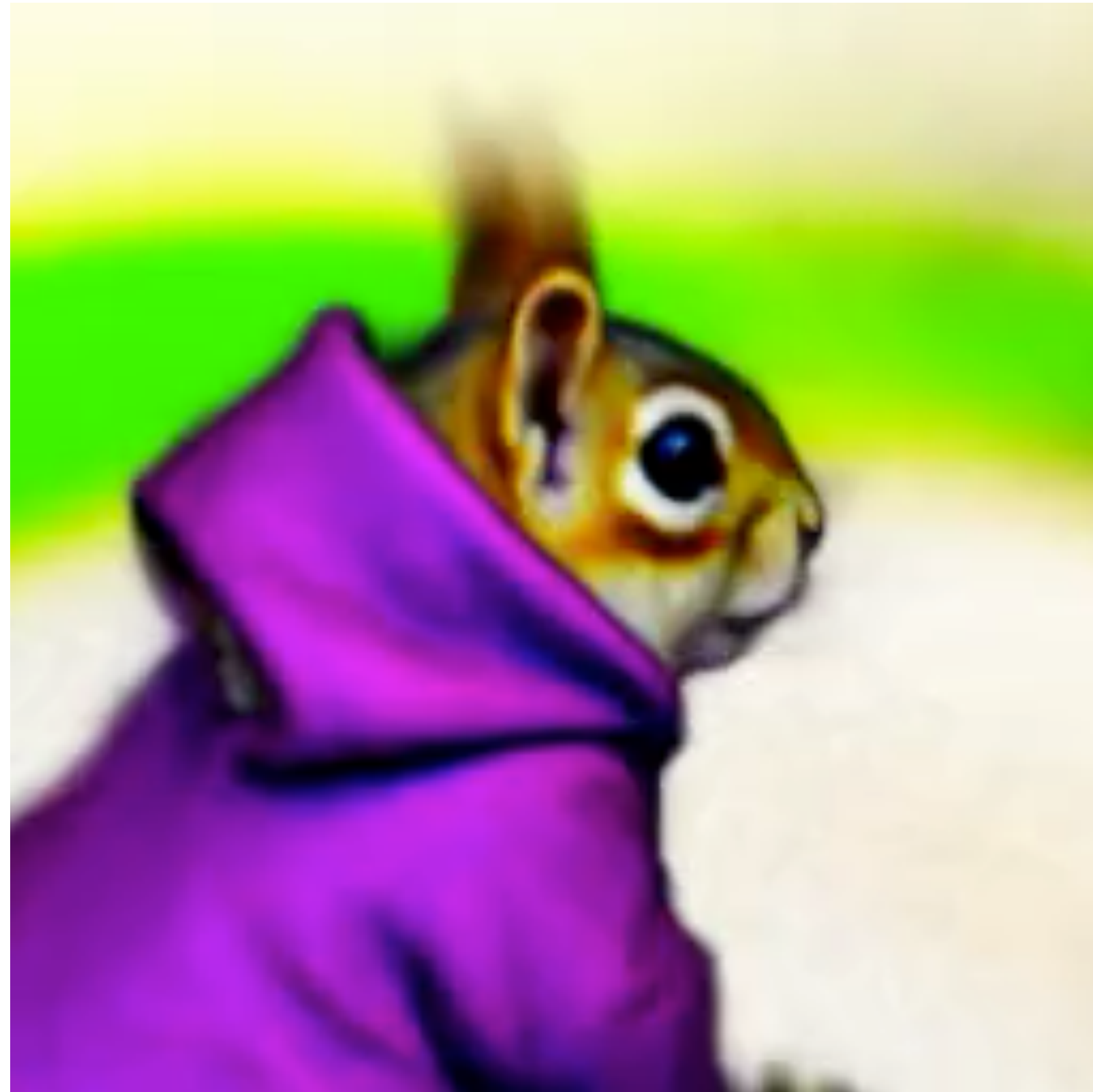


Text-to-image diffusion model



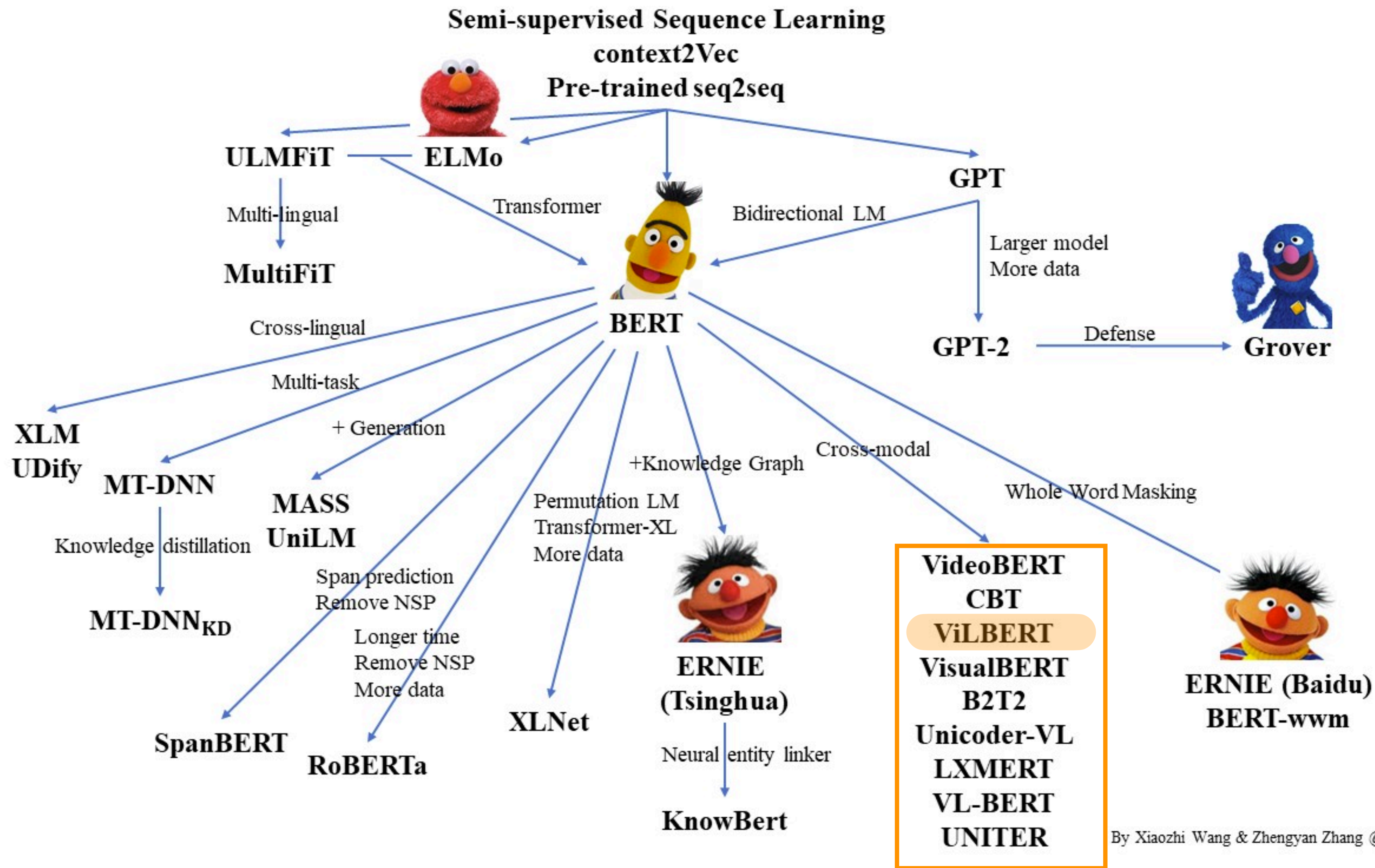
Text-to-3D with diffusion models

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/>

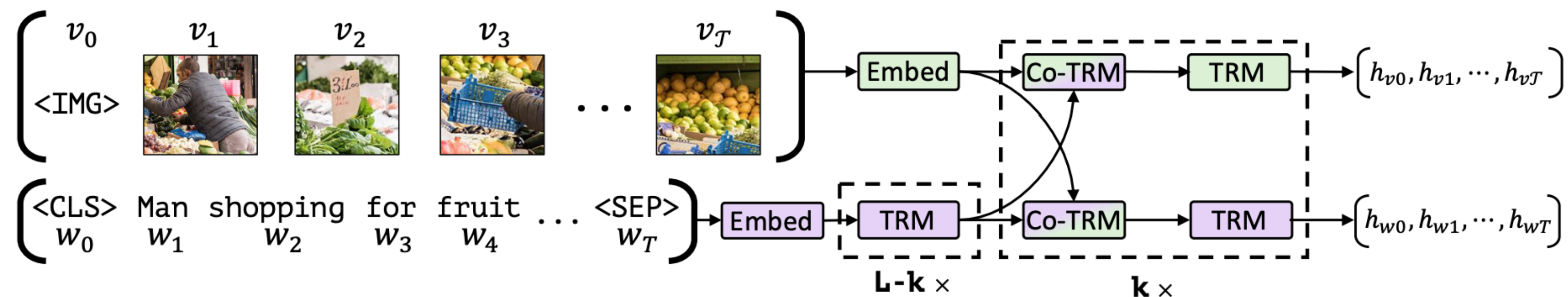
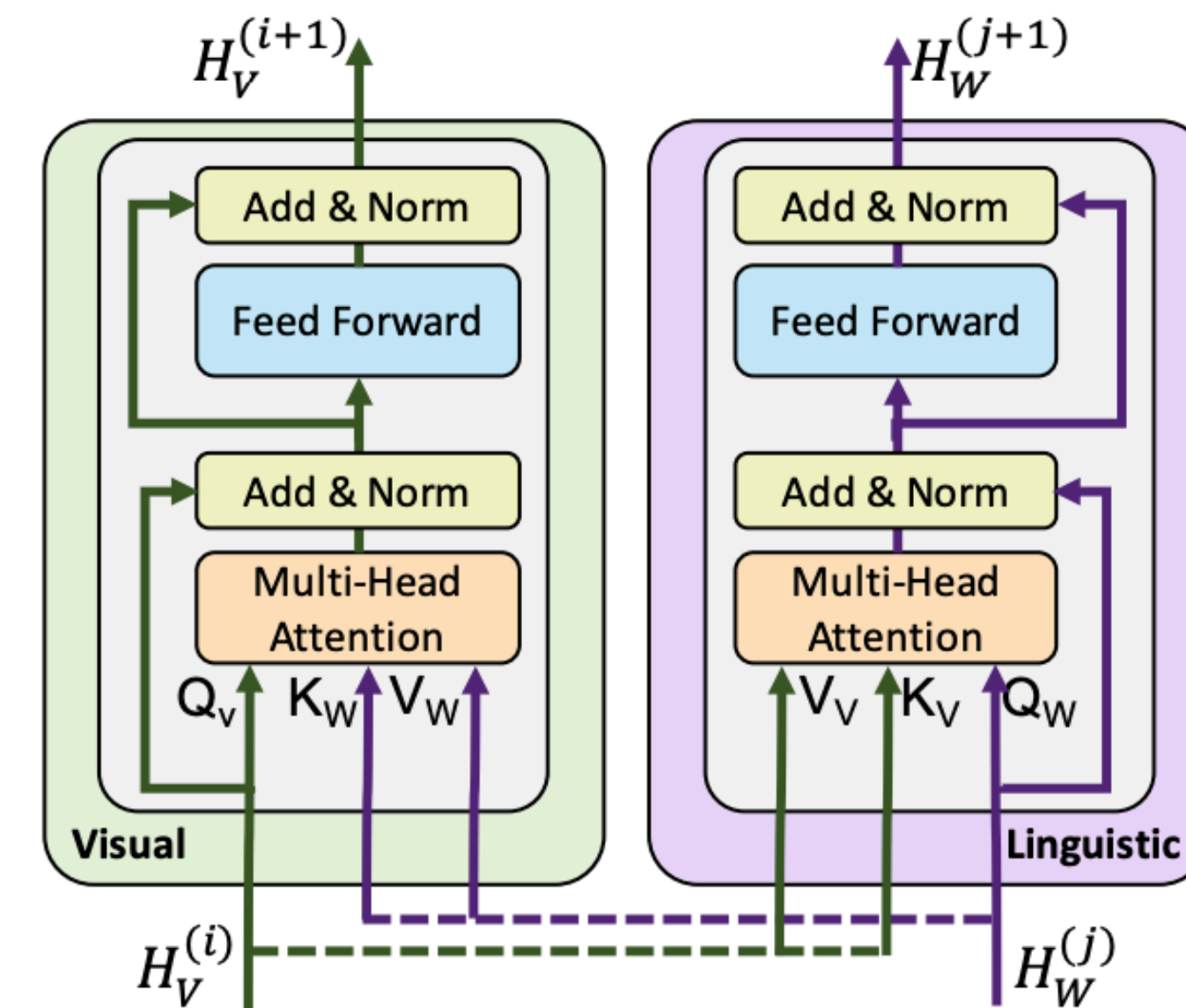
Beyond contrastive loss for multi-modal models



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

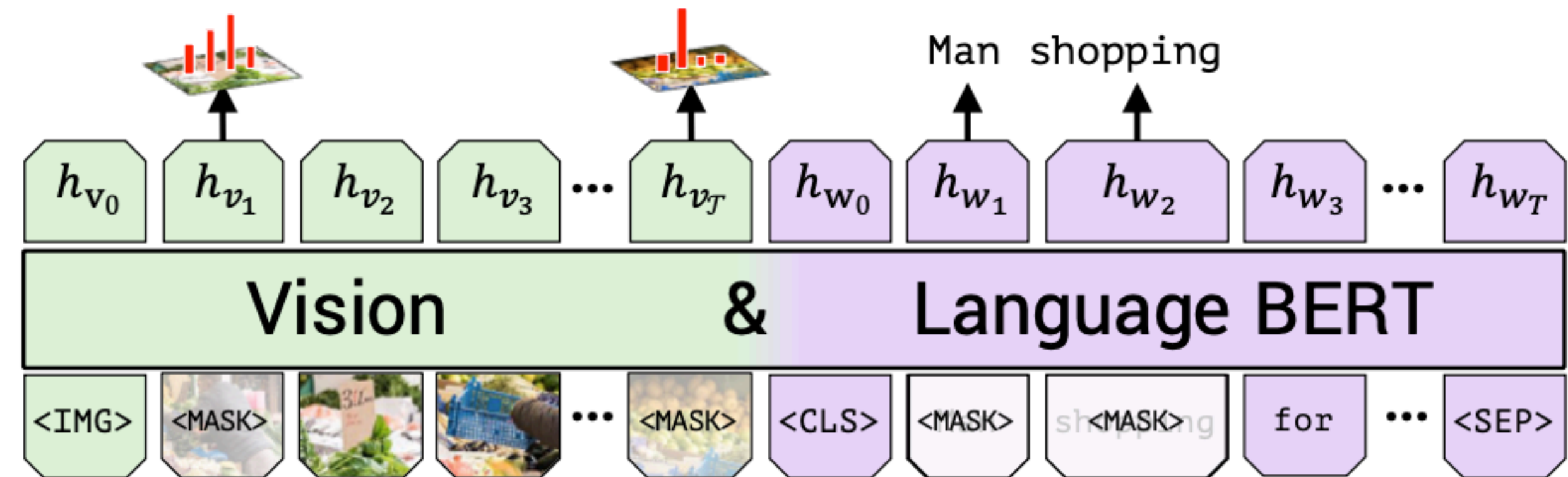


Pretrained representations for vision and language

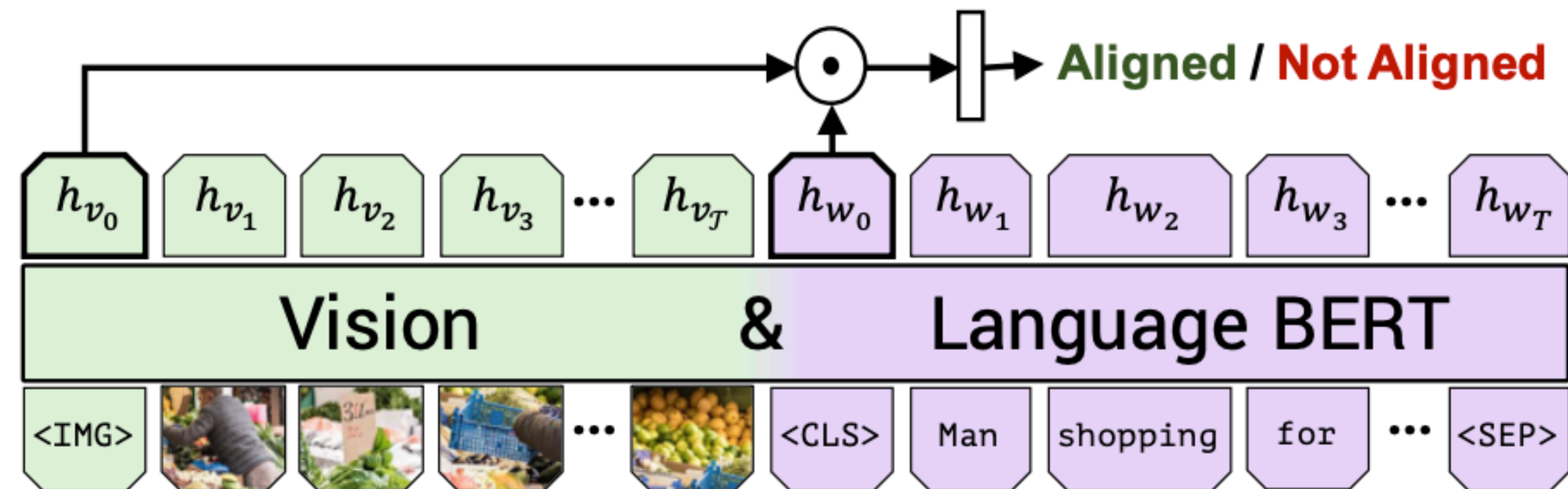
Predict semantic class distribution

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



Pretrained representations for vision and language

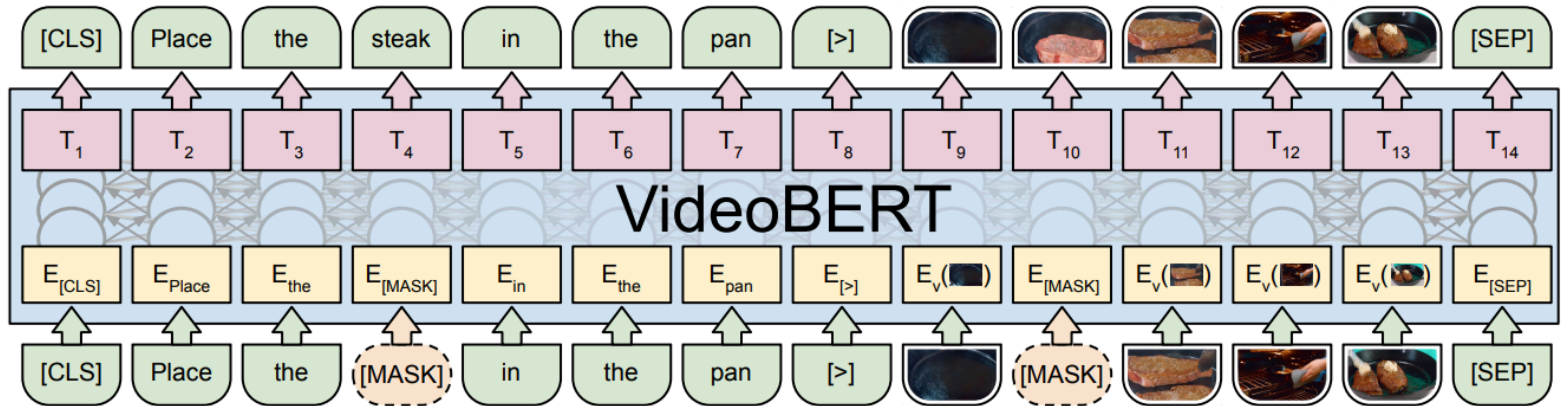
Method	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
	test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	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
	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

Pretraining improves performance on variety of vision+language tasks!



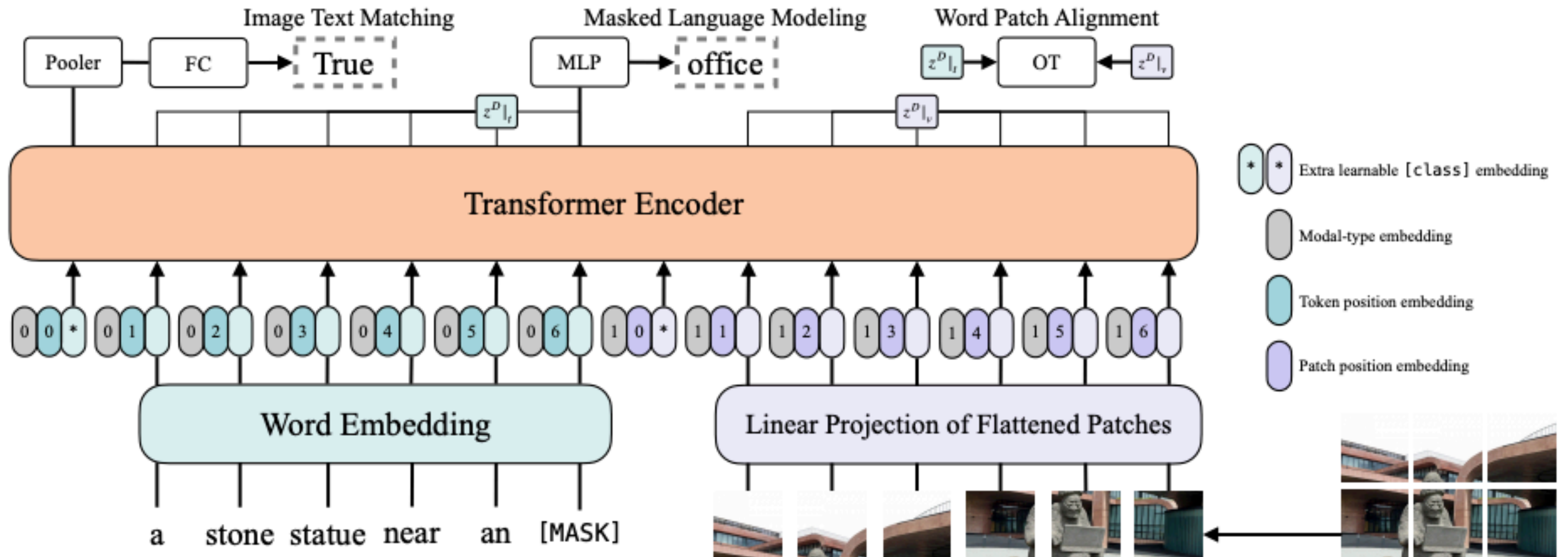
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
[Lu et al 2019, <https://arxiv.org/pdf/1908.02265.pdf>]

Masked modelling for video and language

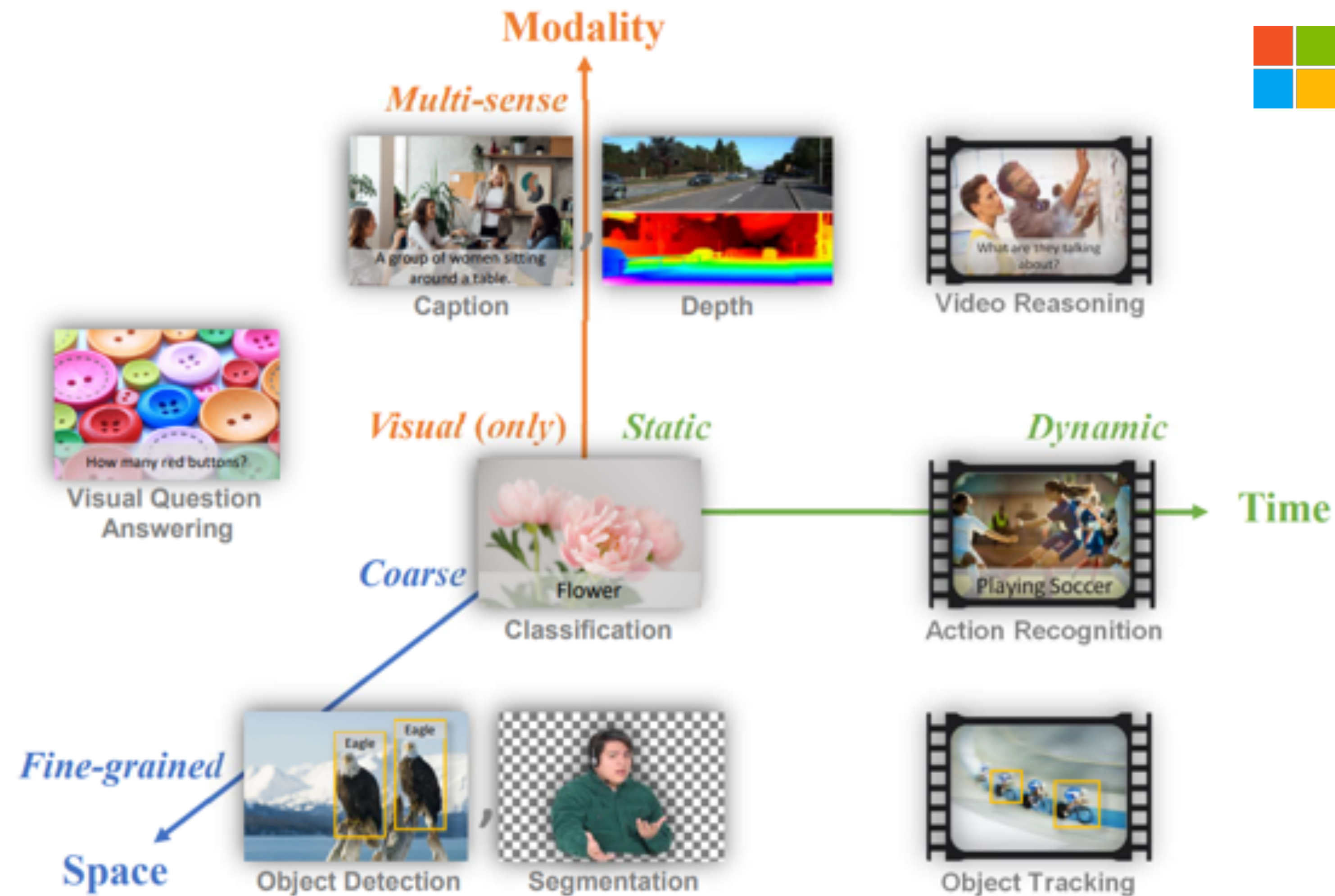


Combining masked modelling with contrastive learning

- Use image patches, no need for object detectors

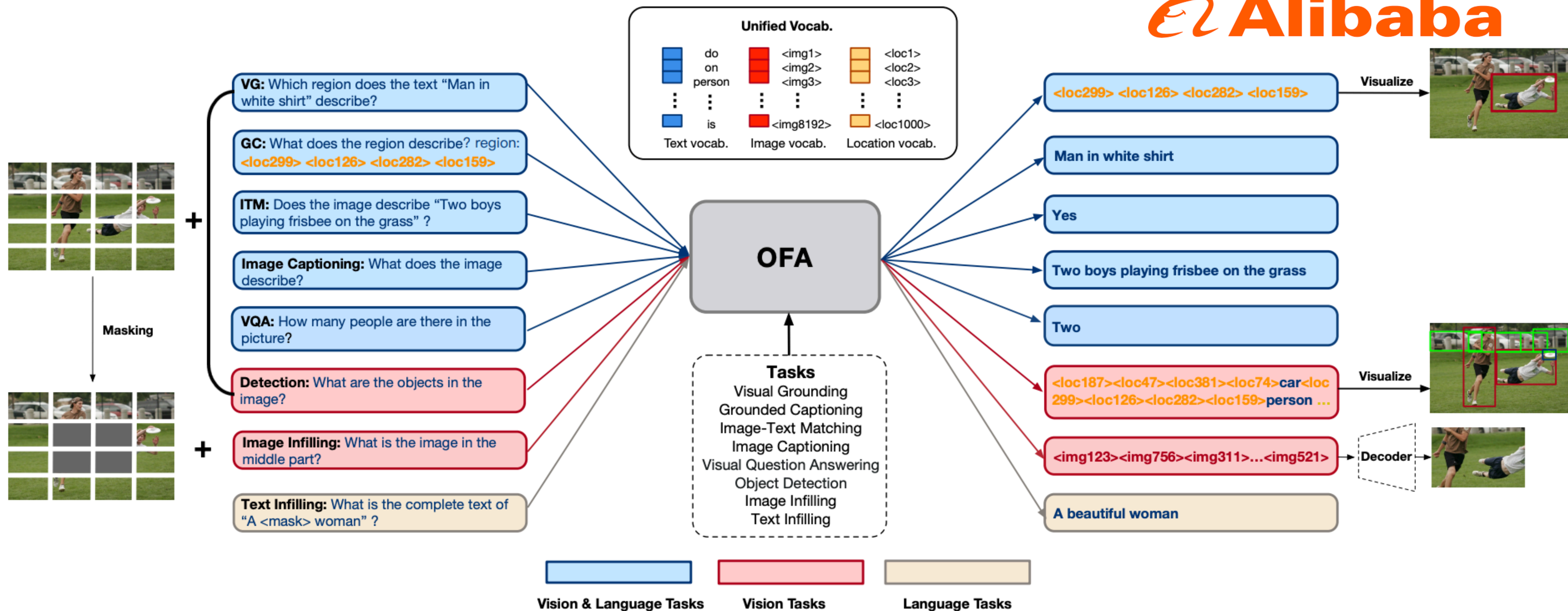


Large multi-modal, multi-lingual models: Florence



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

Large multi-modal models: OFA

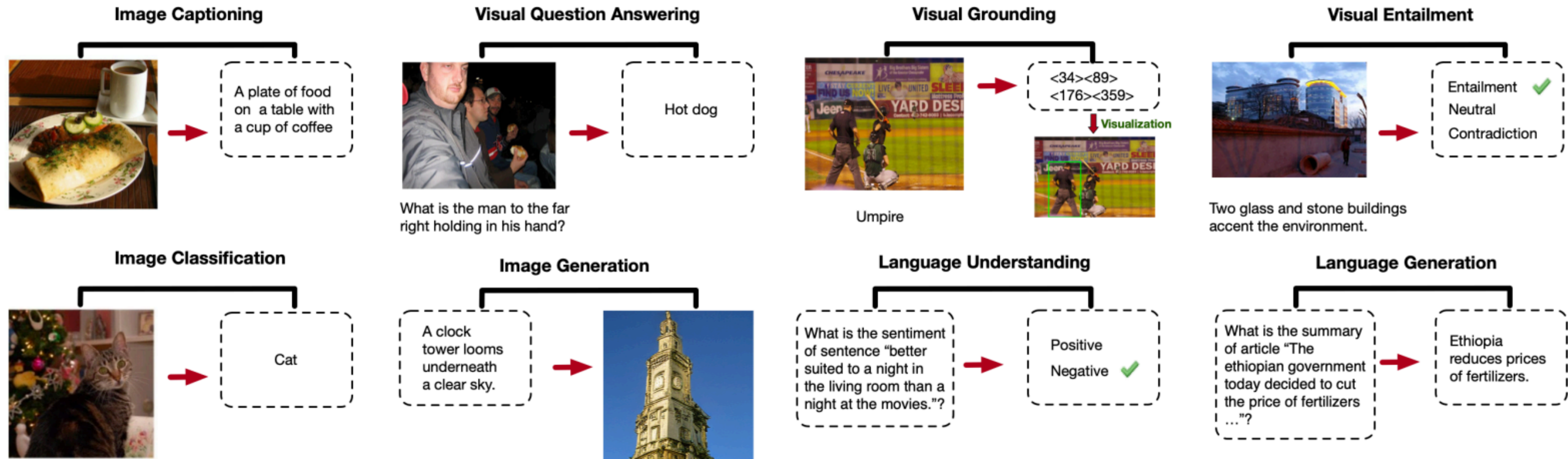


OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework [Wang, et al. ICML 2022] <https://arxiv.org/abs/2202.03052>

Large multi-modal models: OFA

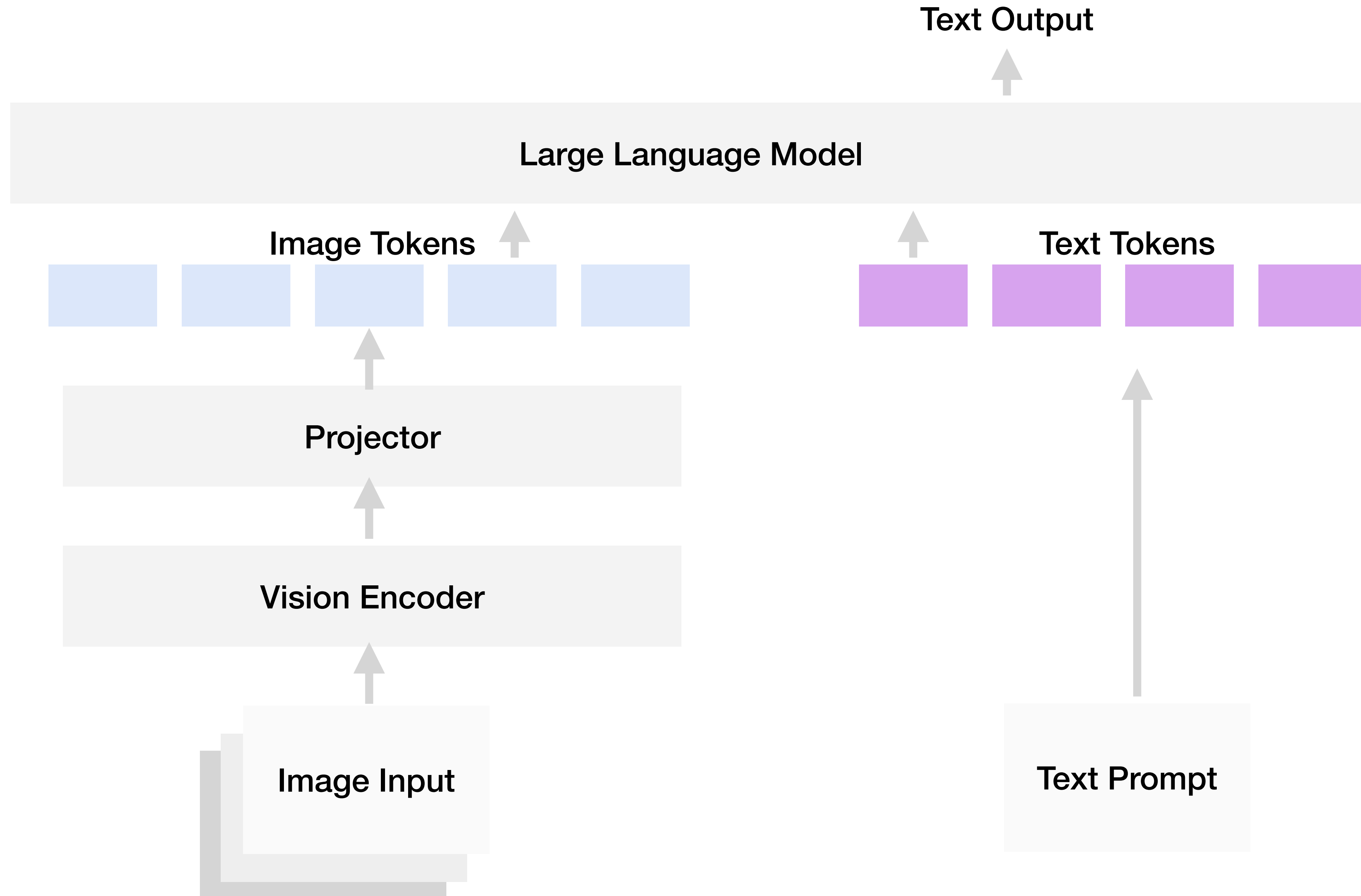


OFA



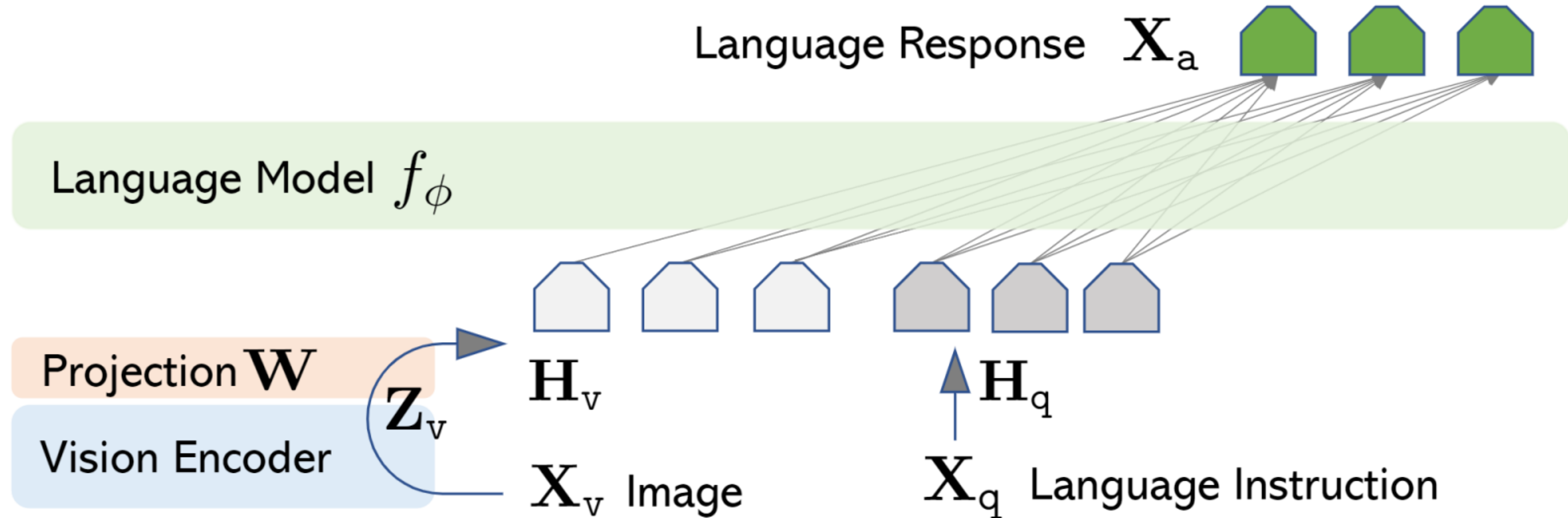
OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework
[Wang, et al. ICML 2022] <https://arxiv.org/abs/2202.03052>

Connecting with pre-trained LLMs



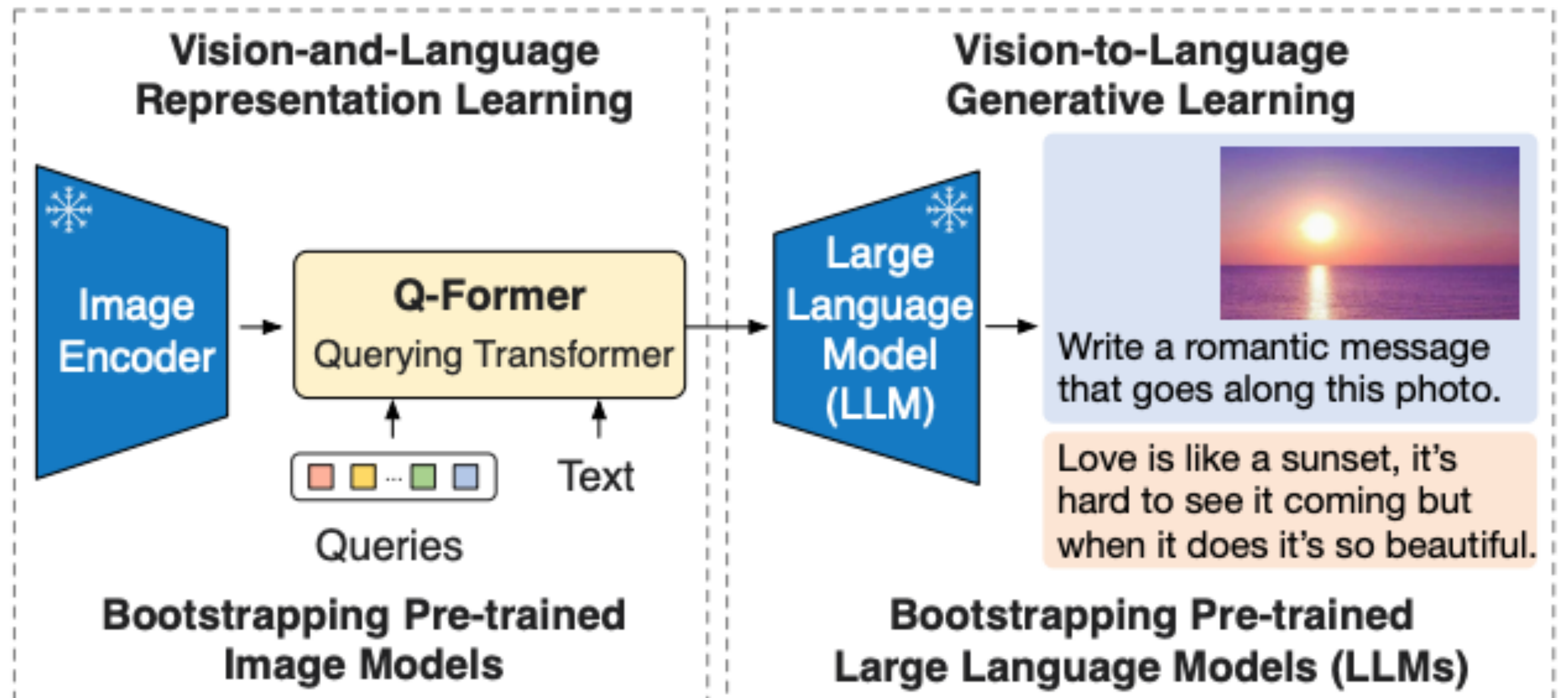
Connecting with pre-trained LLMs

LLaVA: Large Language and Vision Assistant



BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

- Frozen image encoder
- Frozen LLM
- Transformer and FFN to project from image to tokens to LLM

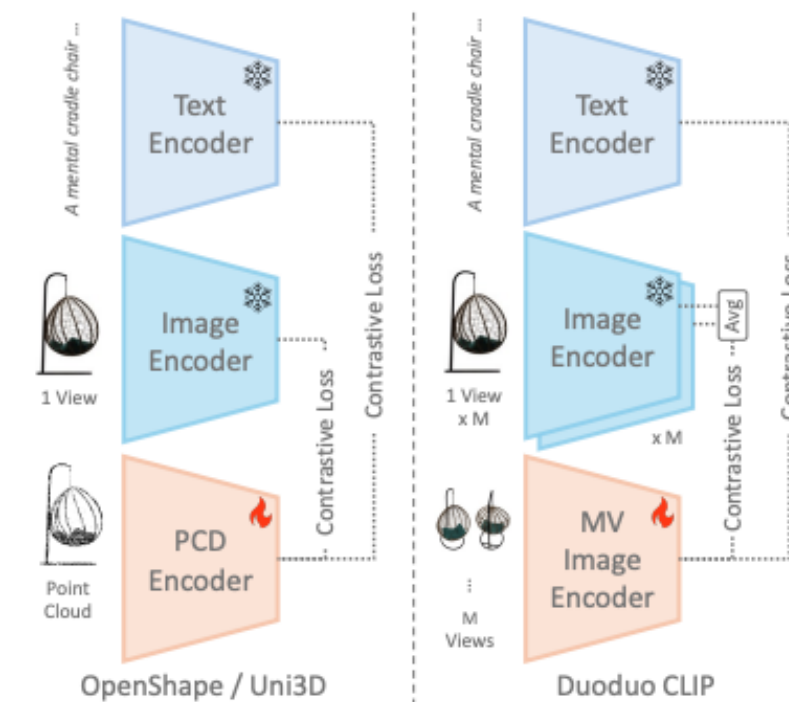


TA presentations

Text and 3D Representation Learning

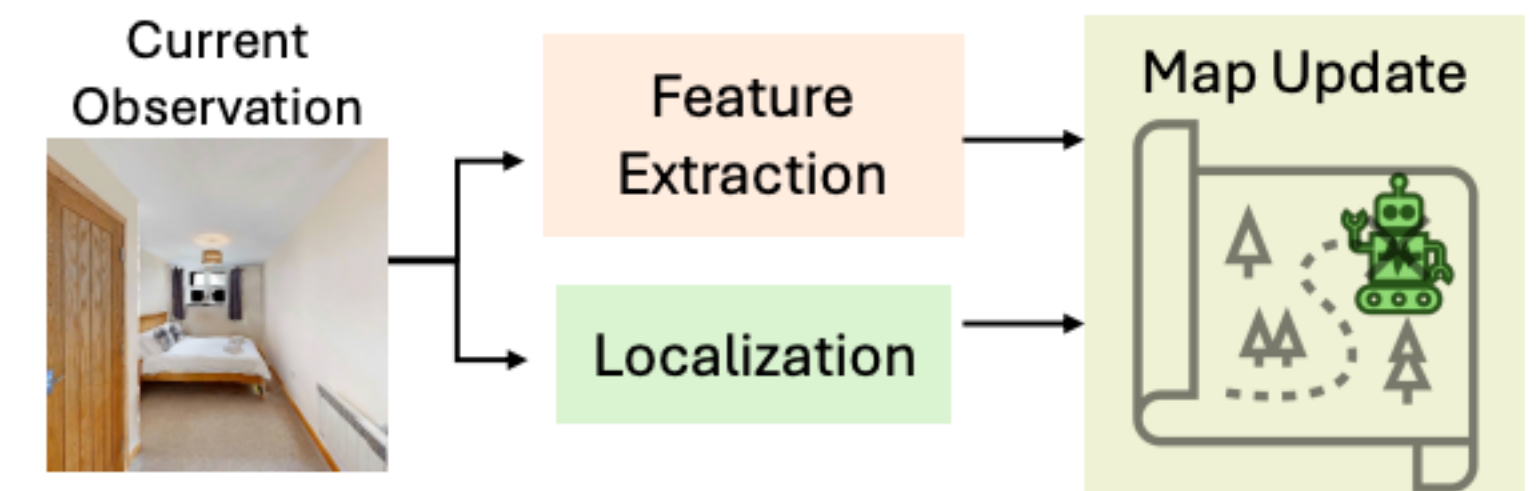


Qirui Wu



Sonia Raychaudhuri

Semantic Representation for Robots

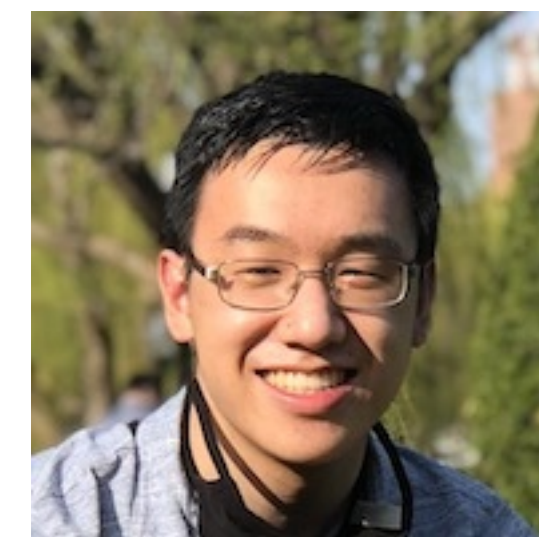


Text-to-Scene Generation

A 1b1b apartment of a researcher who has a cat.



Xiaohao Sun



Austin Wang

LLMs for modeling DNA

