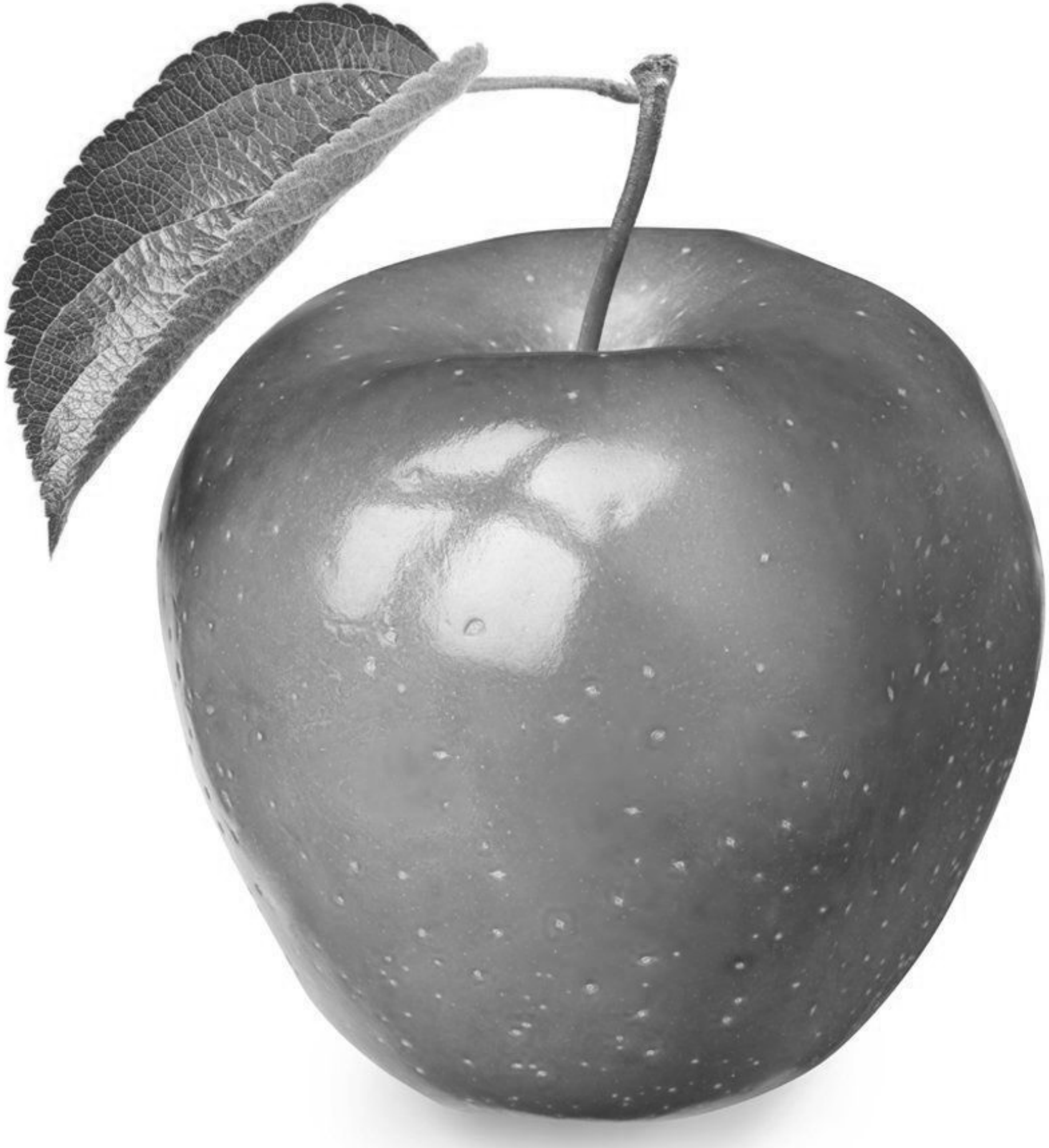
Vision and Language

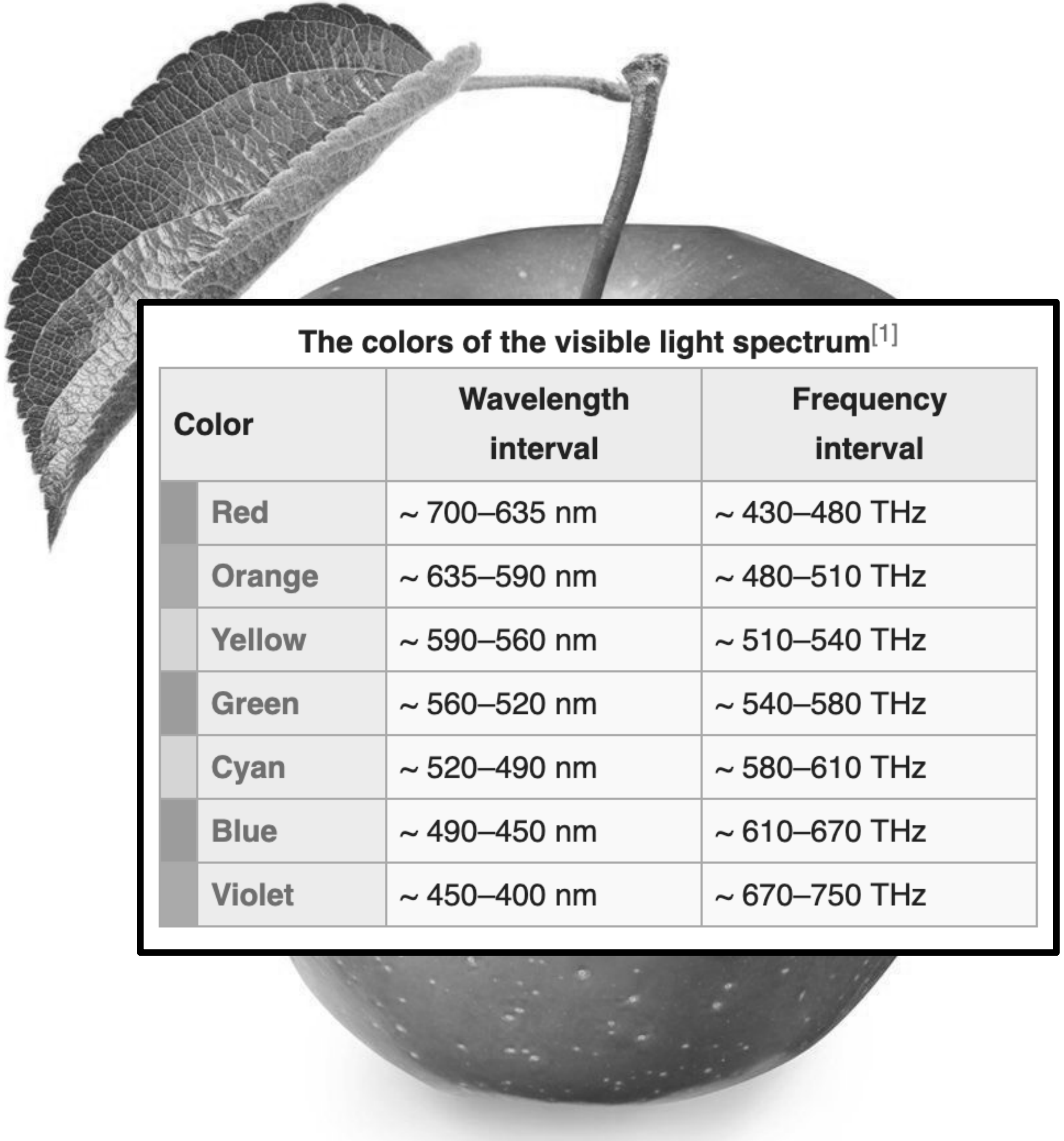
Berkeley NLP

Rodolfo (Rudy) Corona

with thanks to Daniel Fried

CS 288, 4/12/2022





The colors of the visible light spectrum[1]

Color	Wavelength interval	Frequency interval
Red	~ 700–635 nm	~ 430–480 THz
Orange	~ 635–590 nm	~ 480–510 THz
Yellow	~ 590–560 nm	~ 510–540 THz
Green	~ 560–520 nm	~ 540–580 THz
Cyan	~ 520–490 nm	~ 580–610 THz
Blue	~ 490–450 nm	~ 610–670 THz
Violet	~ 450–400 nm	~ 670–750 THz

Source: “Color” in Wikipedia

“Apples are red”

“The numbers this month are in the red”

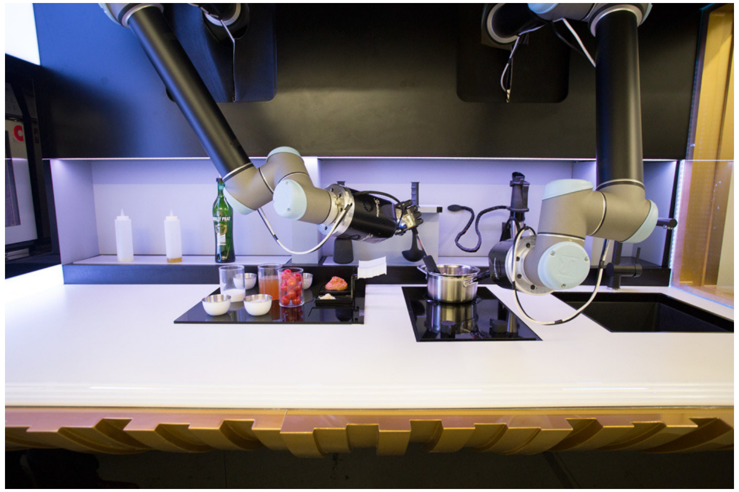
“Red has a wavelength between 635-700nm”

…

“Pixel (1,1) has R=240, pixel (1,2) has …”

What is Language Grounding?

- Tying language to non-linguistic things (e.g. a database in semantic parsing)
- The world only looks like a database some of the time!
- Some settings depend on grounding into *perceptual* or *physical* environments:



"Add the tomatoes and mix"

"Take me to the shop on the corner"

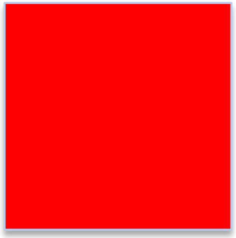
- **Focus today**: Grounding language to *visual perception*.

Grounding

- (Some) possible things to ground into:

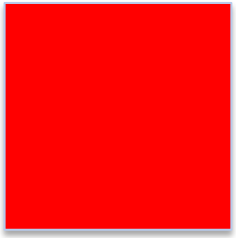
Grounding

- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…



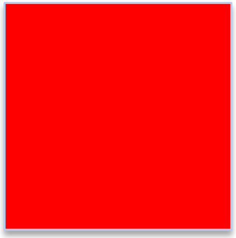
Grounding

- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern



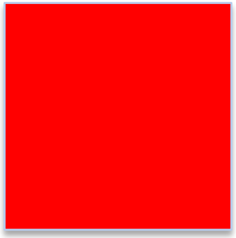
Grounding

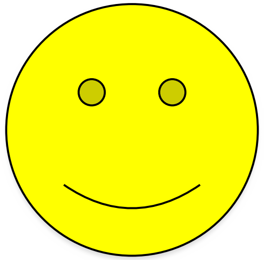
- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern
  - **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation



Grounding

- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern
  - **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation
  - **Social (effects on others)**: polite language is correlated with longer forum discussions





Grounding

- (Some) possible things to ground into:
  - **Low-level percepts**: *red* means this set of RGB values, *loud* means lots of decibels on our microphone, *soft* means these properties on our haptic sensor…
  - **High-level percepts**: *cat* means this type of pattern
  - **Embodiment (effects on the world)**: *go left* means the robot turns left, *speed up* means increasing actuation
  - **Social (effects on others)**: polite language is correlated with longer forum discussions

For a nice taxonomy, related work, and examples, see *Experience Grounds Language* [Bisk et al. 2020]

Grounding

- (Some) key problems:
  - **Representation**: matching low-level percepts to high-level language (pixels vs *cat*)
  - **Abstraction and Composition**: meaning as a combination of parts
  - **Alignment**: aligning parts of language and parts of the world
  - **Content Selection and Context**: what are the important parts of the environment?
  - **Balance**: it's easy for multi-modal models to "cheat", rely on imperfect heuristics, or ignore important parts of the input
  - **Generalization**: to novel world contexts / input combinations

CS294-43: VISION AND LANGUAGE AI SEMINAR

A Gallery of Tasks

Image Captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Microsoft COCO Captions: Chen et al. 2015

Visual Question Answering

What is the dog wearing?

life jacket

collar

How many skiers are there?

2

1

What number is on the train?

7907

8551

What is sitting in the window?

bird

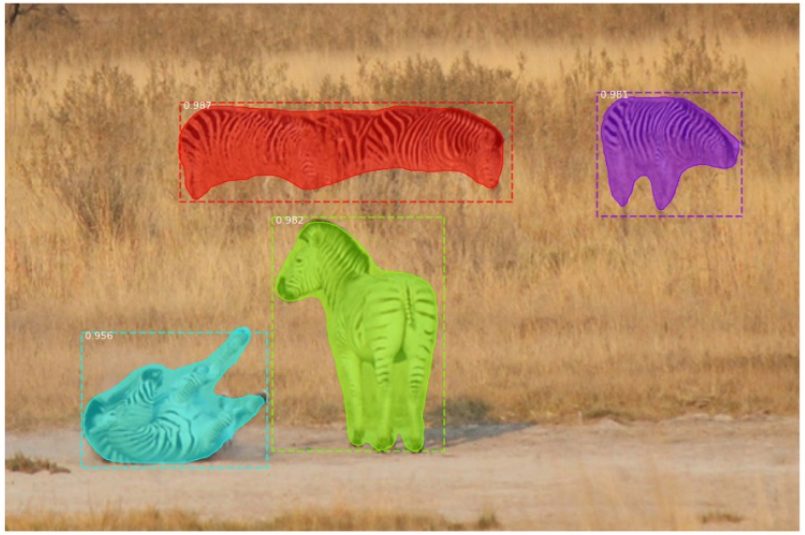
clock

VQA 2.0: Goyal et al. 2017

Object Detection (2D)

(a) Query: “street lamp”

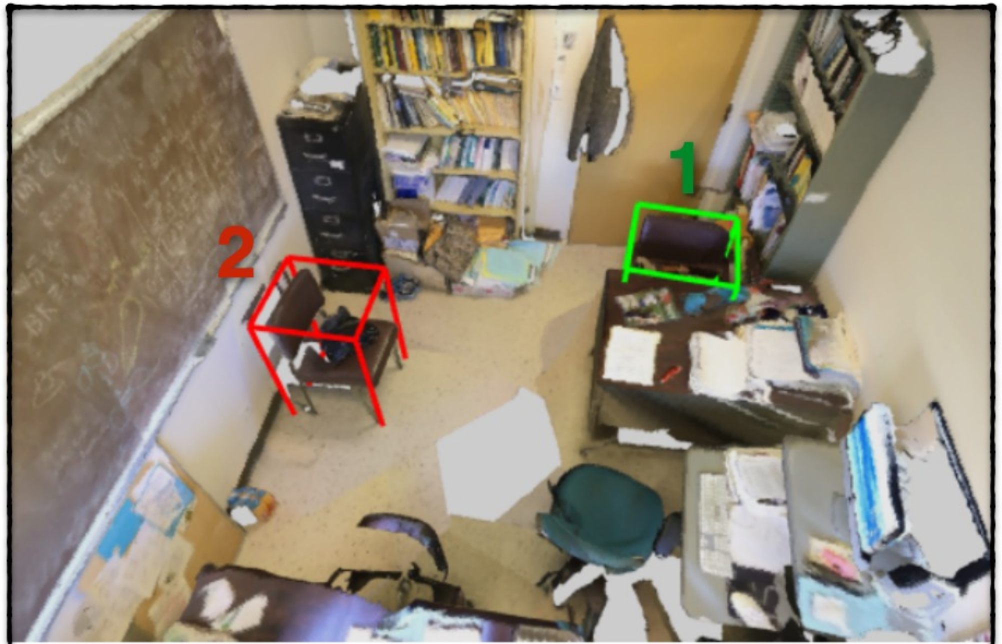
(b) Query: “major league logo”



(c) Query: “zebras on savanna”

MDETR: Kamath et al. 2021

Object Detection (3D)



1. "The chair closest to the door."
2. "The chair under the chalkboard."

1. "The office chair that is green."
2. "Choose the brown office chair pushed under the desk."

ReferIt3D: Achlioptas et al. 2020

Conditional Generation (2D)

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

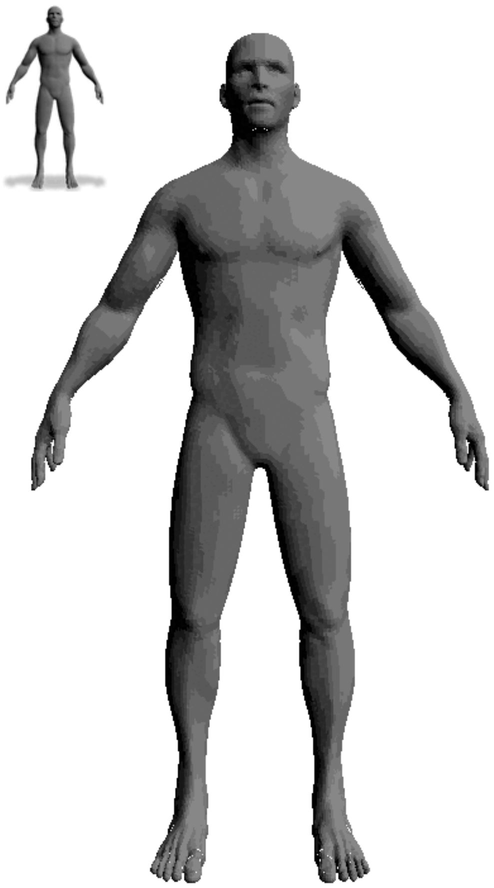
an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

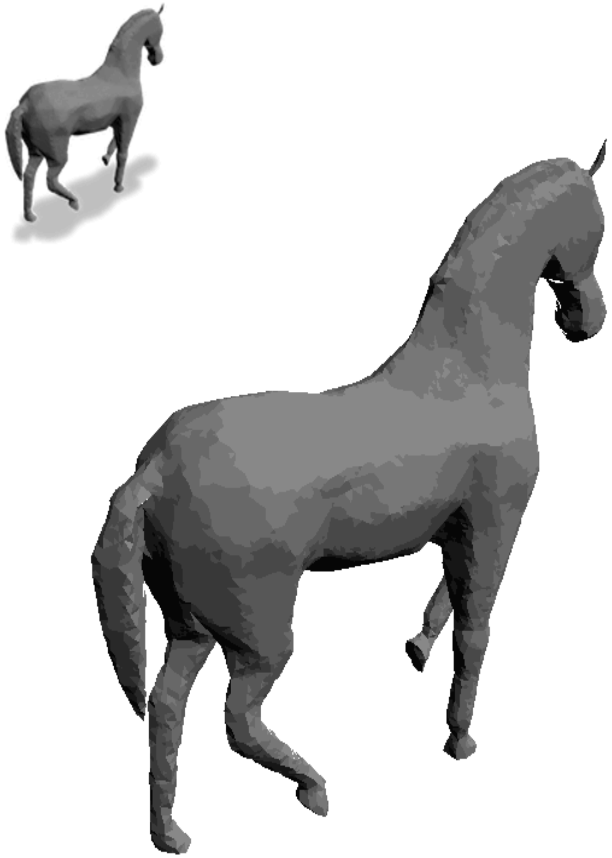
a corgi's head depicted as an explosion of a nebula

DALL-E 2: Ramesh et al. 2022

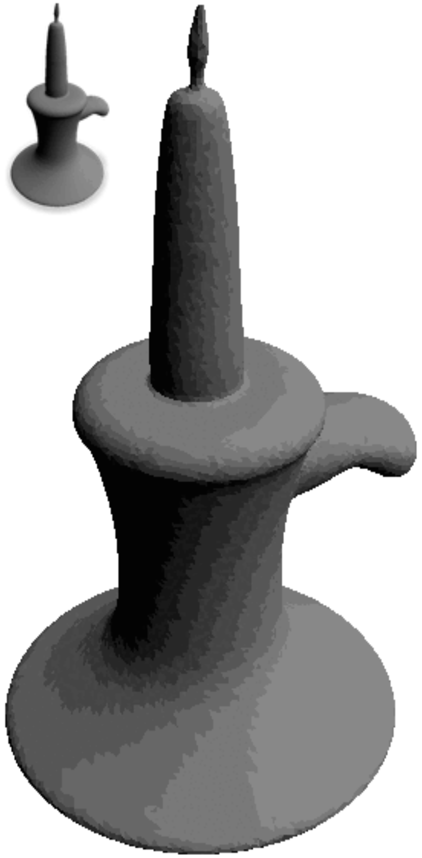
Conditional Generation (3D)



"Iron Man"



"Astronaut Horse"



"Colorful Crochet Candle"

Text2Mesh: Michel et al. 2021

Vision and Language Navigation

"Place a clean ladle on a counter"

ALFRED: Shridhar et al. 2020

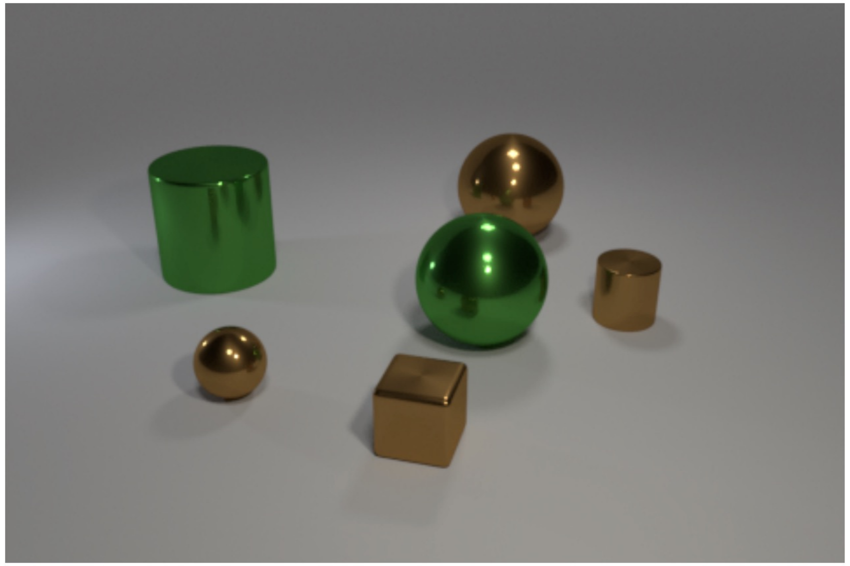
Why Grounded Language?

- Much language refers *to the world*.
- Convenient medium to communicate with machines!
- For many tasks, agents will need perceptual understanding and motor control for this interaction.

Source: Robohub.org

Source: cnn.com

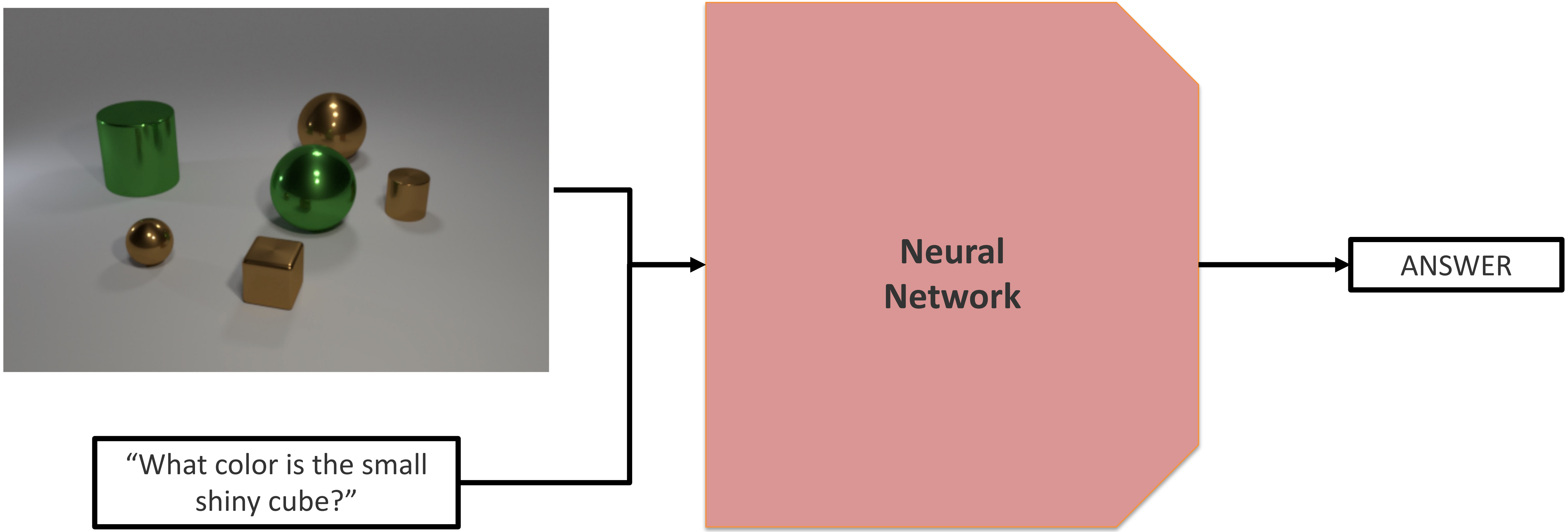
Bottom-Up & Top-Down Reasoning



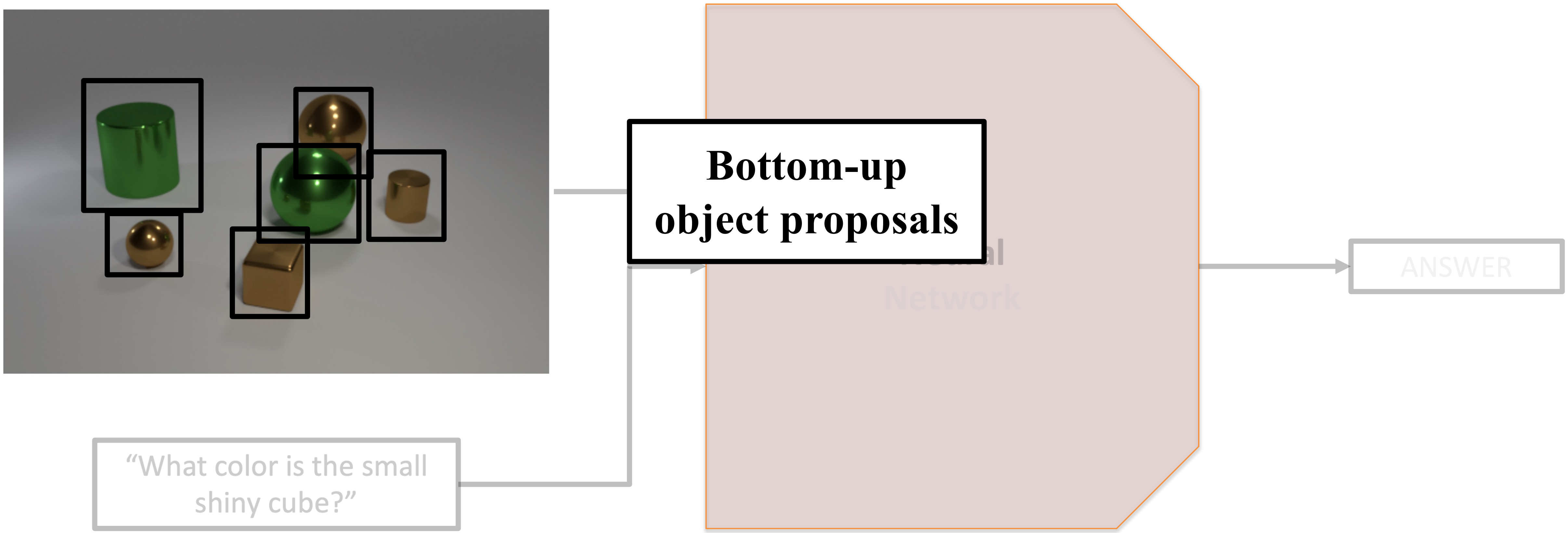
“What color is the small shiny cube?”

CLEVR: Johnson et al. 2016

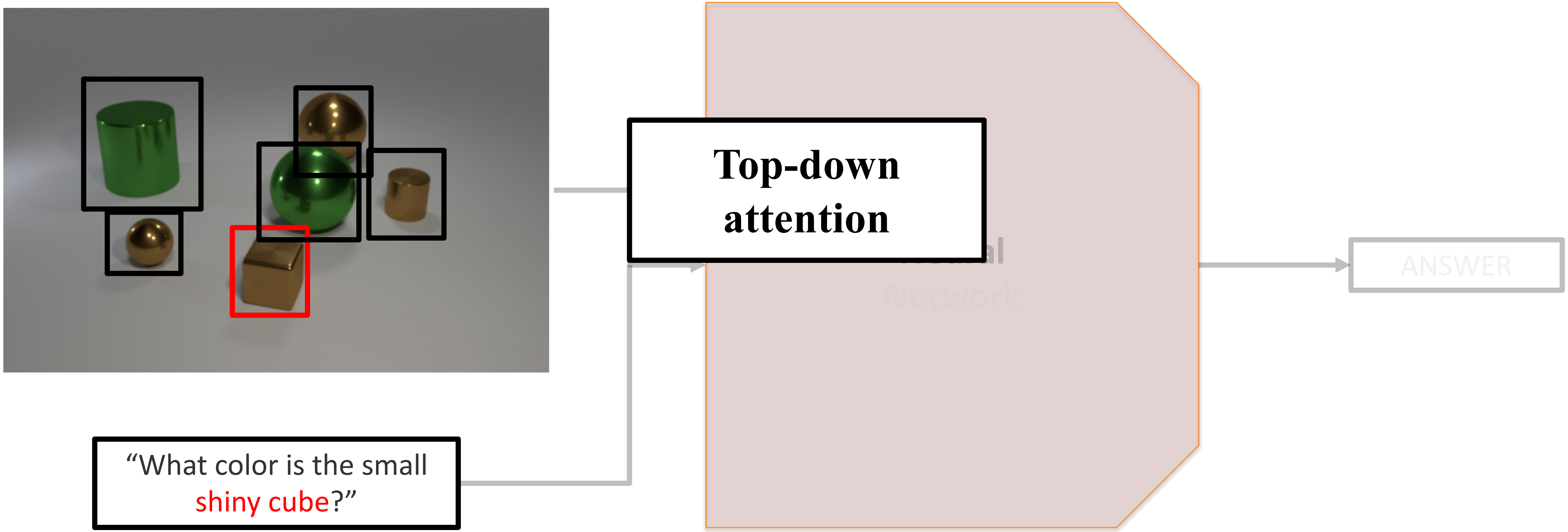
Bottom-Up & Top-Down Reasoning



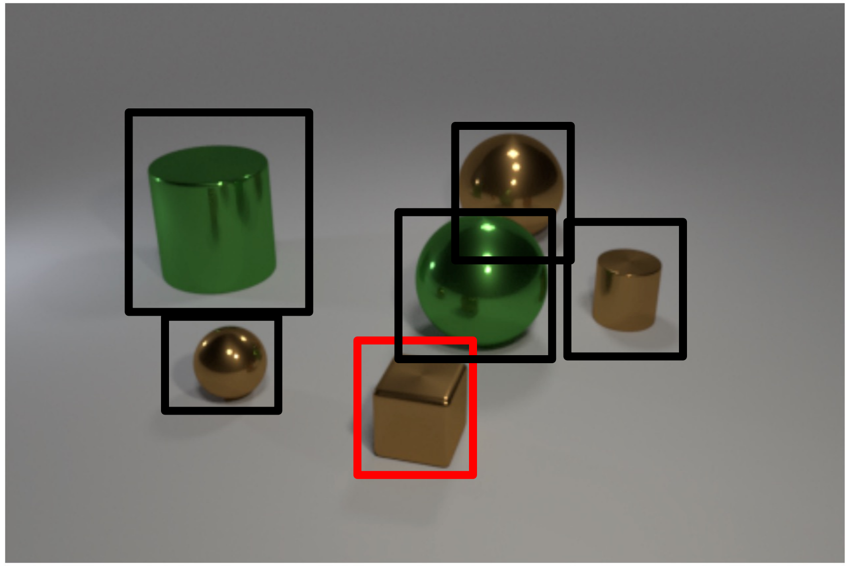
Bottom-Up & Top-Down Reasoning



Bottom-Up & Top-Down Reasoning



Bottom-Up & Top-Down Reasoning



“What color is the small shiny cube?”

Neural Network

ANSWER

Can answer by focusing on single object

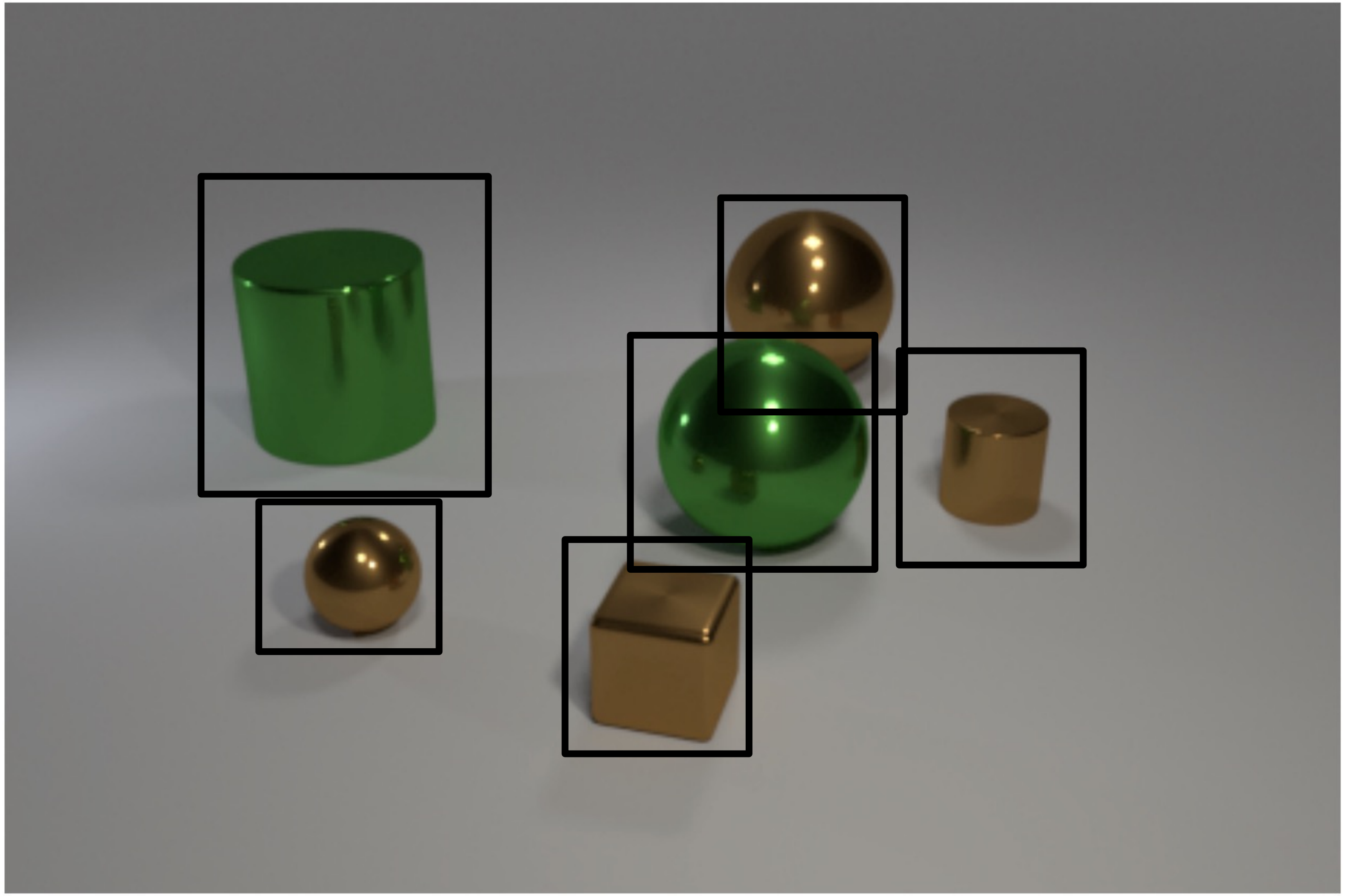
Bottom-Up & Top-Down Reasoning

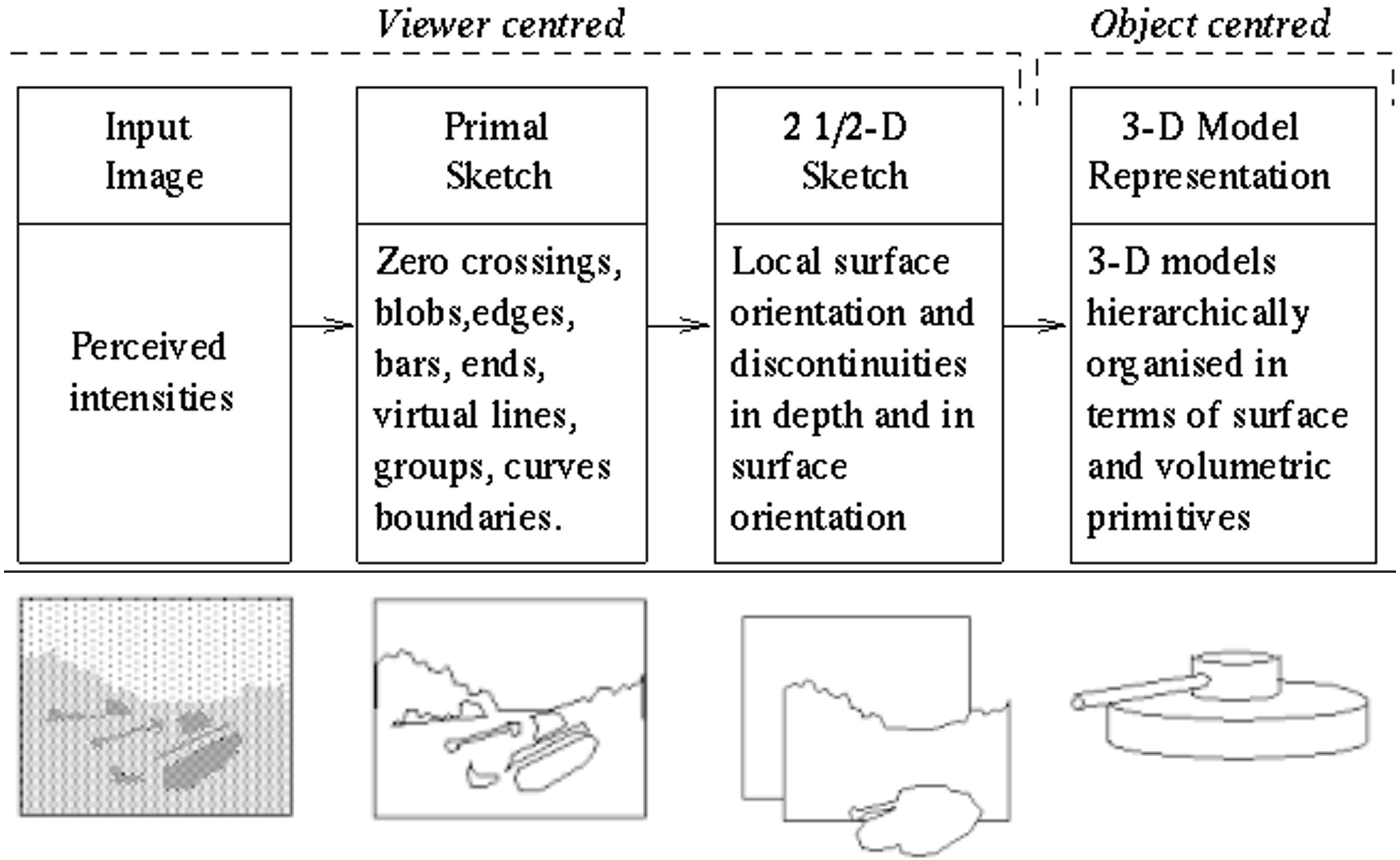
	Yes/No	Number	Other	Overall
Ours: ResNet (1×1)	76.0	36.5	46.8	56.3
Ours: ResNet (14×14)	76.6	36.2	49.5	57.9
Ours: ResNet (7×7)	77.6	37.7	51.5	59.4
Ours: Up-Down	**80.3**	**42.8**	**55.8**	**63.2**
Relative Improvement	3%	14%	8%	6%

Provides inductive bias in both directions!

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering: Anderson et al. 2018

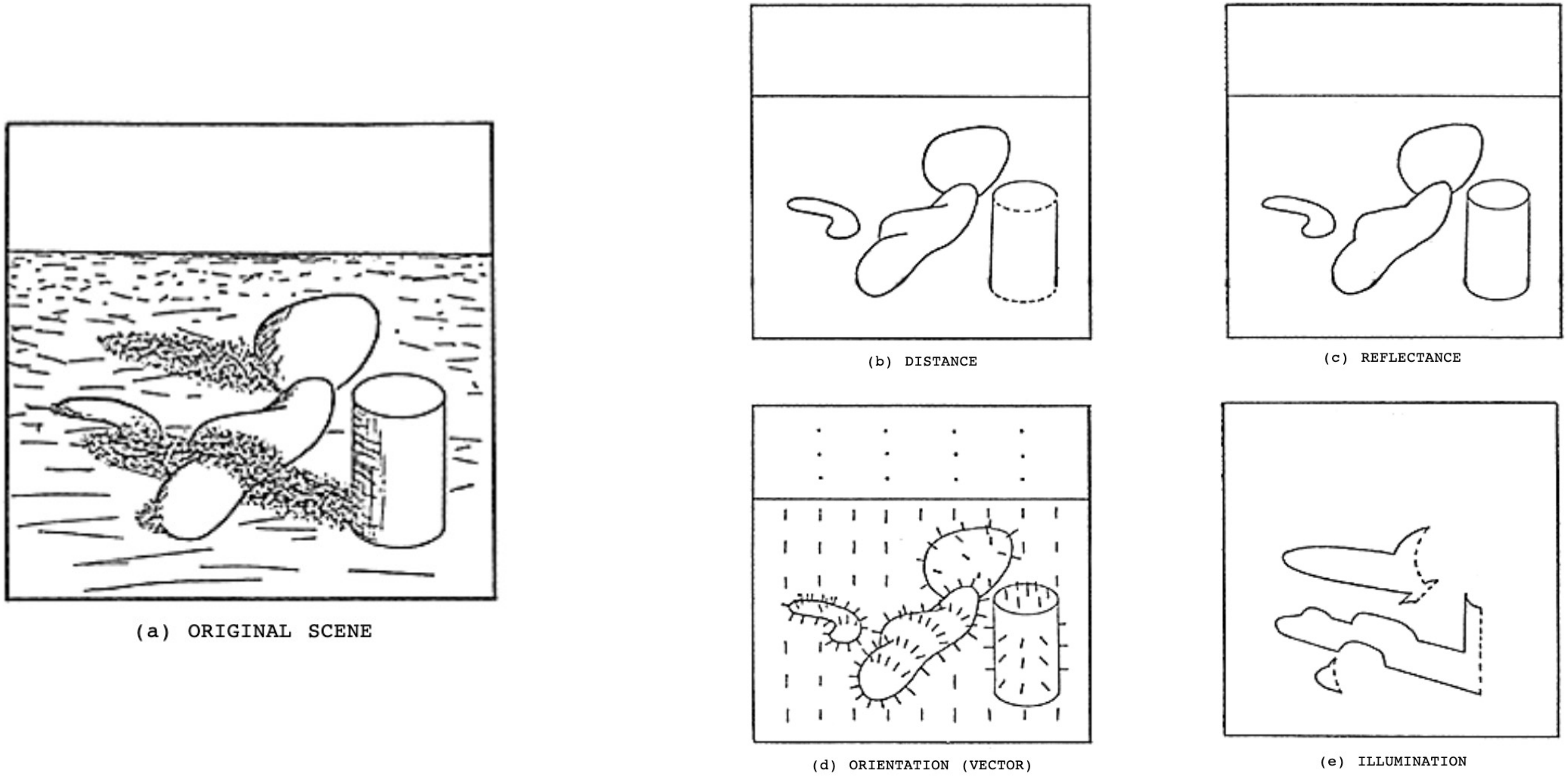
Bottom-Up





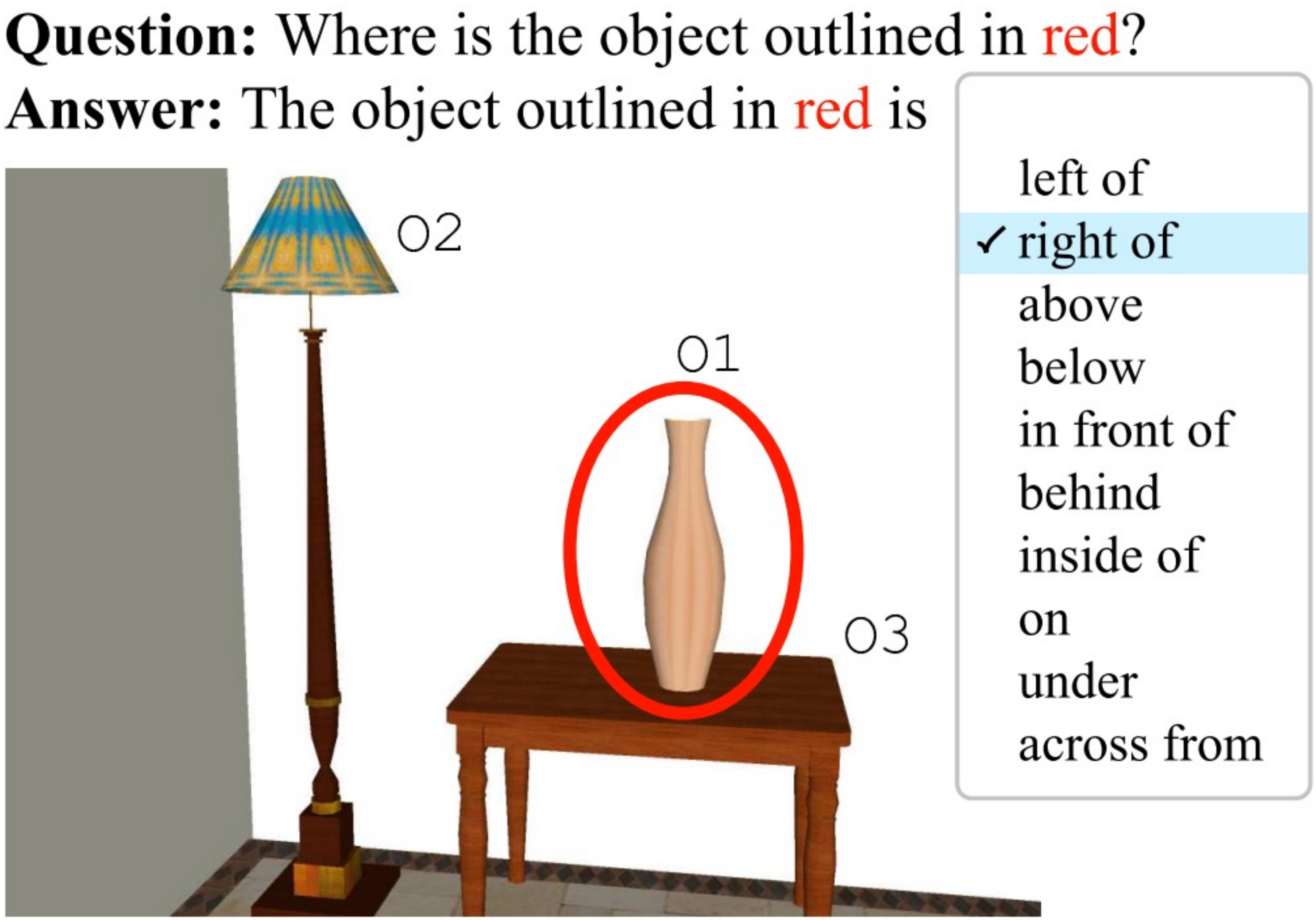
Vision: David Marr 1982

Intrinsic Images



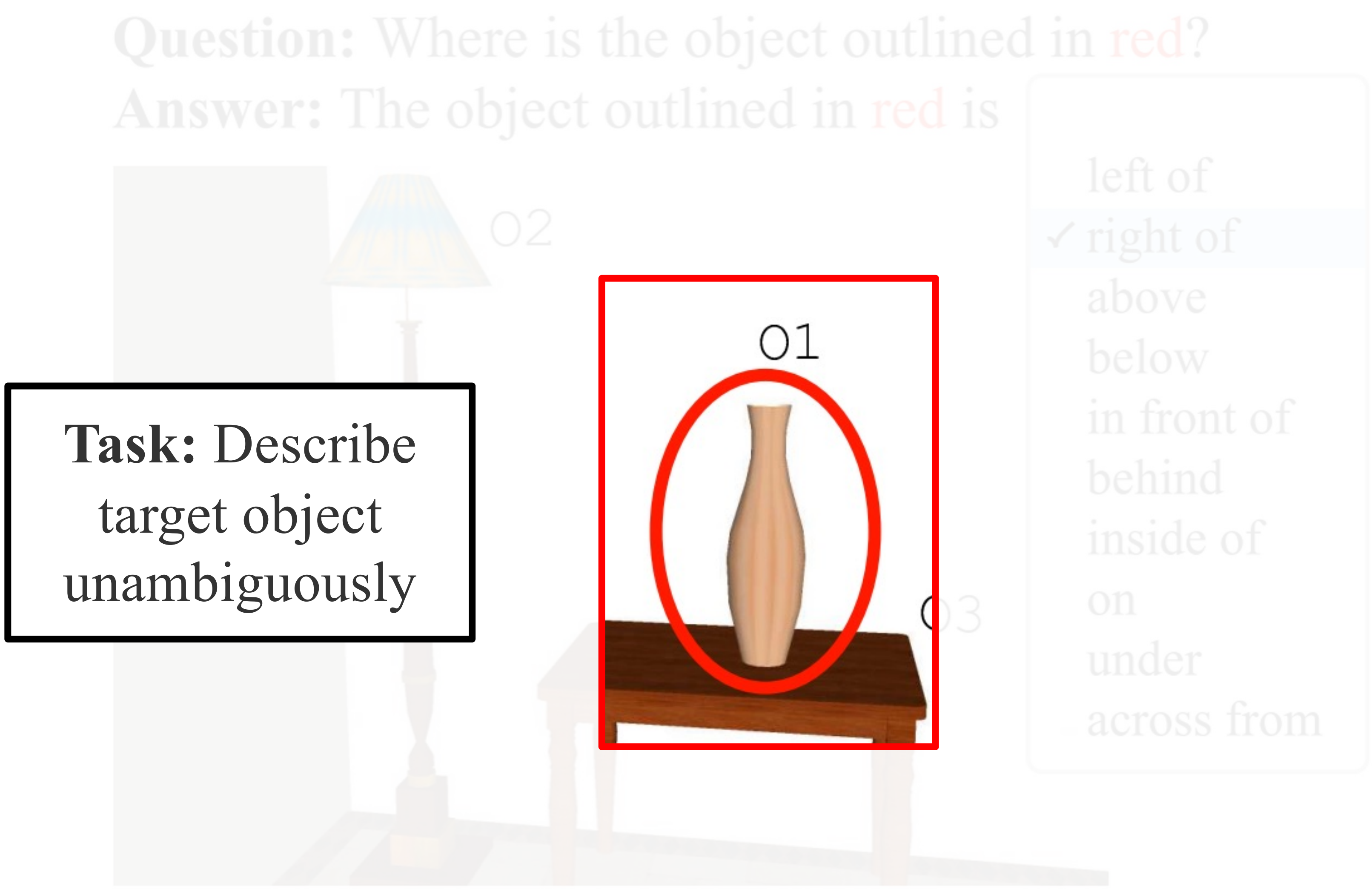
Recovering Intrinsic Scene Characteristics from Images: Barrow and Tenenbaum 1978

“Solved” Perception



A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

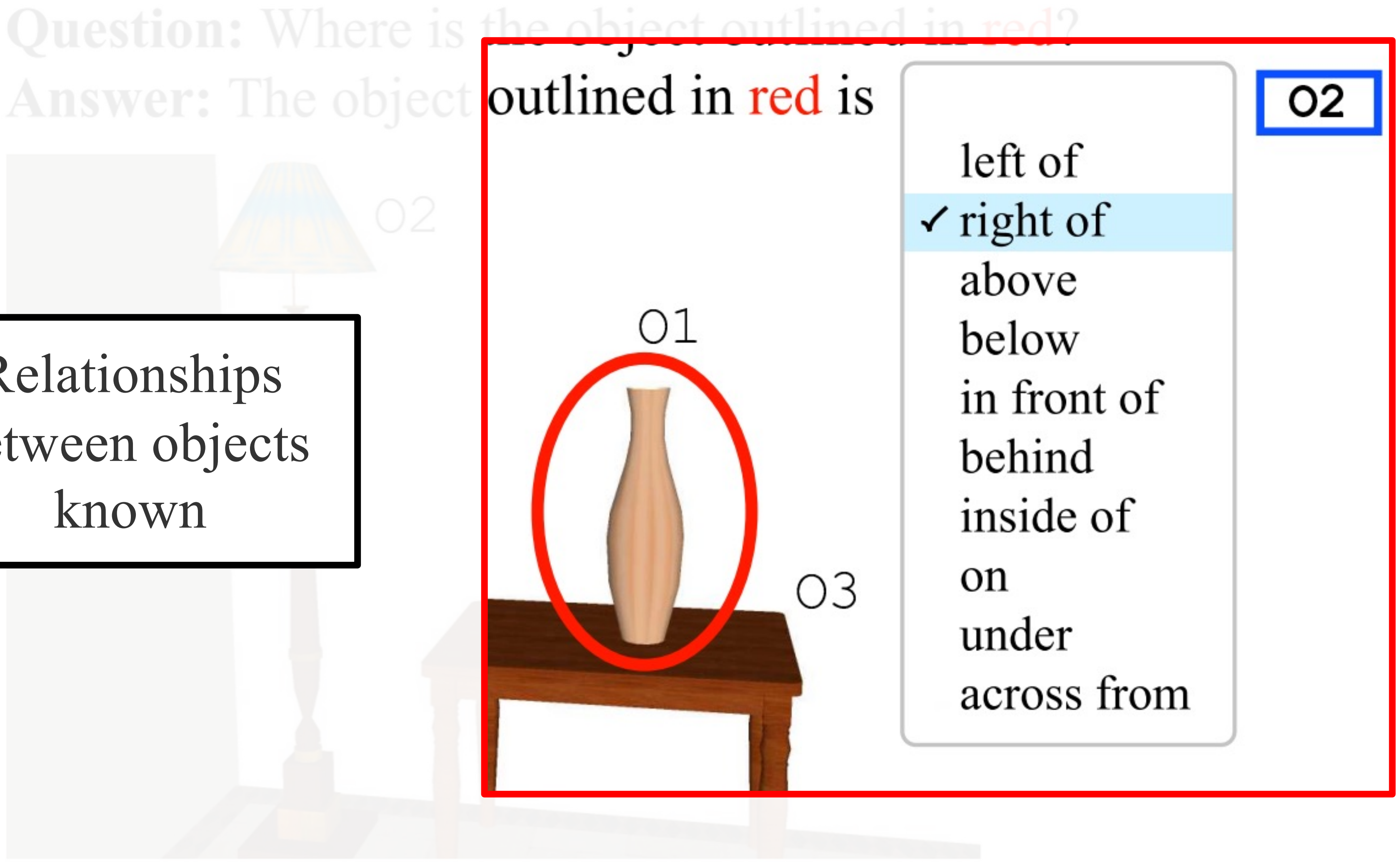
“Solved” Perception



A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

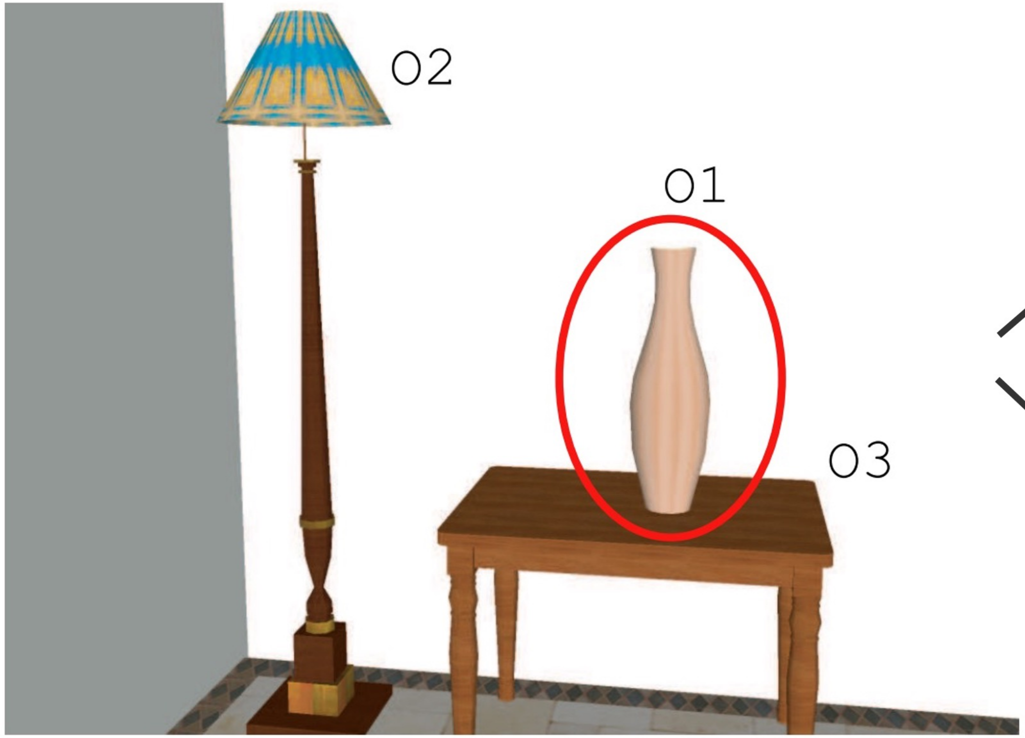
“Solved” Perception

Relationships between objects known



A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

“Solved” Perception



w

S(L):

Right of O2 $\xrightarrow{L}$ O1 (p_L=0.5) ✓ / O3 (p_L=0.5) X

On top of O3 $\xrightarrow{L}$ O1 (p_L=1.0) ✓

$$S(L)(o) = argmax_w p_L(o|w)$$

Problem reduced to pragmatic reasoning

A Game-Theoretic Approach to Generating Spatial Descriptions: Golland et al. 2010

“Solved” Perception

“Go to the last butterfly on the right”

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]

Walk the Talk: MacMahon et al. 2006

“Solved” Perception

“Go to the last butterfly on the right”

What annotators see

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]

Walk the Talk: MacMahon et al. 2006

“Solved” Perception

“Go to the last butterfly on the right”

What agent sees

[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]

Walk the Talk: MacMahon et al. 2006

“Solved” Perception

“Go to the last butterfly on the right”

Reduced to structured prediction problem

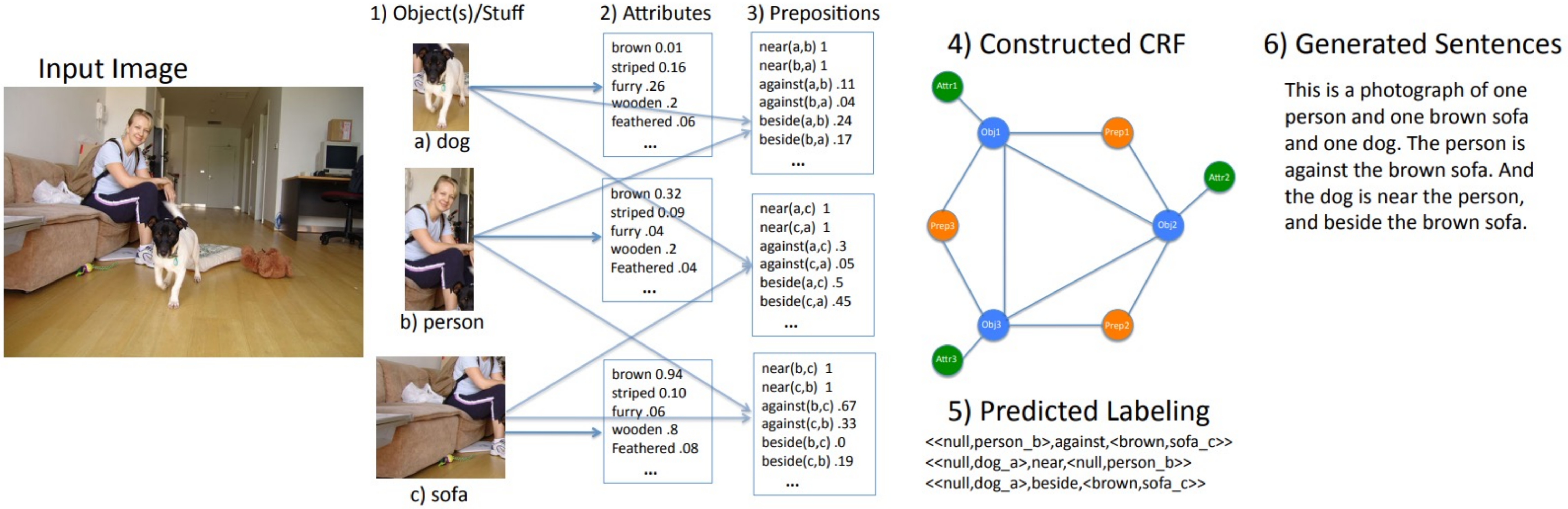
[(Cement, Easel, Cement, Butterfly, Wood, Butterfly),
(Wall, Empty, Wall, Butterfly, Wood, Butterfly),
(Cement, Empty, Wall, End, Wall, End)]

Walk the Talk: MacMahon et al. 2006

"Solved" Perception

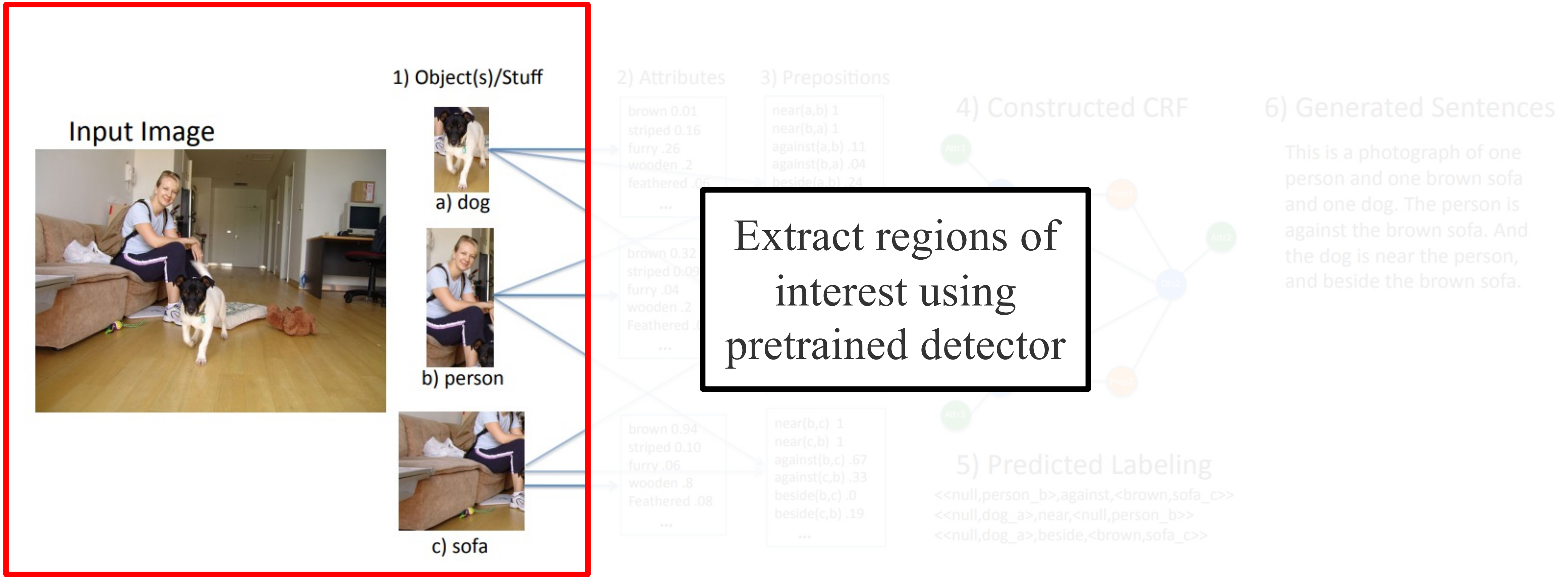
- **Pro:** In early days of vision and language, assuming sub-problems provided traction.
- **Con:** Strong assumptions that don't hold in real world.

Intermediate Representations



BabyTalk: Kulkarni et al. 2013

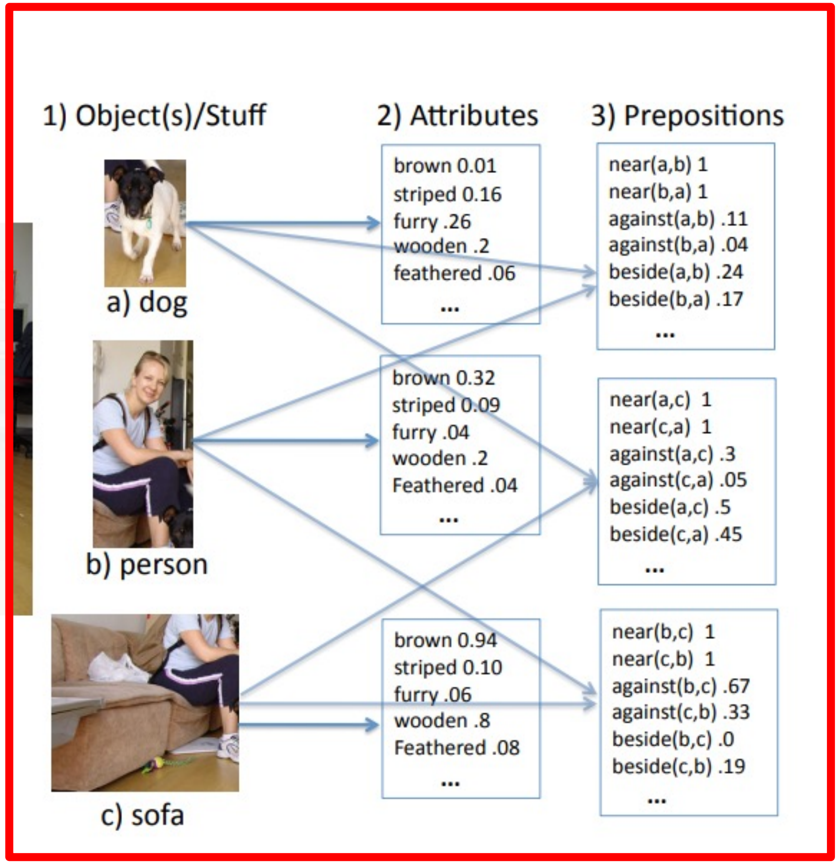
Intermediate Representations



Extract regions of interest using pretrained detector

BabyTalk: Kulkarni et al. 2013

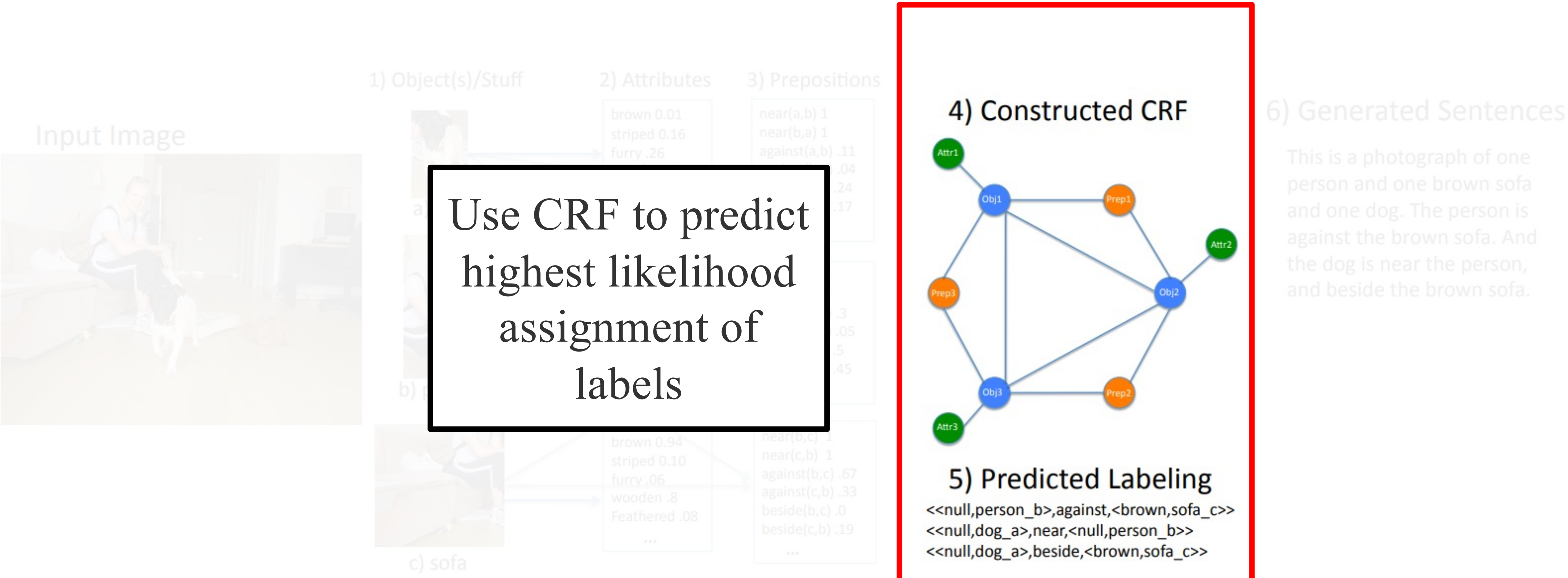
Intermediate Representations



Classifiers score attributes for each region and relationships across them

BabyTalk: Kulkarni et al. 2013

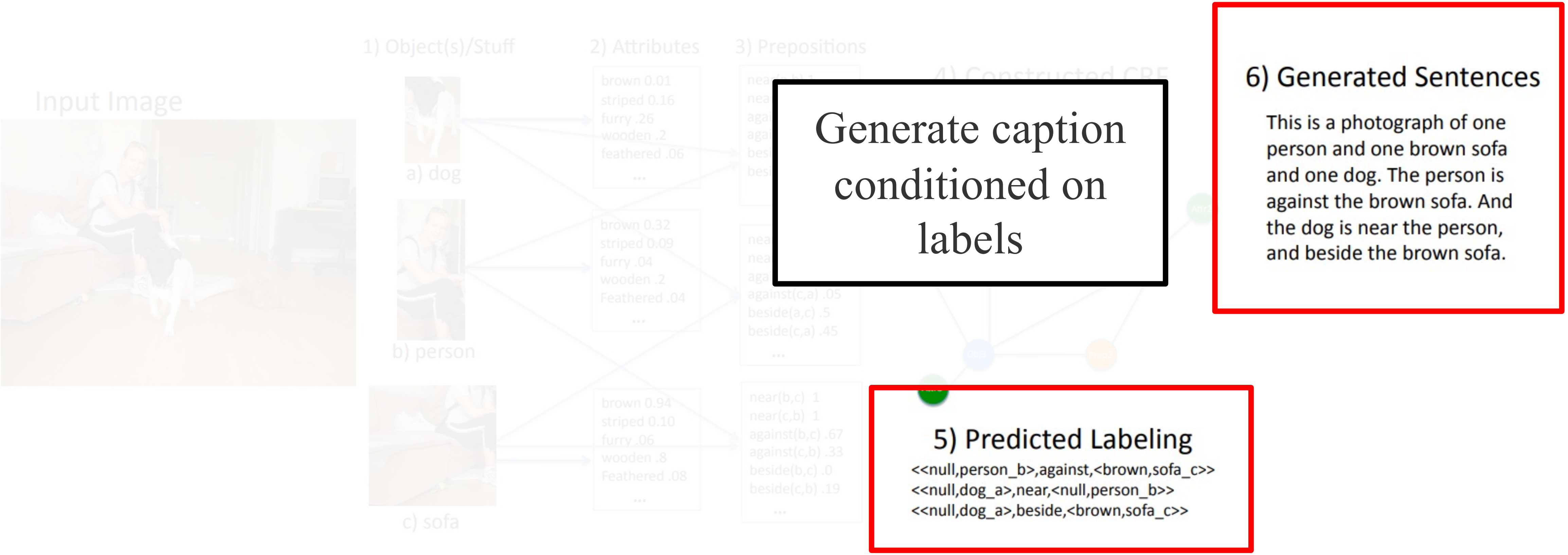
Intermediate Representations



Use CRF to predict highest likelihood assignment of labels

BabyTalk: Kulkarni et al. 2013

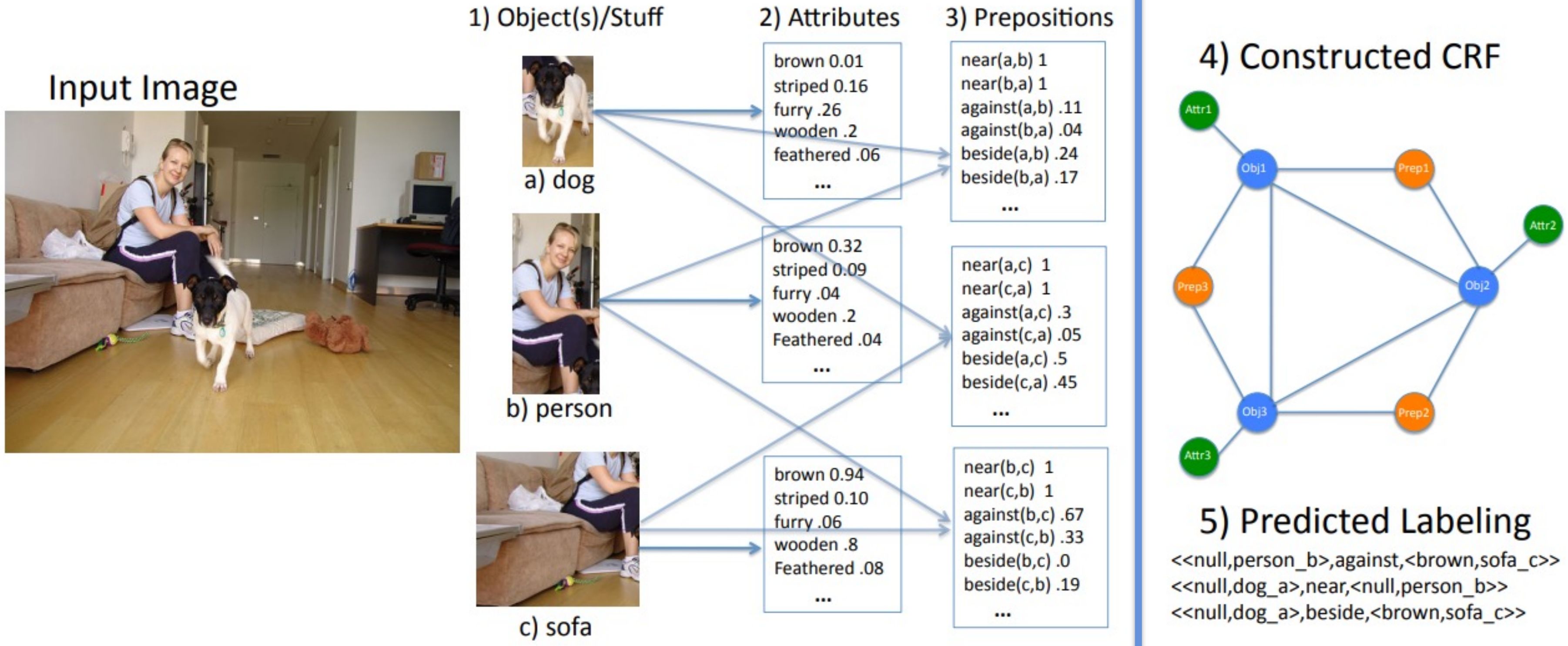
Intermediate Representations



BabyTalk: Kulkarni et al. 2013

Intermediate Representations

Language model never sees pixels!



6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

BabyTalk: Kulkarni et al. 2013

Intermediate Representations

This is a photograph of one sky, one road and one bus. The blue sky is above the gray road. The gray road is near the shiny bus. The shiny bus is near the blue sky.

There are two aeroplanes. The first shiny aeroplane is near the second shiny aeroplane.

There are one cow and one sky. The golden cow is by the blue sky.

There are one dining table, one chair and two windows. The wooden dining table is by the wooden chair, and against the first window, and against the second white window. The wooden chair is by the first window, and by the second white window. The first window is by the second white window.

Here we see one person and one train. The black person is by the train.

This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.

Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

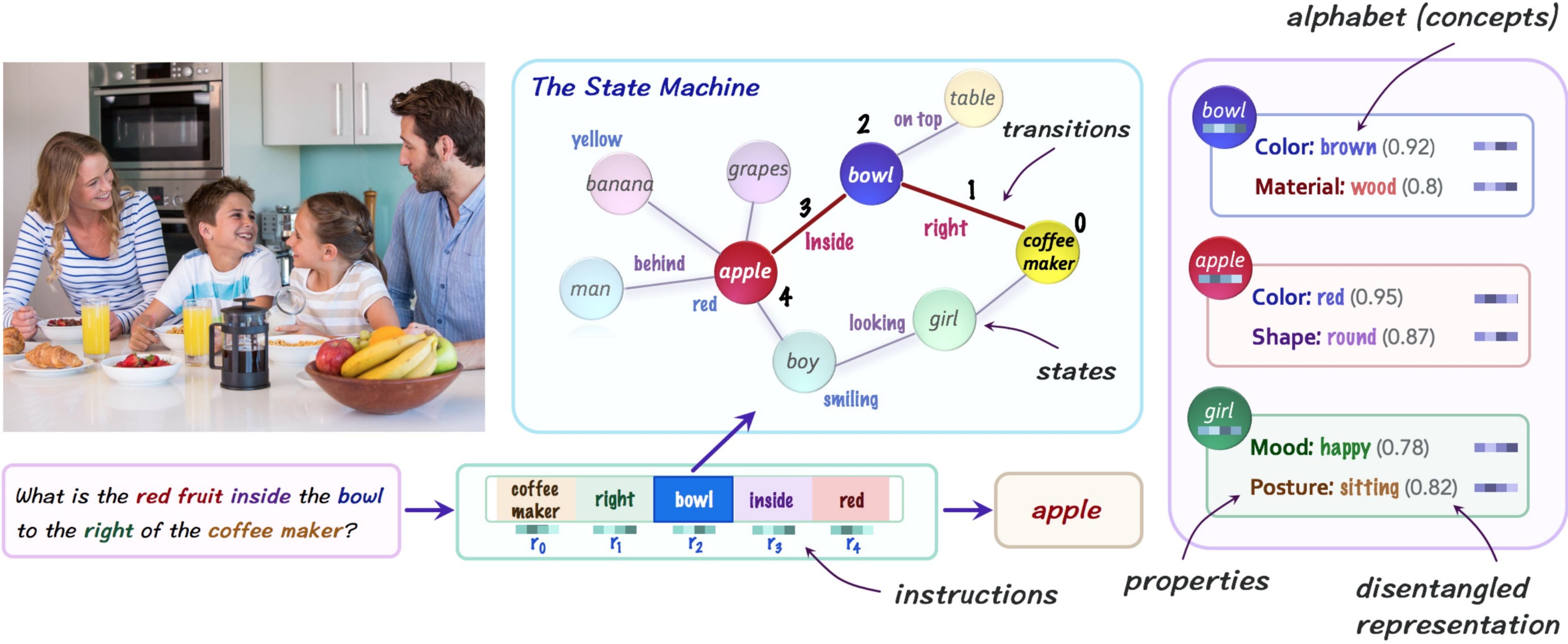
Here we see two persons, one sky and one aeroplane. The first black person is by the blue sky. The blue sky is near the shiny aeroplane. The second black person is by the blue sky. The shiny aeroplane is by the first black person, and by the second black person.

This is a picture of two dogs. The first dog is near the second furry dog.

This is a photograph of two buses. The first rectangular bus is near the second rectangular bus.

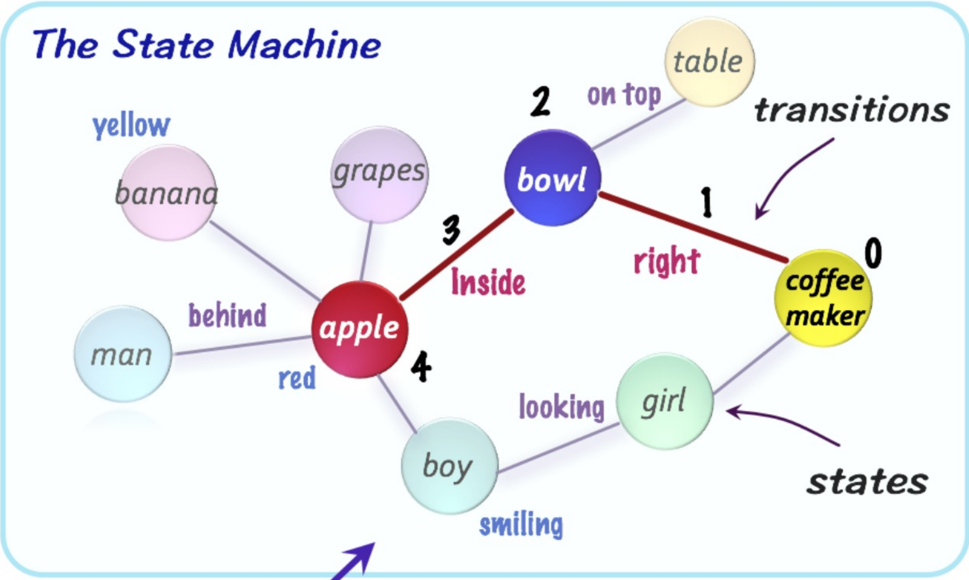
BabyTalk: Kulkarni et al. 2013

Intermediate Representations



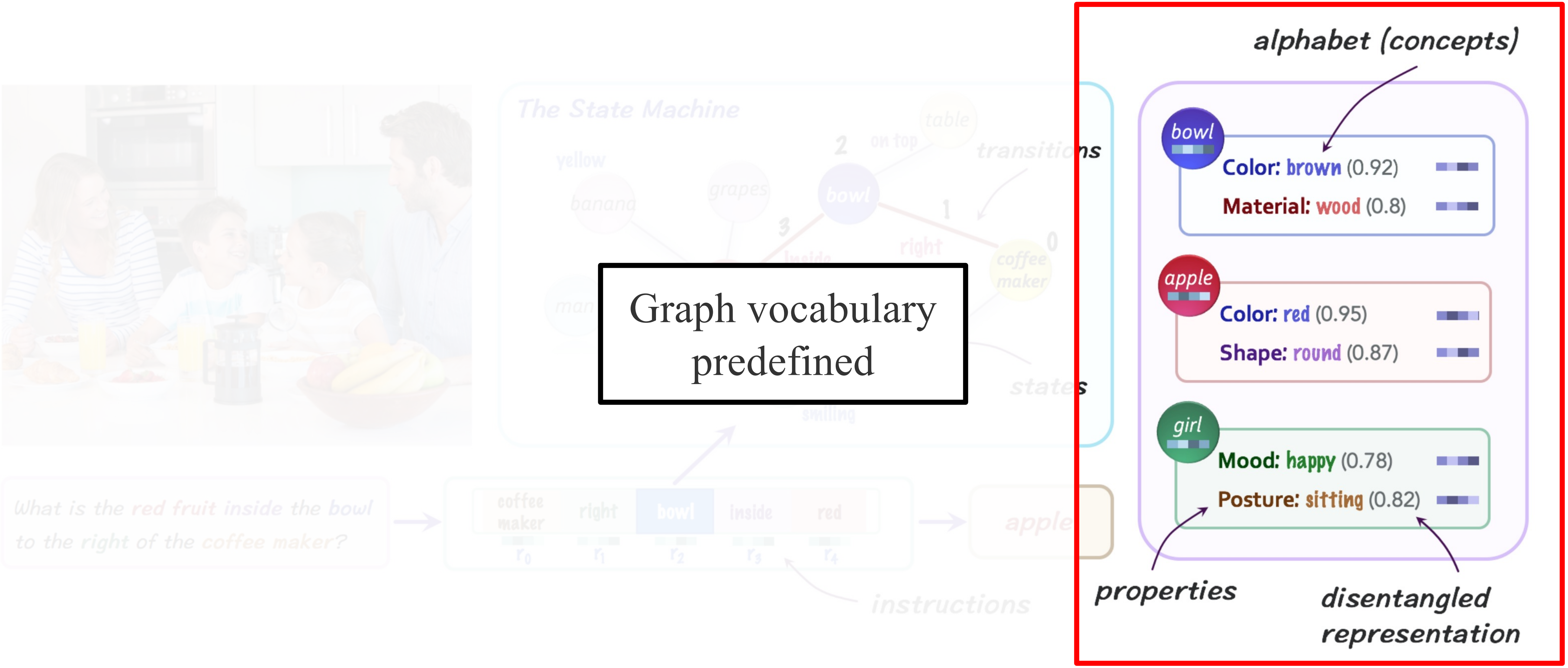
Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

Intermediate Representations



Generate scene graph from image

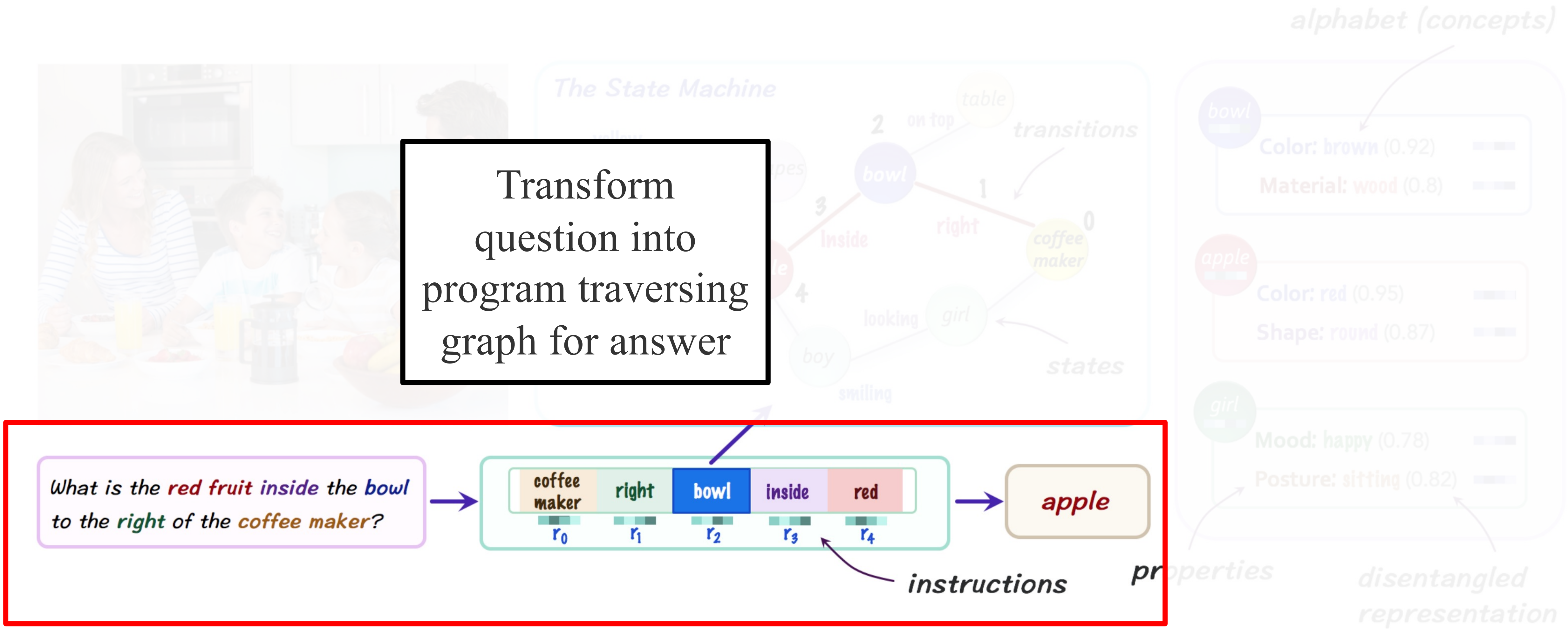
Intermediate Representations



Graph vocabulary predefined

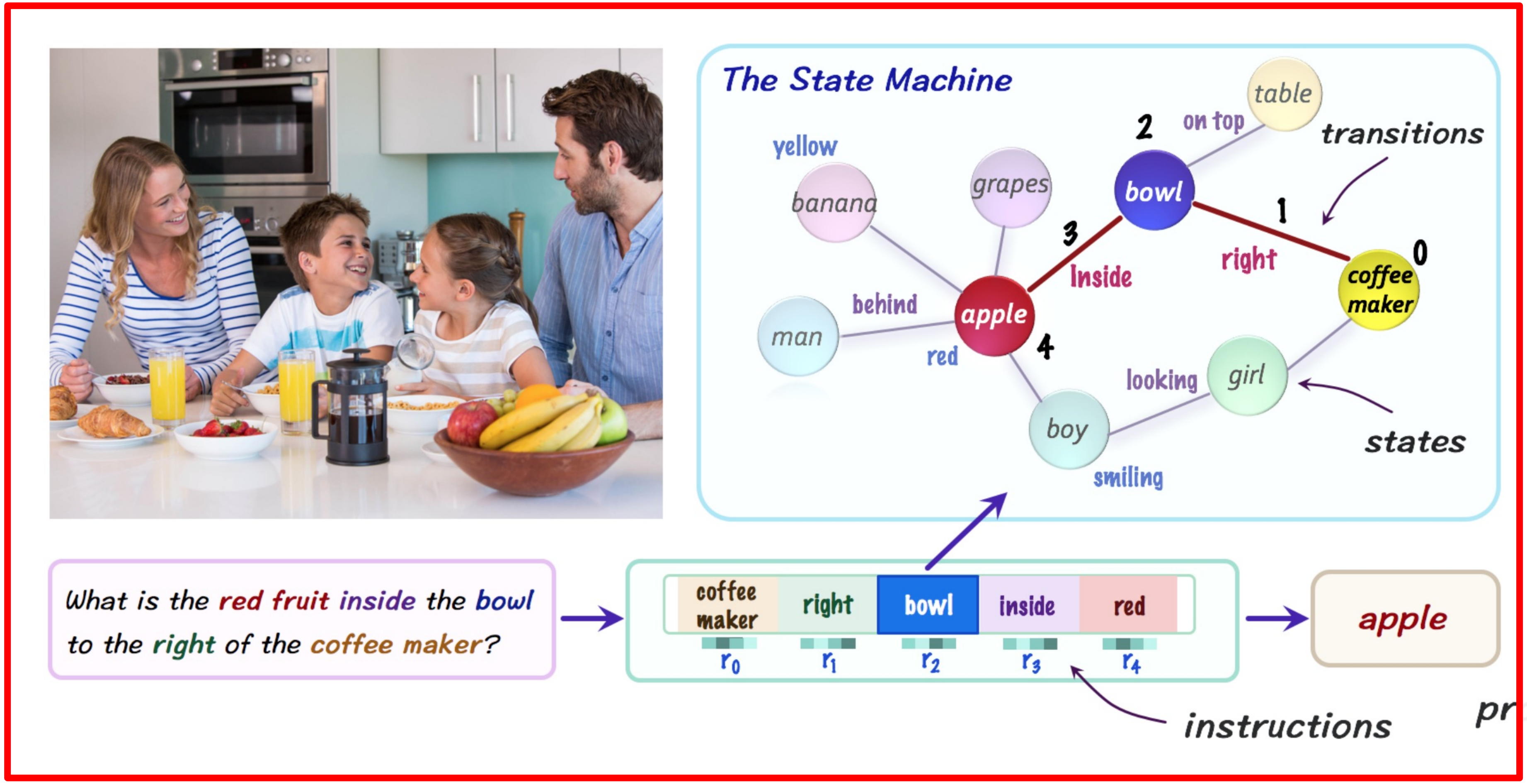
Intermediate Representations

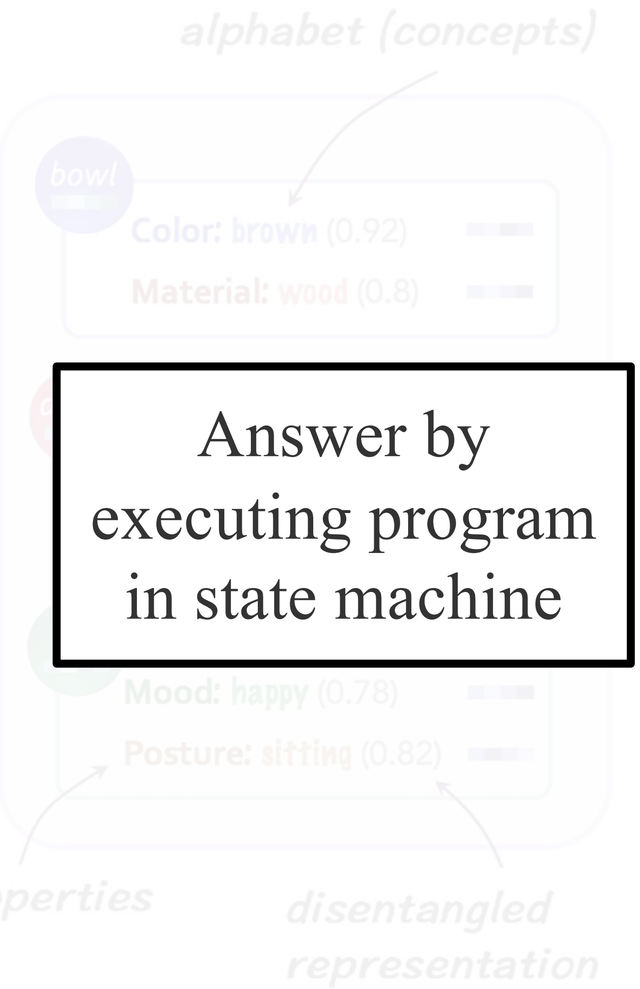
Transform question into program traversing graph for answer



Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

Intermediate Representations

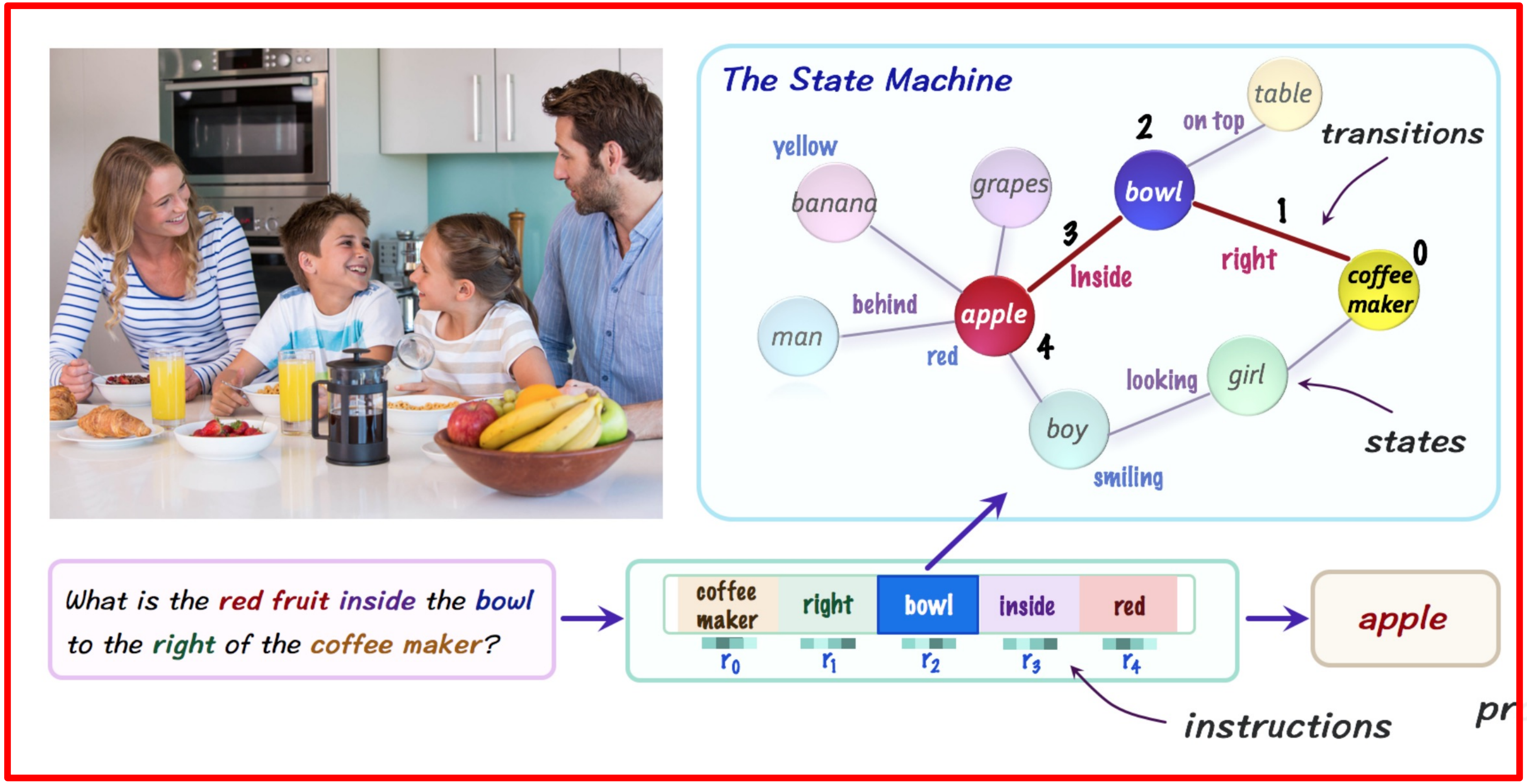




Answer by executing program in state machine

Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

Intermediate Representations



Allows language reasoning to occur solely within abstract structure

Learning by Abstraction: The Neural State Machine: Hudson and Manning 2019

Intermediate Representations

Table 4: GQA generalization

Model	Content	Structure
Global Prior	8.51	14.64
Local Prior	12.14	18.21
Vision	17.51	18.68
Language	21.14	32.88
Lang+Vis	24.95	36.51
BottomUp [5]	29.72	41.83
MAC [40]	31.12	47.27
NSM	**40.24**	**55.72**

TEXT PROMPT

a store front that has the word 'dall-e' written on it. a store front that has the word 'dall-e' written on it. a store front that has the word 'dall-e' written on it. dall-e store front.

AI-GENERATED IMAGES

DALL-E 1: Ramesh et al. 2021

Intermediate Representations

Step 1

Learn Proto-linguistic Code Book

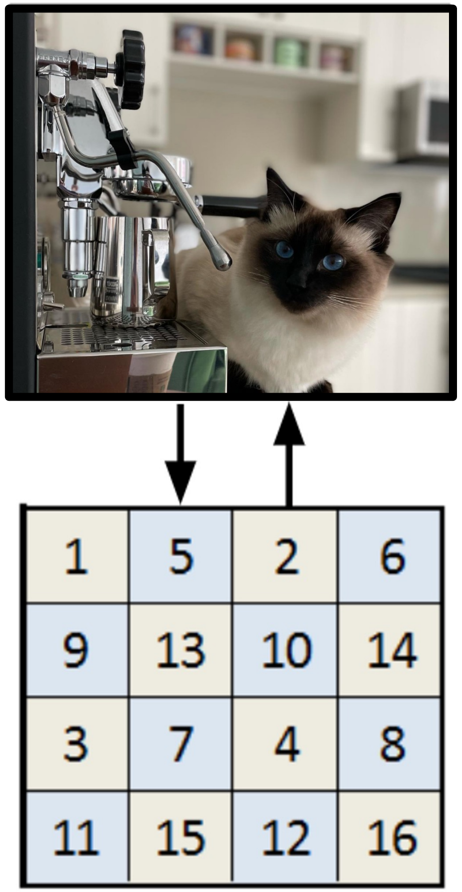
1	5	2	6
9	13	10	14
3	7	4	8
11	15	12	16

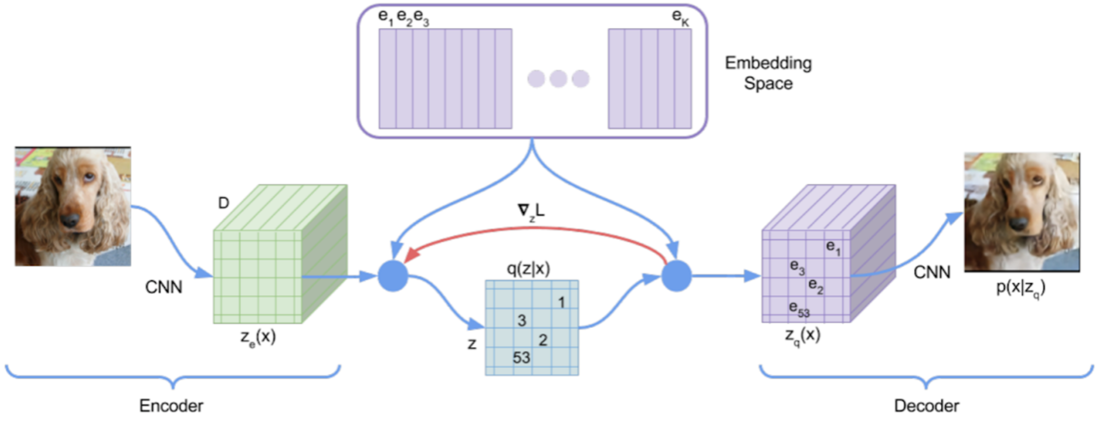
DALL-E 1: Ramesh et al. 2021

Intermediate Representations

<u>Step 1</u>

Learn Proto-linguistic Code Book





Neural Discrete Representation Learning: van Oord et al. 2017

DALL-E 1: Ramesh et al. 2021

Intermediate Representations

Step 2

Learn Joint

Language and Code Distribution

"A kitten with a pink background"

1	5	2	6
9	13	10	14
3	7	4	8
11	15	12	16

DALL-E 1: Ramesh et al. 2021

Intermediate Representations

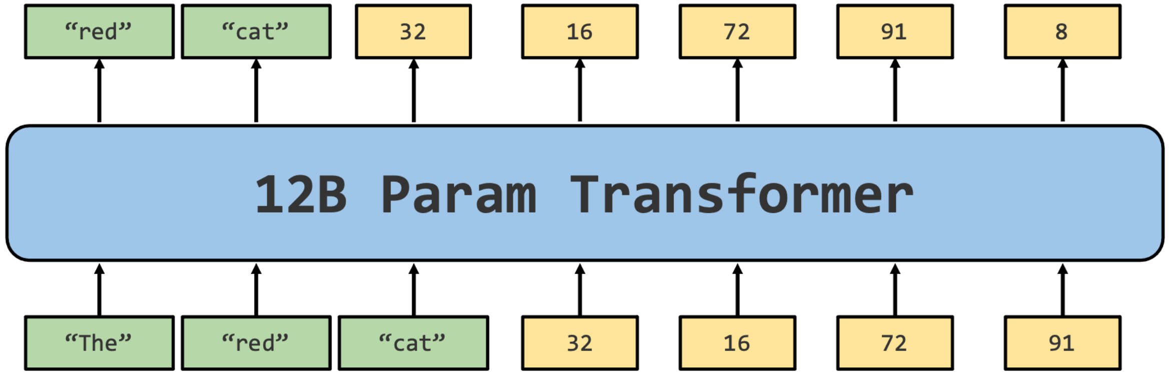
<u>Step 2</u>

Learn Joint

Language and Code Distribution

"A kitten with a pink background"

1	5	2	6
9	13	10	14
3	7	4	8
11	15	12	16



Generating Long Sequences with Sparse Transformers: Child et al. 2019

DALL-E 1: Ramesh et al. 2021

Intermediate Representations

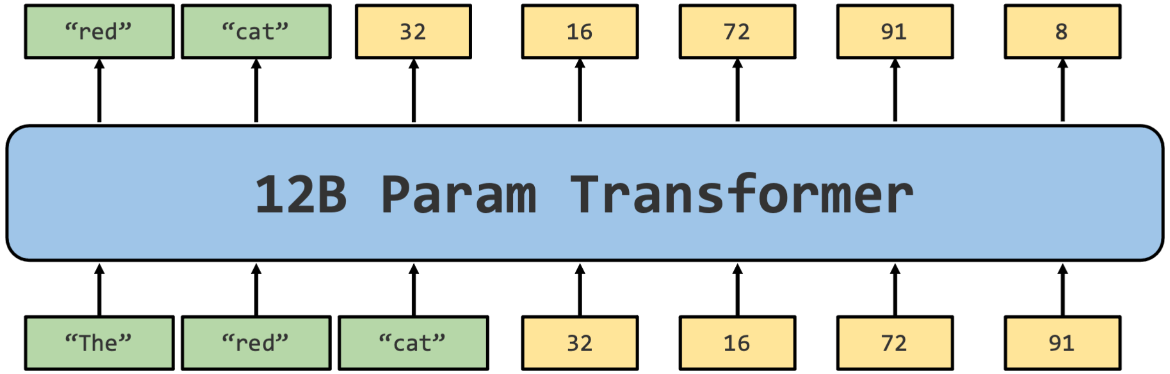
Step 2

Learn Joint

Language and Code Distribution

"A kitten with a pink background"

1	5	2	6
9	13	10	14
3	7	4	8
11	15	12	16



Generating Long Sequences with Sparse Transformers: Child et al. 2019

Reduced to language modeling problem!

DALL-E 1: Ramesh et al. 2021

TEXT PROMPT

an x-ray of a capybara sitting in a forest

AI-GENERATED IMAGES

DALL-E 1: Ramesh et al. 2021

Anchoring to 3D

"The goal of an image understanding system is to transform two-dimensional data into a representation of the three-dimensional spatio-temporal world"

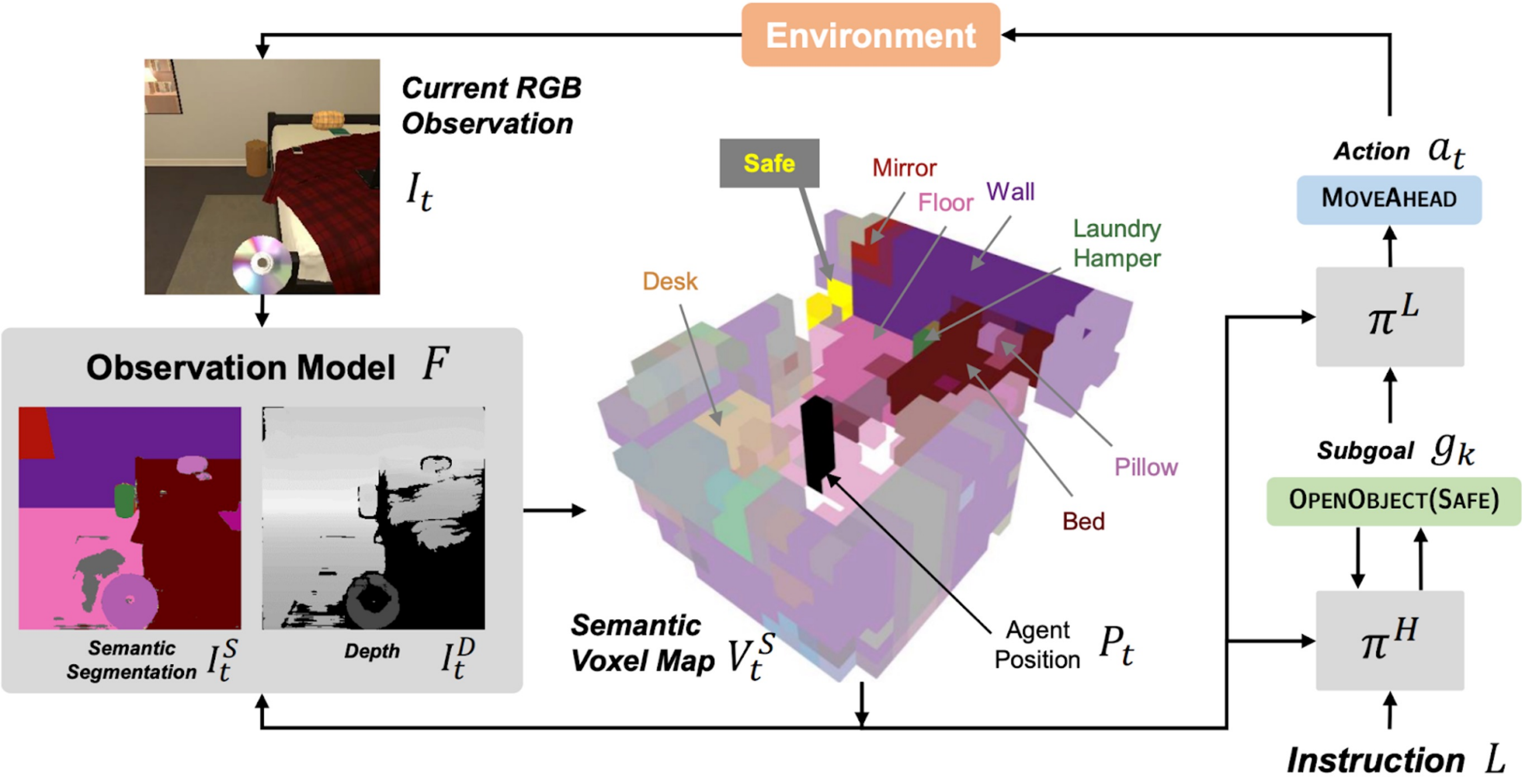
Image Understanding: John Tsotsos 1987

Anchoring to 3D

“Place a clean ladle on a counter”

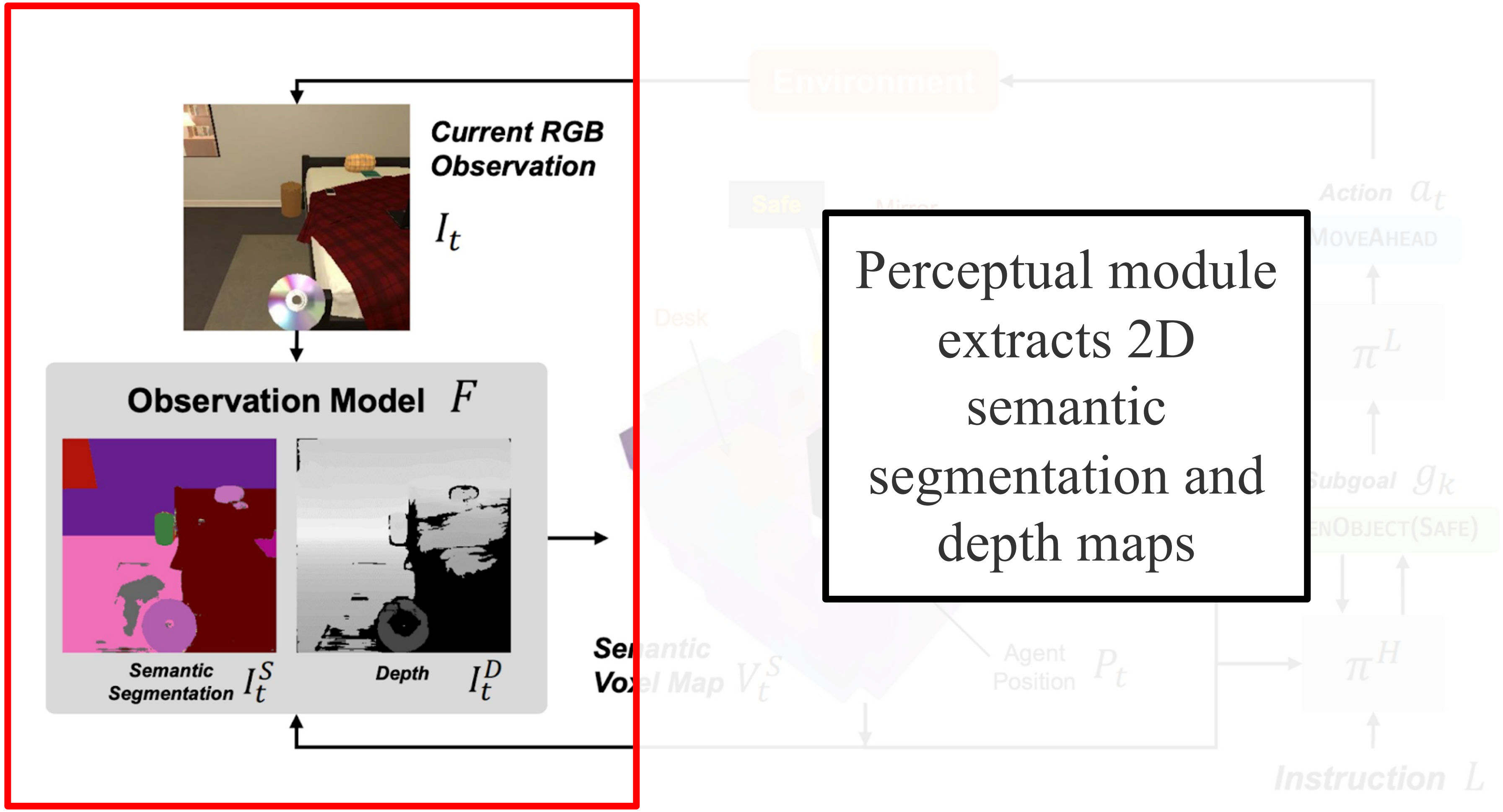
ALFRED: Shridhar et al. 2020

Anchoring to 3D



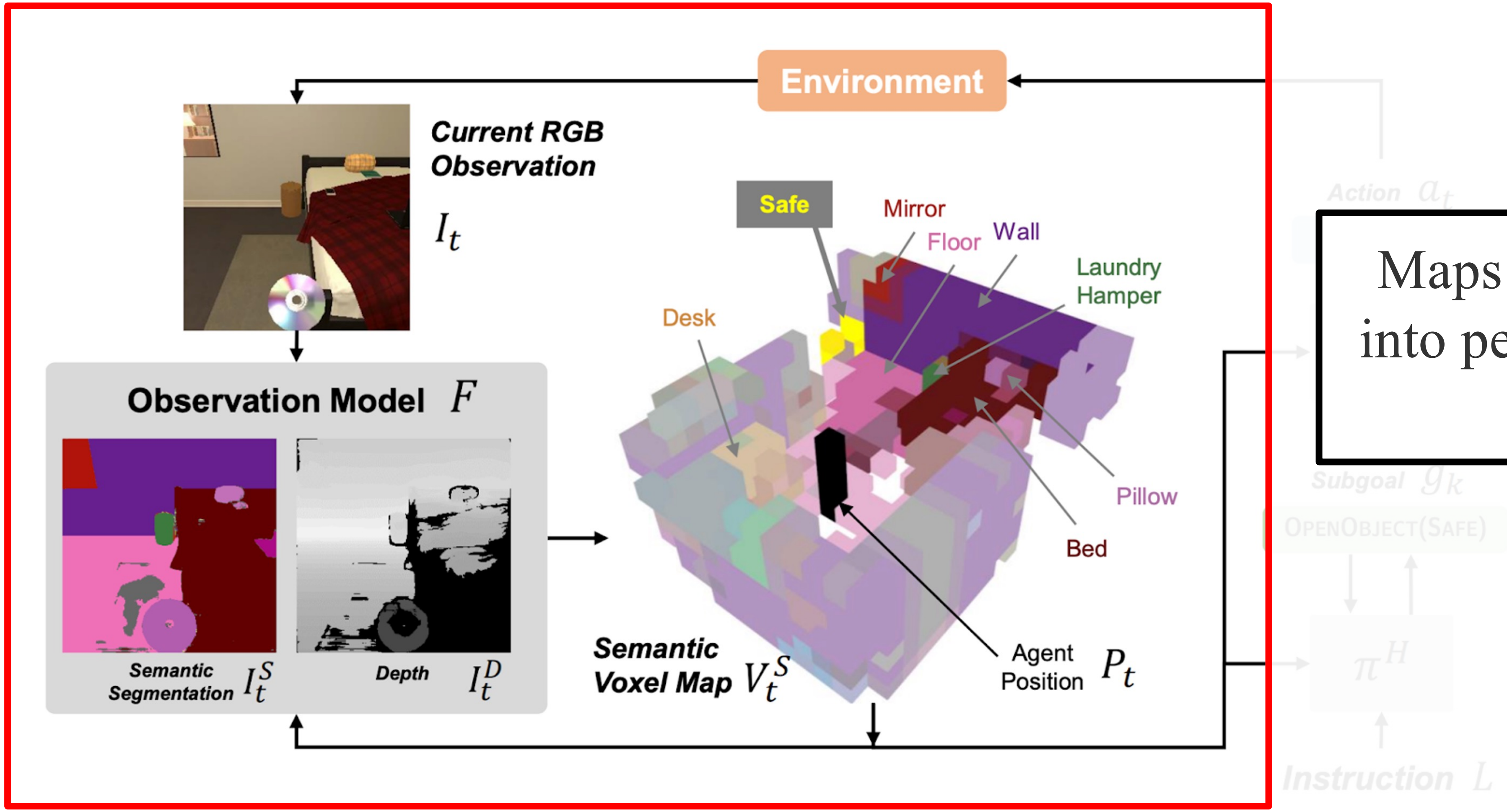
A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

Anchoring to 3D



Perceptual module extracts 2D semantic segmentation and depth maps

Anchoring to 3D

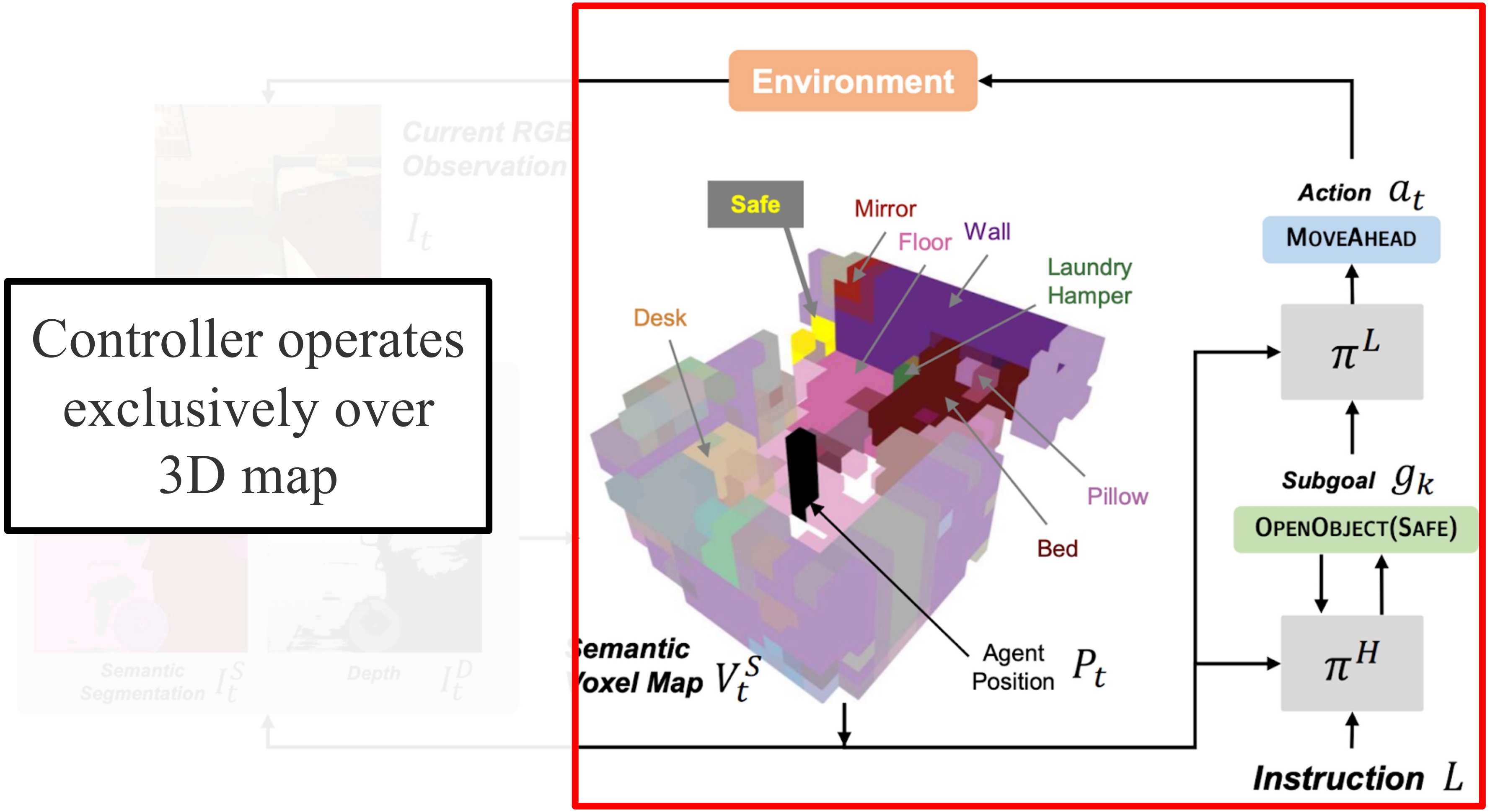


Maps integrated into persistent 3D map

A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

Anchoring to 3D

Controller operates exclusively over 3D map



A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

Anchoring to 3D

Method	Validation Seen SR	Validation Seen GC	Validation Unseen SR	Validation Unseen GC
HLSM	29.6	38.8	18.3	**31.2**
+ gt depth	29.6	40.5	20.1	33.7
+ gt depth, gt seg.	40.7	50.4	40.2	52.2
+ gt seg.	36.2	47.0	34.7	47.8
w/o language enc.	0.9	8.6	0.2	7.5
w/o subg. hist. enc.	29.4	38.5	16.6	29.2
w/o state repr enc.	30.0	40.6	**18.9**	30.8

3D Map useful for improving performance

A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

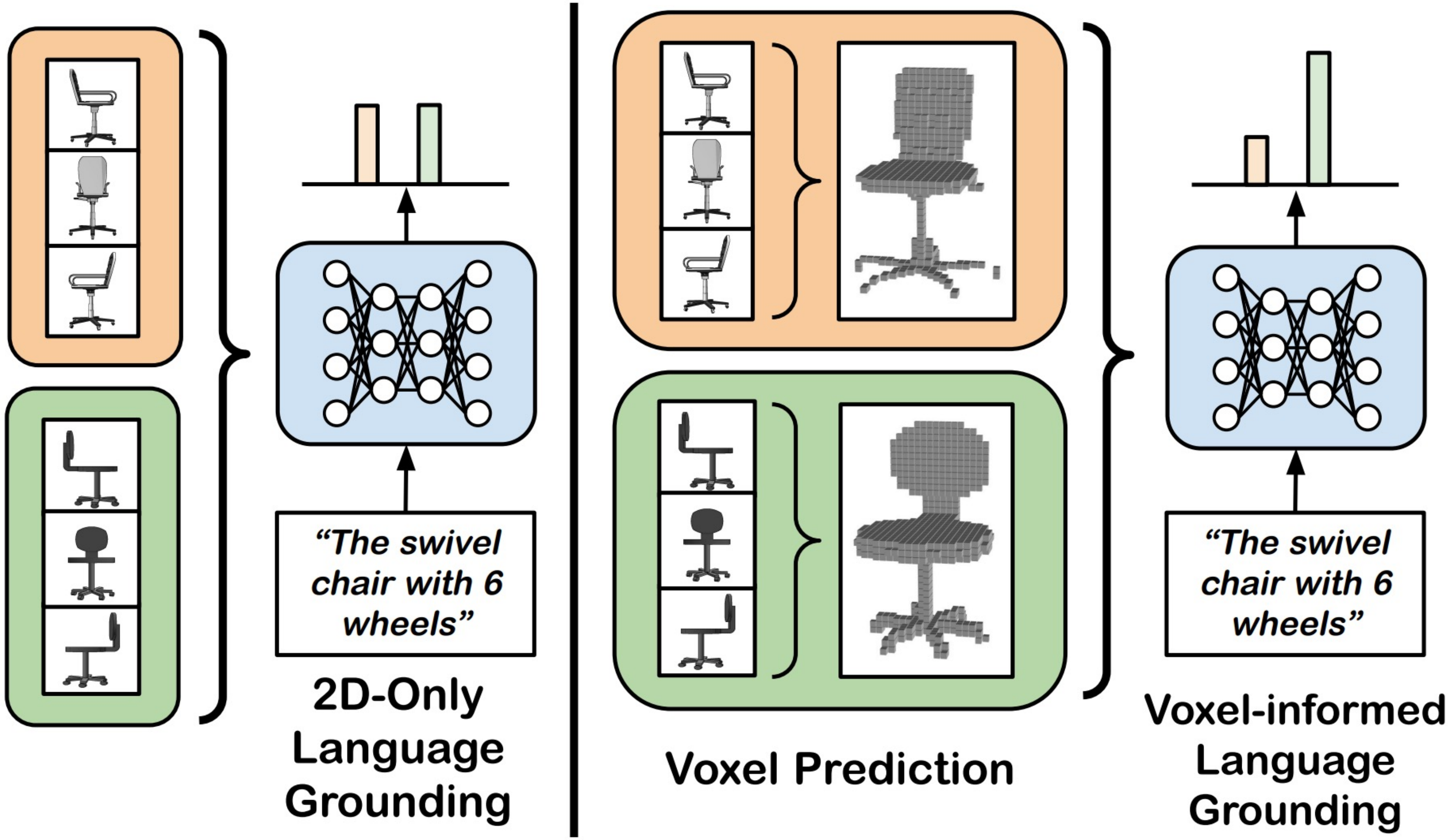
Anchoring to 3D

Method	Validation Seen SR	Validation Seen GC	Validation Unseen SR	Validation Unseen GC
HLSM	29.6	38.8	18.3	**31.2**
+ gt depth	29.6	40.5	20.1	33.7
+ gt depth, gt seg.	40.7	50.4	40.2	52.2
+ gt seg.	36.2	47.0	34.7	47.8
w/o language enc.	0.9	8.6	0.2	7.5
w/o subg. hist. enc.	29.4	38.5	16.6	29.2
w/o state repr enc.	30.0	40.6	**18.9**	30.8

However, benefits held back by cascading errors

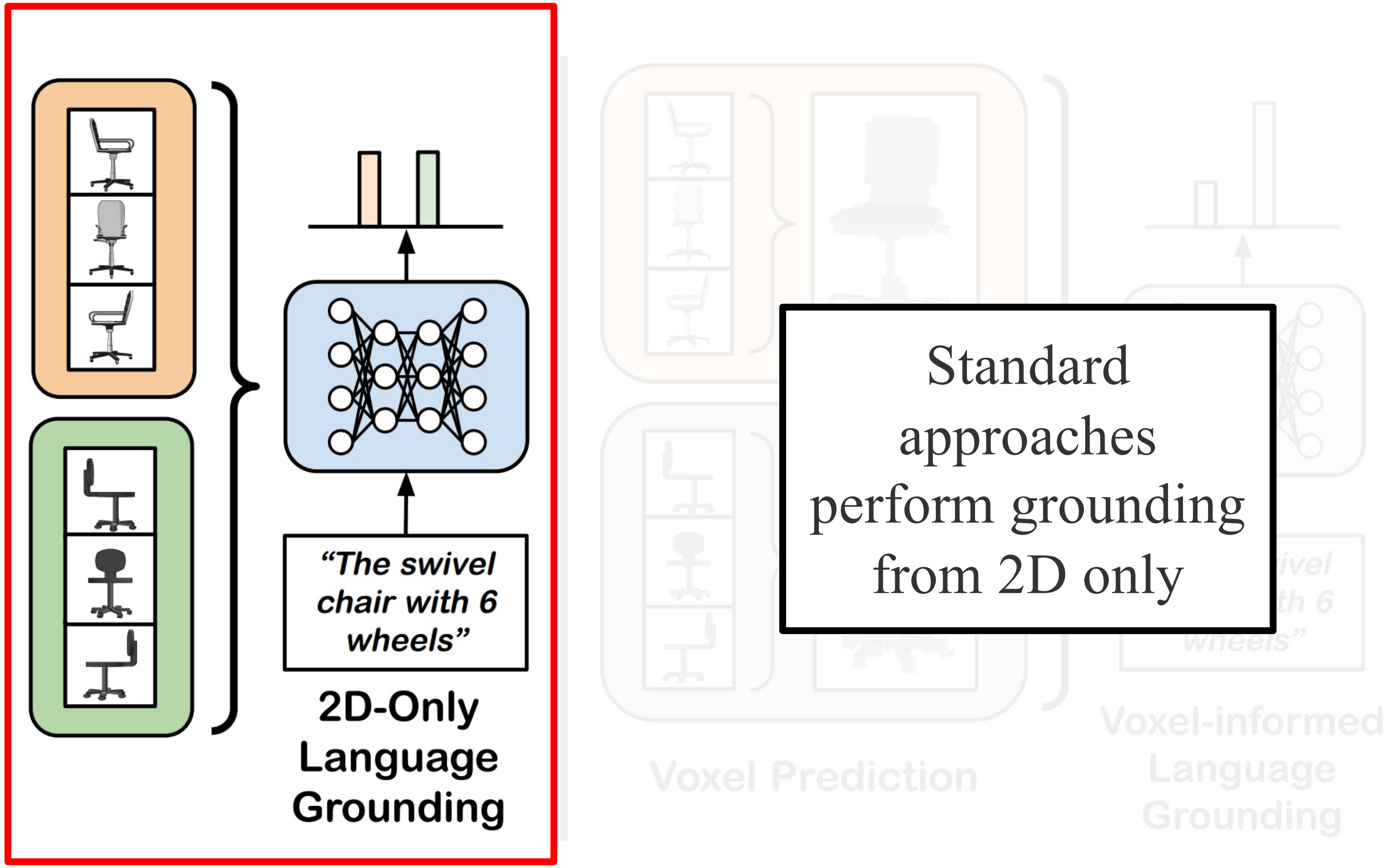
A Persistent Spatial Semantic Representation for High-Level Natural Language Instruction Execution: Blukis et al. 2021

Anchoring to 3D



Voxel-informed Language Grounding: Corona et al. 2022

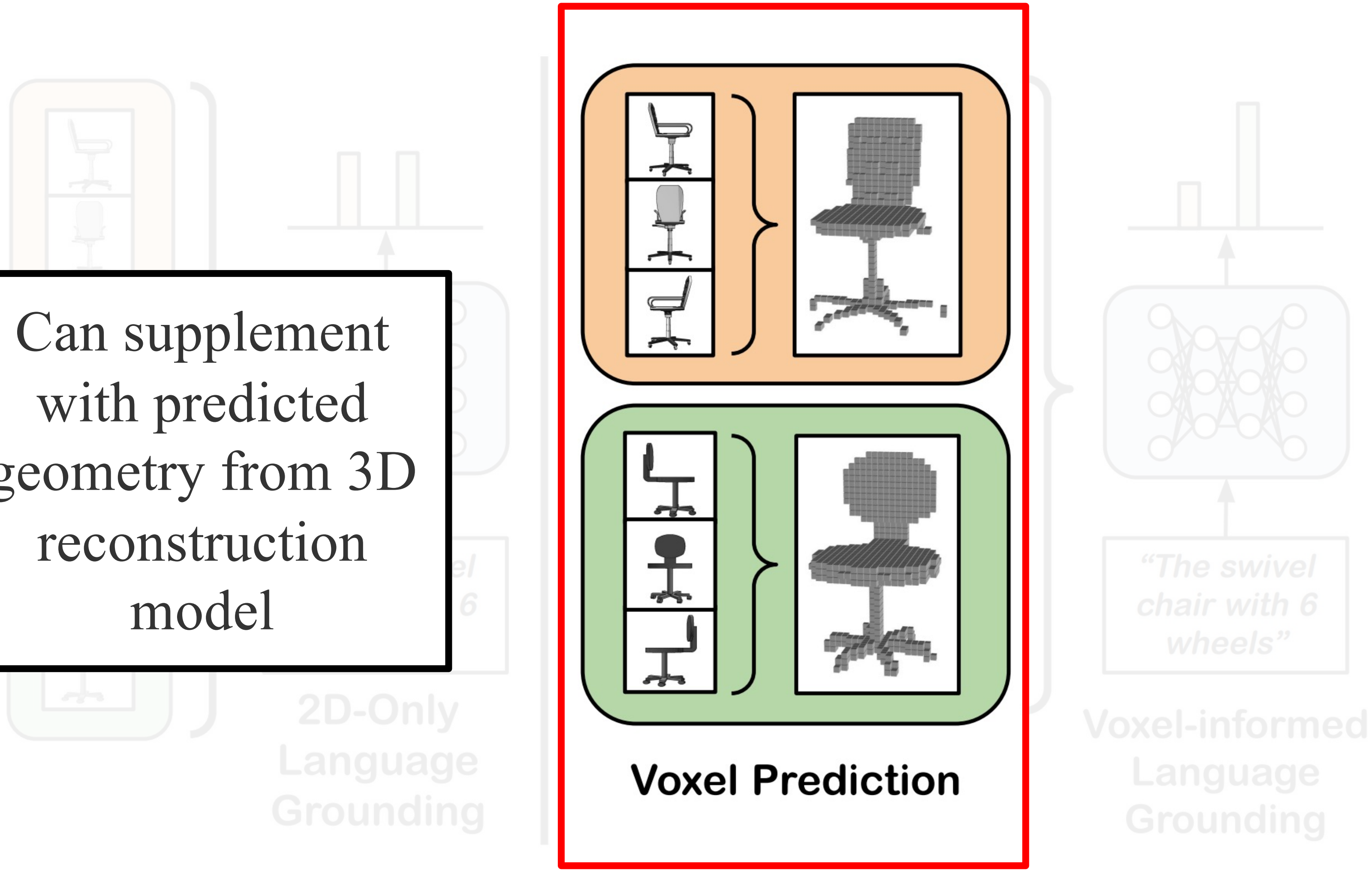
Anchoring to 3D



Standard approaches perform grounding from 2D only

Voxel-informed Language Grounding: Corona et al. 2022

Anchoring to 3D



Can supplement with predicted geometry from 3D reconstruction model

Anchoring to 3D

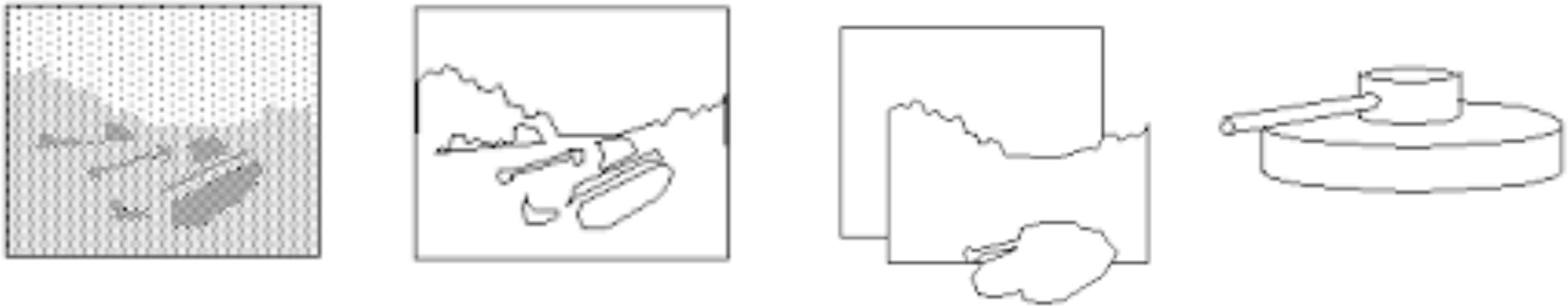
Model	VALIDATION Visual	Blind	All	TEST Visual	Blind	All
ViLBERT	89.5	76.6	83.1	80.2	**73.0**	76.6
MATCH	89.2 (0.9)	75.2 (0.7)	82.2 (0.4)	83.9 (0.5)	68.7 (0.9)	76.5 (0.5)
MATCH*	90.6 (0.4)	75.7 (1.2)	83.2 (0.8)	-	-	-
LAGOR	89.8 (0.4)	75.3 (0.7)	82.6 (0.4)	84.3 (0.4)	69.4 (0.5)	77.0 (0.5)
LAGOR*	89.8 (0.5)	75.0 (0.4)	82.5 (0.1)	-	-	-
VLG (Ours)	**91.2** (0.4)	**78.4**†(0.7)	**84.9**†(0.3)	**86.0**	71.7	**79.0**

Improves performance over 2D-only approaches

Voxel-informed Language Grounding: Corona et al. 2022

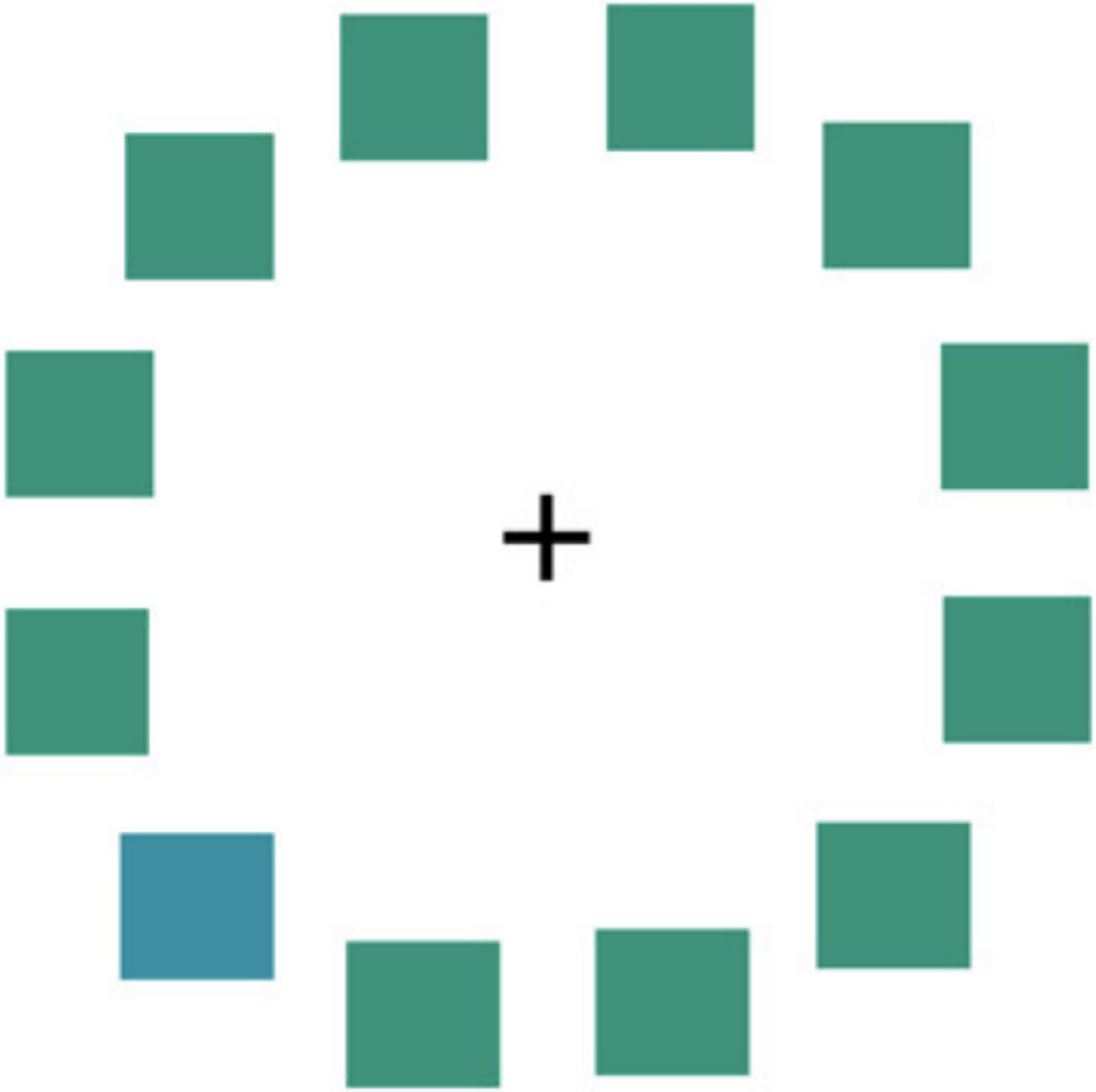
Bottom-Up Takeaways

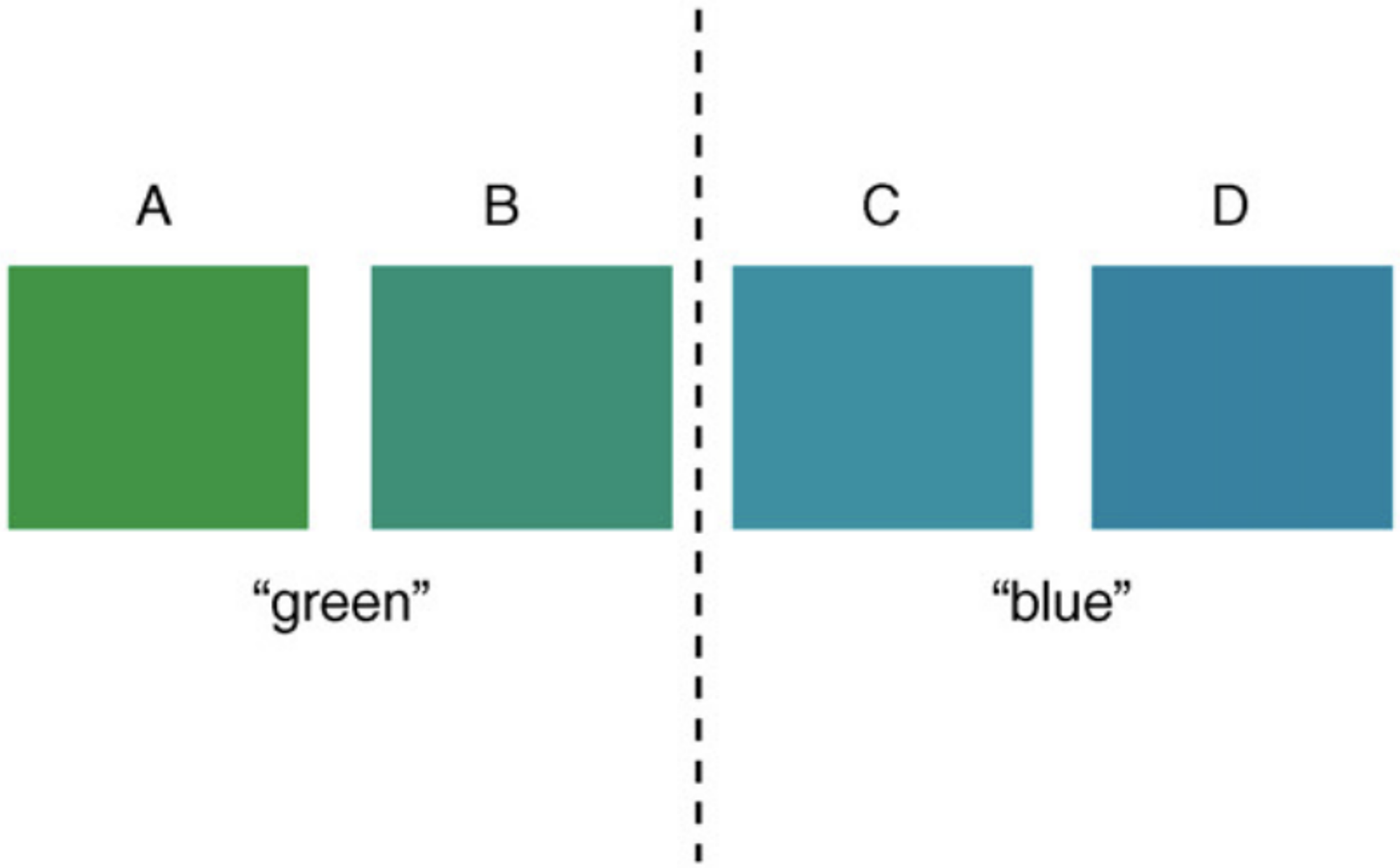
- Grounding to intermediate representations more tractable than grounding directly to pixels.
- Constrains the space of things to ground to.
- **Limitation:**
  - May suffer from cascading error.
  - Not always informed by language.



Top-Down

“What color is the small shiny cube?”





Language, thought, and color: Regier and Kay 2009

ImageNet: Deng et al. 2009

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: wordnet Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

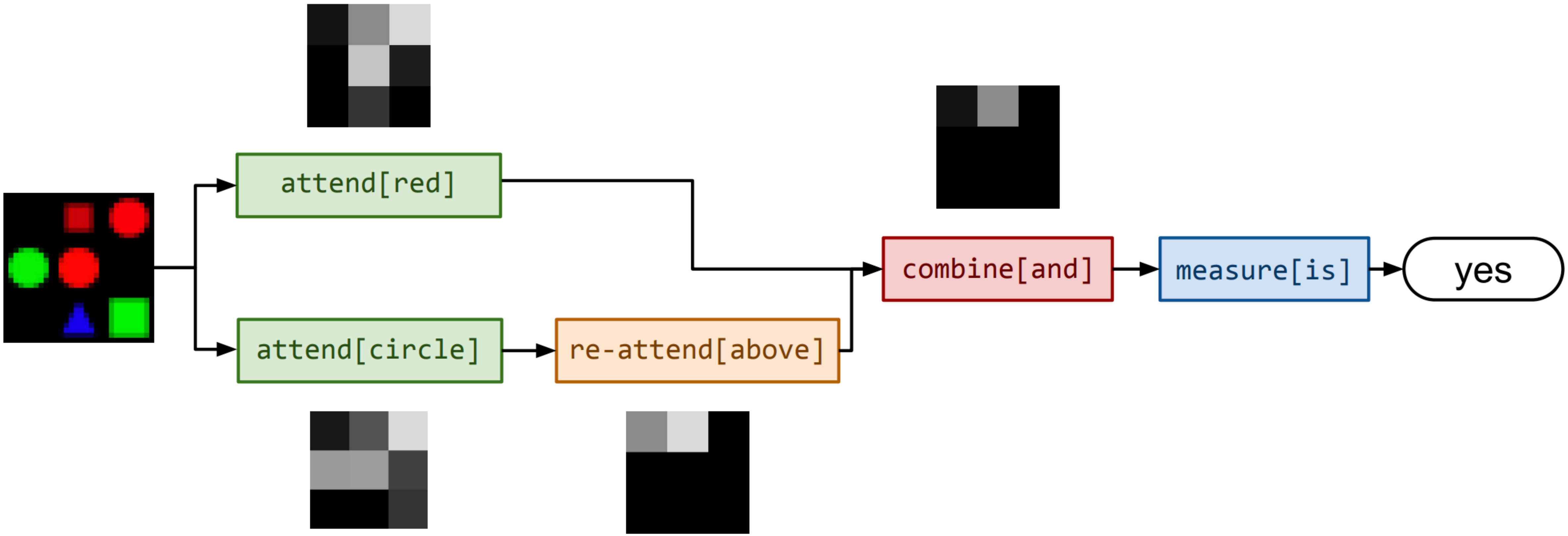
Noun

- S: (n) **wordnet** (any of the machine-readable lexical databases modeled after the Princeton WordNet)
- S: (n) **WordNet**, Princeton WordNet (a machine-readable lexical database organized by meanings; developed at Princeton University)

WordNet: Miller 1995

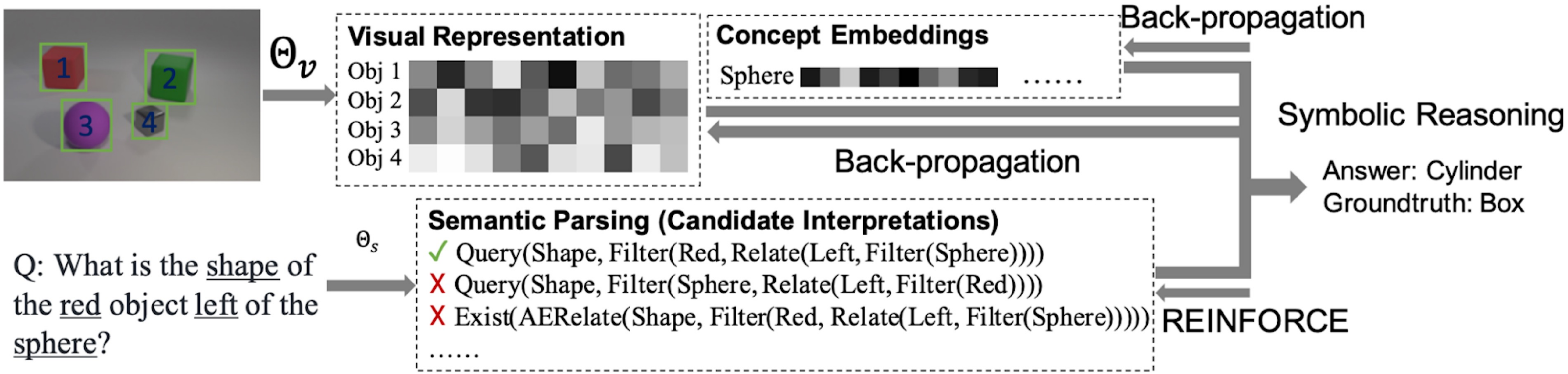
Modular Systems

"Is there a red sphere above a circle?"



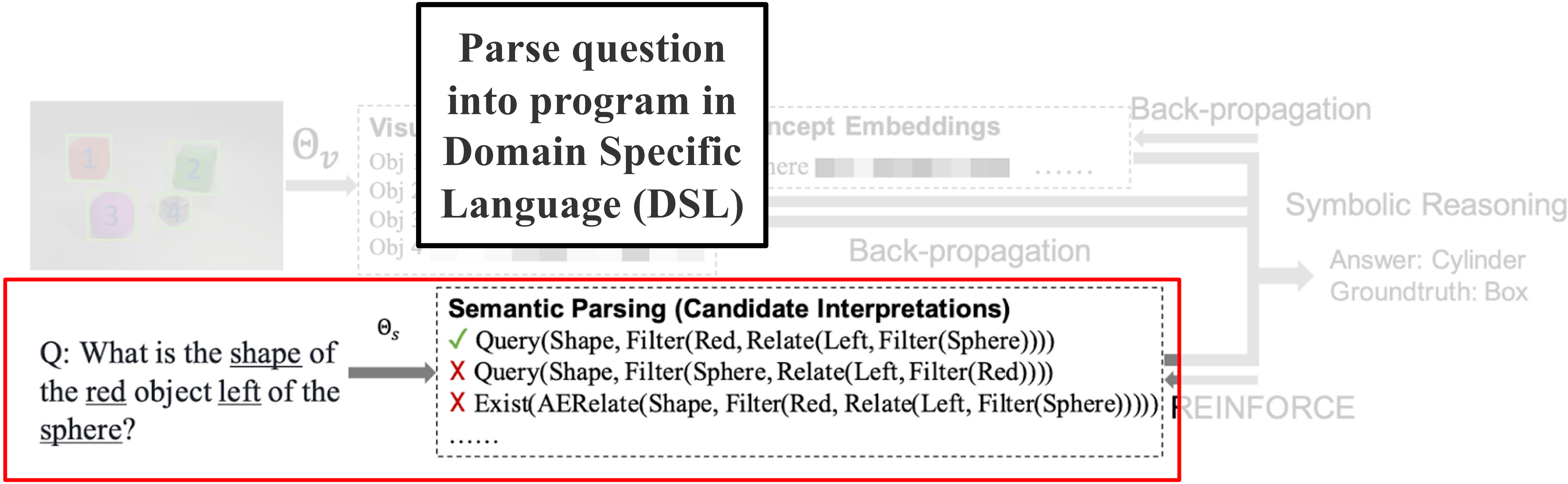
Neural Module Networks: Andreas et al. 2016

Modular Systems



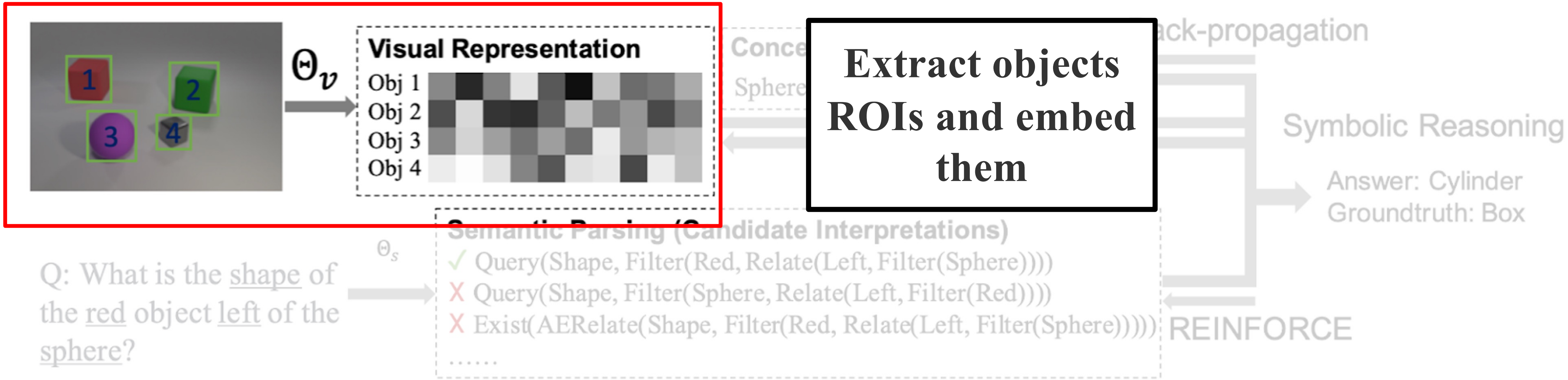
The Neuro-Symbolic Concept Learner: Mao et al. 2020

Modular Systems



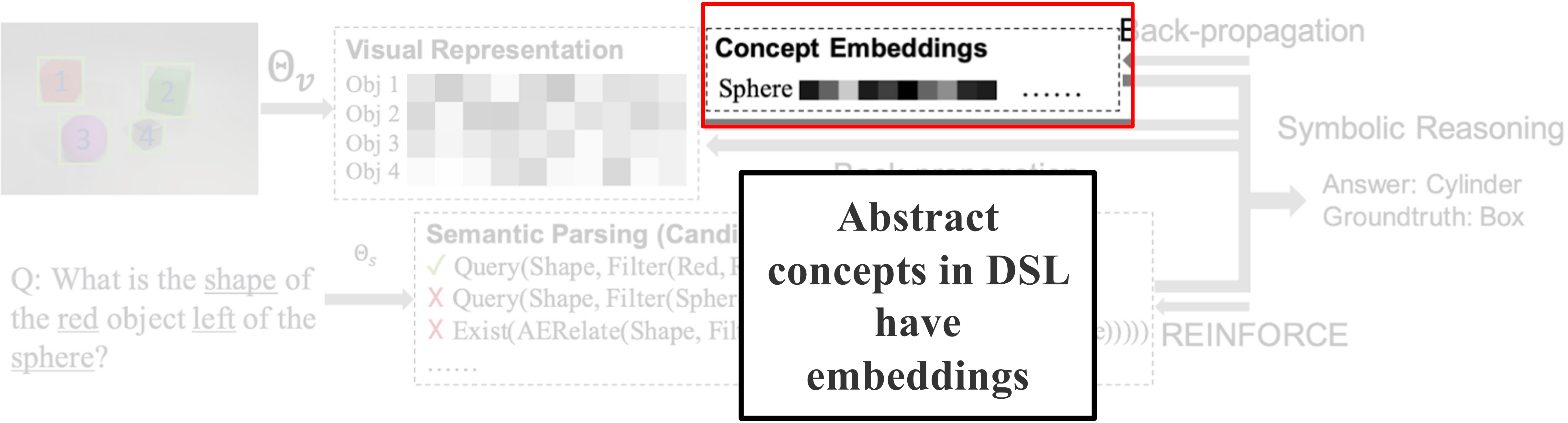
The Neuro-Symbolic Concept Learner: Mao et al. 2020

Modular Systems



Extract objects ROIs and embed them

Modular Systems



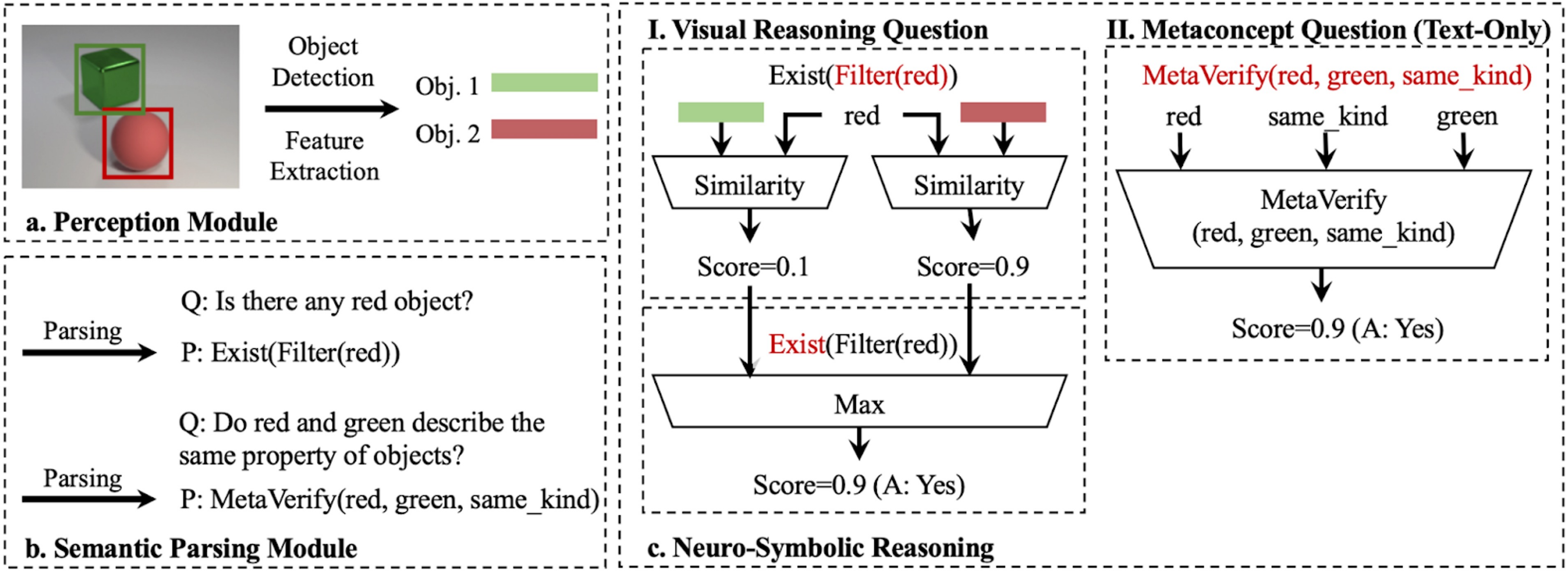
The Neuro-Symbolic Concept Learner: Mao et al. 2020

Modular Systems

Signature	Implementation
`Scene`() $\rightarrow$ *out*: ObjectSet	$out_i := 1$
`Filter`(*in*: ObjectSet, *oc*: ObjConcept) $\rightarrow$ *out*: ObjectSet	$out_i := \min(in_i, \text{ObjClassify}(oc)_i)$
`Relate`(*in*: Object, *rc*: RelConcept) $\rightarrow$ *out*: ObjectSet	$out_i := \sum_j (in_j \cdot \text{RelClassify}(rc)_{j,i}))$
`AERelate`(*in*: Object, *a*: Attribute) $\rightarrow$ *out*: ObjectSet	$out_i := \sum_j (in_j \cdot \text{AEClassify}(a)_{j,i}))$
`Intersection`($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) $\rightarrow$ *out*: ObjectSet	$out_i := \min(in_i^{(1)}, in_i^{(2)})$
`Union`($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) $\rightarrow$ *out*: ObjectSet	$out_i := \max(in_i^{(1)}, in_i^{(2)})$
`Query`(*in*: Object, *a*: Attribute) $\rightarrow$ *out*: ObjConcept	$\Pr[out = oc] := \sum_i in_i \cdot \frac{\text{ObjClassify}(oc)_i \cdot b_a^{oc}}{\sum_{oc'} \text{ObjClassify}(oc')_i \cdot b_a^{oc'}}$
`AEQuery`($in^{(1)}$: Object, $in^{(2)}$: Object, *a*: Attribute) $\rightarrow$ *b*: Bool	$b := \sum_i \sum_j (in_i^{(1)} \cdot in_j^{(2)} \cdot \text{AEClassify}(a)_{j,i}))$
`Exist`(*in*: ObjectSet) $\rightarrow$ *b*: Bool	$b := \max_i in_i$
`Count`(*in*: ObjectSet) $\rightarrow$ *i*: Integer	$i := \sum_i in_i$
`CLessThan`($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) $\rightarrow$ *b*: Bool	$b := \sigma\big((\sum_i in_i^{(2)} - \sum_i in_i^{(1)} - 1 + \gamma_c)/\tau_c\big)$
`CGreaterThan`($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) $\rightarrow$ *b*: Bool	$b := \sigma\big((\sum_i in_i^{(1)} - \sum_i in_i^{(2)} - 1 + \gamma_c)/\tau_c\big)$
`CEqual`($in^{(1)}$: ObjectSet, $in^{(2)}$: ObjectSet) $\rightarrow$ *b*: Bool	$b := \sigma\big((-\lvert\sum_i in_i^{(1)} - \sum_i in_i^{(2)}\rvert + \gamma_c)/(\gamma_c \cdot \tau_c)\big)$

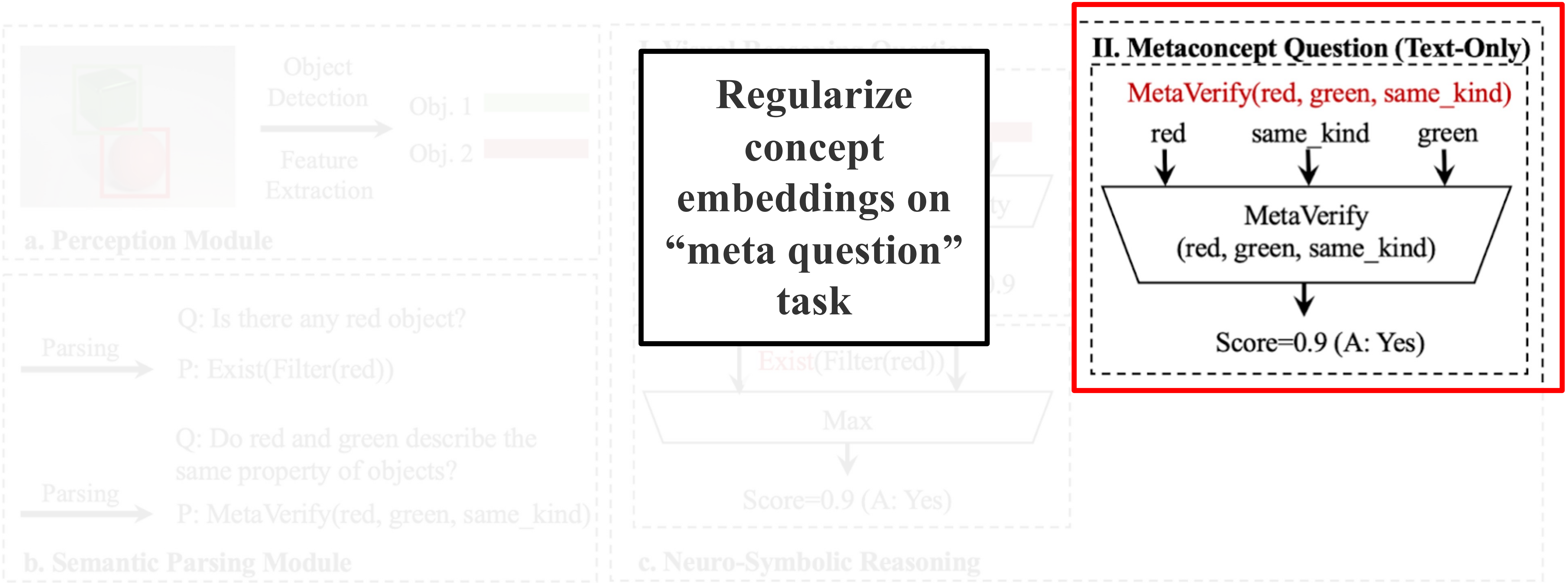
All operations deterministic and pre-defined!

Modular Systems



Visual Concept-Metaconcept Learning: Han et al. 2019

Modular Systems



Regularize concept embeddings on “meta question” task

Modular Systems

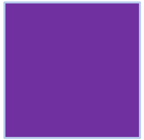
“block” == “square”

	GRU-CNN	MAC	NS-CL	VCML
CLEVR	$50.0_{\pm 0.0}$	$68.7_{\pm 3.8}$	$80.2_{\pm 3.1}$	$\mathbf{94.1_{\pm 4.6}}$
GQA	$50.0_{\pm 0.0}$	$49.5_{\pm 0.2}$	$49.3_{\pm 0.6}$	$\mathbf{50.5_{\pm 0.1}}$

Learning *synonyms* helps zero-shot generalization

Visual Concept-Metaconcept Learning: Han et al. 2019

Modular Systems

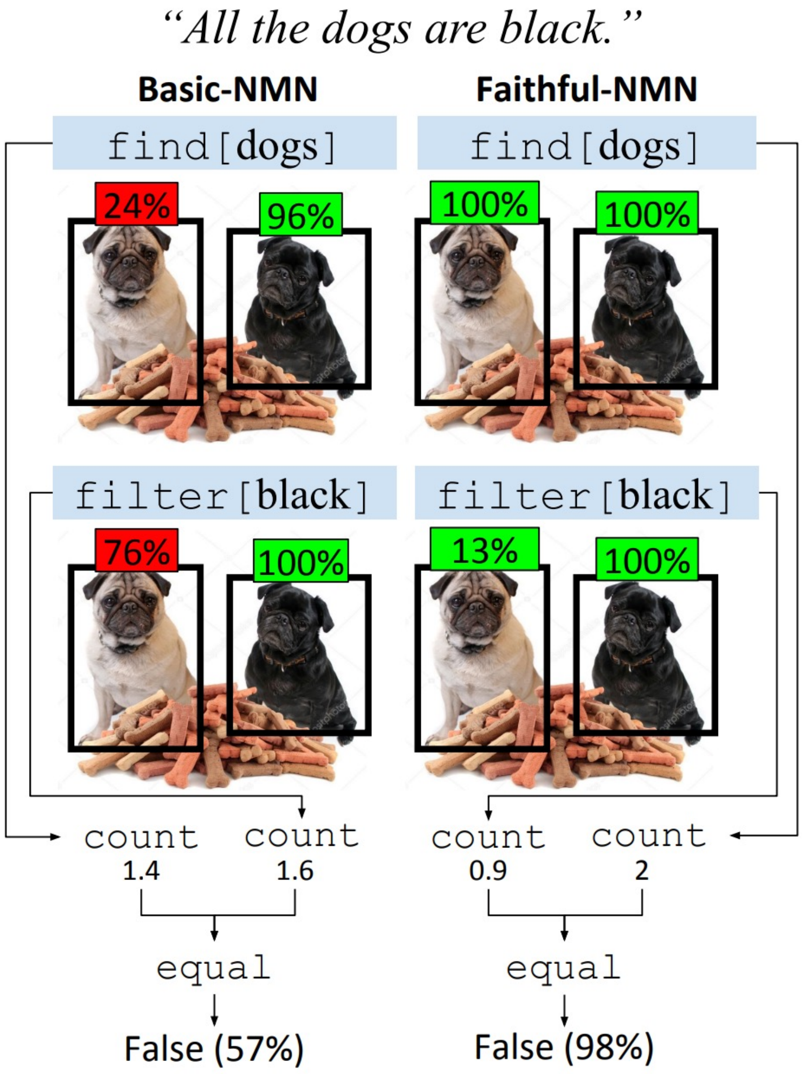
 == ”purple” + “square”

	GRU-CNN	MAC	NS-CL	VCML
CLEVR-200	$50.0_{\pm 0.0}$	$94.2_{\pm 3.3}$	$98.5_{\pm 0.3}$	$\mathbf{98.9}_{\pm \mathbf{0.2}}$
CLEVR-20	$50.0_{\pm 0.0}$	$79.7_{\pm 2.6}$	$\mathbf{95.7}_{\pm \mathbf{0.0}}$	$95.1_{\pm 1.6}$

Learning *same kind* helps compositional generalization

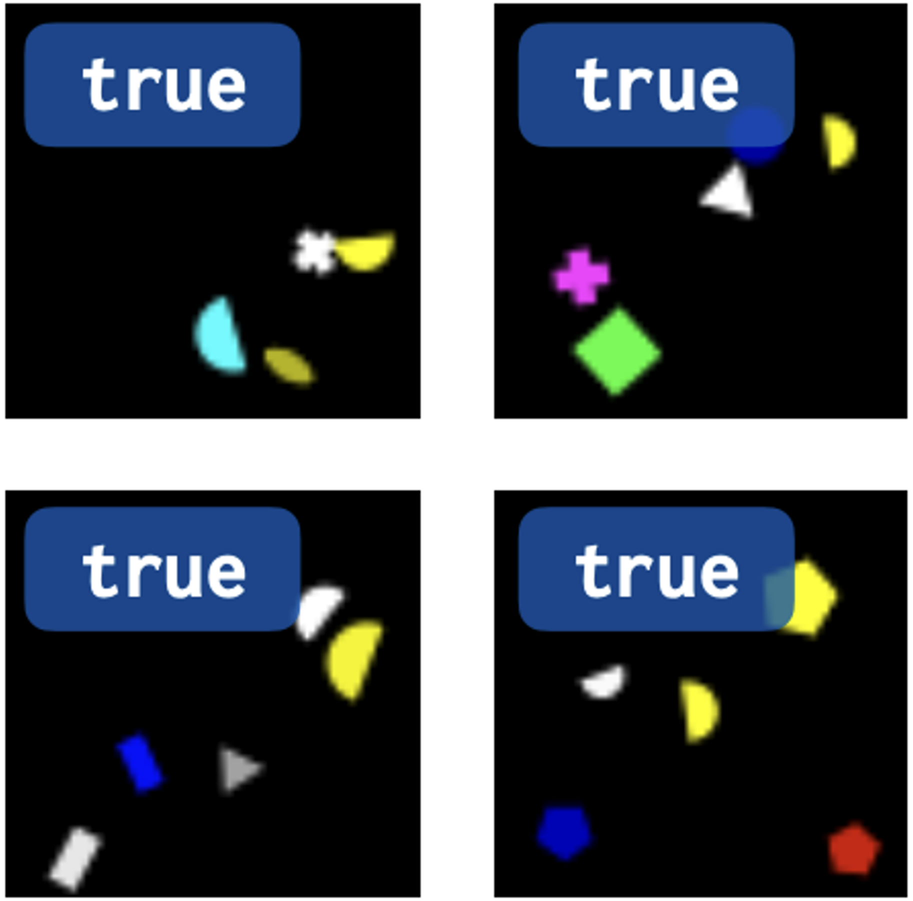
Visual Concept-Metaconcept Learning: Han et al. 2019

Modular Systems



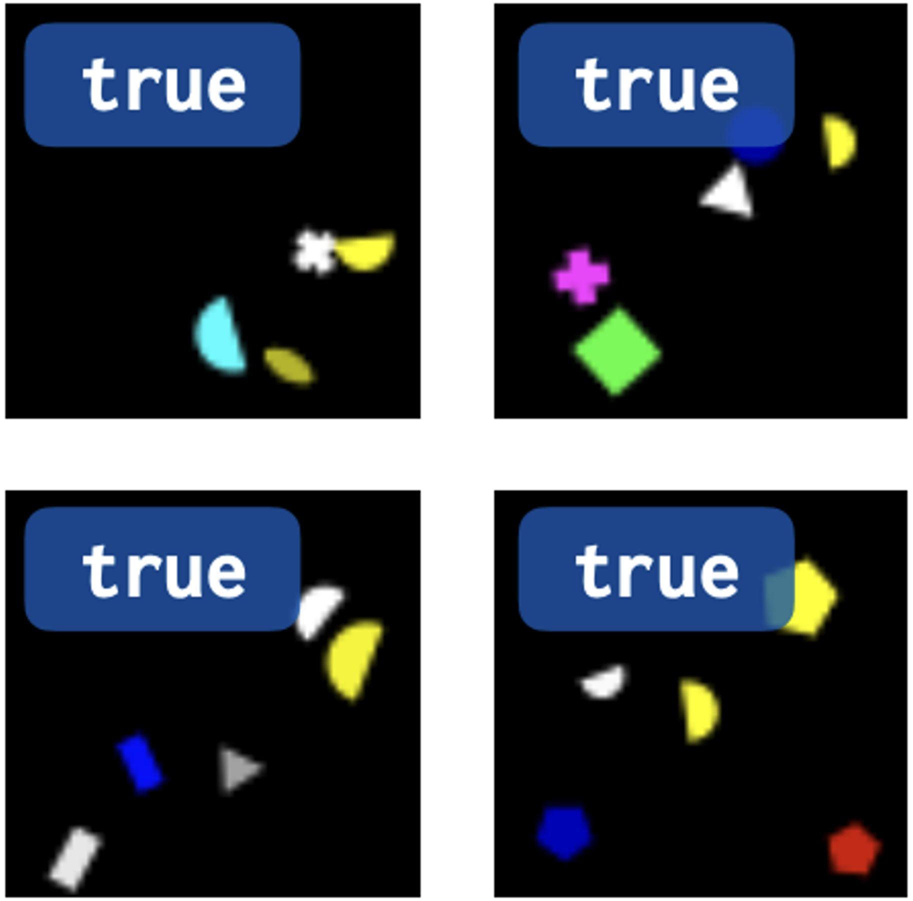
Obtaining Faithful Interpretations from Compositional Neural Networks: Subramanian et al. 2020

Language as Signal for Abstractions



true

Language as Signal for Abstractions

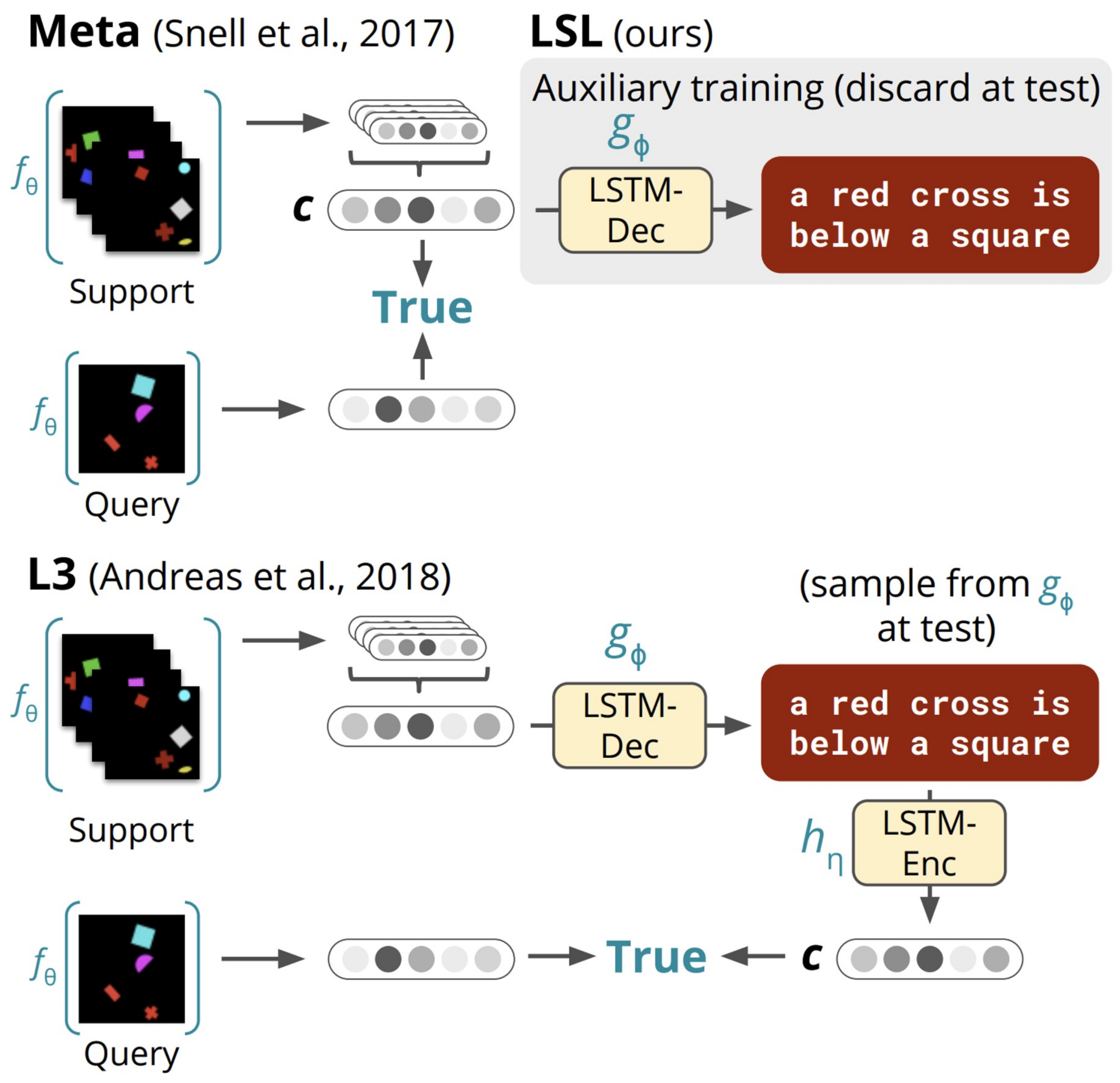


Available at Training

a white shape is left of a yellow semicircle

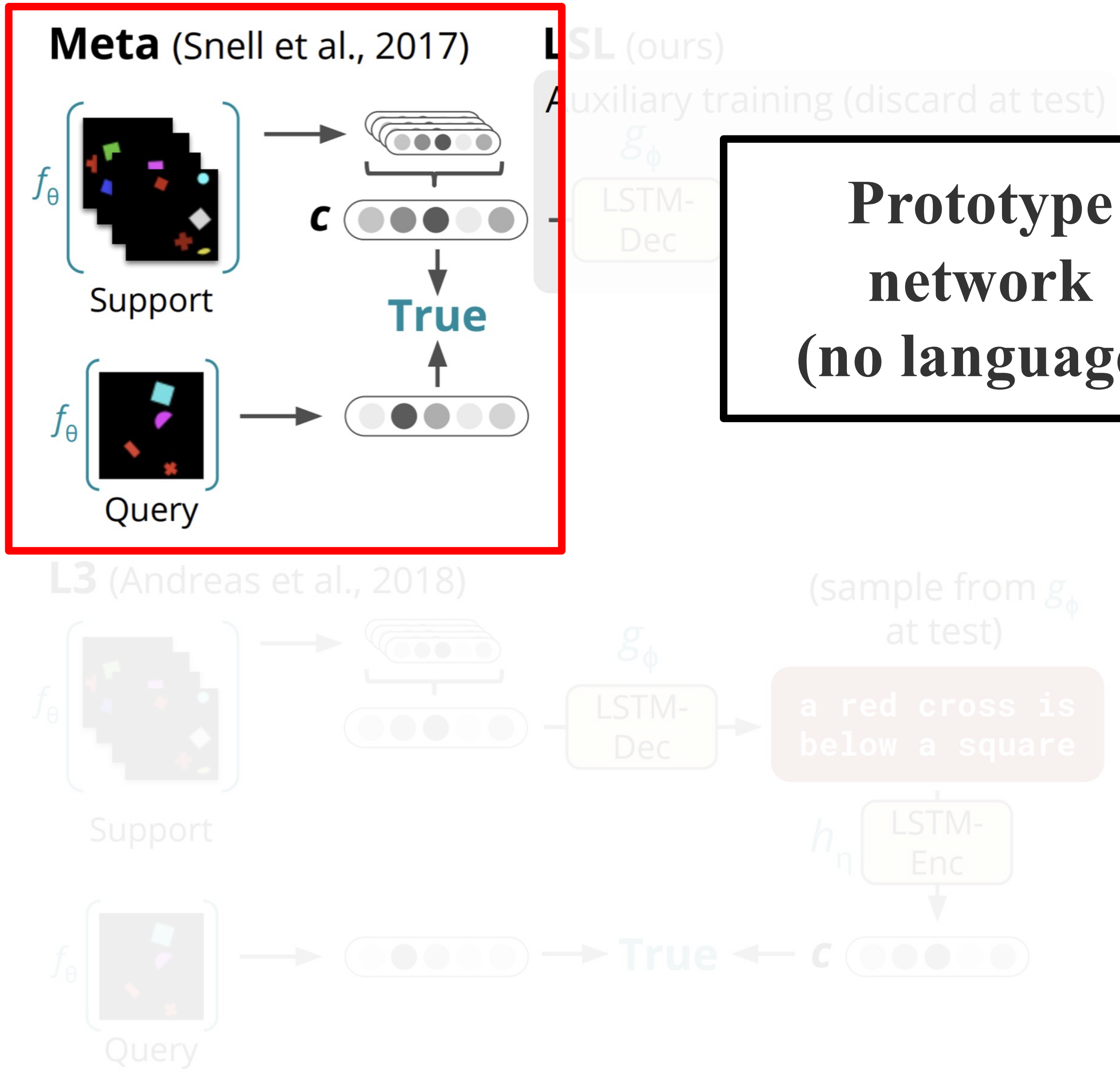
true

Language as Signal for Abstractions



Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020

Language as Signal for Abstractions

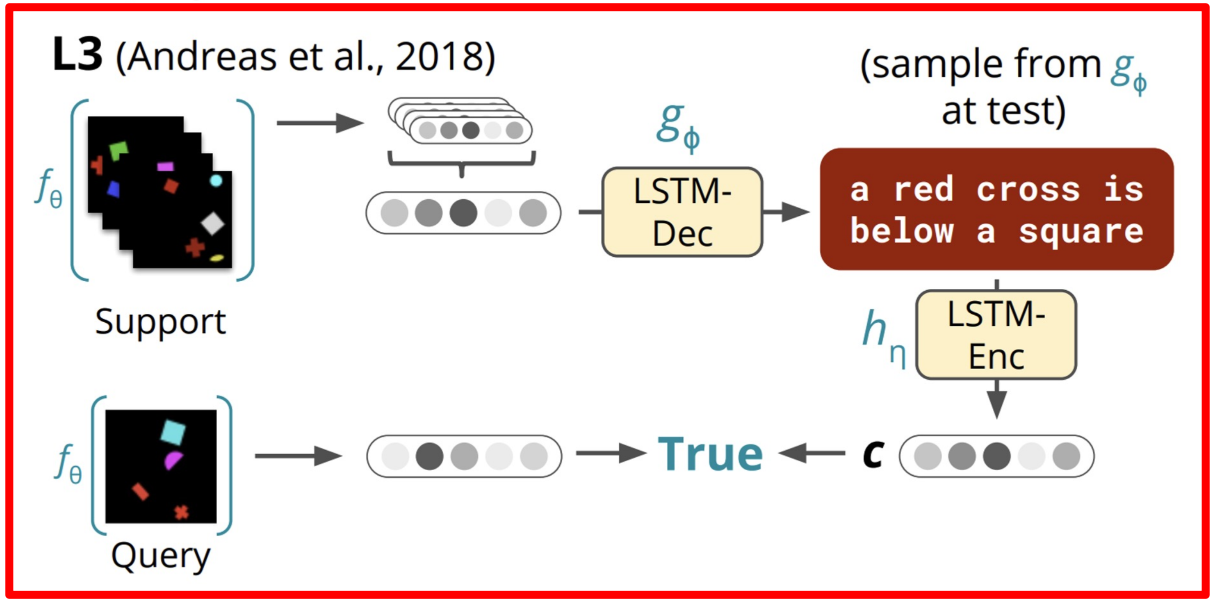


Prototype network (no language)

Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020

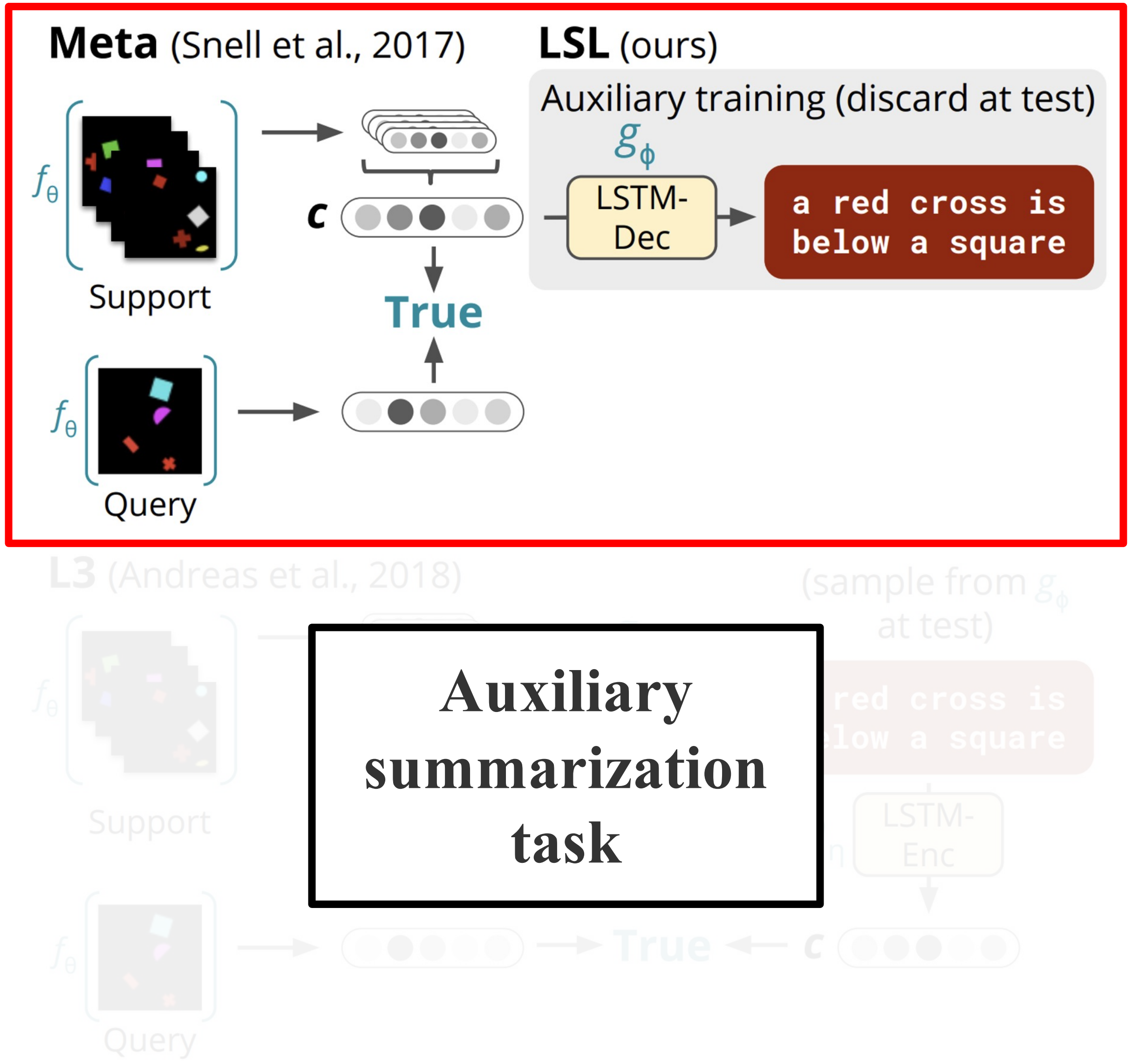
Language as Signal for Abstractions

Discrete language bottleneck



Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020

Language as Signal for Abstractions



Auxiliary summarization task

Language as Signal for Abstractions

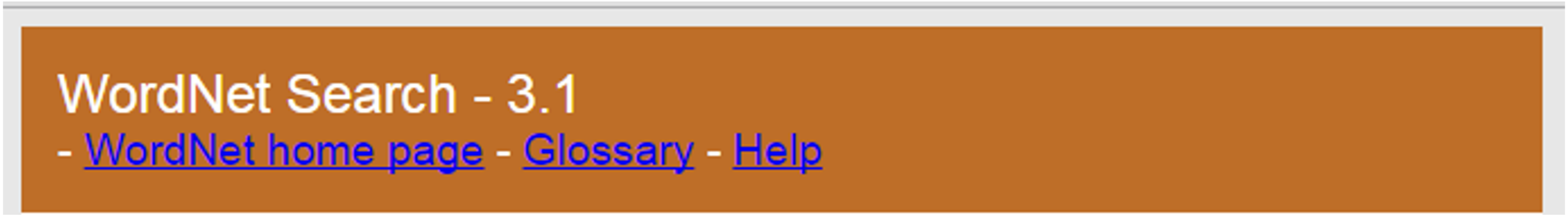
Test Set Accuracy

	ShapeWorld	Birds ($D = 20$)
Meta	60.59 ± 1.07	57.97 ± 0.96
L3	66.60 ± 1.18	53.96 ± 1.06
LSL	$\mathbf{67.29} \pm \mathbf{1.03}$	$\mathbf{61.24} \pm \mathbf{0.96}$

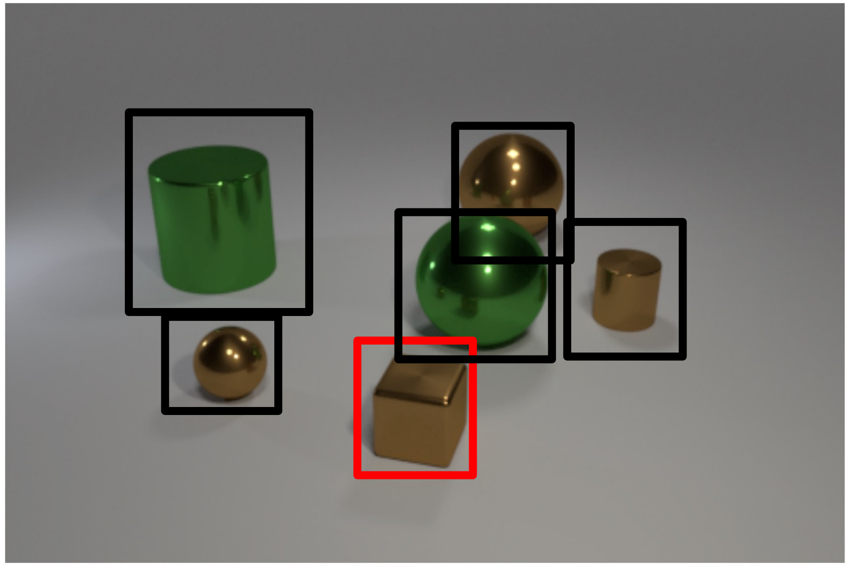
Shaping Visual Representations with Language for Few-Shot Classification: Mu et al. 2020

Top-Down Takeaways

- Language provides labels for supervised learning of perceptual systems.
- Can provide powerful inductive biases in computational structure ***if known***.
- Serves as signal for useful perceptual abstractions to learn either as bottleneck or auxiliary signal.

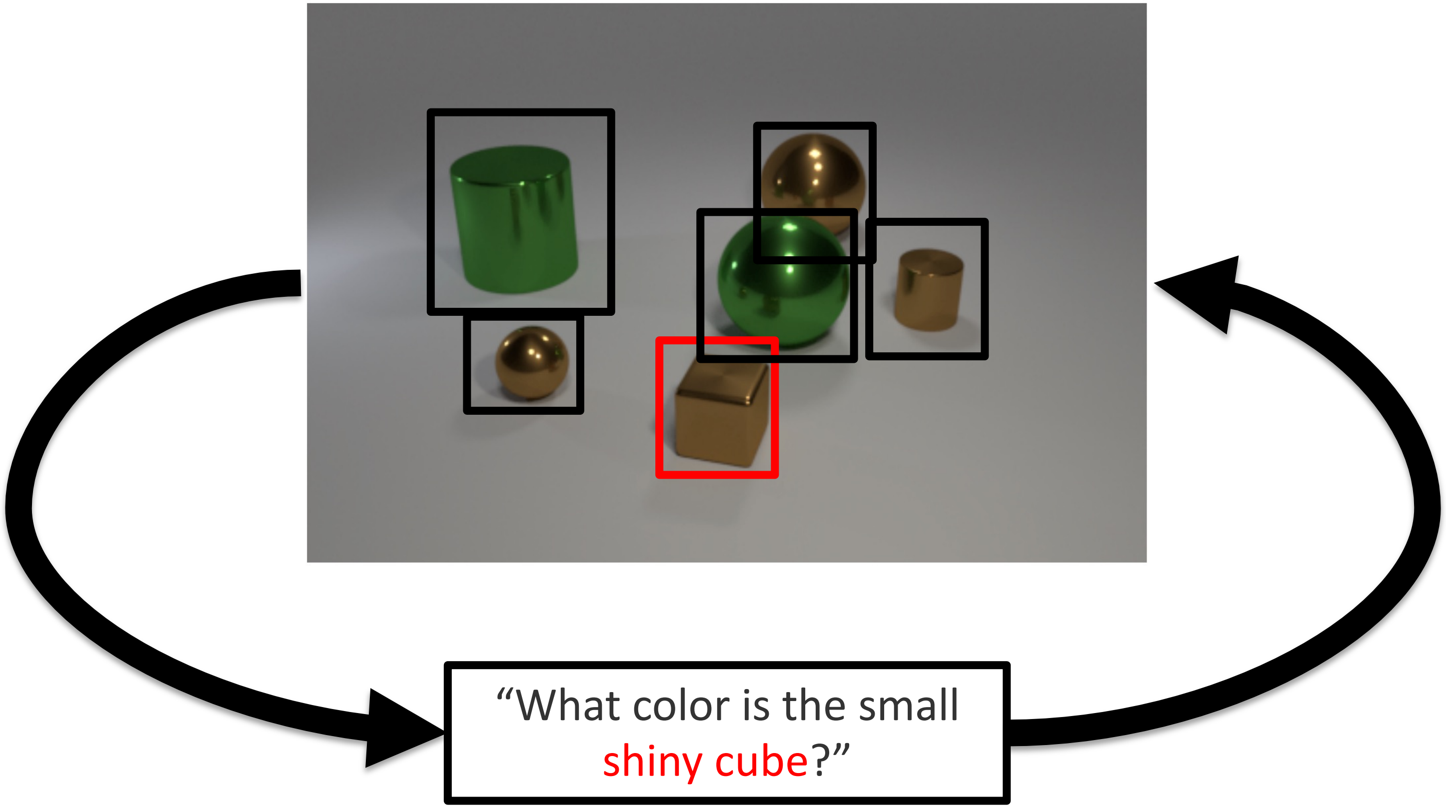


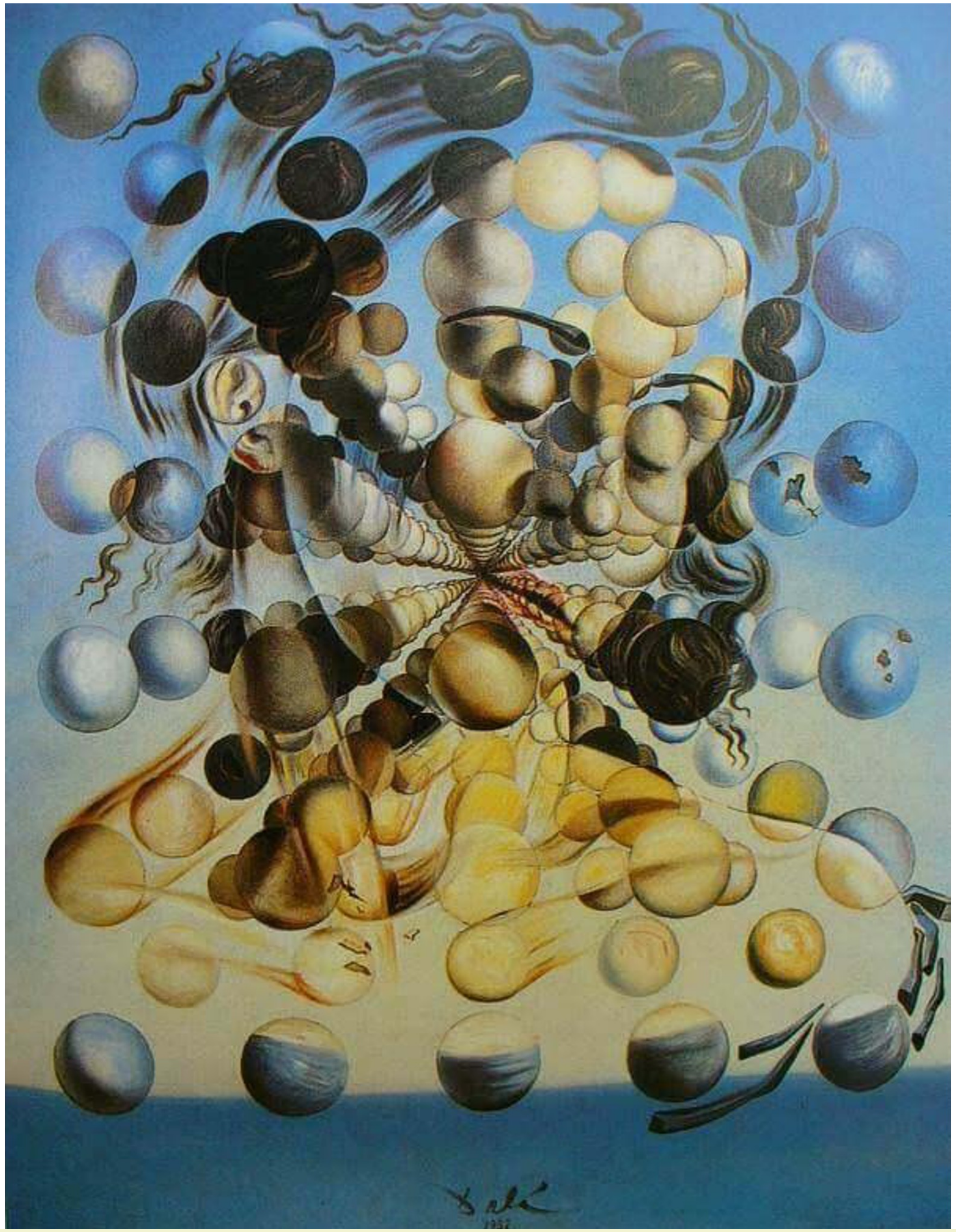
Bottom-Up & Top-Down Reasoning



"What color is the small shiny cube?"

Bottom-Up & Top-Down Reasoning





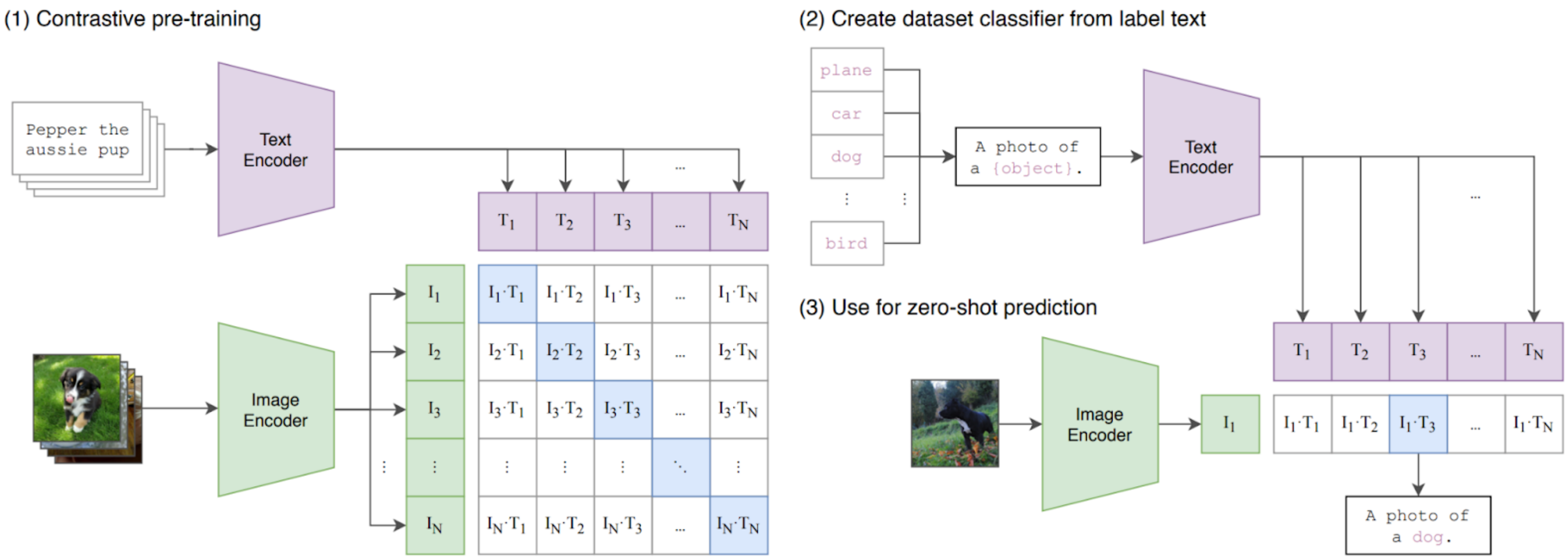
Galatea of the Spheres, Salvador Dali 1952

Extra Slides

Open-Set Models

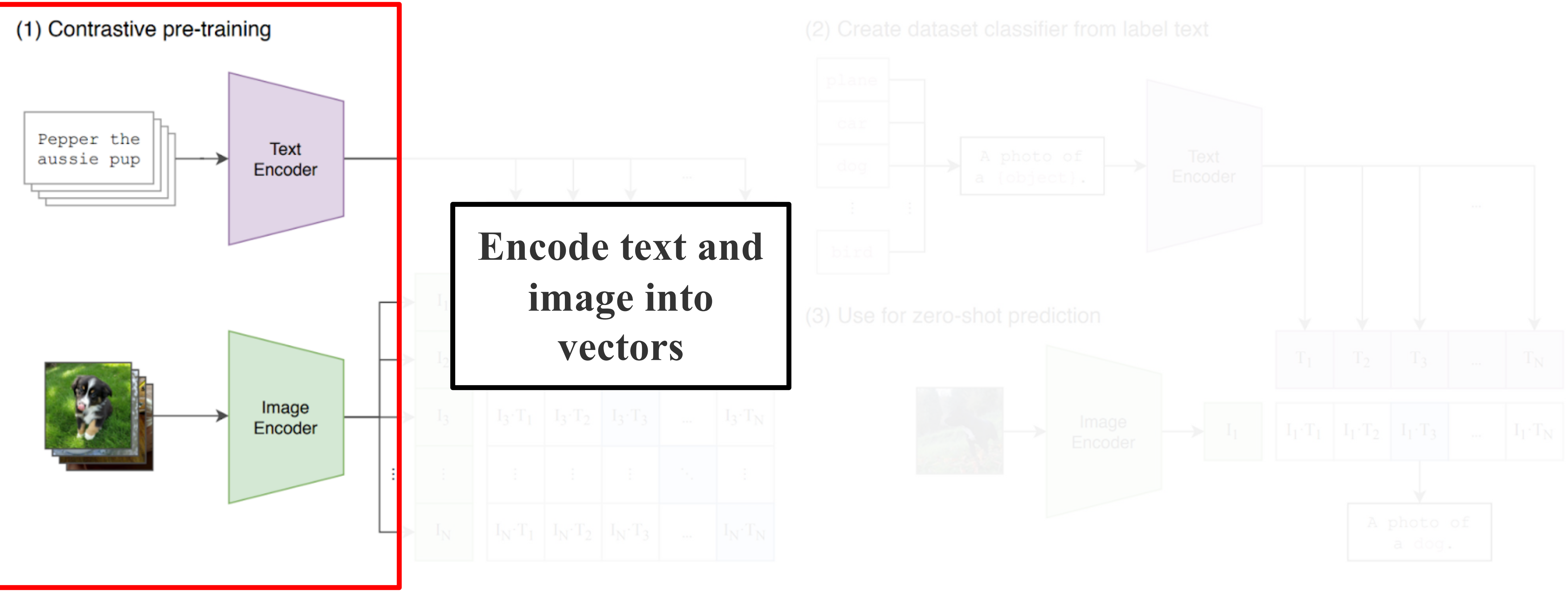
Models which leverage the open-vocabulary of language to enjoy a practically open set of labels!

Open-Set Models



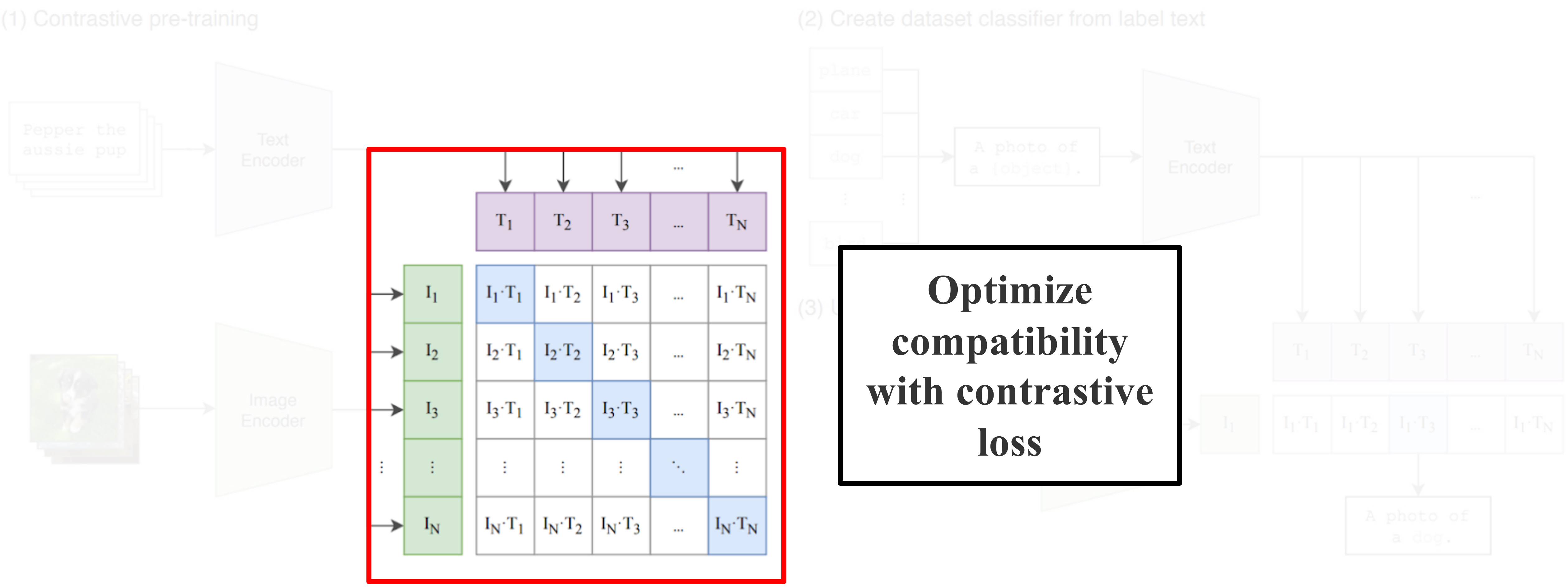
CLIP: Radford et al. 2021

Open-Set Models



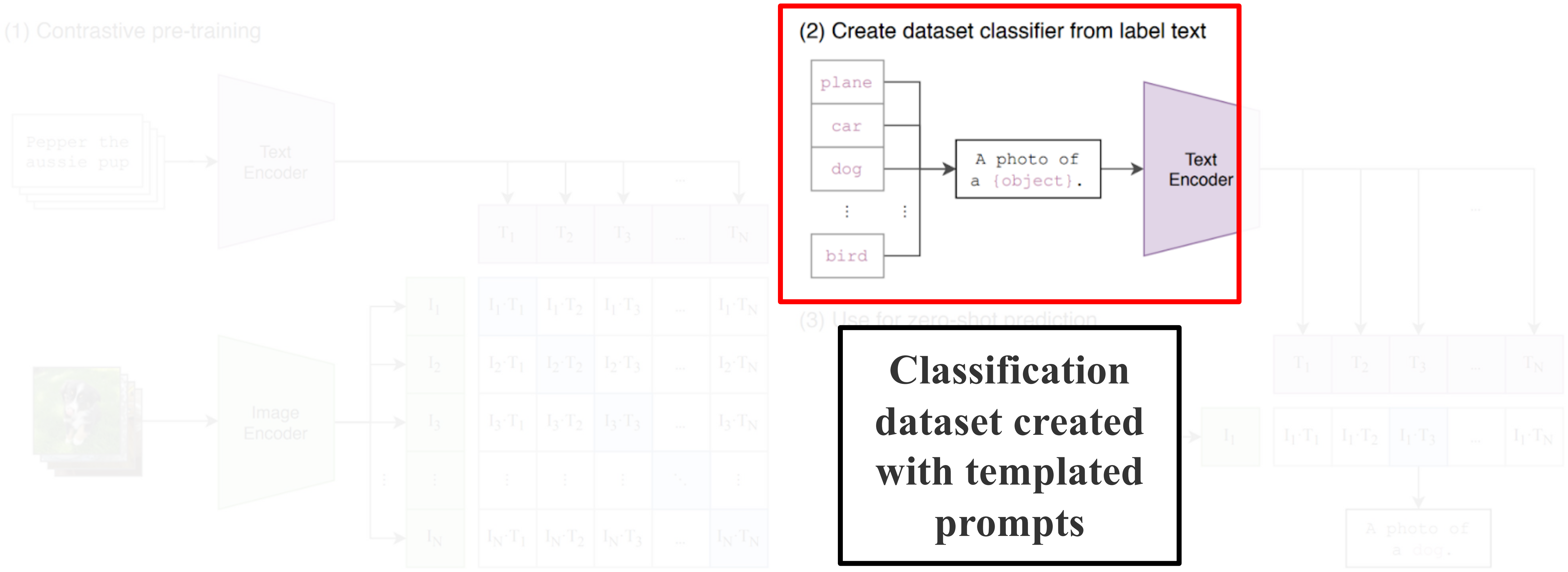
CLIP: Radford et al. 2021

Open-Set Models



CLIP: Radford et al. 2021

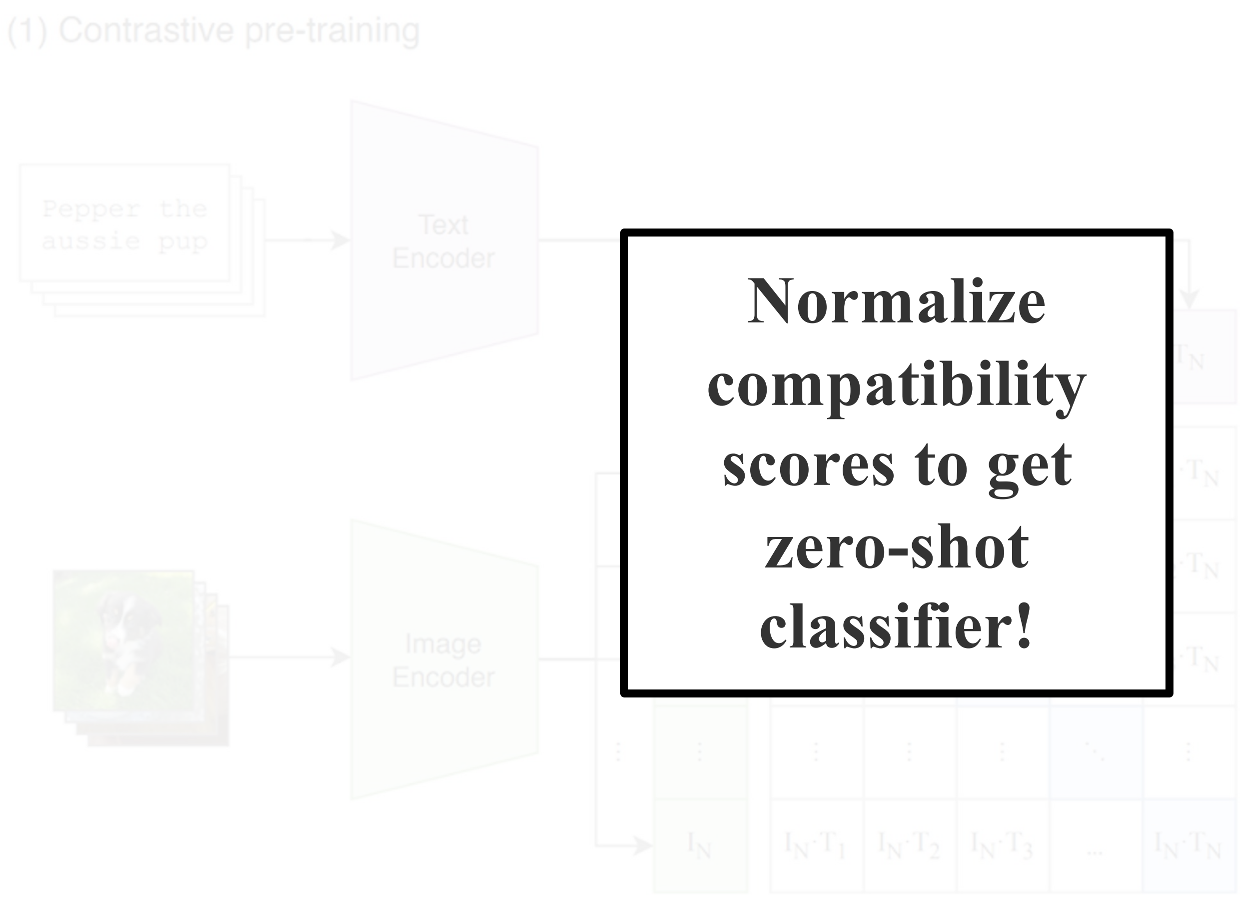
Open-Set Models



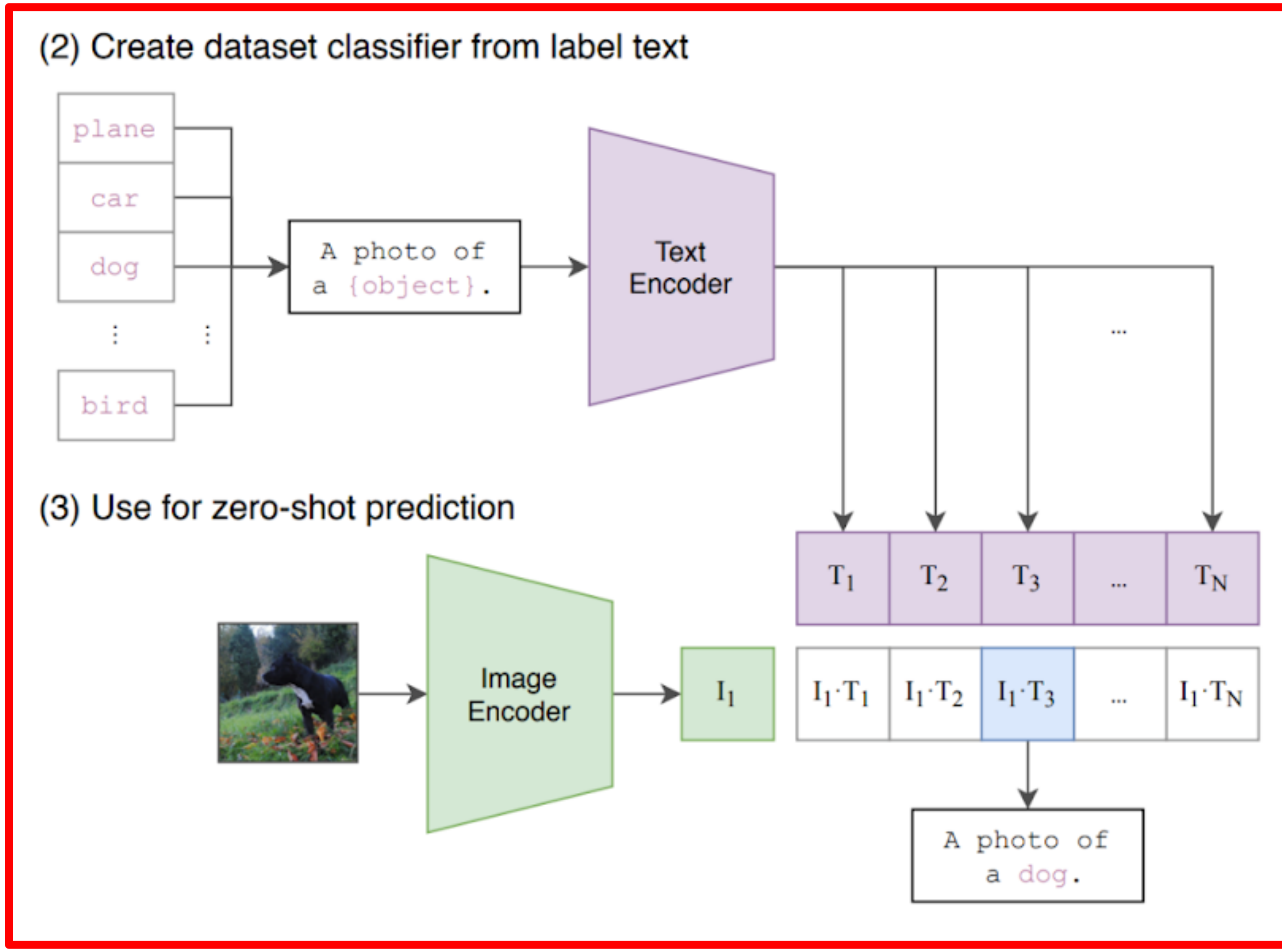
Classification dataset created with templated prompts

CLIP: Radford et al. 2021

Open-Set Models

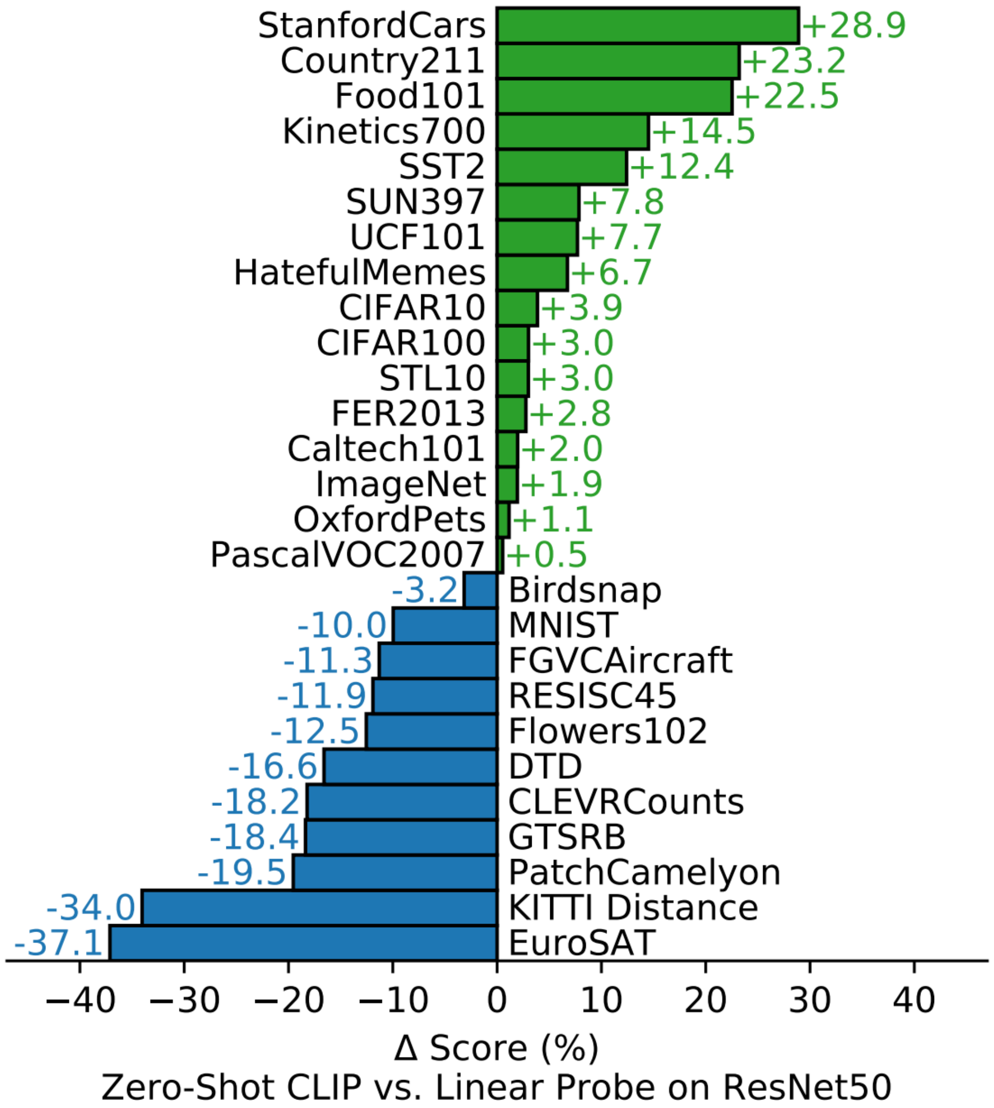


Normalize compatibility scores to get zero-shot classifier!



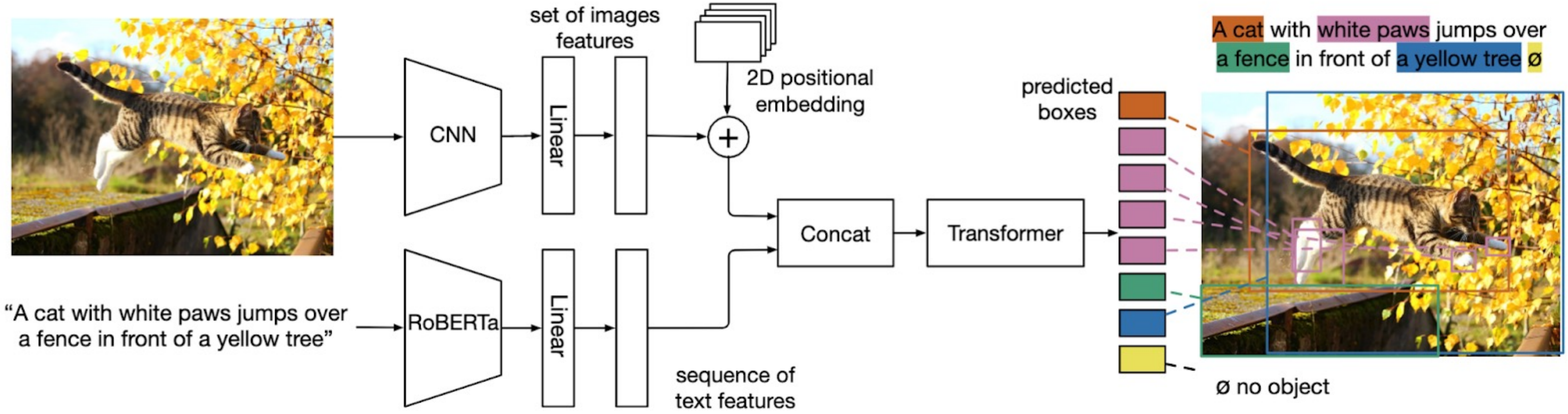
CLIP: Radford et al. 2021

Open-Set Models



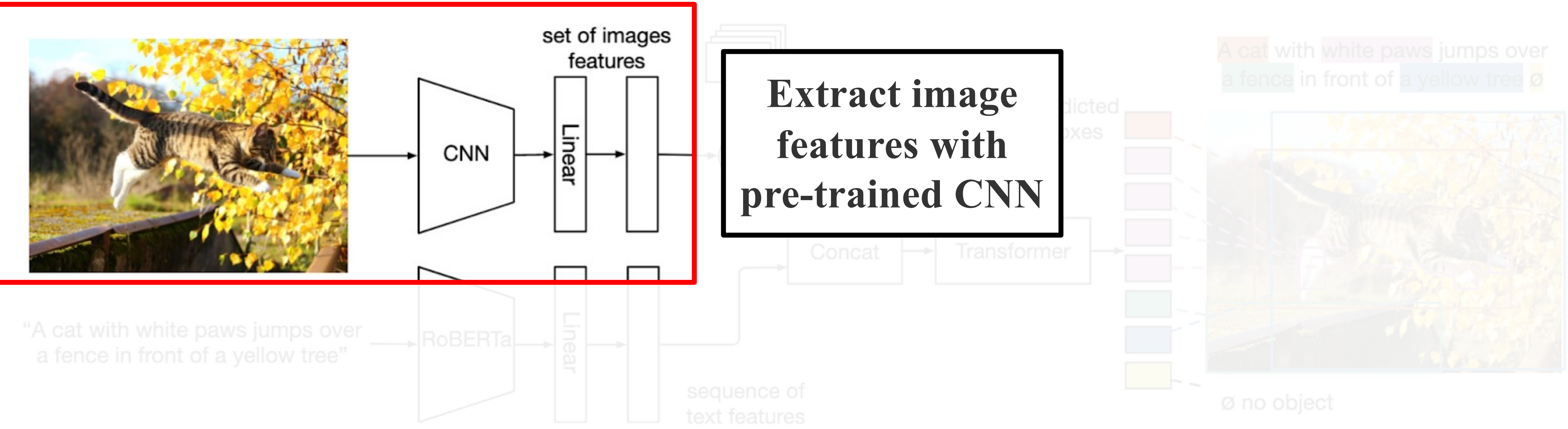
CLIP: Radford et al. 2021

Open-Set Models



MDETR: Kamath et al. 2021

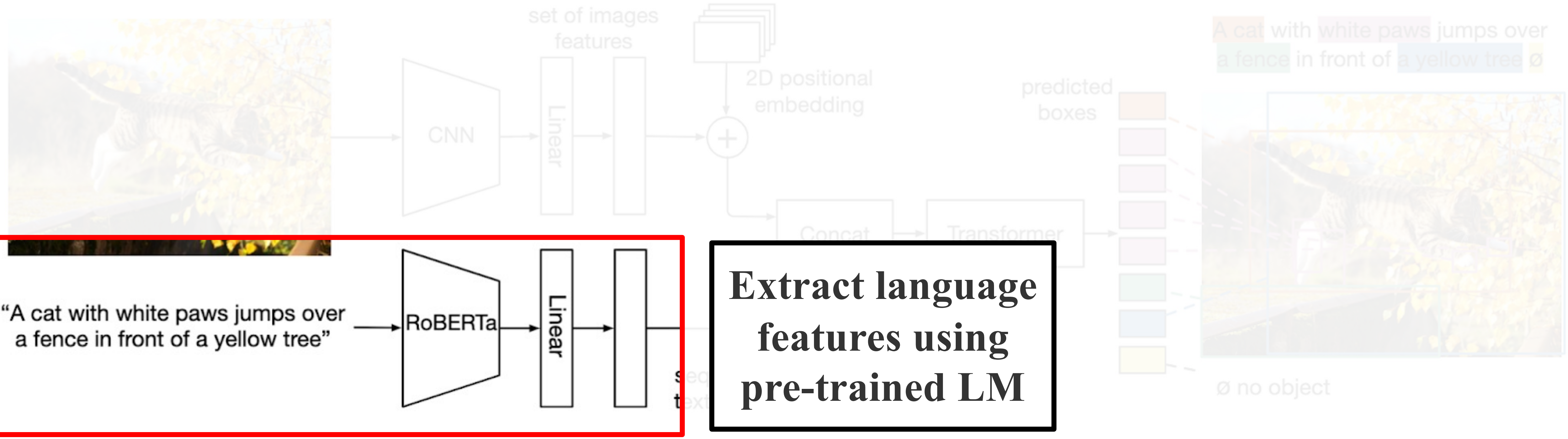
Open-Set Models



Extract image features with pre-trained CNN

MDETR: Kamath et al. 2021

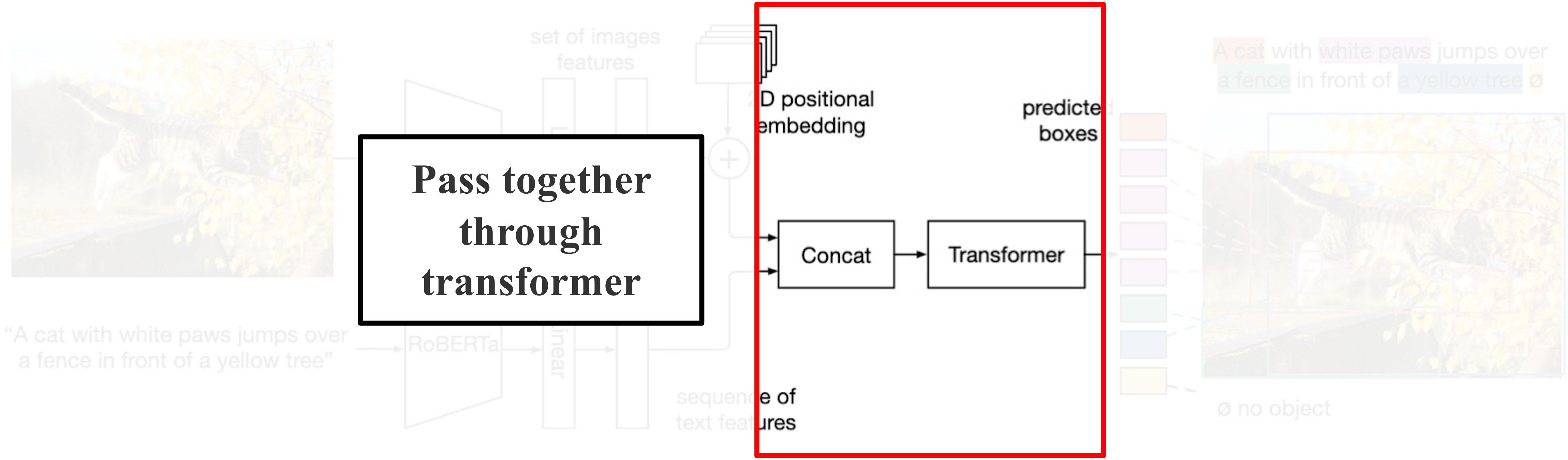
Open-Set Models



Extract language features using pre-trained LM

MDETR: Kamath et al. 2021

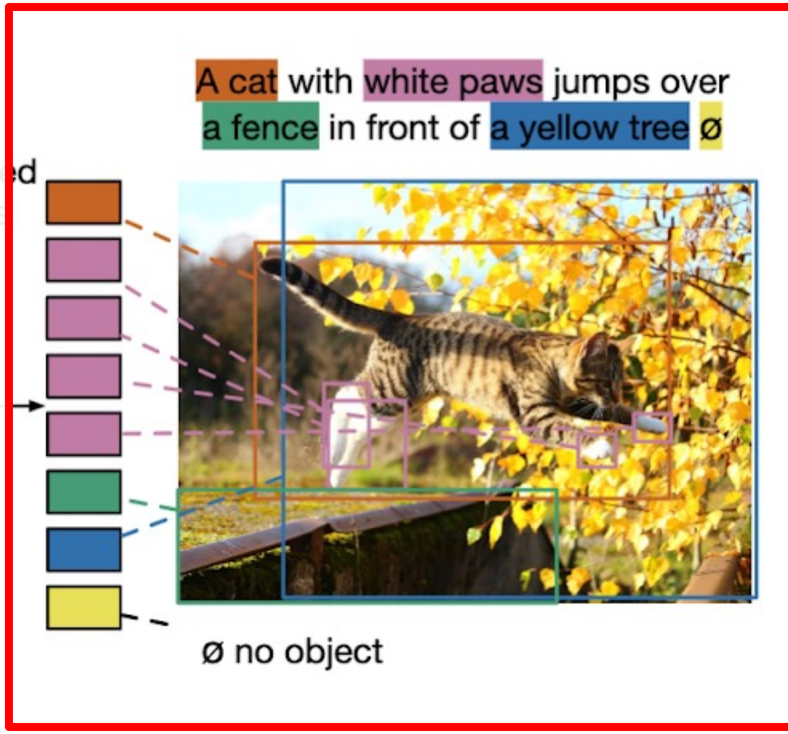
Open-Set Models



MDETR: Kamath et al. 2021

Open-Set Models

Learned embedding "queries" tied to tokens in input text/image regions



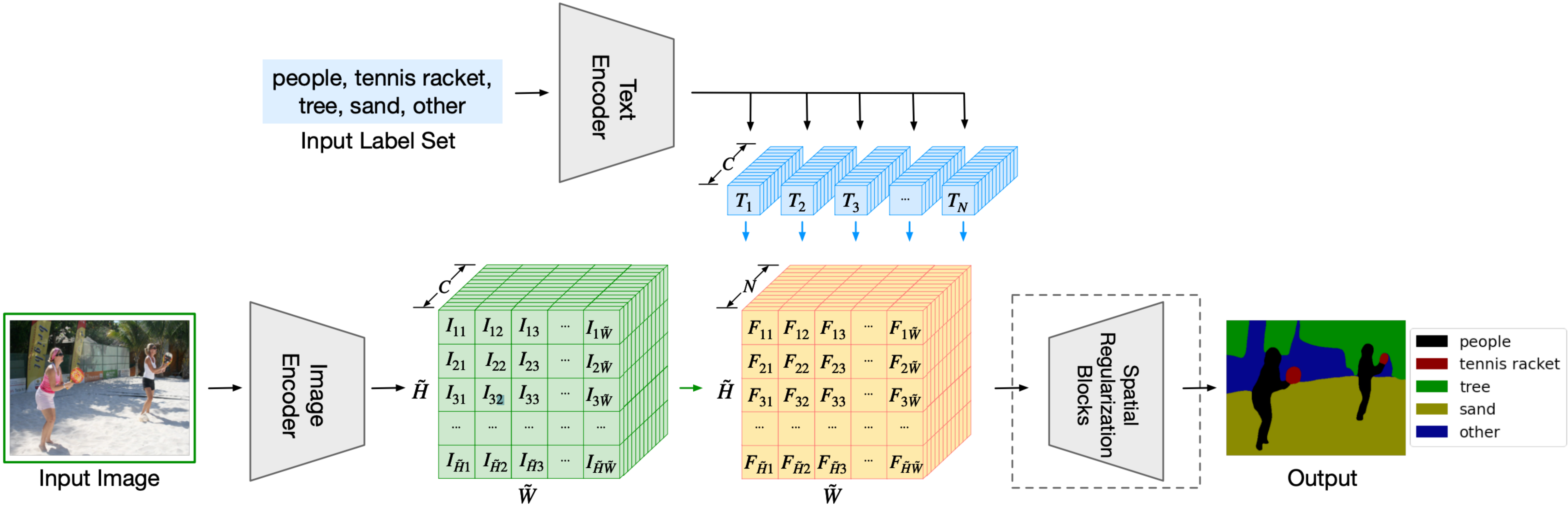
MDETR: Kamath et al. 2021

Open-Set Models

(a) "one small boy climbing a pole with the help of another boy on the ground" **(b)** "A man talking on his cellphone next to a jewelry store"

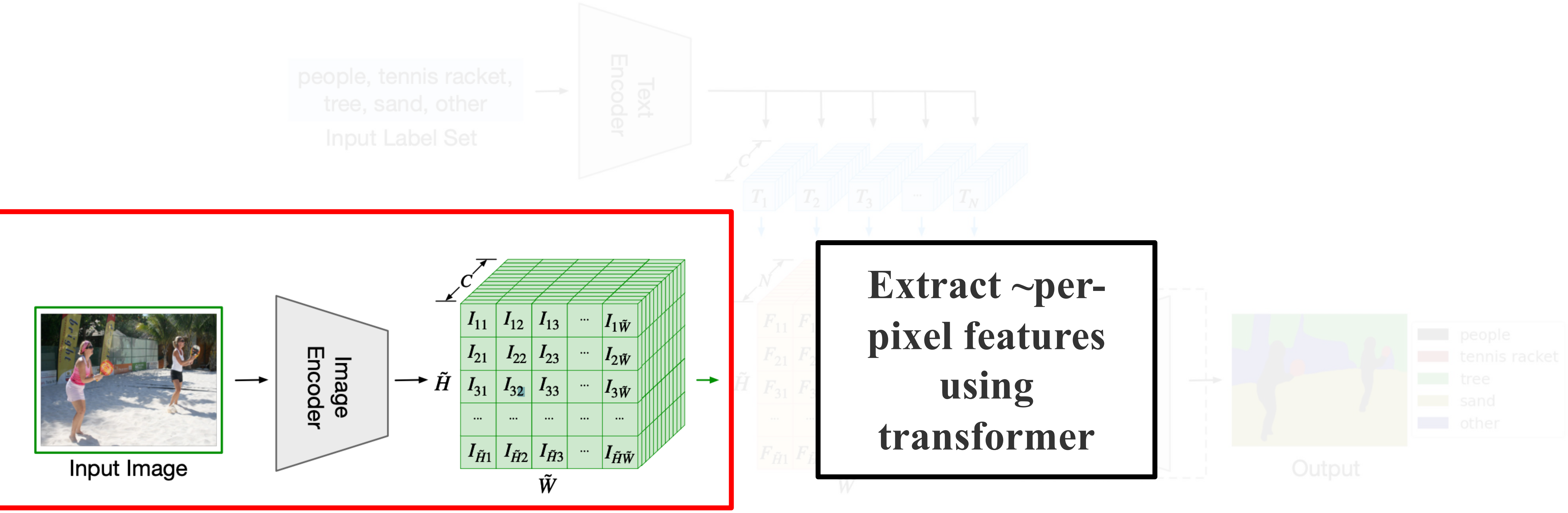
MDETR: Kamath et al. 2021

Open-Set Models



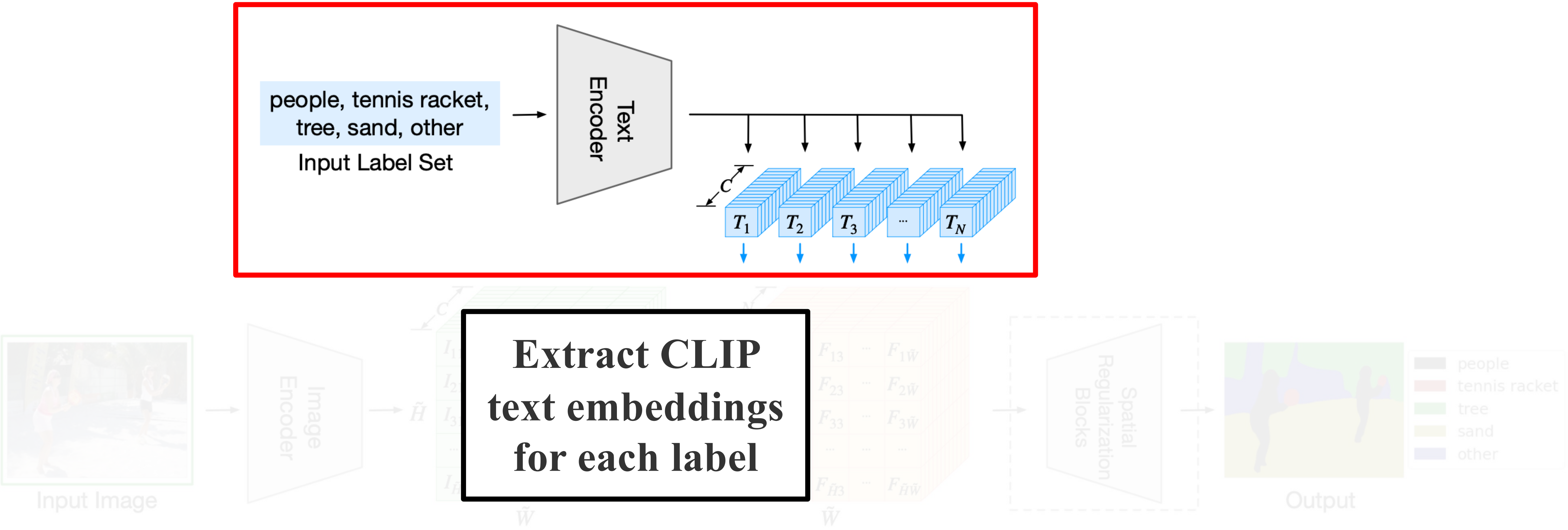
Language-Driven Semantic Segmentation: Li et al. 2022

Open-Set Models



Extract ~per-pixel features using transformer

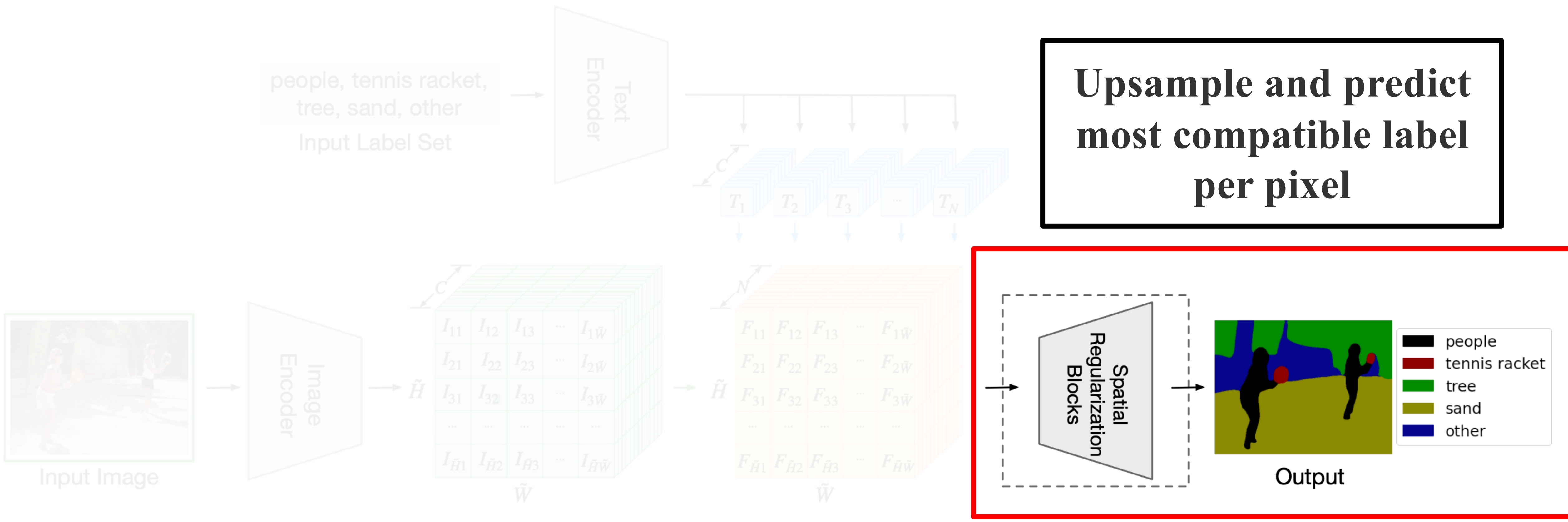
Open-Set Models



Extract CLIP text embeddings for each label

Language-Driven Semantic Segmentation: Li et al. 2022

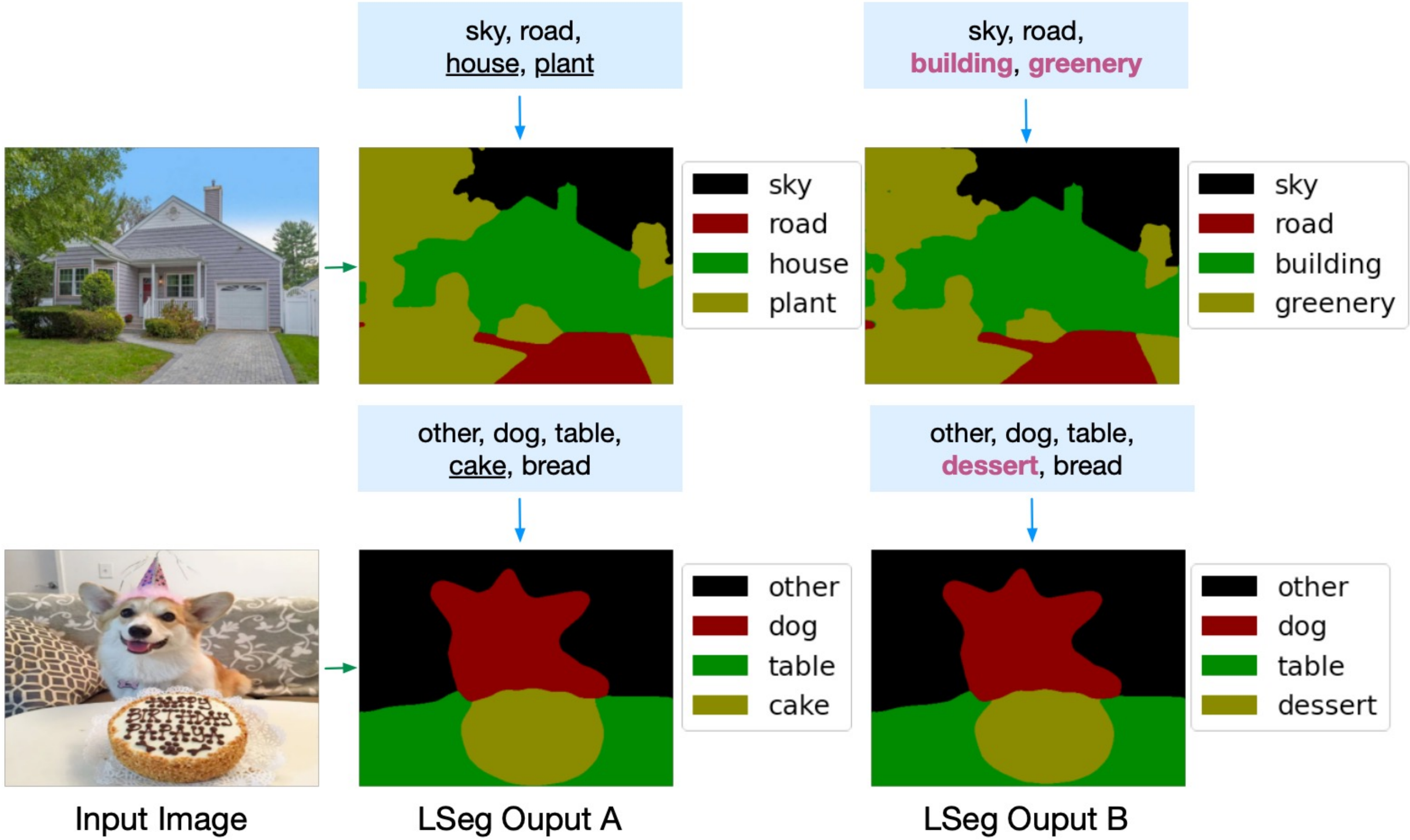
Open-Set Models



Upsample and predict most compatible label per pixel

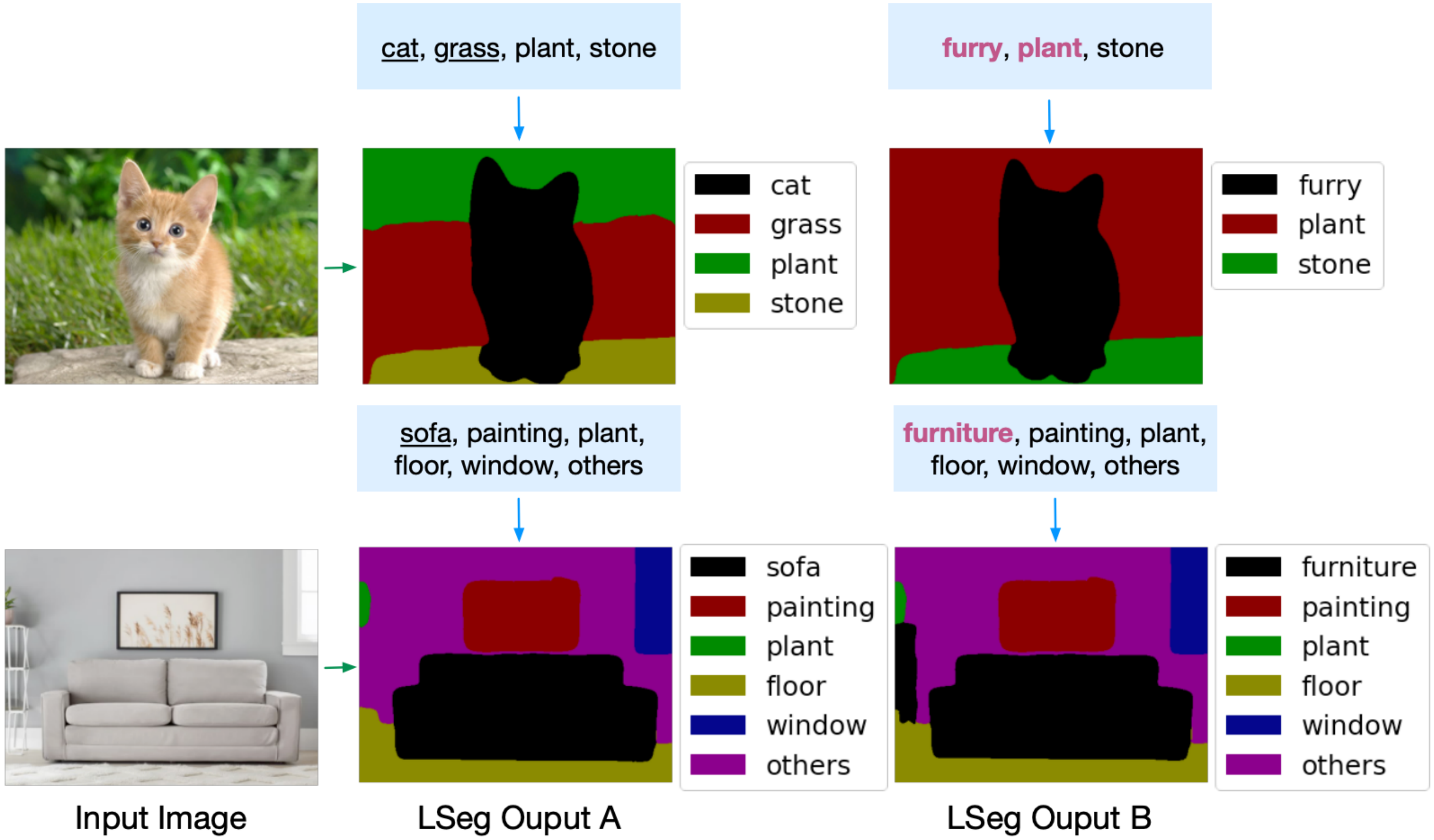
Language-Driven Semantic Segmentation: Li et al. 2022

Open-Set Models



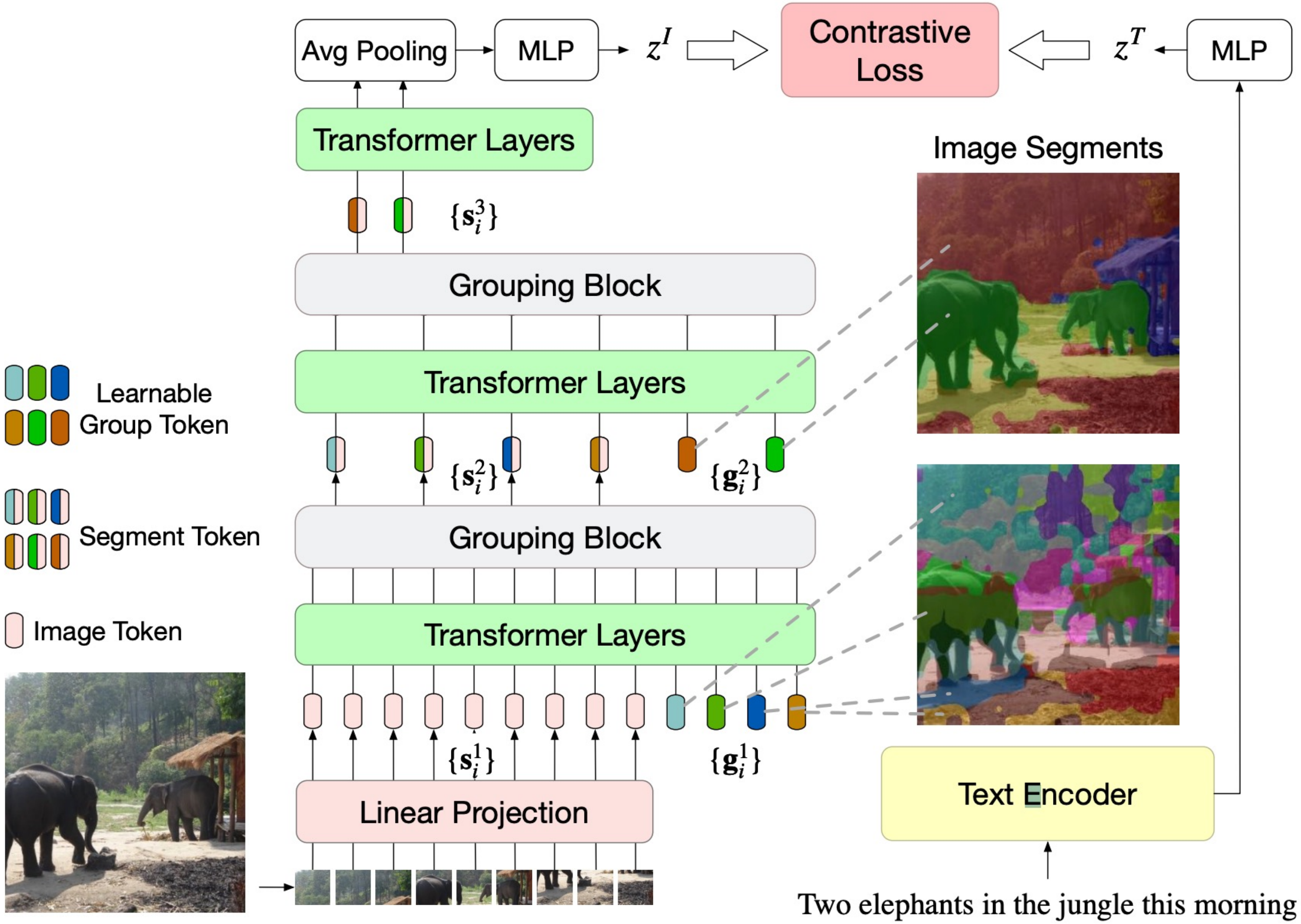
Language-Driven Semantic Segmentation: Li et al. 2022

Open-Set Models



Language-Driven Semantic Segmentation: Li et al. 2022

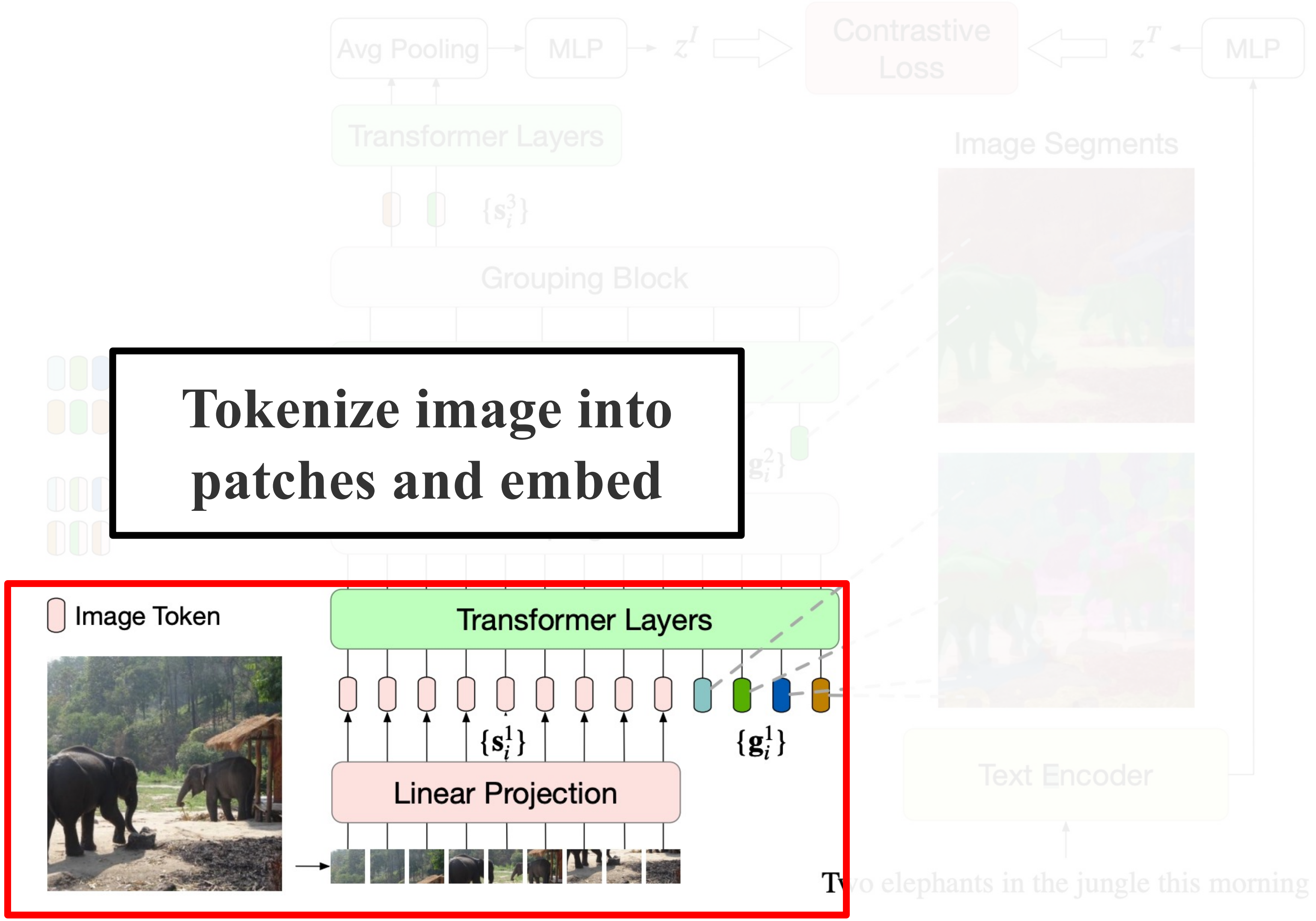
Open-Set Models



GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

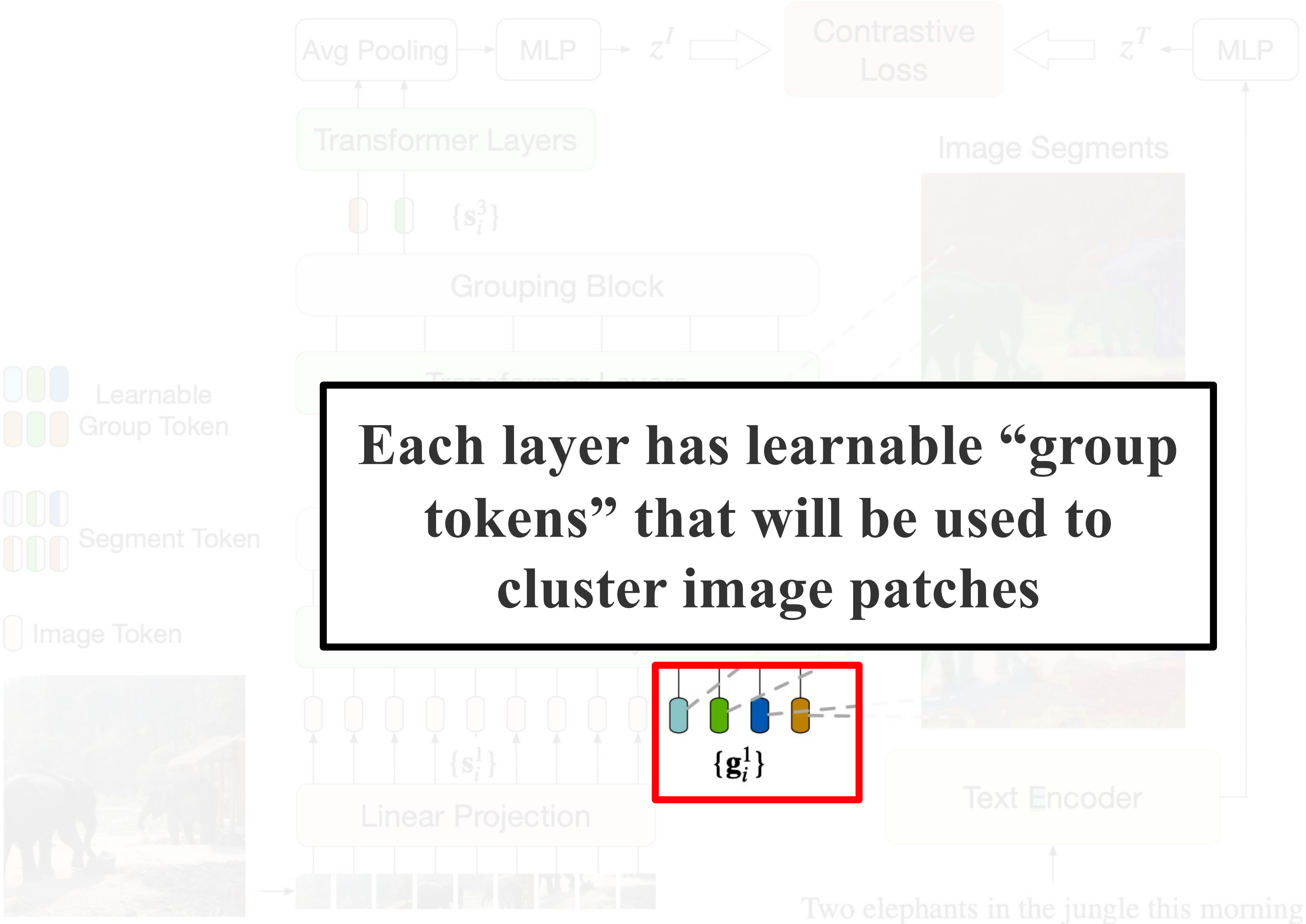
Open-Set Models

Tokenize image into patches and embed



GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models

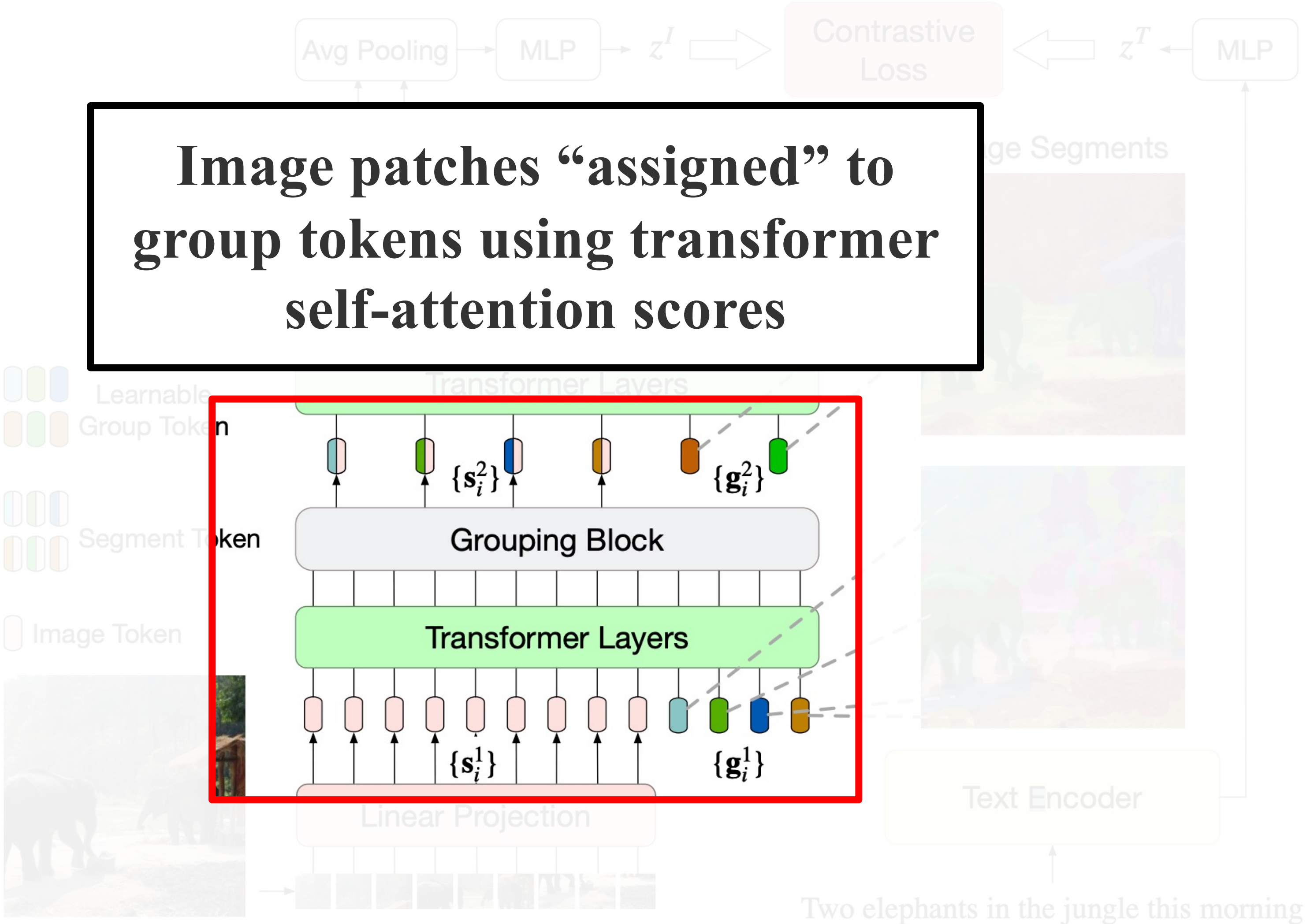


Each layer has learnable "group tokens" that will be used to cluster image patches

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

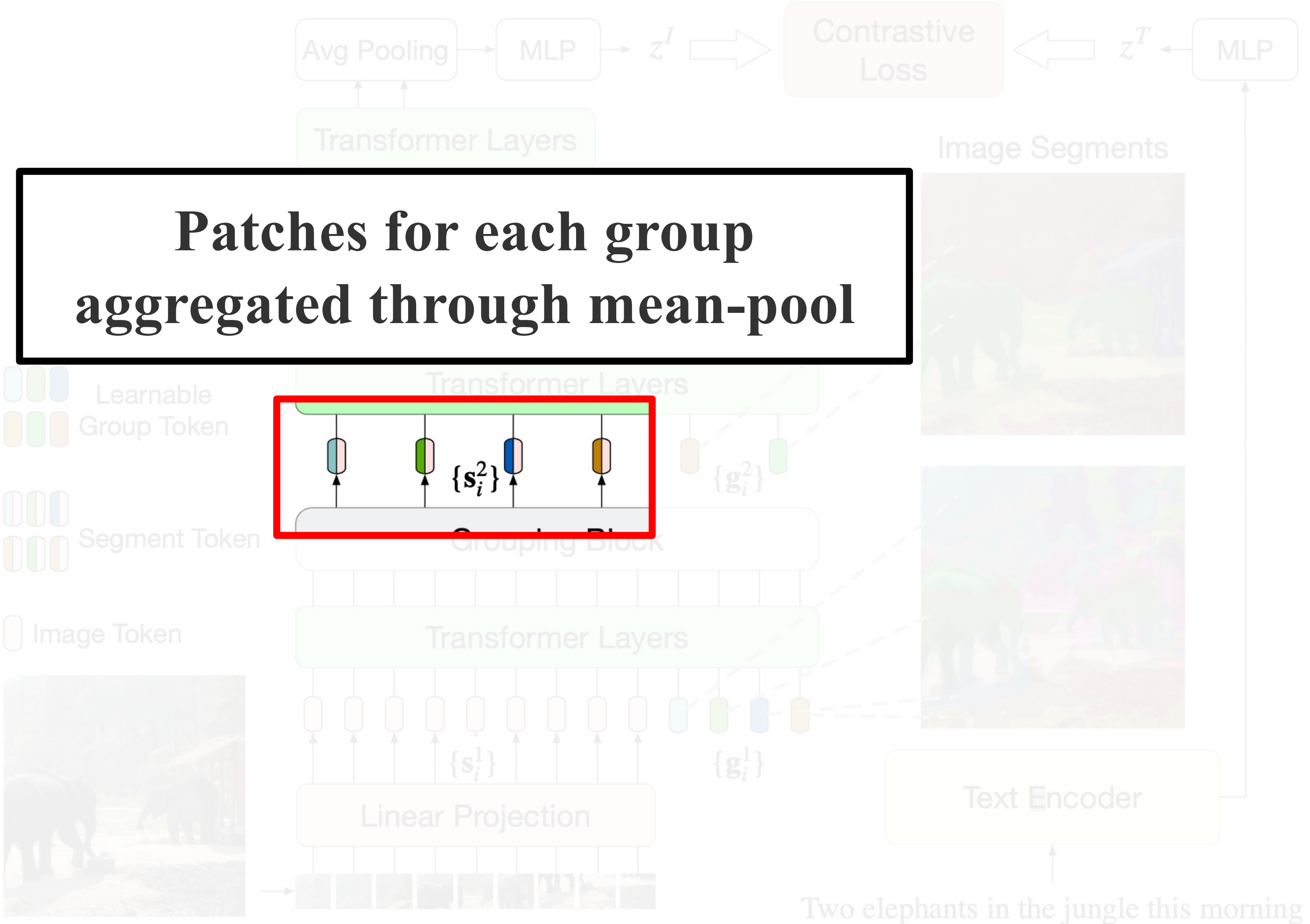
Open-Set Models

Image patches "assigned" to group tokens using transformer self-attention scores



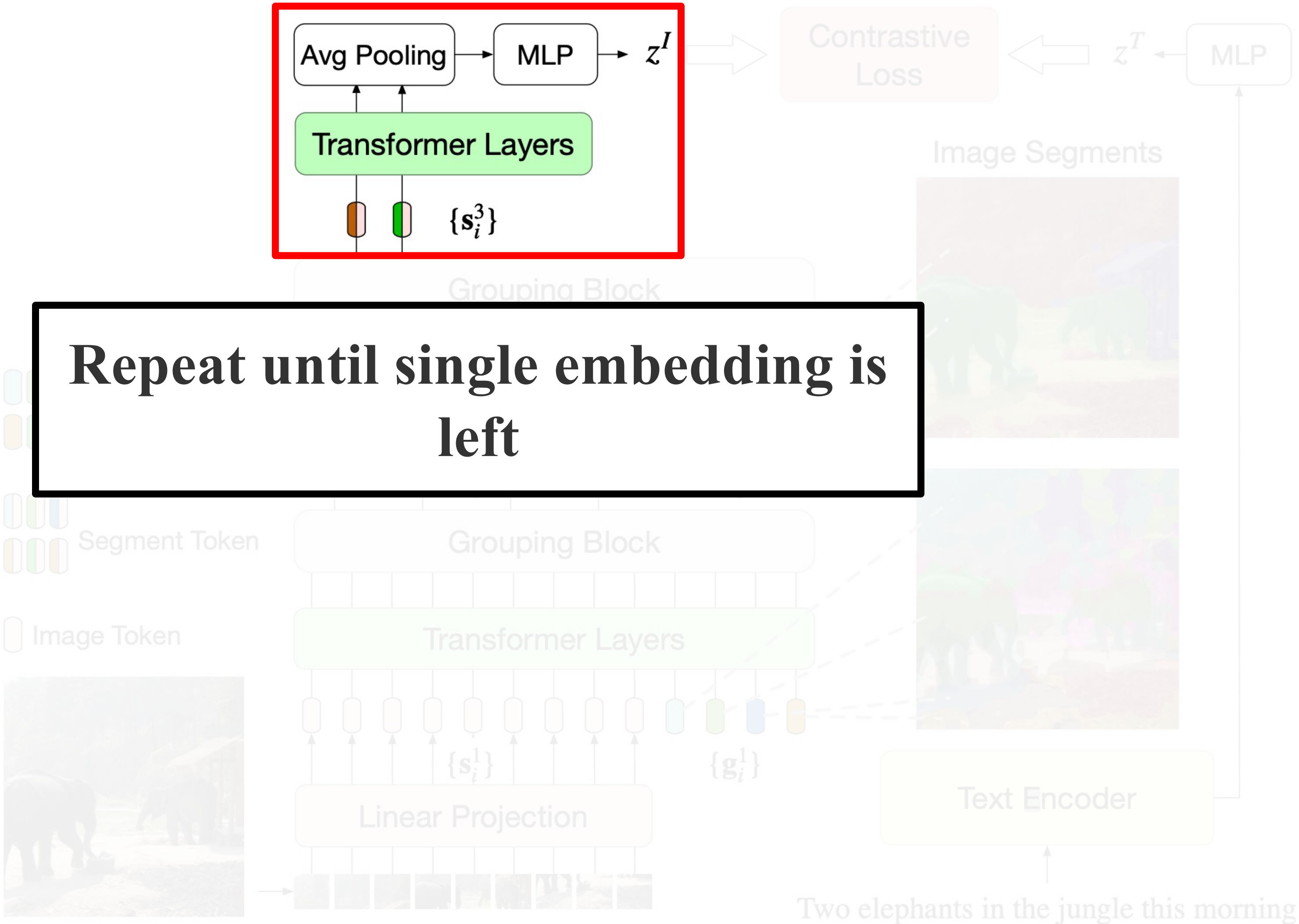
GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models



GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

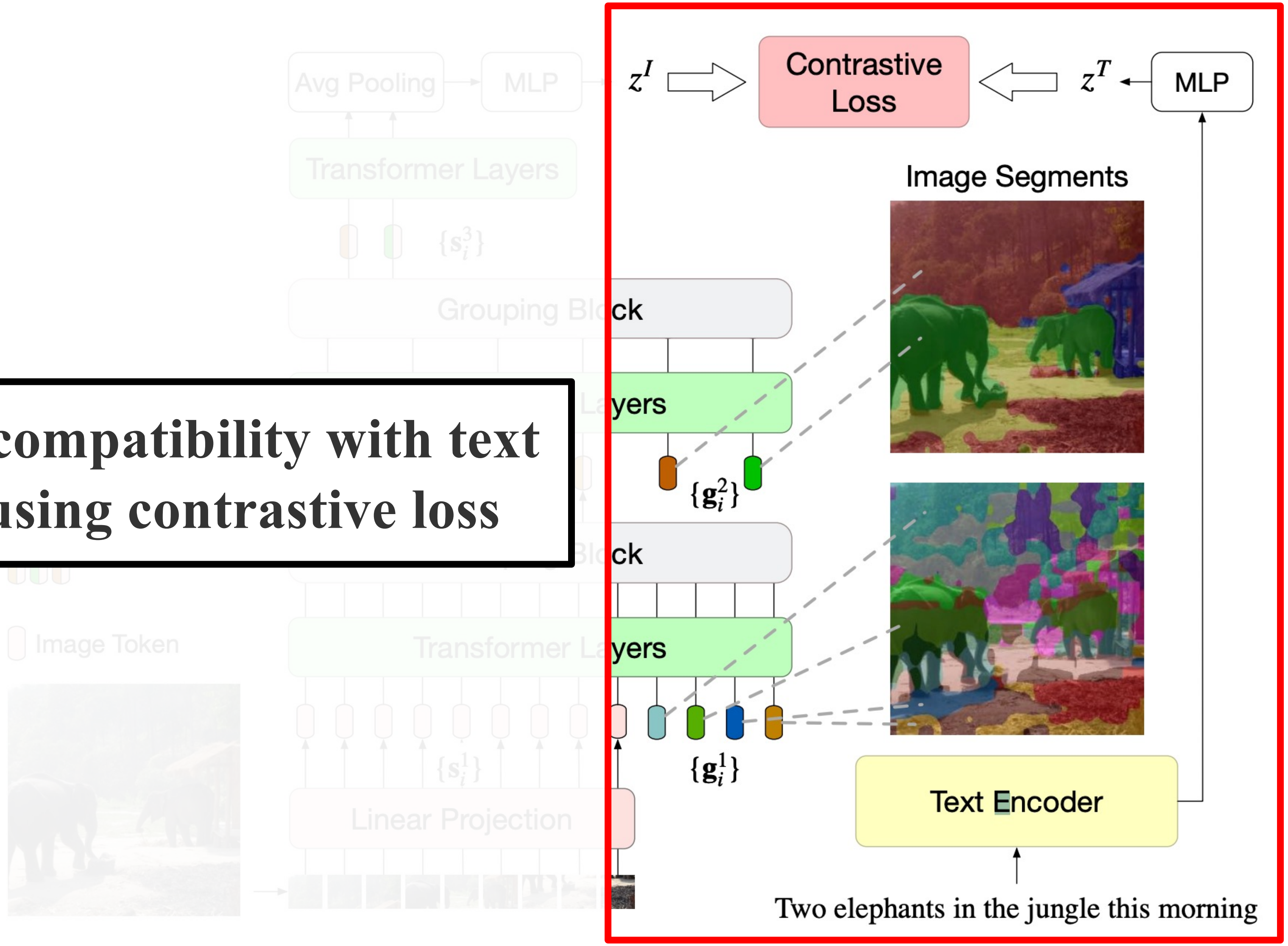
Open-Set Models



Repeat until single embedding is left

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models



Optimize compatibility with text caption using contrastive loss

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models

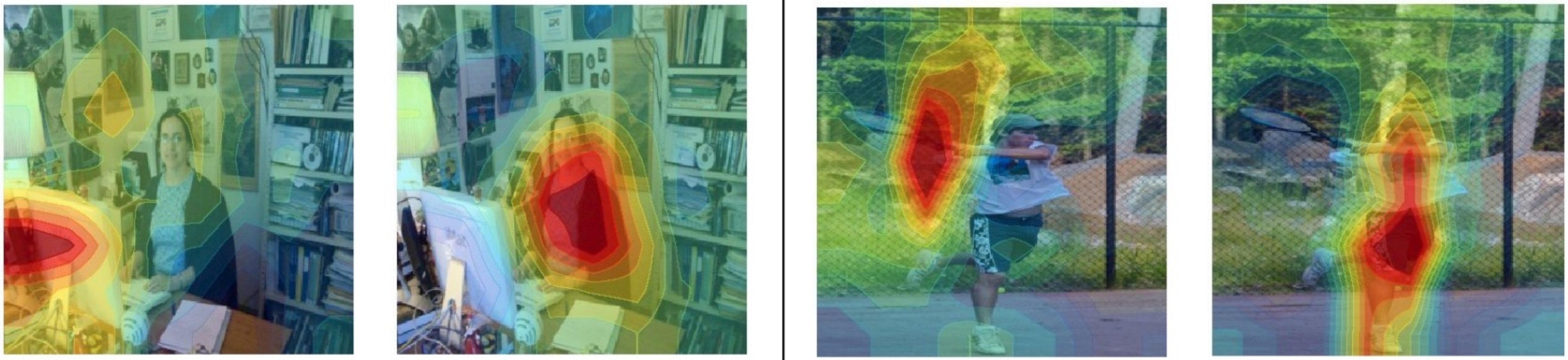
GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Open-Set Models

GroupViT: Semantic Segmentation Emerges from Text Supervision: Xu et al. 2022

Bias in Vision and Language Models

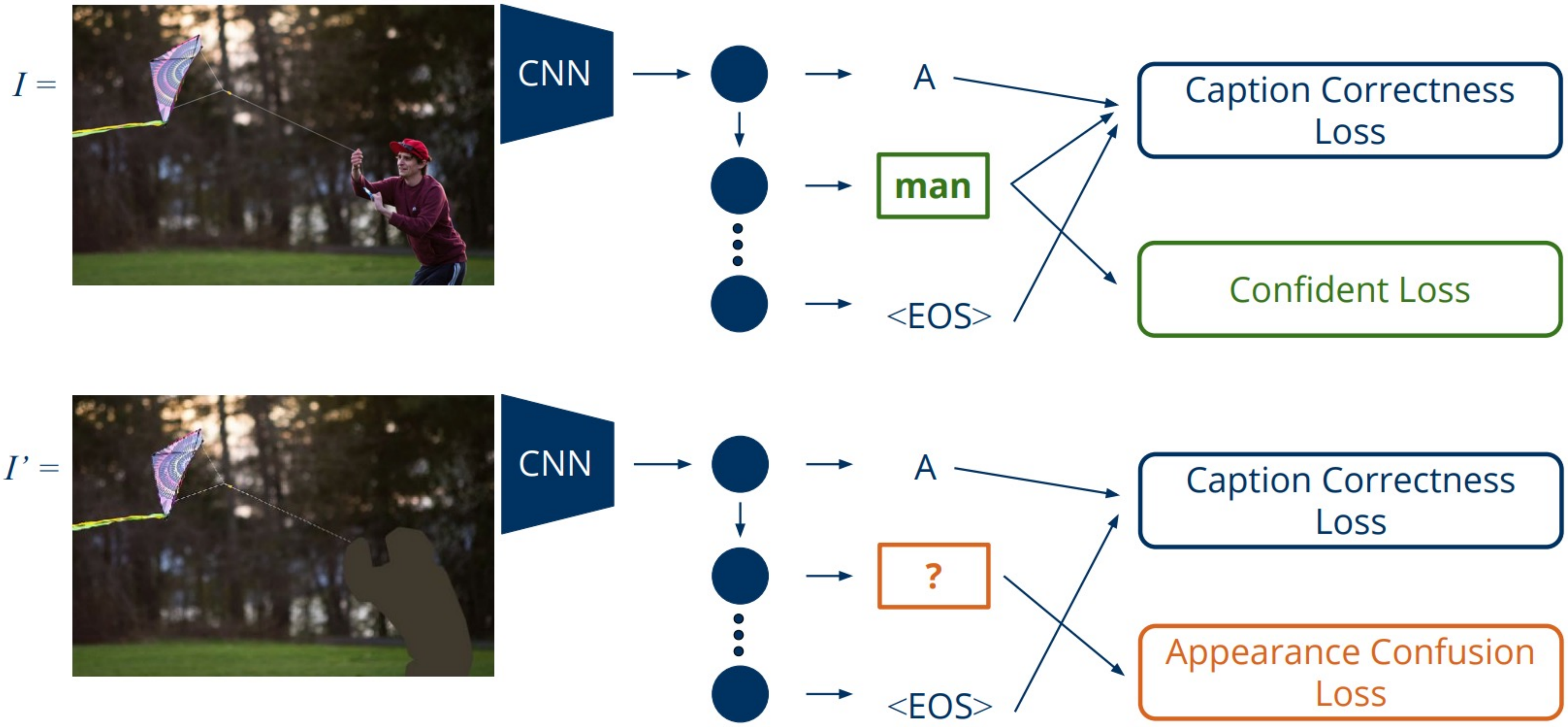
Wrong	Right for the Right Reasons	Right for the Wrong Reasons	Right for the Right Reasons



Baseline: *A **man** sitting at a desk with a laptop computer.*	Our Model: *A **woman** sitting in front of a laptop computer.*	Baseline: *A **man** holding a tennis racquet on a tennis court.*	Our Model: *A **man** holding a tennis racquet on a tennis court.*

Women also Snowboard: Overcoming Bias in Caption Models: Burns et al. 2019

Bias in Vision and Language Models



Women also Snowboard: Overcoming Bias in Caption Models: Burns et al. 2019

Bias in Vision and Language Models

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications: Agarwal et al. 2021

Bias in Vision and Language Models

Neurons work

Multimiodal Neurons in Artificial Neural Networks: Goh et al. 2021

Bias in Vision and Language Models

Prompt: a photo of a personal assistant;
Date: April 1, 2022

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

Bias in Vision and Language Models

Prompt: lawyer;
Date: April 6, 2022

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

Bias in Vision and Language Models

Prompt: nurse;
Date: April 6, 2022

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022

Bias in Vision and Language Models

DALL-E 2 Preview – Risk and Limitations: Mishkin et al. 2022