Efficient and Less-biased Visual Learning

Leonid Sigal

Professor of Computer Science

NSERC Canada Research Chair (CRC) in Computer Vision and ML
CIFAR Artificial Intelligence (AI) Chair, Vector Institute

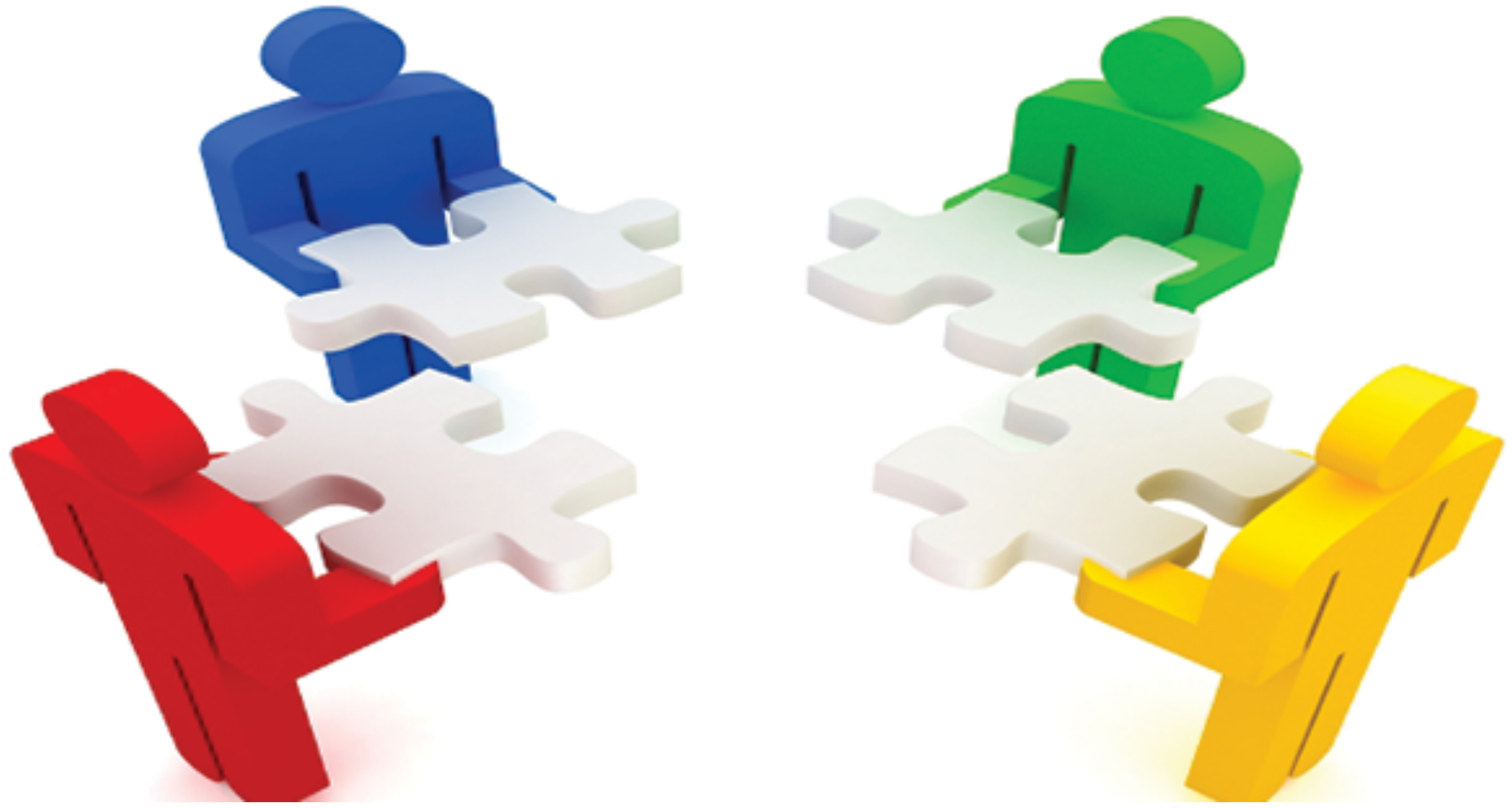
THE UNIVERSITY OF BRITISH COLUMBIA

VECTOR INSTITUTE

1st workshop on Vision-based InduStrial InspectiON

1st workshop on Vision-based InduStrial InspectiON

1st workshop on Vision-based InduStrial InspectiON

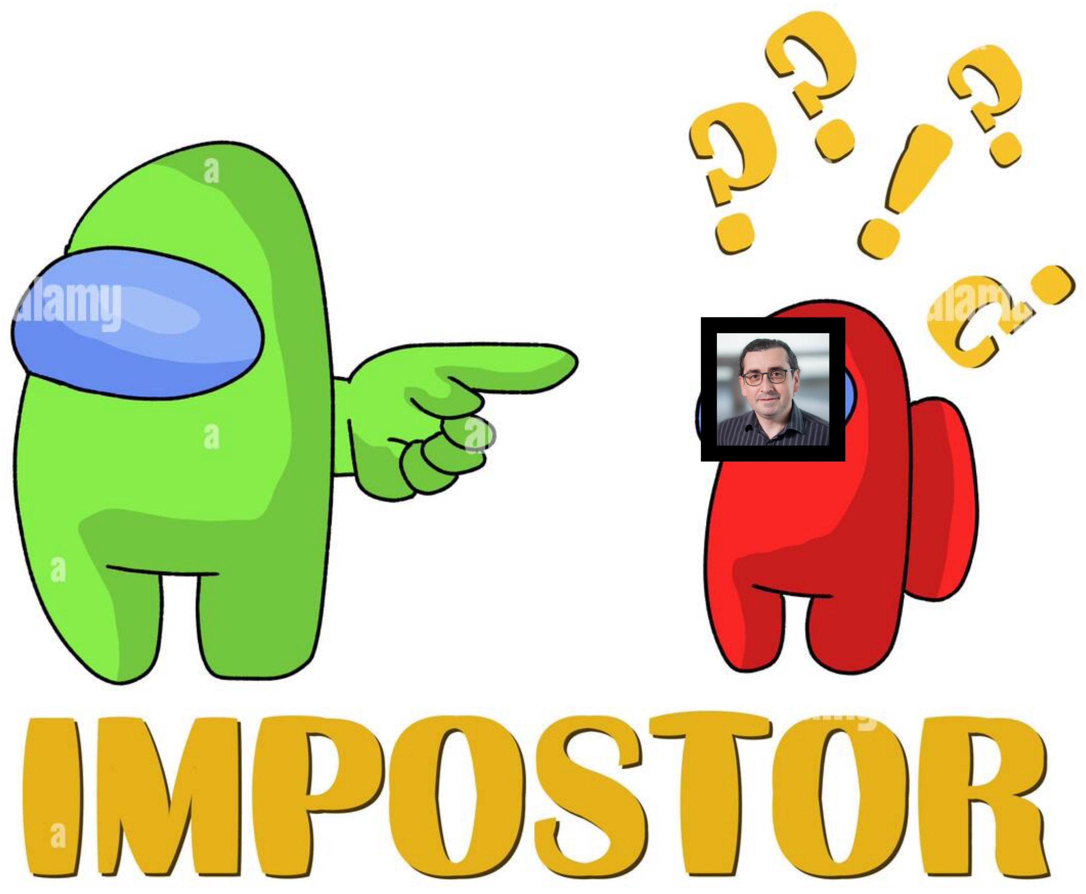


Huge potential for real-world impact! … by bringing together vision **researchers** and **industrial practitioners**

Track	Description	Make a Challenge Submission
Challenge 1	Data-efficient Defect Detection	
Challenge 2	Data-generation for Defect Detection	

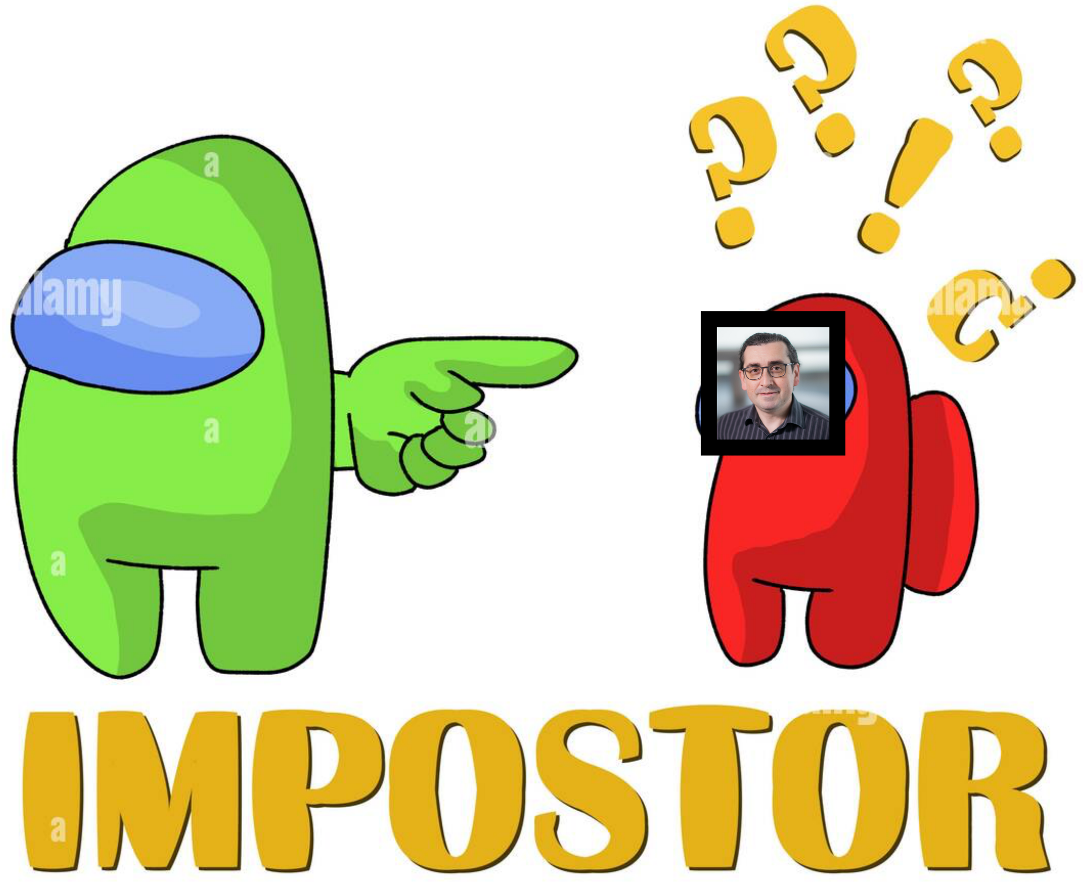
… and formalizing challenges with supporting **data**

1st workshop on Vision-based InduStrial InspectiON



1st workshop on Vision-based InduStrial InspectiON

Self Disclosure: I do not work on InduStrial InspectiON applications, but the high-level ideas and methods we are developing in other vision domains may be useful in these applications

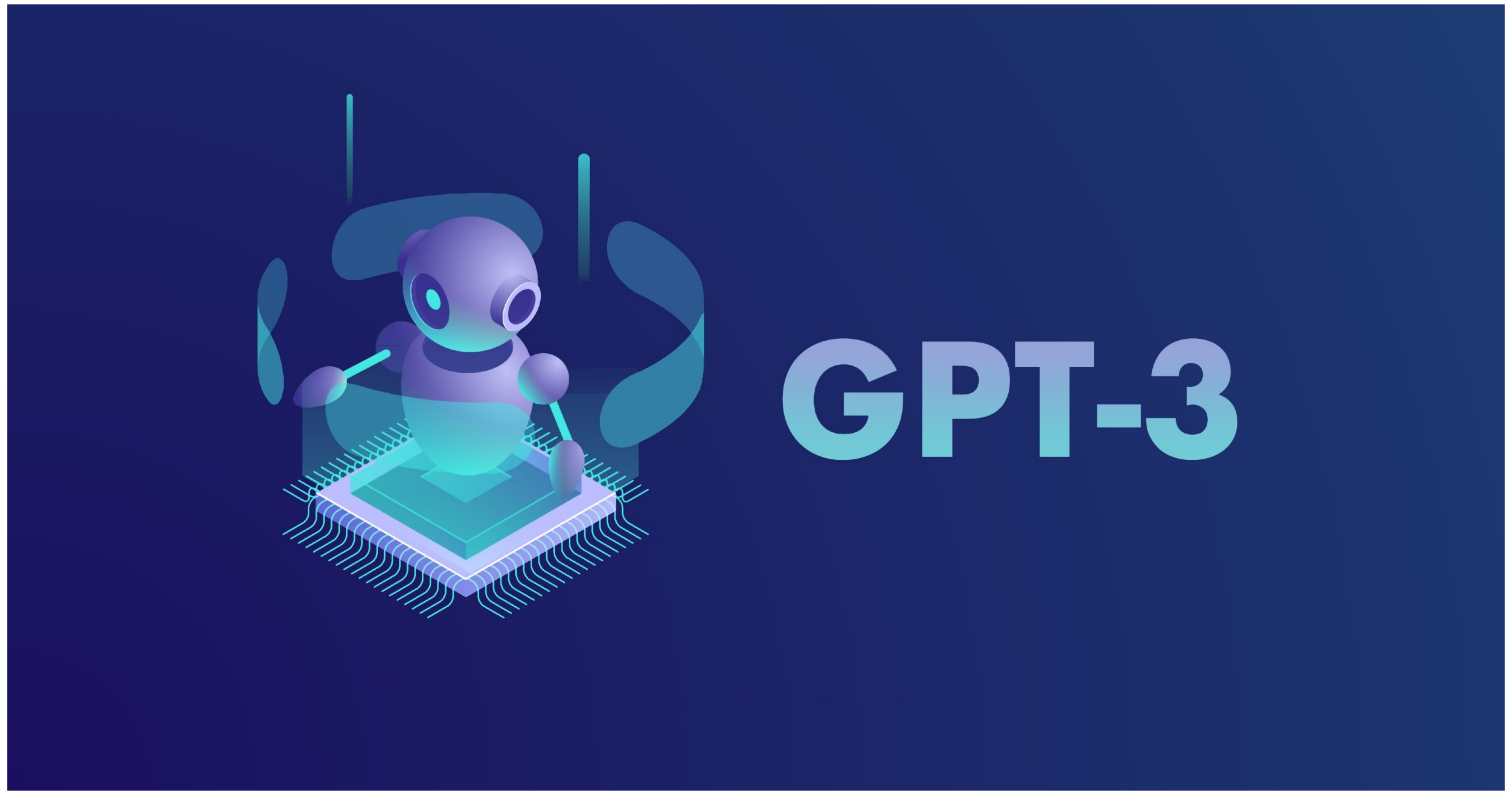


Efficient and Less-biased Visual Learning

Efficient and Less-biased Visual Learning

Why **Data-efficient** Learning?

- **Scientific curiosity**
  Most current neural network architectures are not nearly as efficient as human learners (e.g., GPT-3 is trained on 400 billion words, which would take a human 400 years of continuous reading [1])



https://decemberlabs.com/blog/openai-gpt3-the-new-ai-that-will-blow-your-mind-might-also-be-a-little-overrated/

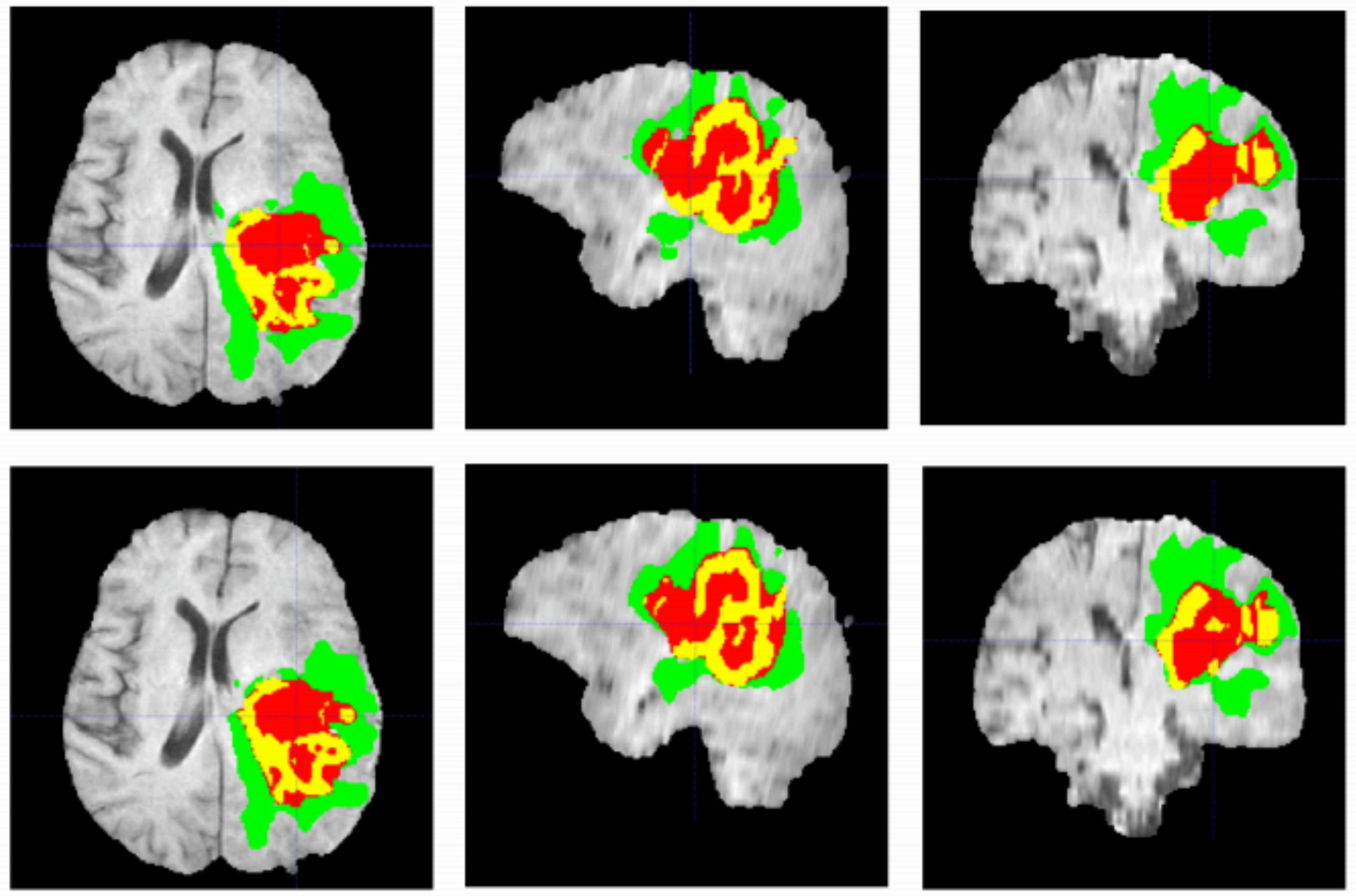
https://www.scientificamerican.com/article/are-there-too-many-neuroscientists/

[1] https://theconversation.com/were-told-ai-neural-networks-learn-the-way-humans-do-a-neuroscientist-explains-why-thats-not-the-case-183993

Why **Data-efficient** Learning?

- **Scientific curiosity**
  Most current neural network architectures are not nearly as efficient as human learners (e.g., GPT-3 is trained on 400 billion words, which would take a human 400 years of continuous reading [1])

- **Inherent inability to large-scale label data**
  For some domains / problems there may not be enough data to label (e.g., Adamantinoma — a rare bone cancer — may have as few as 300 reported cases)



https://developer.nvidia.com/blog/automatically-segmenting-brain-tumors-with-ai/

Why **Data-efficient** Learning?

- **Scientific curiosity**
  Most current neural network architectures are not nearly as efficient as human learners (e.g., GPT-3 is trained on 400 billion words, which would take a human 400 years of continuous reading[1])

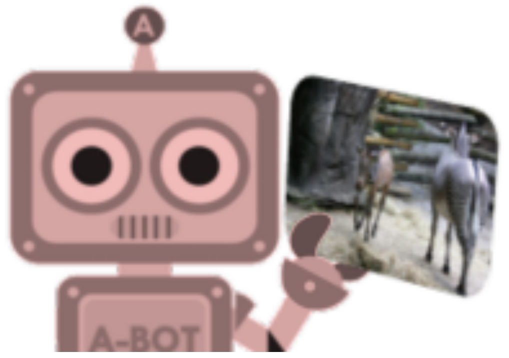
- **Inherent inability to large-scale label data**
  For some domains / problems there may not be enough data to label (e.g., Adamantinoma — a rare bone cancer — may have as few as 300 reported cases)

- **Scaling and granularity of vision tasks**
  As we attempt to scale vision systems to address more challenging inference tasks, we will not be able to get away with exhaustive data labeling

Granularity of the task vs. annotation **cost** …

Image-level
Classification

Man, Woman, Horse



Granularity of the task vs. annotation **cost** …

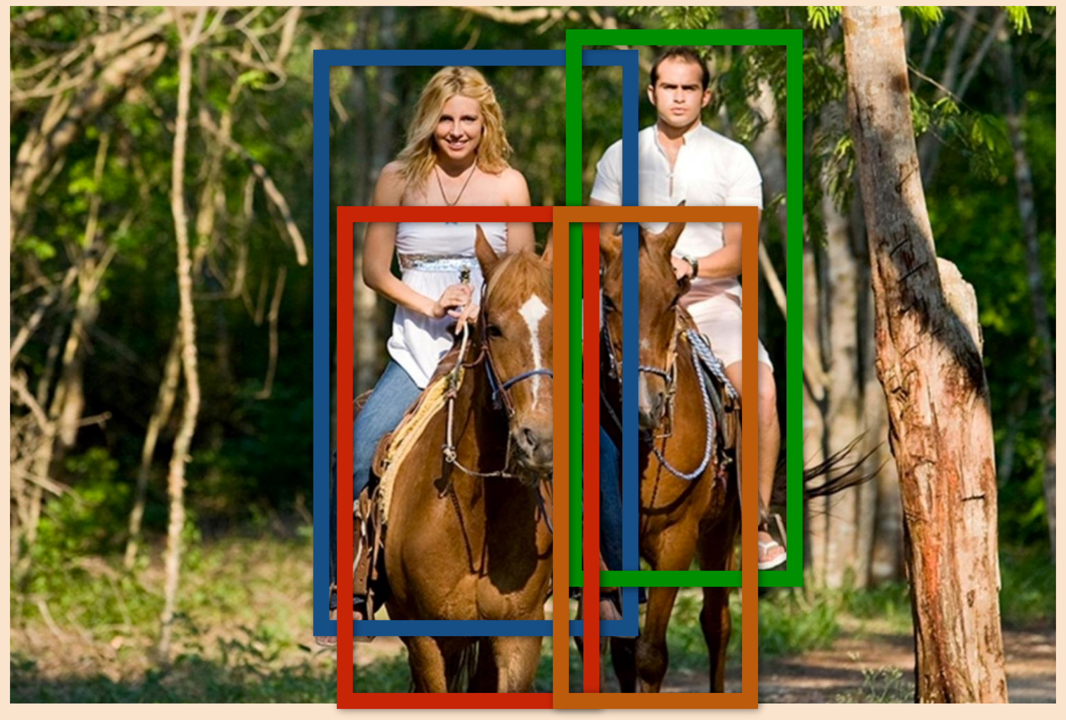
Image-level
Classification

Man, Woman, Horse

Instance-level
Detection

Instance-level
Segmentation

Man, Woman, Horse, Horse



Granularity of the task vs. annotation **cost** …

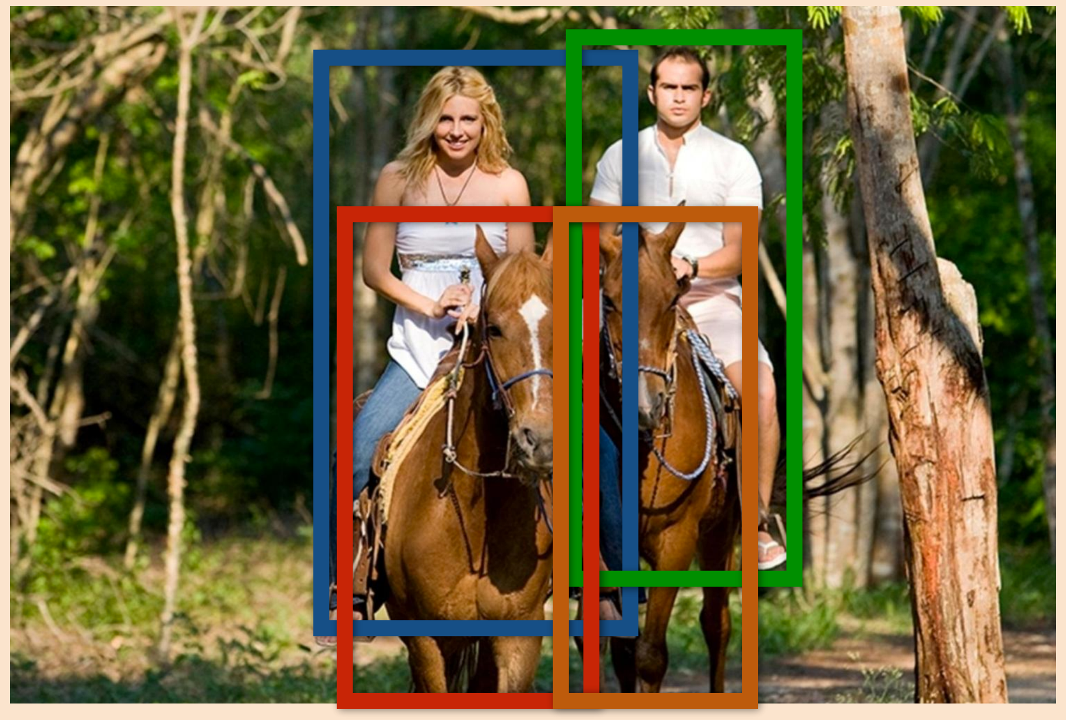
Image-level
Classification

Man, Woman, Horse

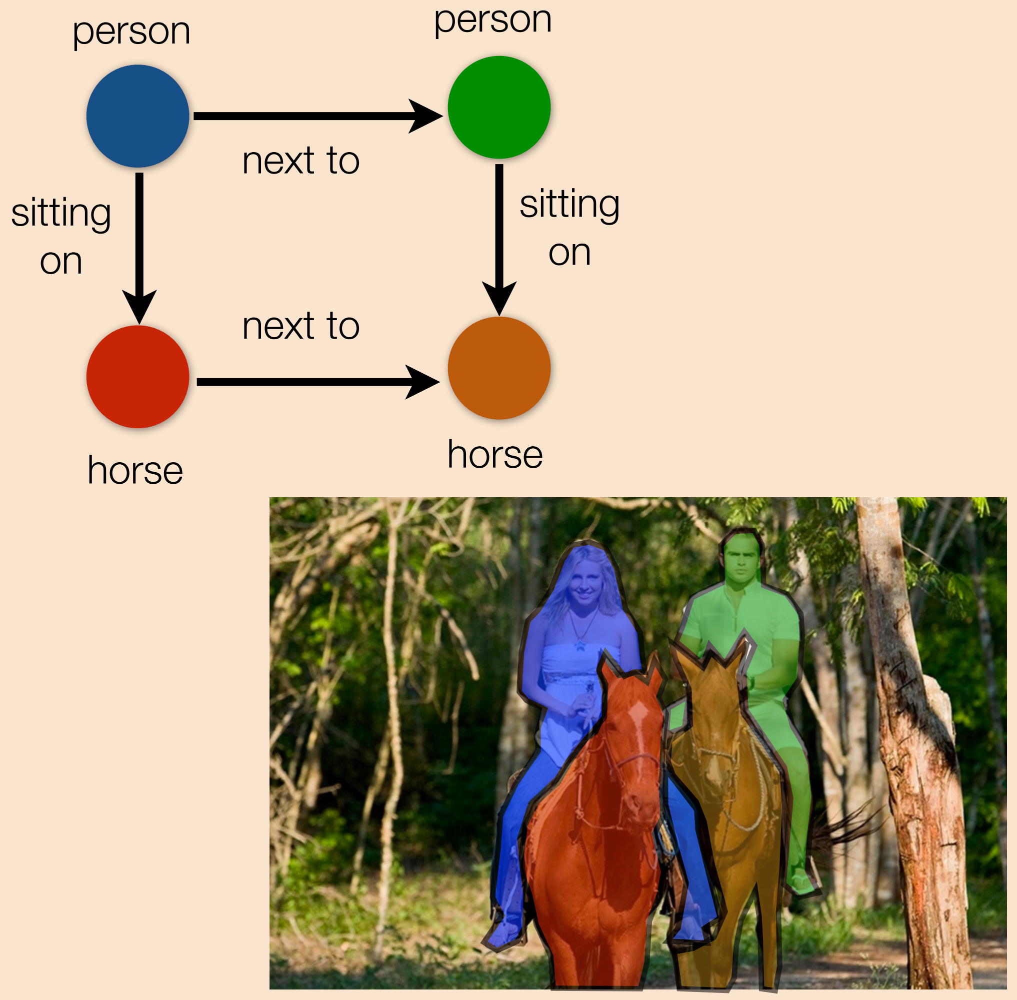
Instance-level
Detection

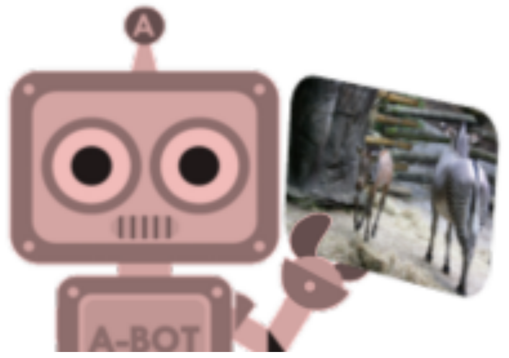
Instance-level
Segmentation

Man, **Woman**, **Horse**, **Horse**



Scene-graph
Generation



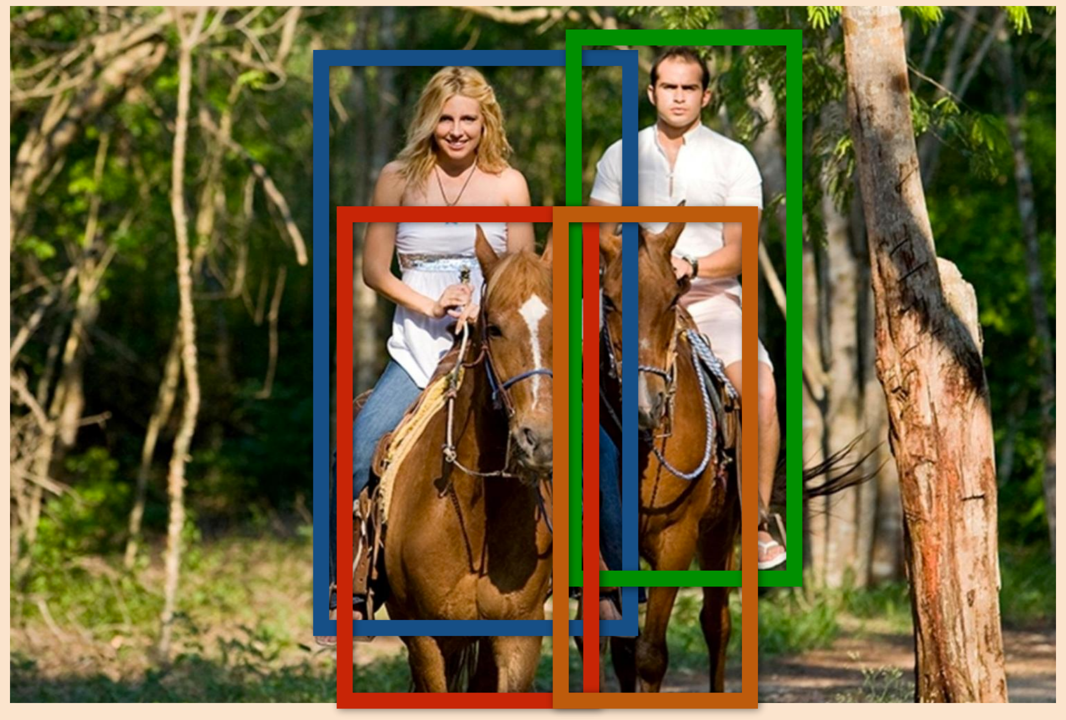


Granularity of the task vs. annotation **cost** …

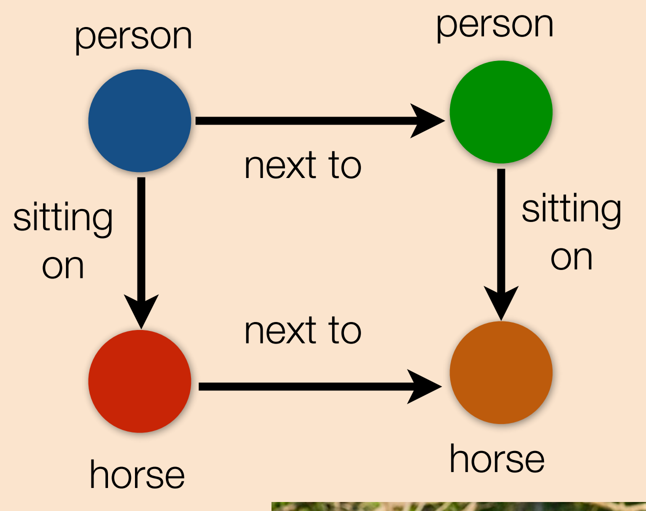
Instance-level
Detection

Instance-level
Segmentation

Man, **Woman**, **Horse**, **Horse**

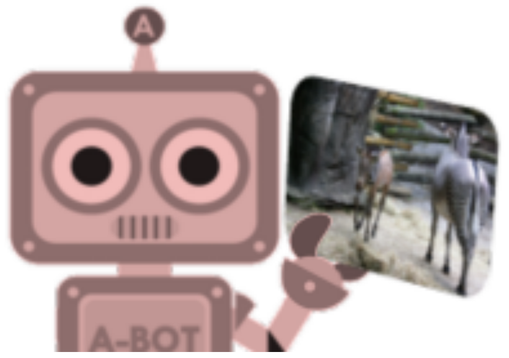


Scene-graph
Generation



Question
Answering

Q: What are people doing?
Q: What time of the year is it?
Q: Are the people married?



Why **Compute-efficient** Learning?

- **Ability to run on low-compute devices**
  Most current neural network architectures are not able to run on mobile or embedded devices

- **Low-latency inference**
  Ability to run with low-latency, means high throughput for the system

- **High adaptability of the model**
  If both learning and inference are compute-efficient, we can potentially adopt models mode easily with incoming data

Why **Compute-efficient** Learning?

- **Low-latency inference**
  Ability to run with low-latency, means high throughput for the system
- **High adaptability of the model**
  If both learning and inference are compute-efficient, we can potentially adopt models mode easily with incoming data

Inspection Applications:

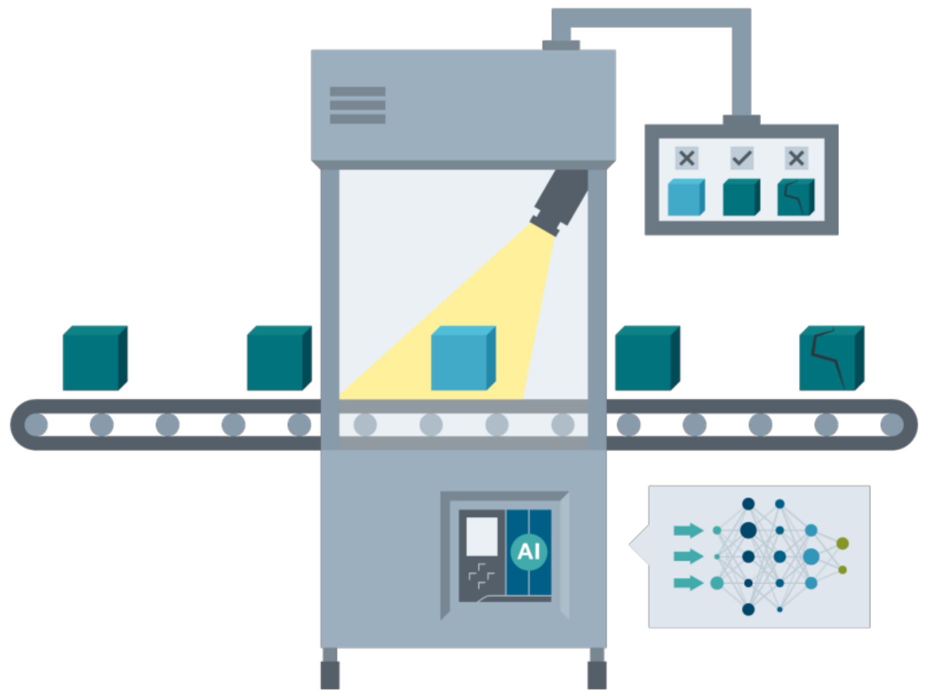


Image from CameraLyze

Efficient and Less-biased Visual Learning

Why **Less-biased** Learning?

- **Biases in ML models have been shown and are concerning**
  Existing models are excellent in picking up, modeling (and in some cases) even amplifying (human) biases available in the data

Why **Less-biased** Learning?

- **Biases in ML models have been shown and are concerning**
  Existing models are excellent in picking up, modeling (and in some cases) even amplifying (human) biases available in the data

Language Model (trained to complete analogies)

Testing:

Input: Man to computer programmer as woman to ???
Output: Homemaker

Input: Man to doctor as woman to ???
Output: Nurse

["Man is to Computer Programmer as Woman is to Homemaker? Debasing Word Embeddings", Bolukbasi, Chang, Zou, Saligrama, KalaiNeurIPS, 2016]

Why **Less-biased** Learning?

- **Biases in ML models have been shown and are concerning**
  Existing models are excellent in picking up, modeling (and in some cases) even amplifying (human) biases available in the data

Language Model (trained to complete analogies)

Testing:

Input: Man to computer programmer as woman to ???
Output: Homemaker

Input: Man to doctor as woman to ???
Output: Nurse

["Man is to Computer Programmer as Woman is to Homemaker? Debasing Word Embeddings", Bolukbasi, Chang, Zou, Saligrama, KalaiNeurIPS, 2016]

Prompt: "A photo of a **doctor**"

Prompt: "A photo of a **nurse**"

DALL-E Generated: 2.35 male doctors for every 1 female

US Empirical Statistics: 1.78 male doctors for every 1 female

https://cornell-data.medium.com/how-biased-are-text-to-image-models-99e8fdb8c5ab

Why **Less-biased** Learning?

- **Biases in ML models have been shown and are concerning**
  Existing models are excellent in picking up, modeling (and in some cases) even amplifying (human) biases available in the data

Language Model (trained to complete analogies)

Testing:

Input: Man to computer programmer as woman to ???
Output: Homemaker

Input: Man to doctor as woman to ???
Output: Nurse

["Man is to Computer Programmer as Woman is to Homemaker? Debasing Word Embeddings", Bolukbasi, Chang, Zou, Saligrama, KalaiNeurIPS, 2016]

DALL-E 2 is less biased

Prompt: "A photo of a **doctor**"

Prompt: "A photo of a **nurse**"

DALL-E Generated: 2.35 male doctors for every 1 female

US Empirical Statistics: 1.78 male doctors for every 1 female

https://cornell-data.medium.com/how-biased-are-text-to-image-models-99e8fdb8c5ab

Inspection, **Defect** and **Anomaly** Detection

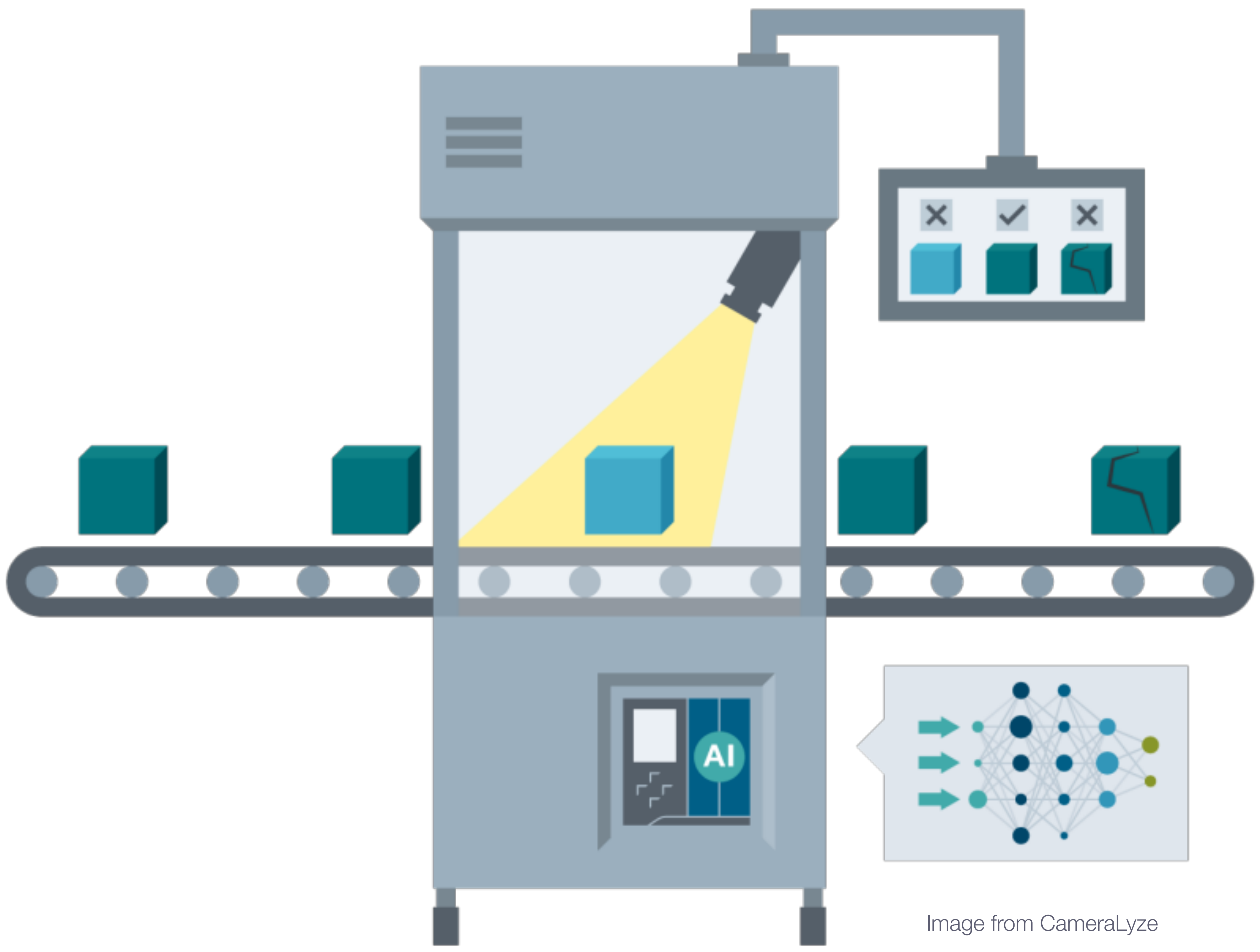


Image from CameraLyze

Inspection, **Defect** and **Anomaly** Detection

- **Large-number of defect-free images may not be available**
  (e.g., new product lines starting to be manufactured)

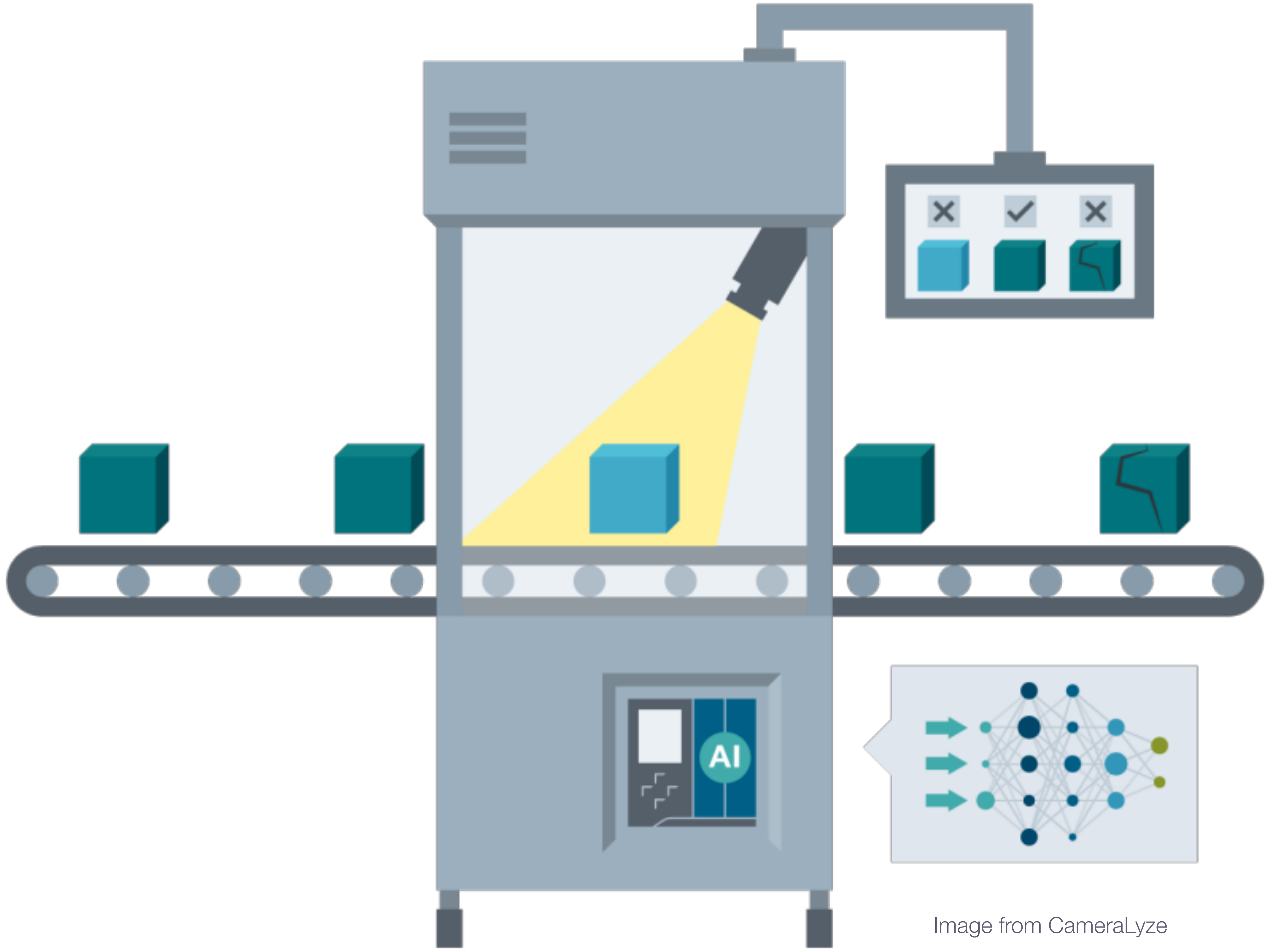


Image from CameraLyze

[“A hierarchical transformation-discriminating generative model for few shot anomaly detection”, Sheynin, Benaim, Wolf, ICCV, 2021.]
[“Registration based few-shot anomaly detection”, Huang, Guan, Jiang, Zhang, Spratling, Wang. ECCV, 2022.]
[“Anomaly detection via few-shot learning on normality”, Ando,Yamamoto. ECML PKDD, 2022.]
[“Same same but differnet: Semi-supervised defect detection with normalizing flows”, Rudolph, Wandt, Rosenhahn, WACV, 2021]

Inspection, **Defect** and **Anomaly** Detection

- **Large-number of defect-free images may not be available**
  (e.g., new product lines starting to be manufactured)
- **There will be few defect images if any**
  (e.g., leading to huge class imbalance)

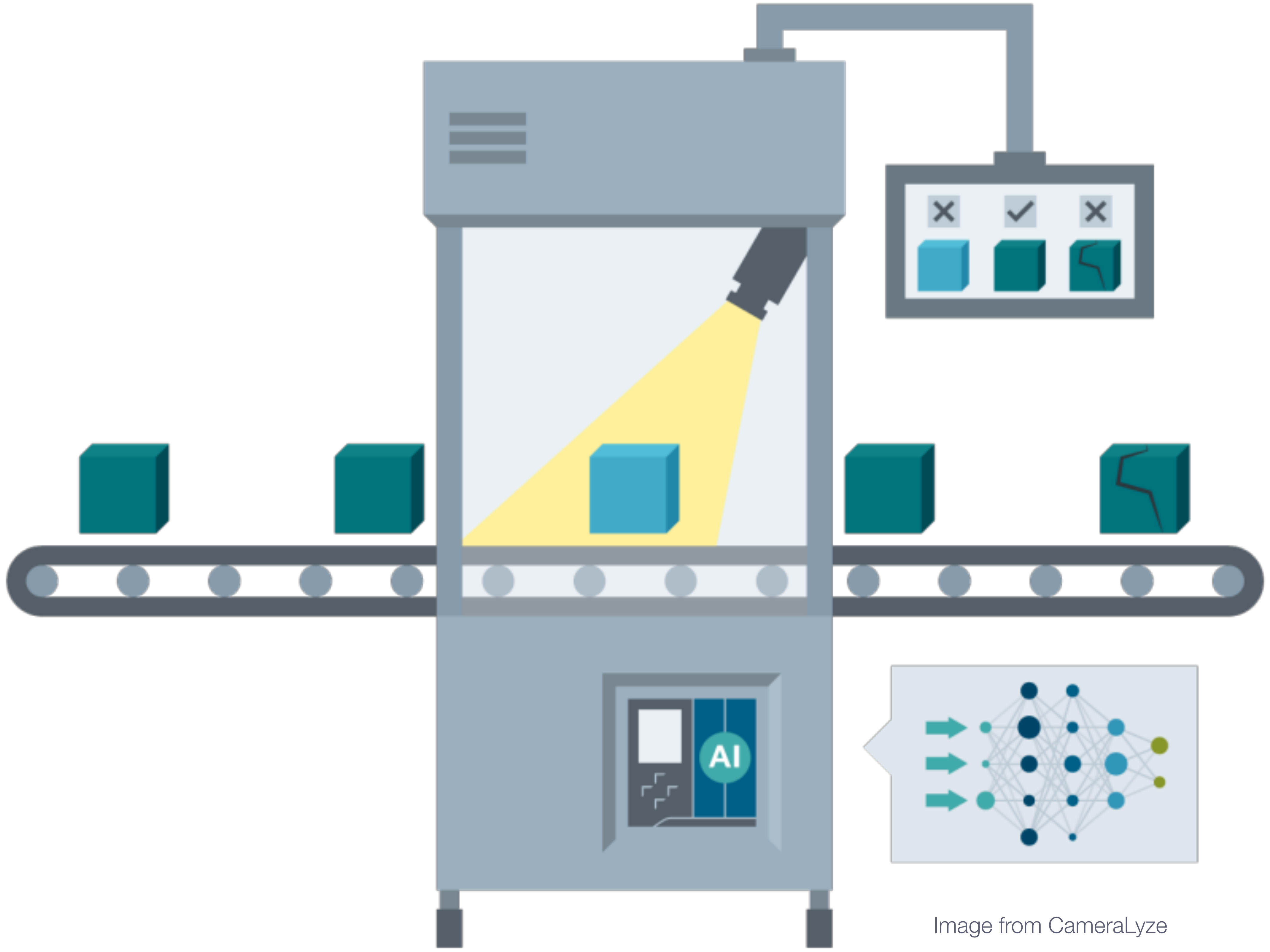
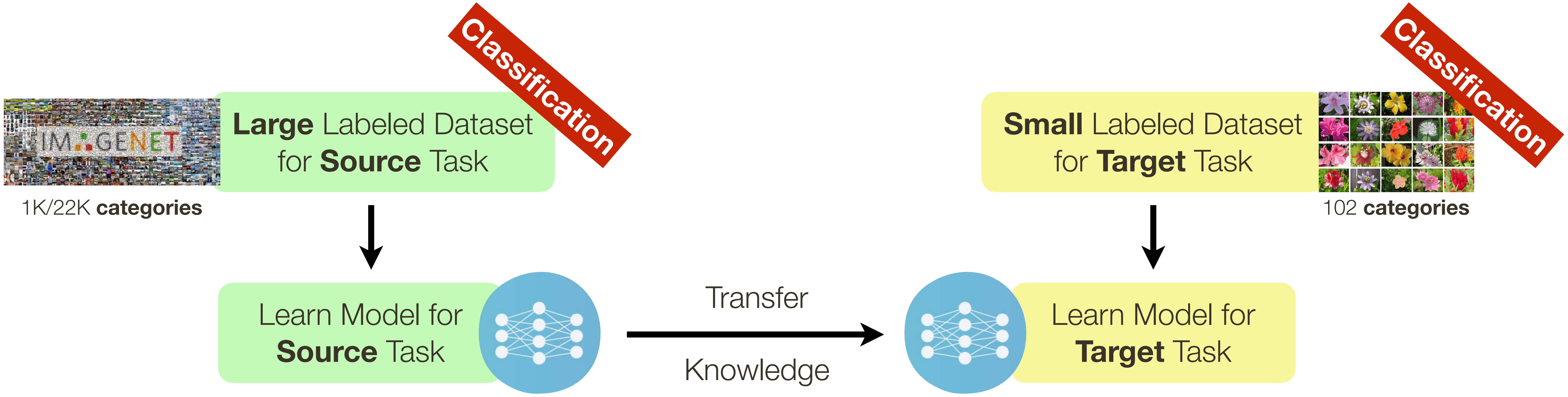


Image from CameraLyze

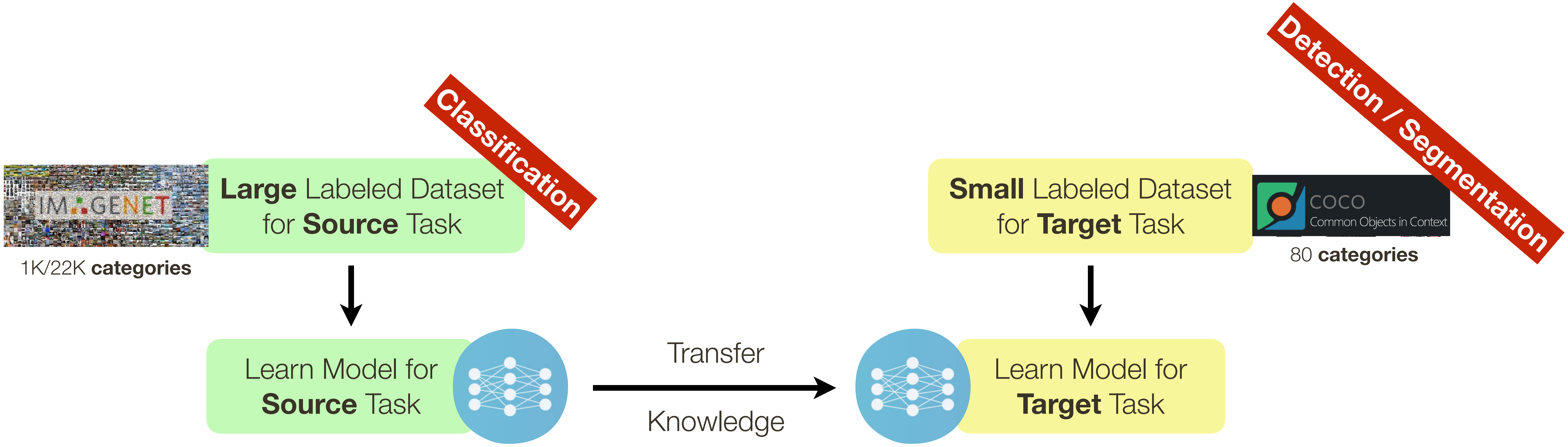
Data Efficiency, Strategy 1:

Large Model + Transfer Learning



Data Efficiency, Strategy 1:

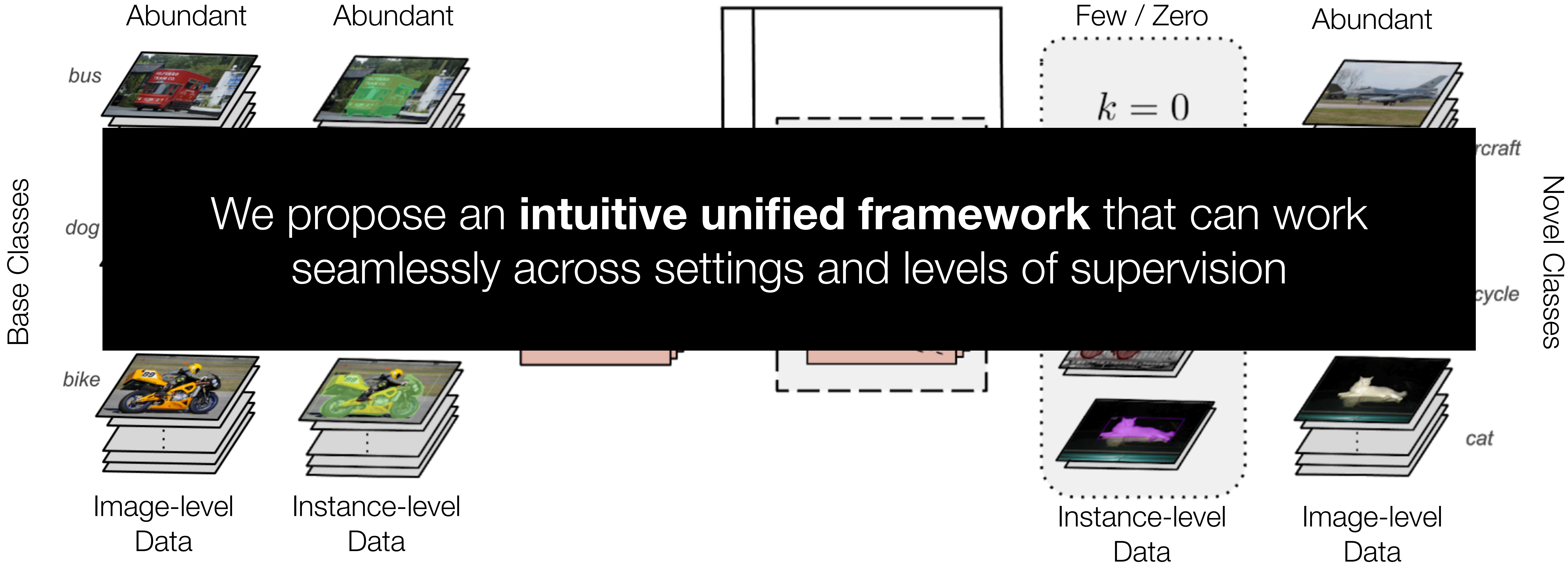
Large Model + Transfer Learning



UniT: Unified Knowledge Transfer for Any-shot Detection

There is no single unified solution that is applicable to a wide range of supervision: from zero to a few instance-level samples for *novel* classes

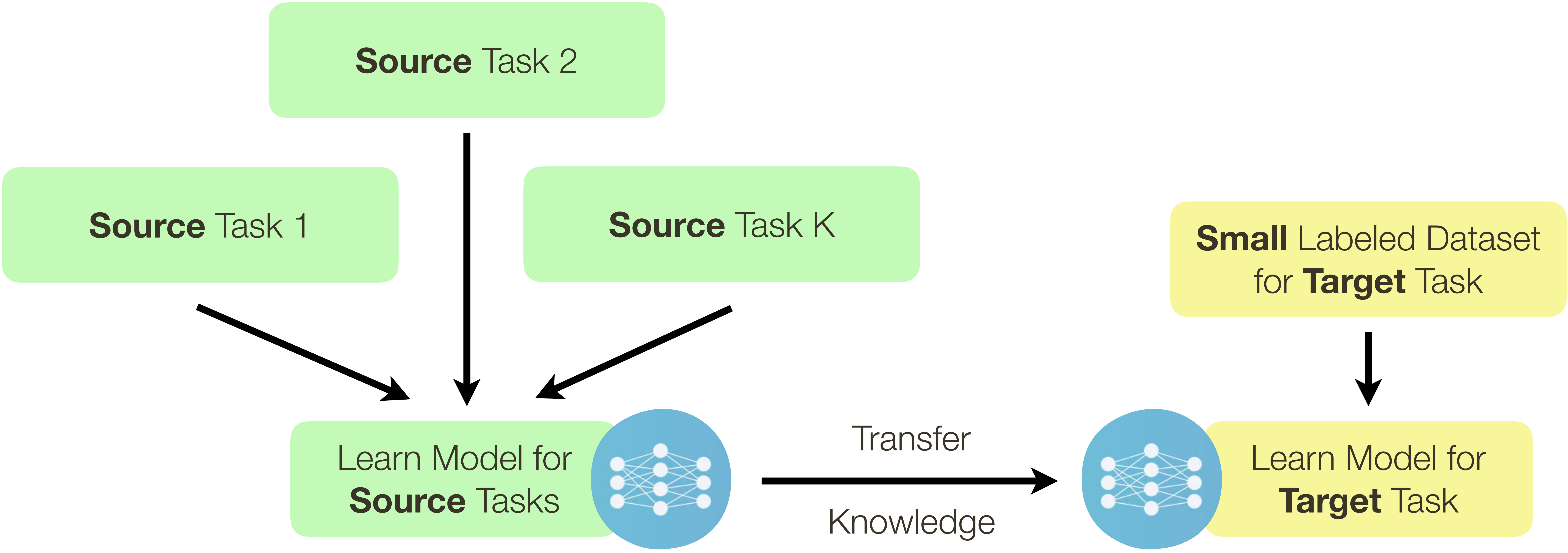
Siddhesh Khandelwal* Raghav Goyal*



We propose an **intuitive unified framework** that can work seamlessly across settings and levels of supervision

Data Efficiency, **Strategy 2**:

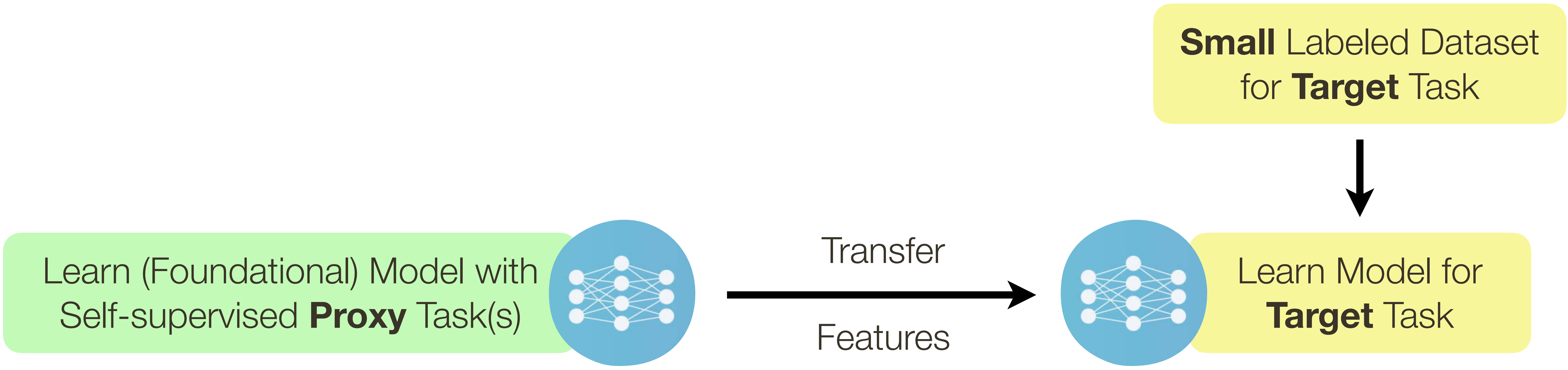
Multi-task + Transfer Learning



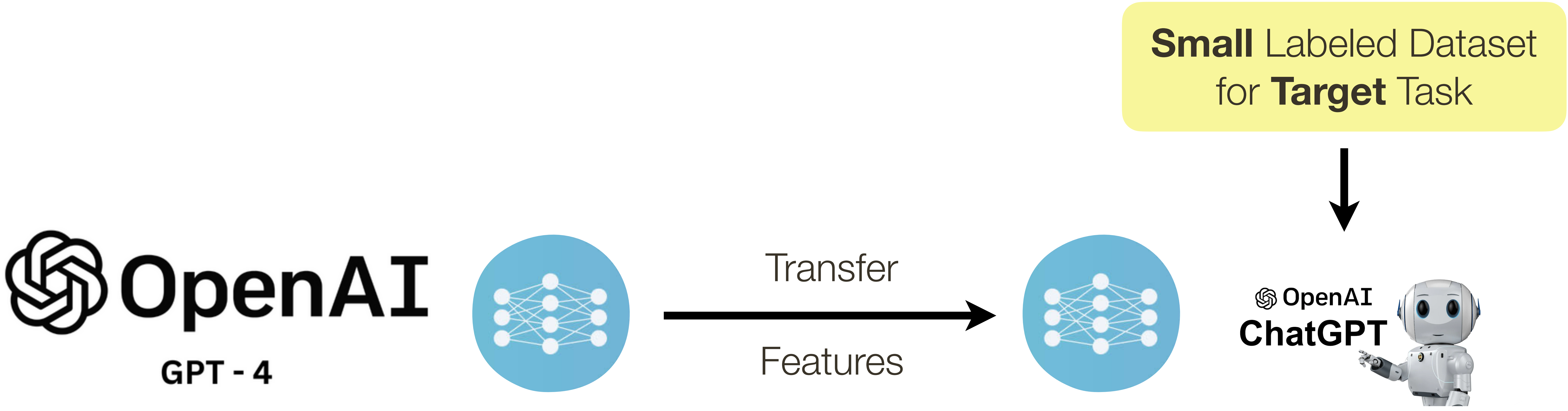
Multi-task **Video Understanding**

Data Efficiency, Strategy 3:

Foundational Model

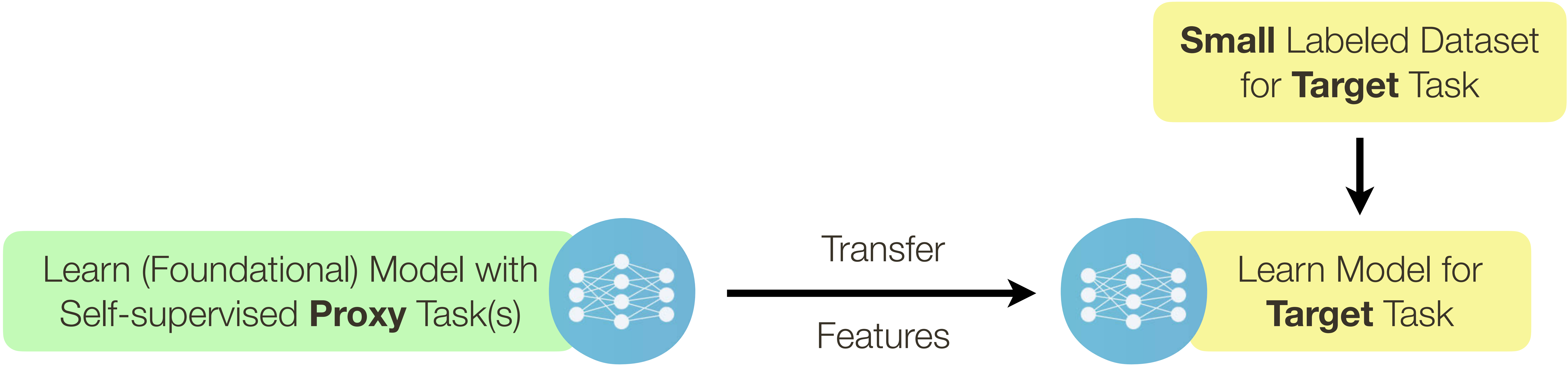


Data Efficiency, Strategy 3: Foundational Model



Data Efficiency, Strategy 3:

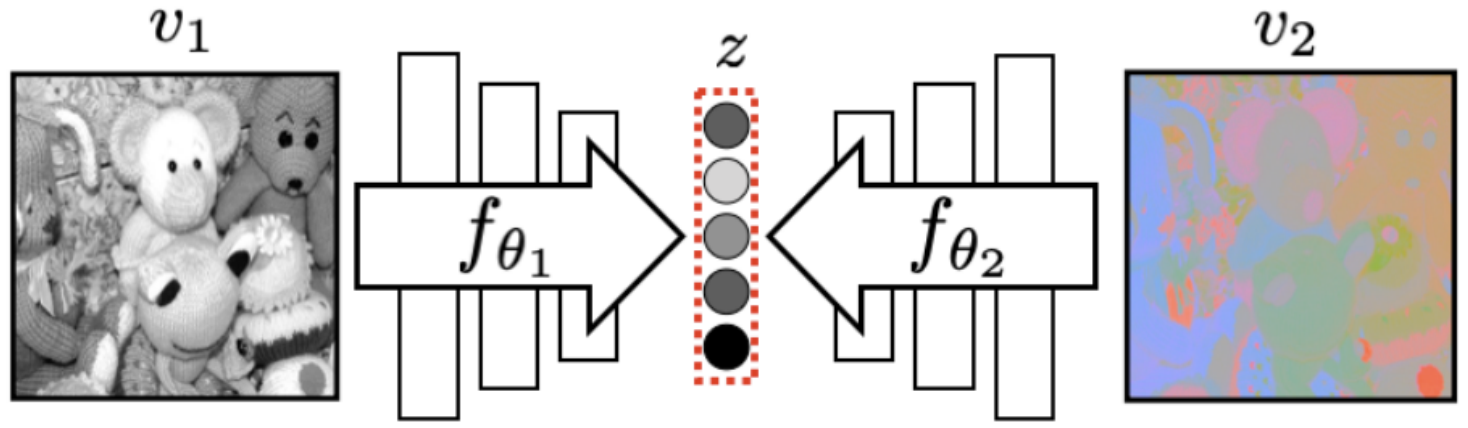
Foundational Model



Self-supervised Learning

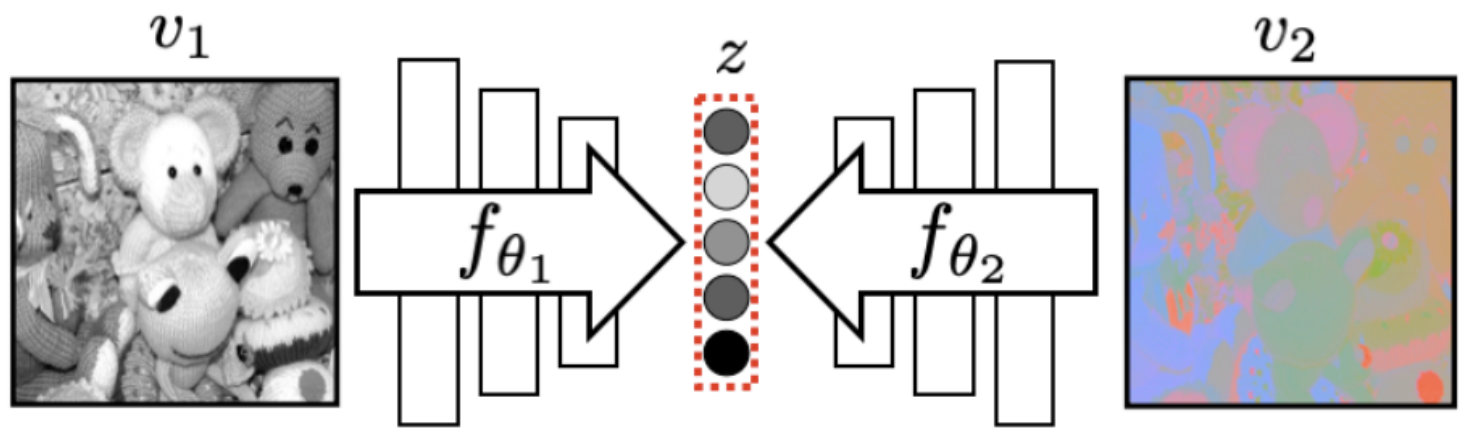
Self-supervised Learning

- **Contrastive / Discriminative** Learning (introduce transformations and learn invariant representation)
  - With negative samples (e.g., SimCLR [Chen *et al.*, ICML'20], MoCo [He *et al.*, CVPR'20])
  - Without negative samples (e.g., BYOL [Grill *et al.*, NeurIPS'20], DINO [Caron *et al.*, ICCV'21])

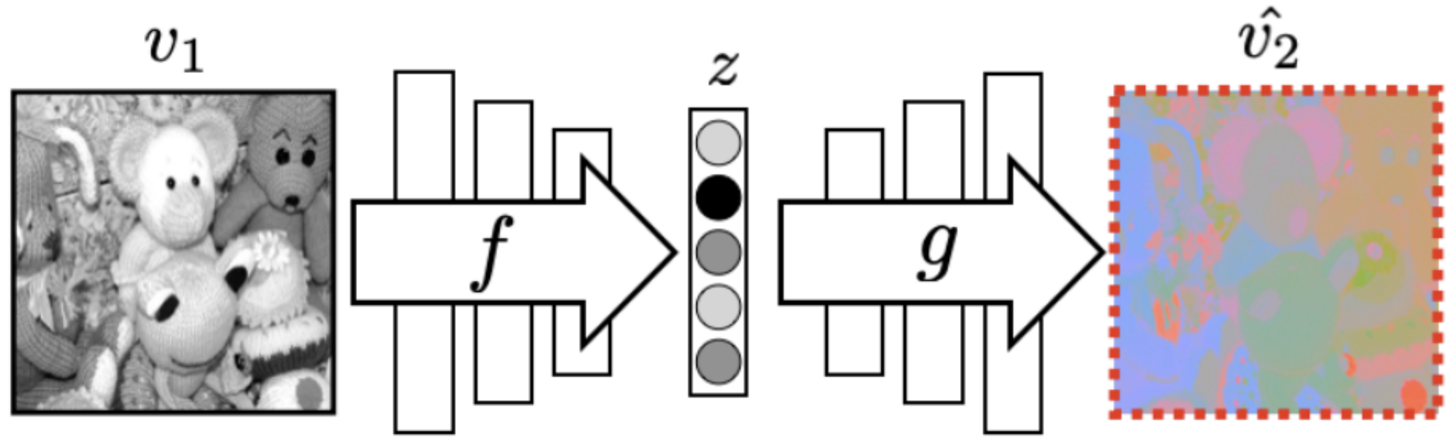


Self-supervised Learning

- **Contrastive / Discriminative** Learning (introduce transformations and learn invariant representation)
  - With negative samples (e.g., SimCLR [Chen *et al.*, ICML'20], MoCo [He *et al.*, CVPR'20])
  - Without negative samples (e.g., BYOL [Grill *et al.,* NeurIPS'20], DINO [Caron *et al.*, ICCV'21])



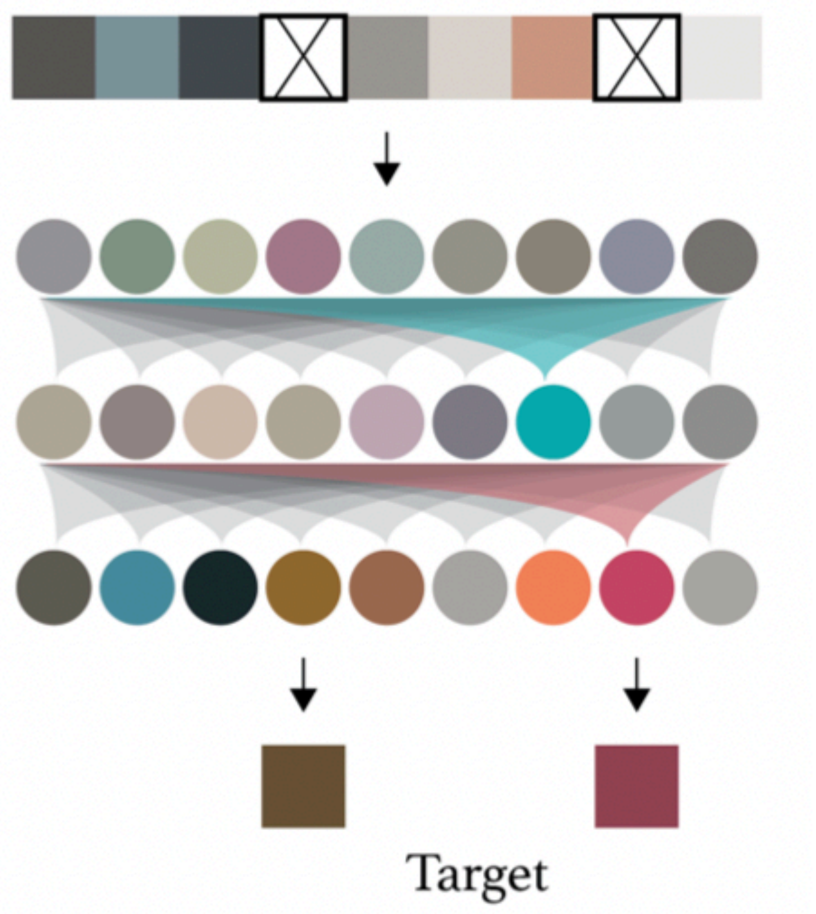
- **Predictive / Generative** Learning (predict what is missing and/or what comes next)
  - Bert-style masked image modeling (e.g., BEiT [Bao *et al.*, ICLR'22], MAE [He *et al.*, CVPR'22])
  - GPT-style autoregressive image modeling (e.g., iGPT [Chen *et al.*, ICML'20])



Self-supervised Learning

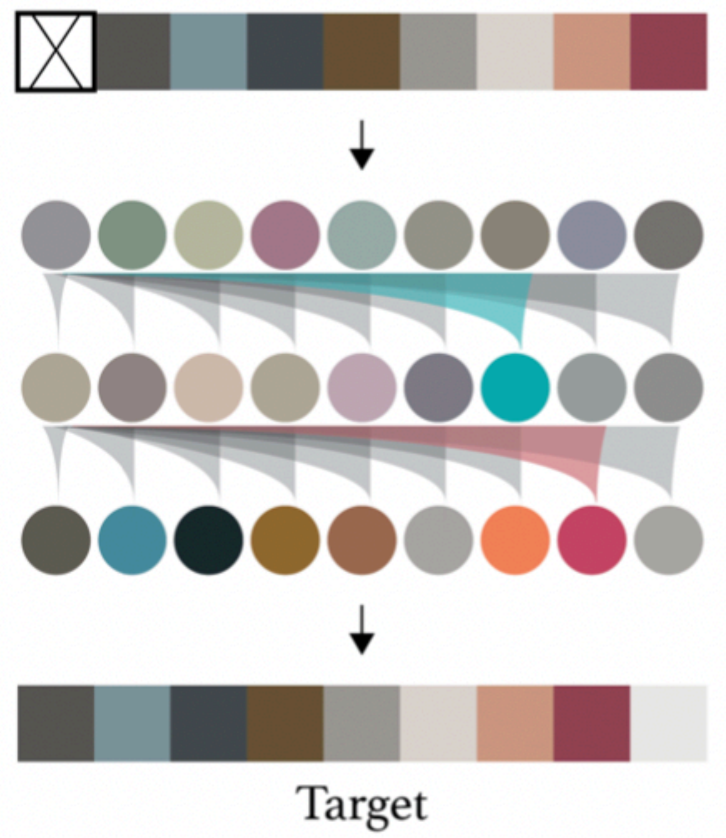
Bert-style

Task: predict missing pixels / patches

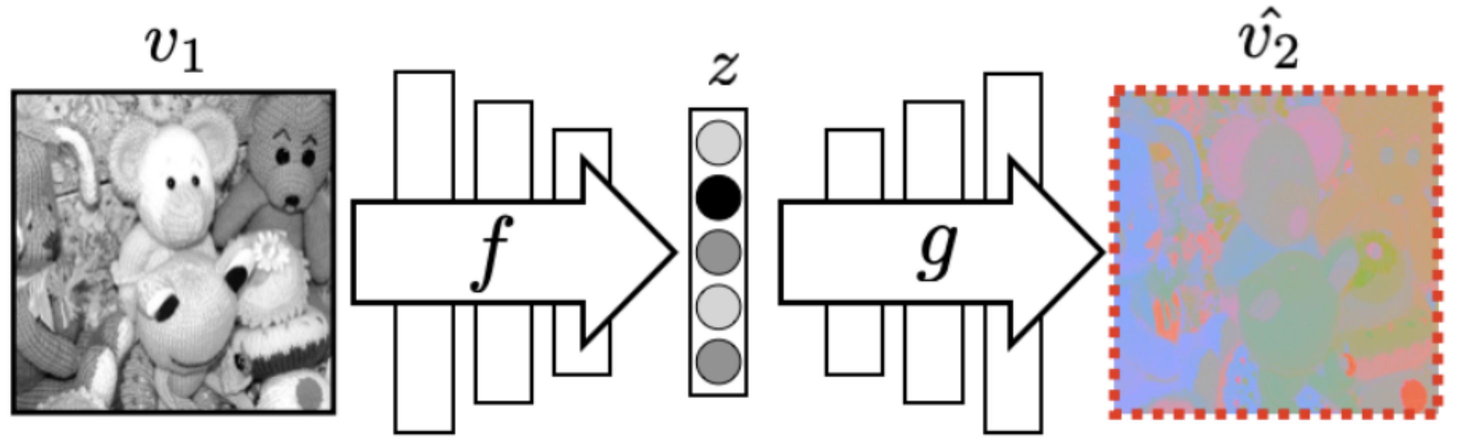


GPT-style

Task: iteratively predict the next pixel / patch



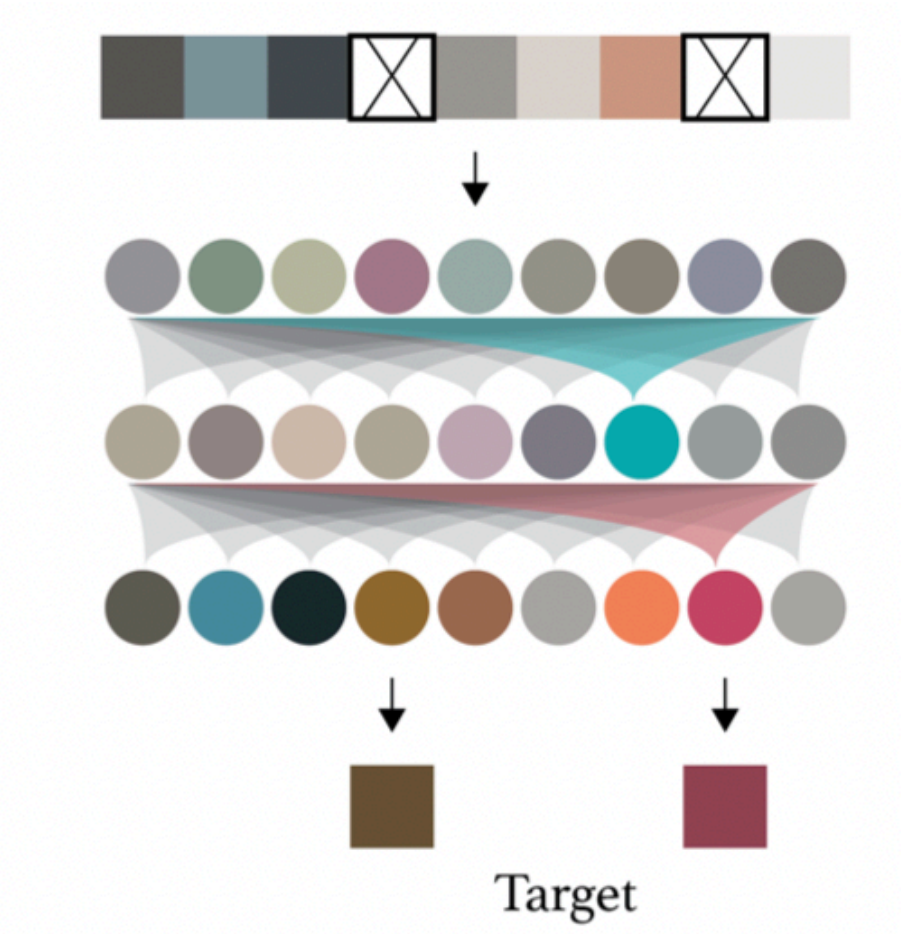
- **Predictive / Generative** Learning (predict what is missing and/or what comes next)
  - Bert-style masked image modeling (e.g., BEiT [Bao *et al.*, ICLR'22], MAE [He *et al.*, CVPR'22])
  - GPT-style autoregressive image modeling (e.g., iGPT [Chen *et al.*, ICML'20])



Self-supervised Learning

Bert-style

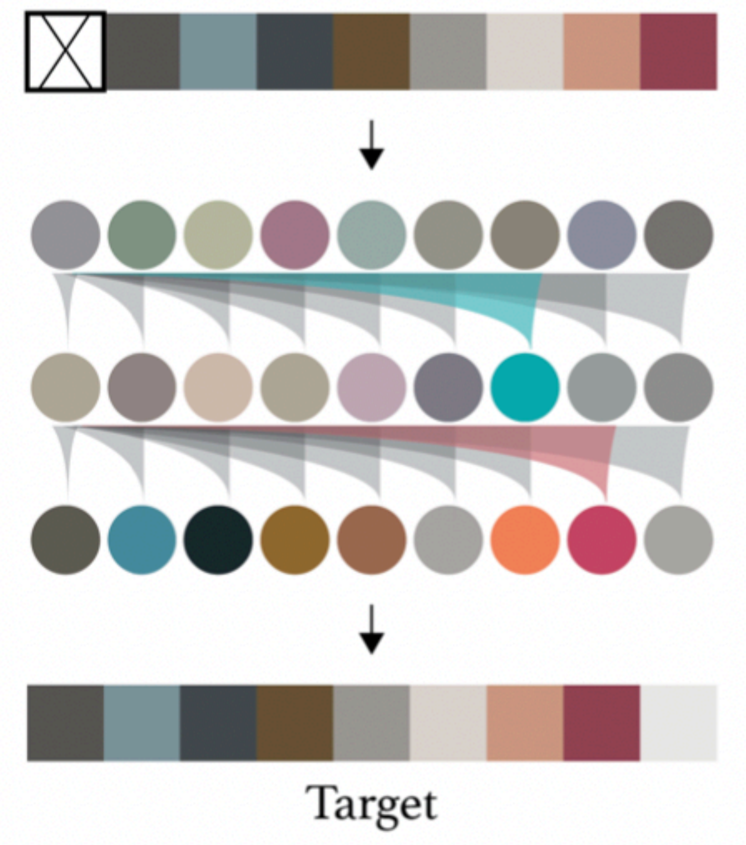
Task: predict missing pixels / patches



Language ✓
Vision ✓

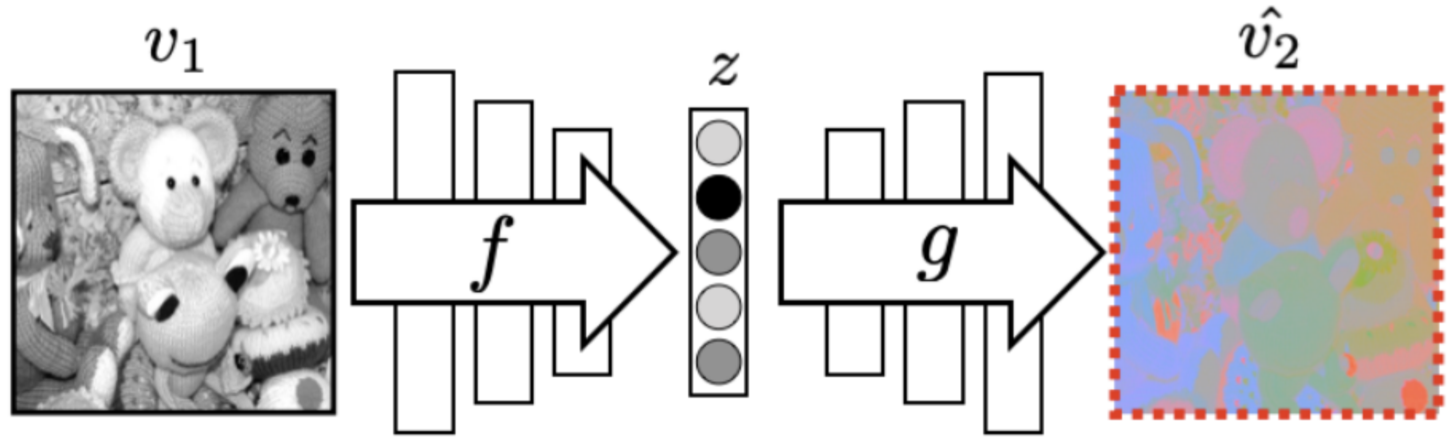
GPT-style

Task: iteratively predict the next pixel / patch

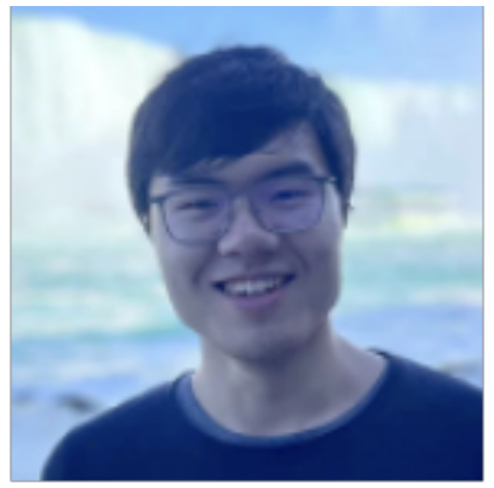


Language ✓
Vision ✗

- **Predictive / Generative** Learning (predict what is missing and/or what comes next)
  - Bert-style masked image modeling (e.g., BEiT [Bao *et al.*, ICLR'22], MAE [He *et al.*, CVPR'22])
  - GPT-style autoregressive image modeling (e.g., iGPT [Chen *et al.*, ICML'20])

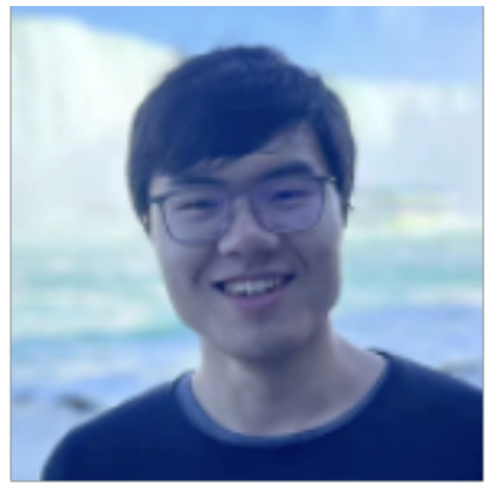


Visual “Language” Model



Tianyu Hua
(MSc, UBC)

Visual “Language” Model



Tianyu Hua
(MSc, UBC)

How do we partition an image into “**words**”?

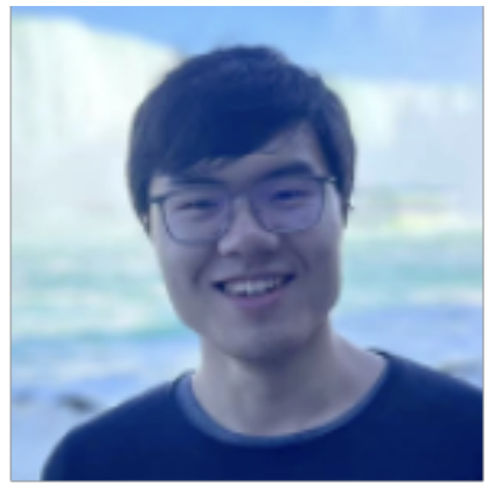
How do we serialize an image into a **sequence** of these words?

How do we **formalize the prediction** for the next likely word?

Self-supervision through Random Segments with Autoregressive Coding (RandSAC), Hua, Tian, Ren, Raptis, Zhao, **Sigal,** *(2023)*

ICLR

Visual "Language" Model

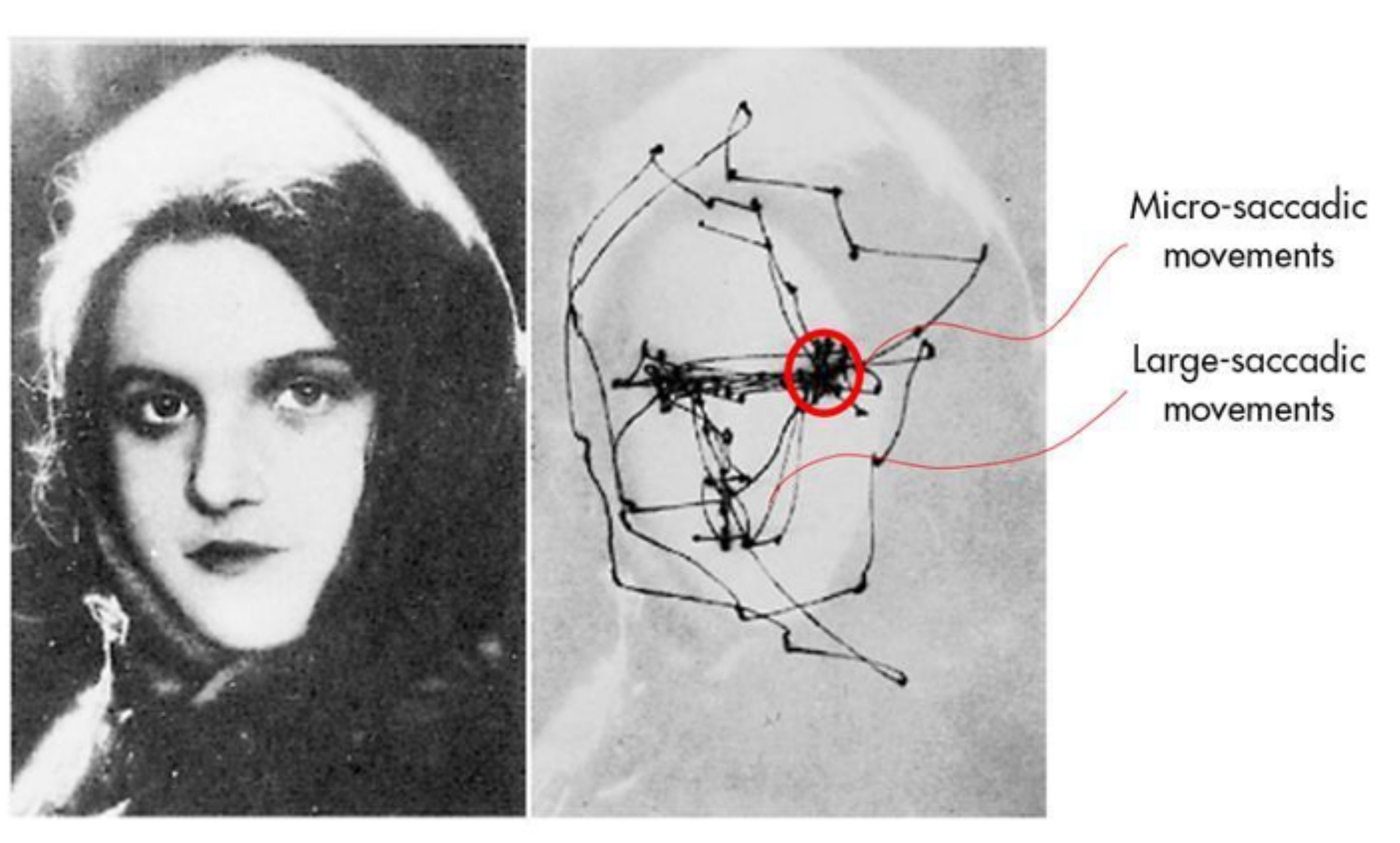


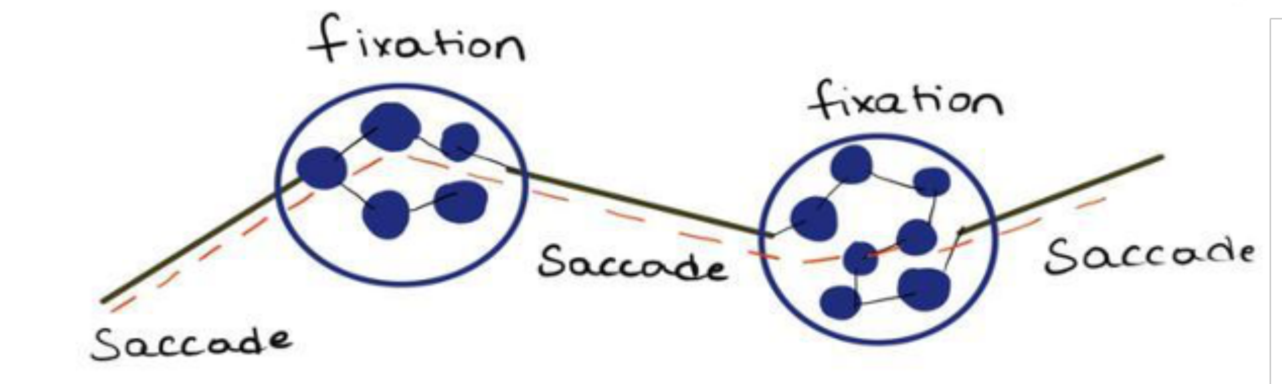
Tianyu Hua
(MSc, UBC)

How do we partition an image into "**words**"?

How do we serialize an image into a **sequence** of these words?

How do we **formalize the prediction** for the next likely word?

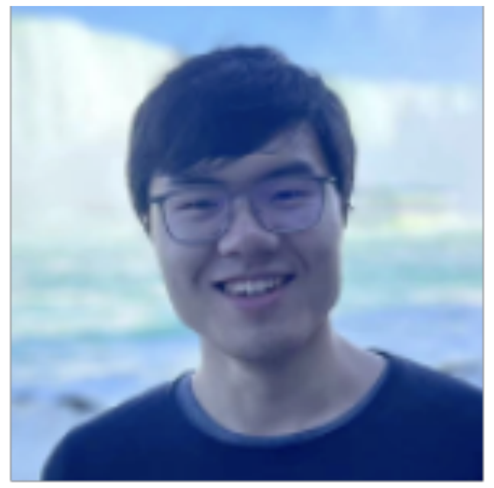




Self-supervision through Random Segments with Autoregressive Coding (RandSAC), Hua, Tian, Ren, Raptis, Zhao, **Sigal,** *(2023)*

ICLR

Random Segments with **Autoregressive Coding**



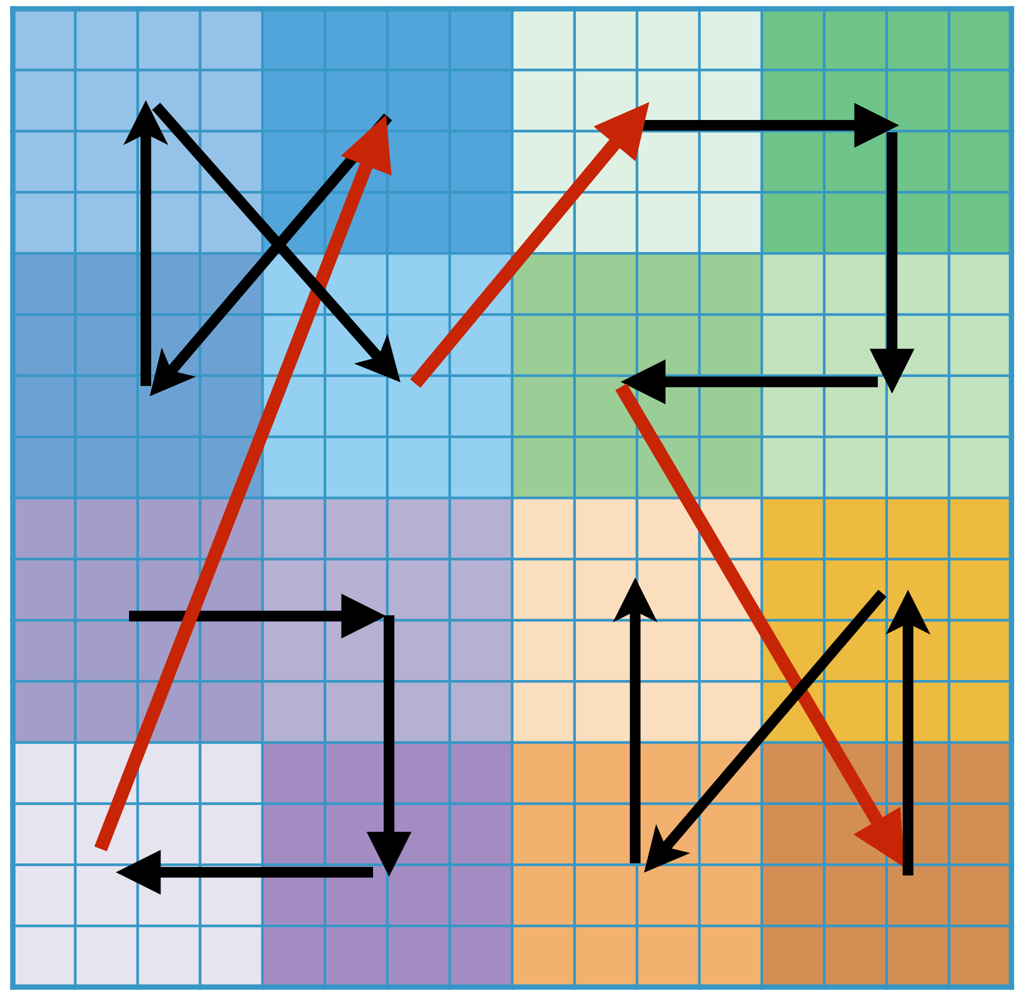
Tianyu Hua
(MSc, UBC)

Group **pixels** into **patches** (visual words)

Group **images patches** (words) into **hierarchically arranged segments** (phrases and sentences)

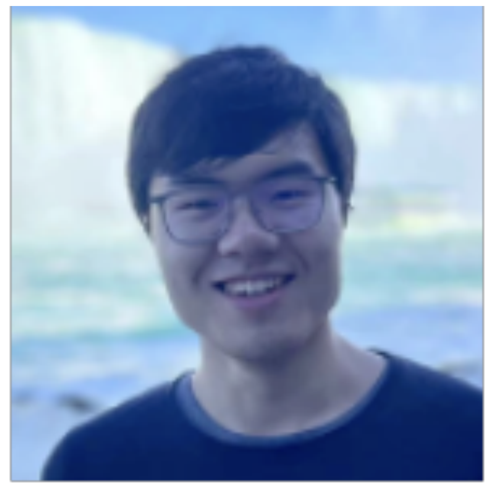
— Within each segment, predictions are made in parallel

— Across segments, predictions are made sequentially



Randomized serialization strategy to account for different order of visual traversal

Random Segments with **Autoregressive Coding**



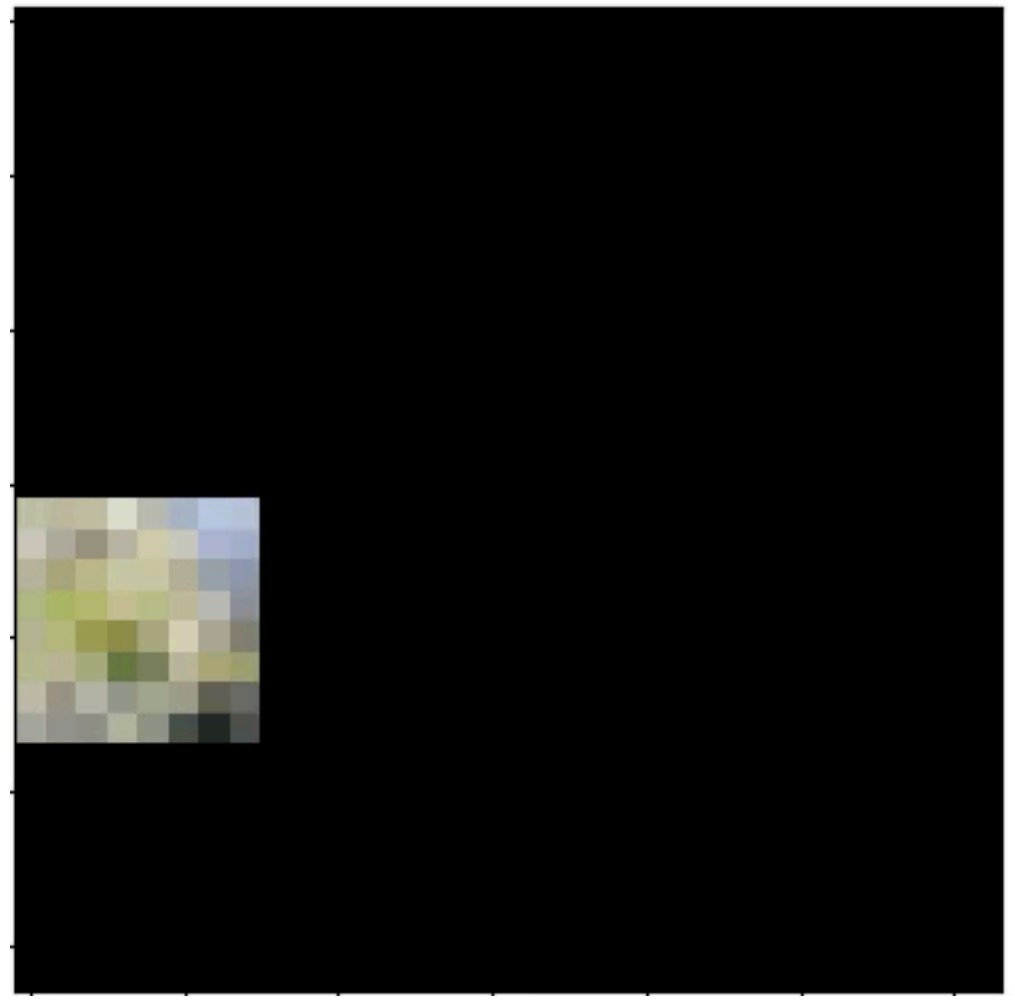
Tianyu Hua
(MSc, UBC)

Group **pixels** into **patches** (visual words)

Group **images patches** (words) into **hierarchically arranged segments** (phrases and sentences)

— Within each segment, predictions are made in parallel

— Across segments, predictions are made sequentially

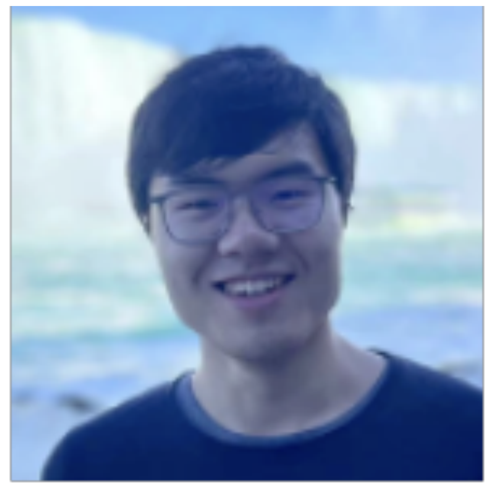


Randomized serialization strategy to account for different order of visual traversal

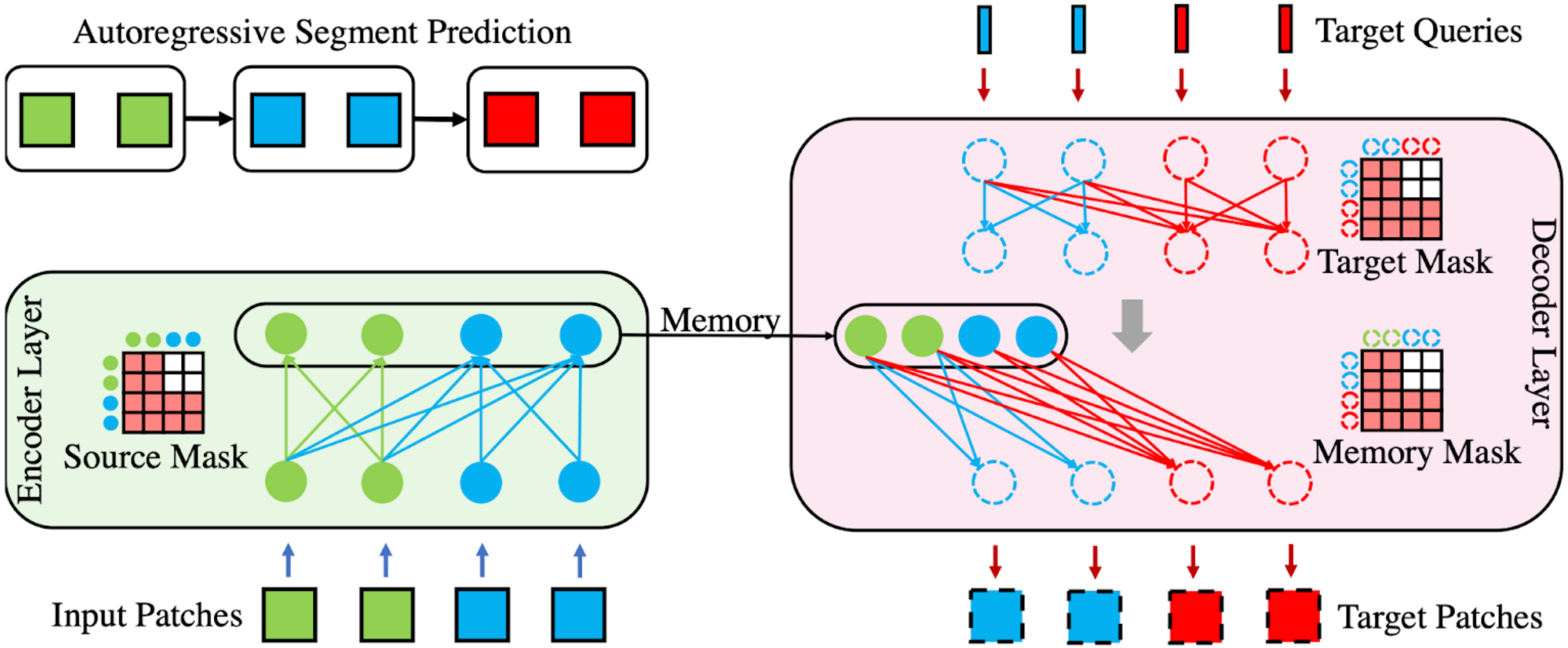
Self-supervision through Random Segments with Autoregressive Coding (RandSAC), Hua, Tian, Ren, Raptis, Zhao, **Sigal,** *(2023)*

ICLR

Random Segments with **Autoregressive Coding**



Tianyu Hua
(MSc, UBC)



Self-supervision through Random Segments with Autoregressive Coding (RandSAC), Hua, Tian, Ren, Raptis, Zhao, **Sigal,** *(2023)*

ICLR

Image **Tokenization**

CIFAR10

Pixel-raster

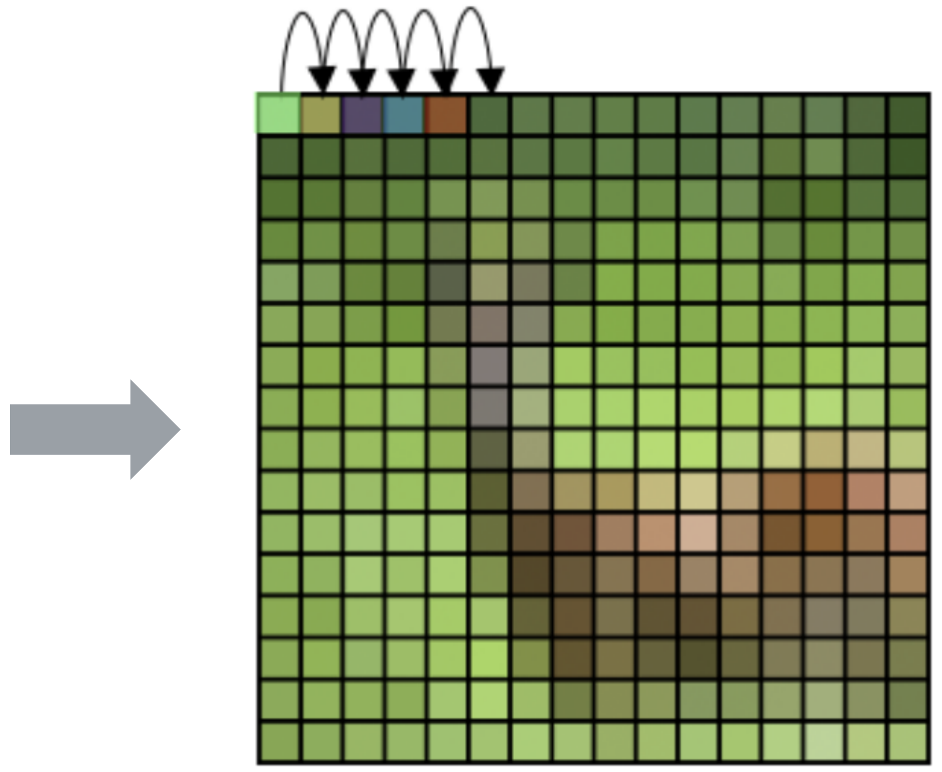
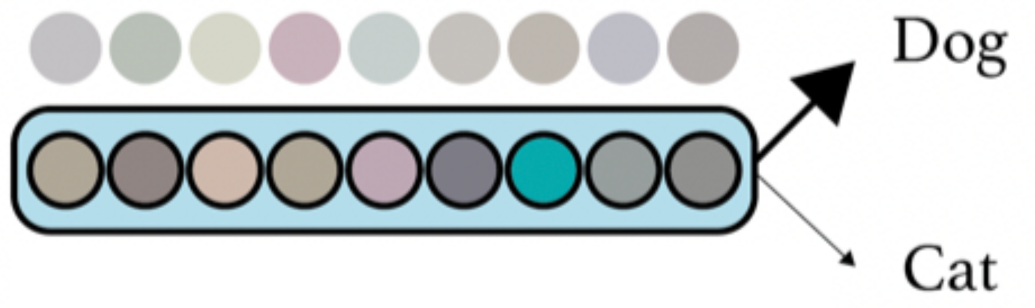


Image **Tokenization**

Linear Probing ↑



CIFAR10

Pixel-raster

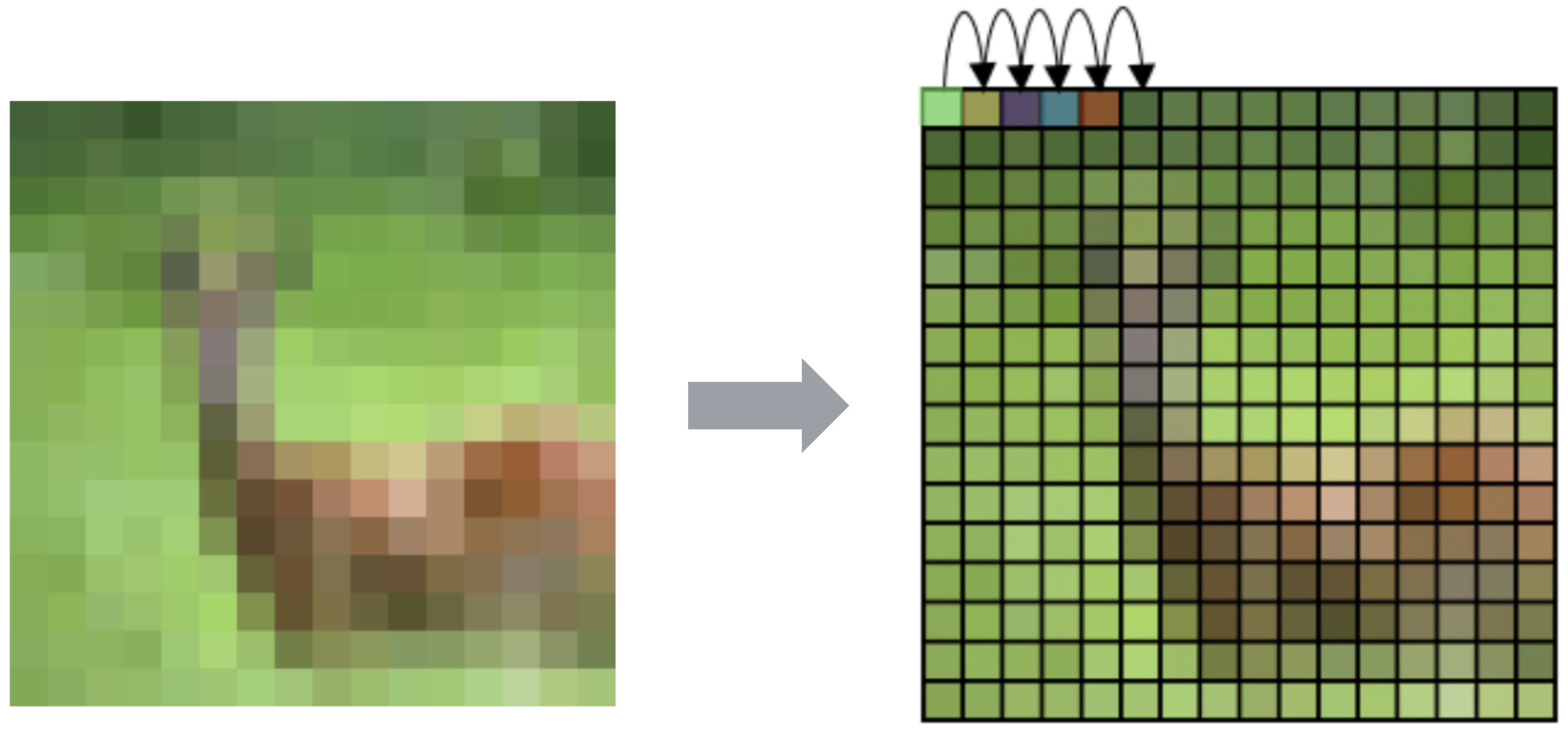
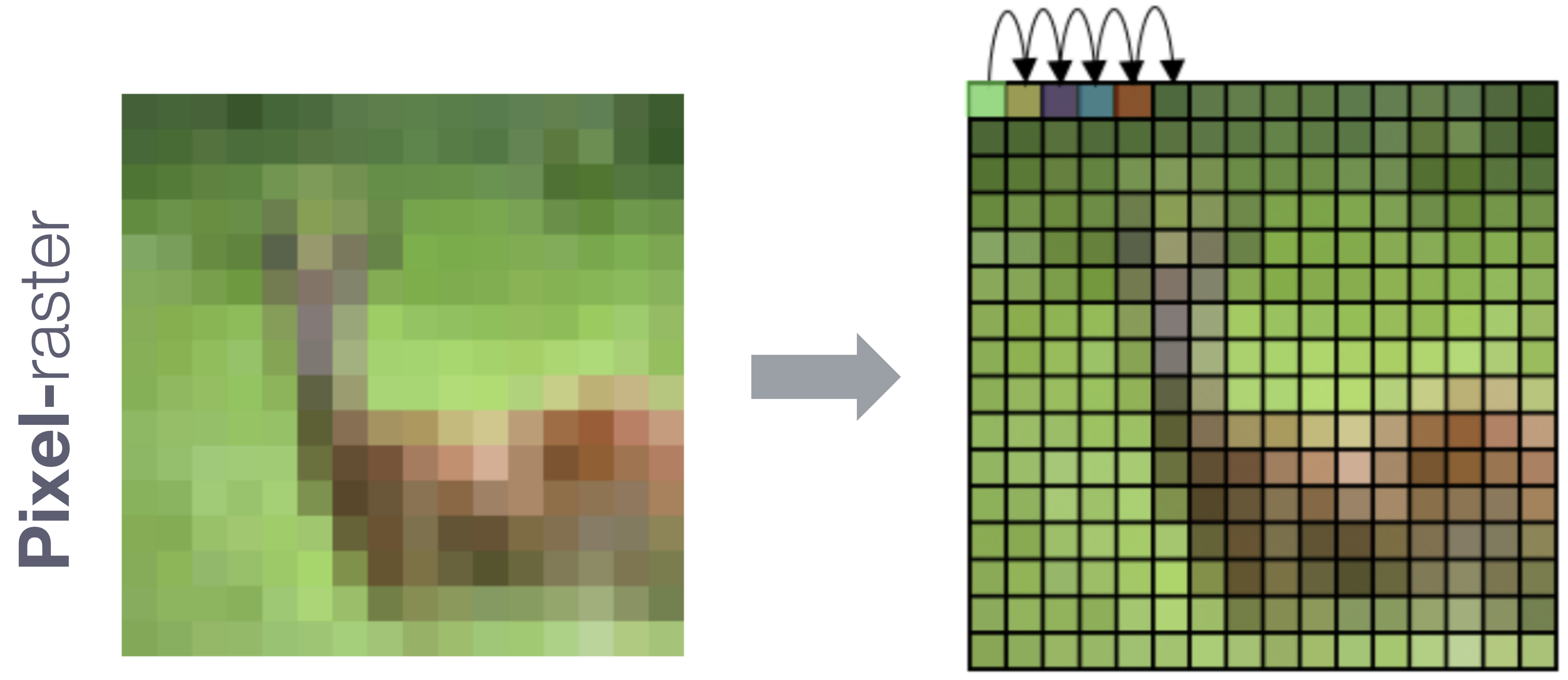
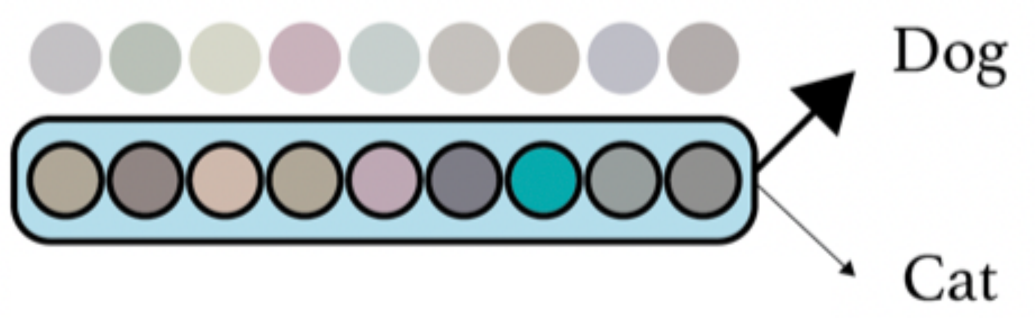


Image **Tokenization**

CIFAR10



Linear Probing ↑



Finetuning ↑

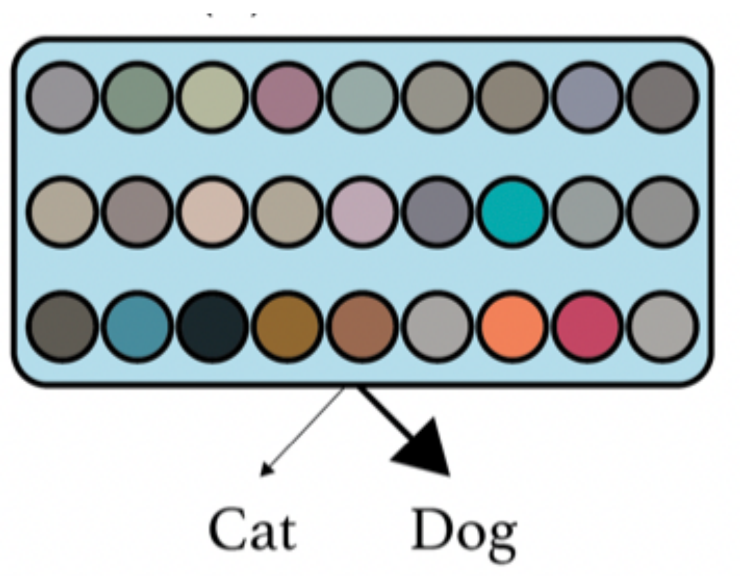
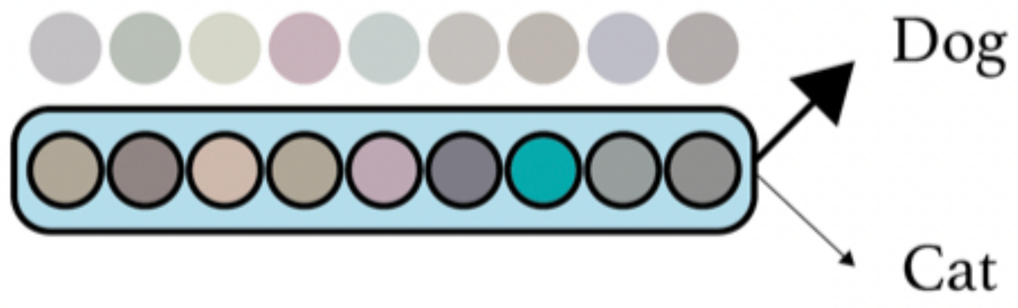
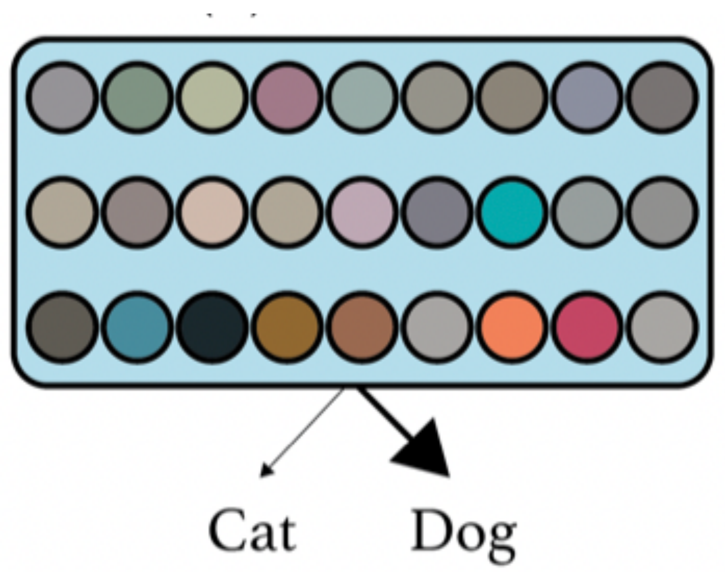
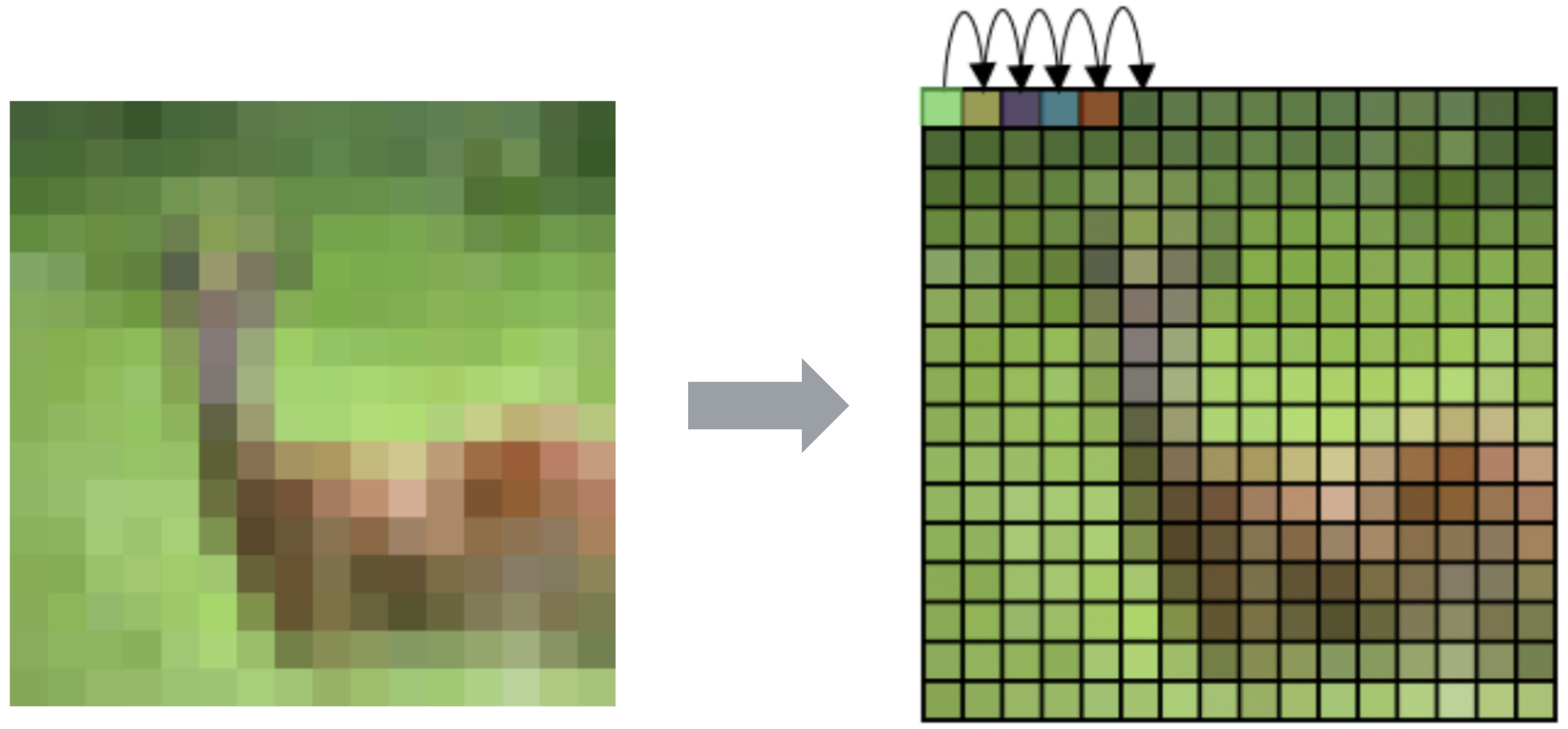


Image **Tokenization**

	Linear Probing ↑	Finetuning ↑
		

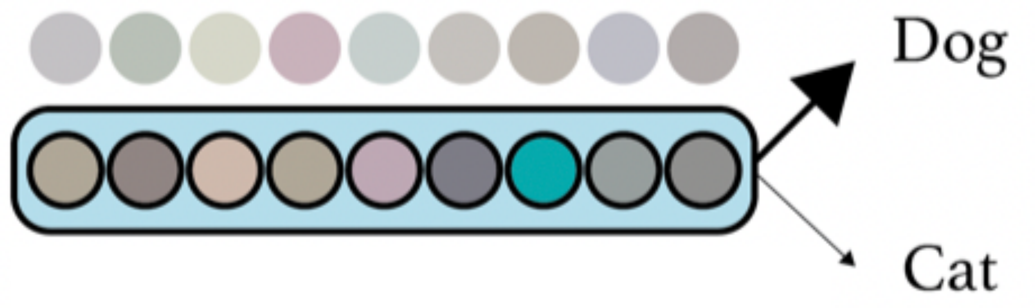
CIFAR10

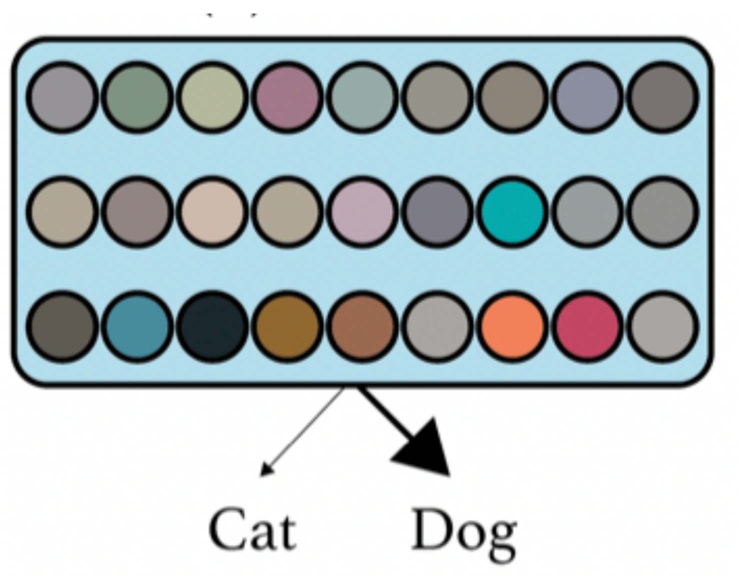


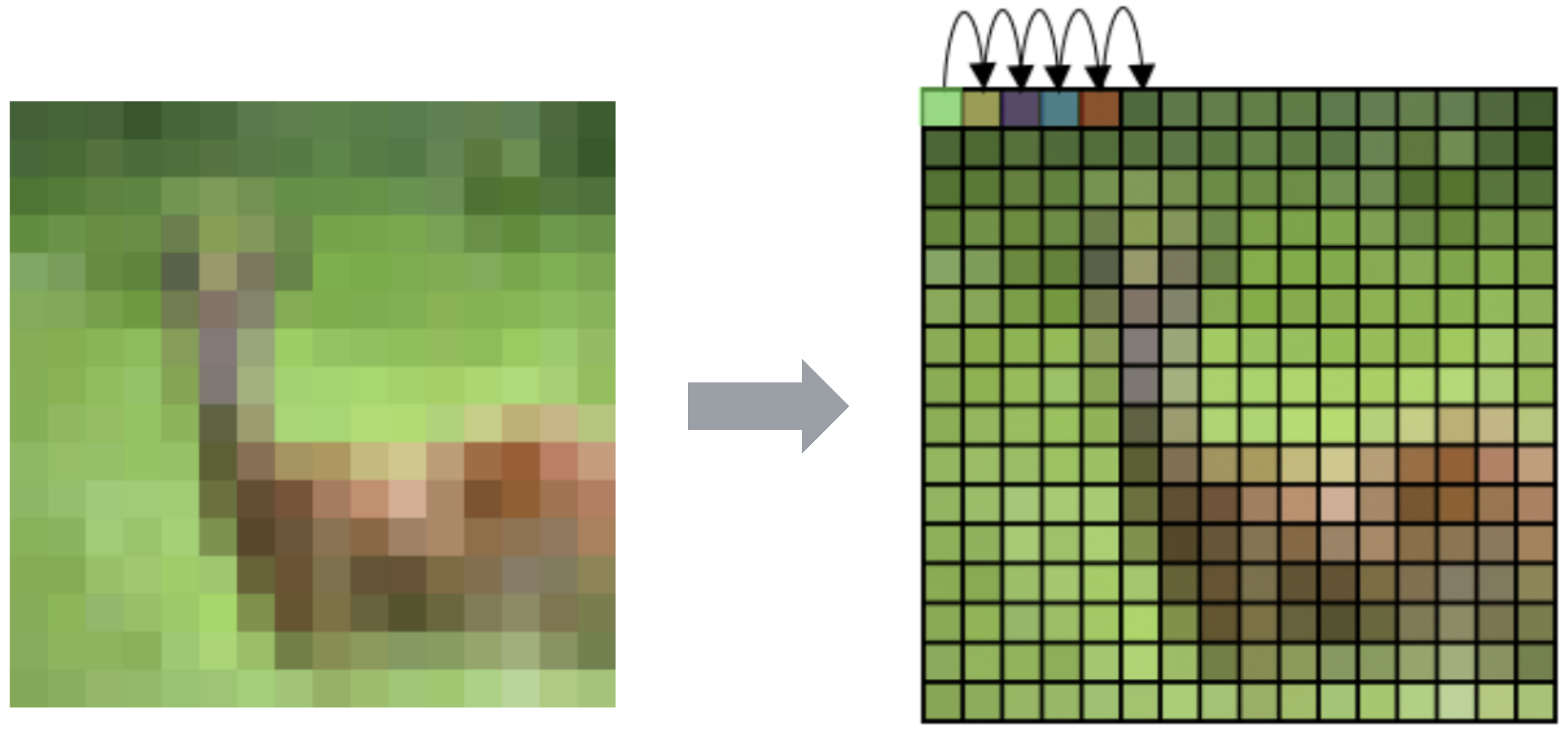
	Linear Probing ↑	Finetuning ↑
Pixel-raster	41.7	59.4

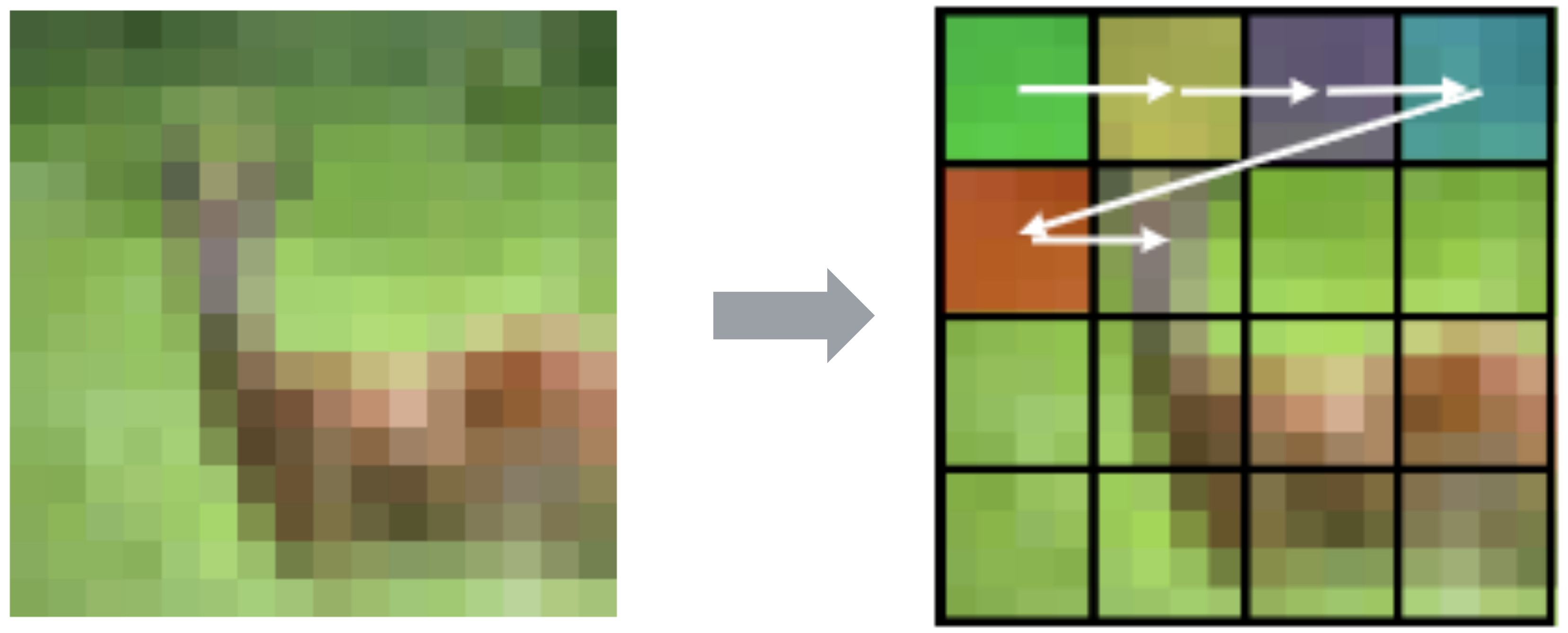
Image **Tokenization**

CIFAR10





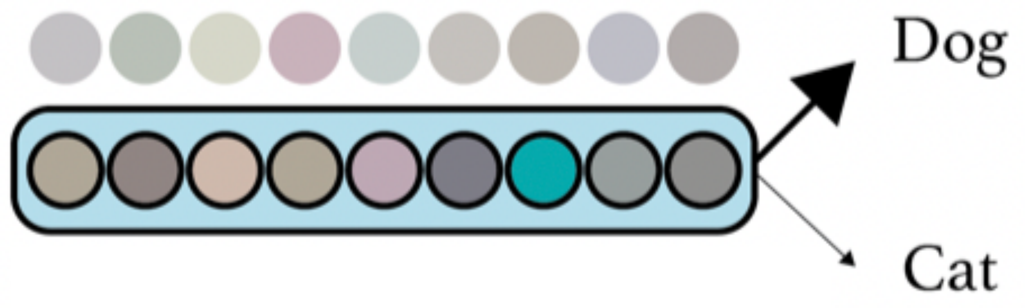


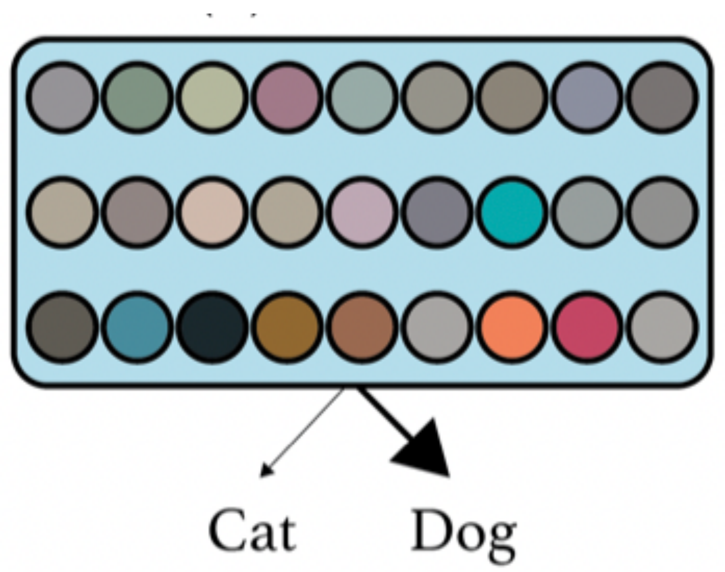


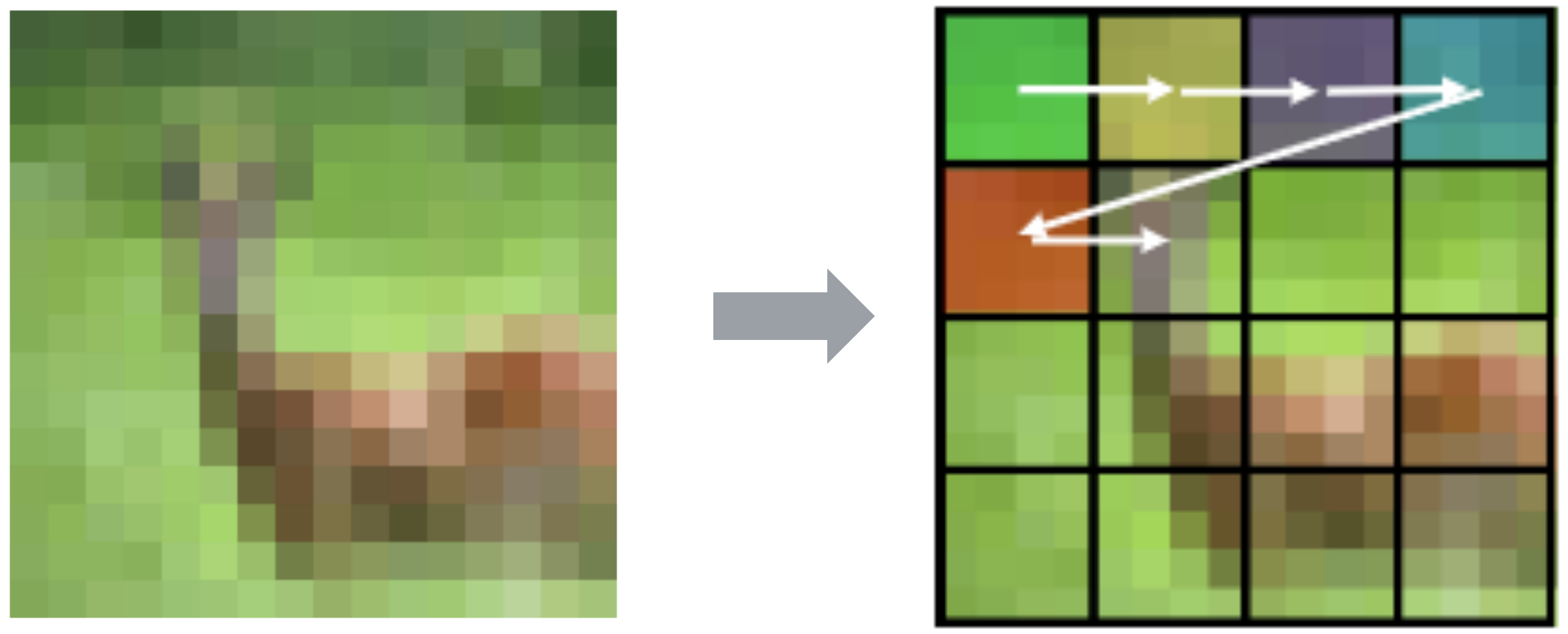
	Linear Probing ↑	Finetuning ↑
Pixel-raster	41.7	59.4
Patch-raster	**55.5**	**78.7**

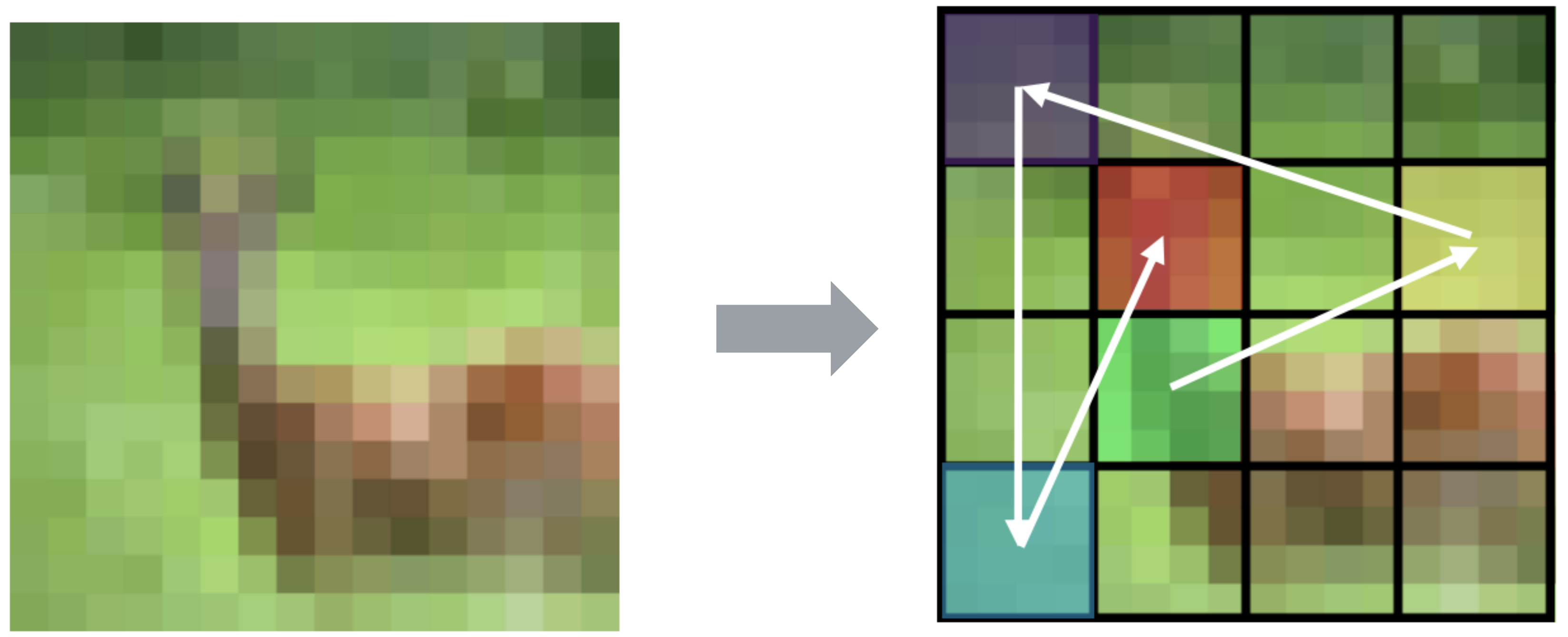
Image **Serialization**

CIFAR10





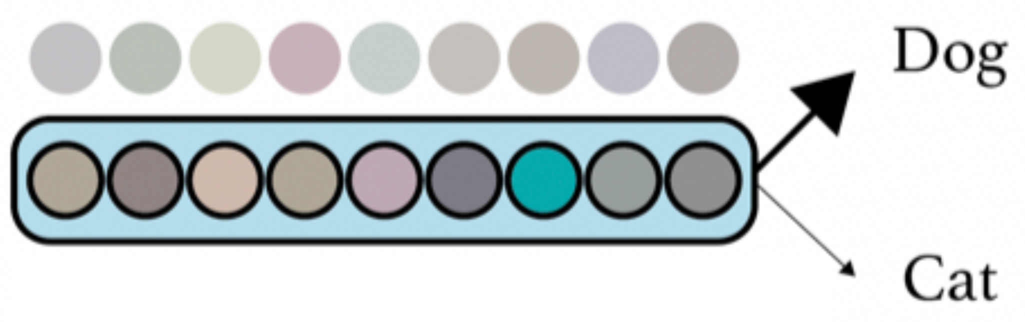


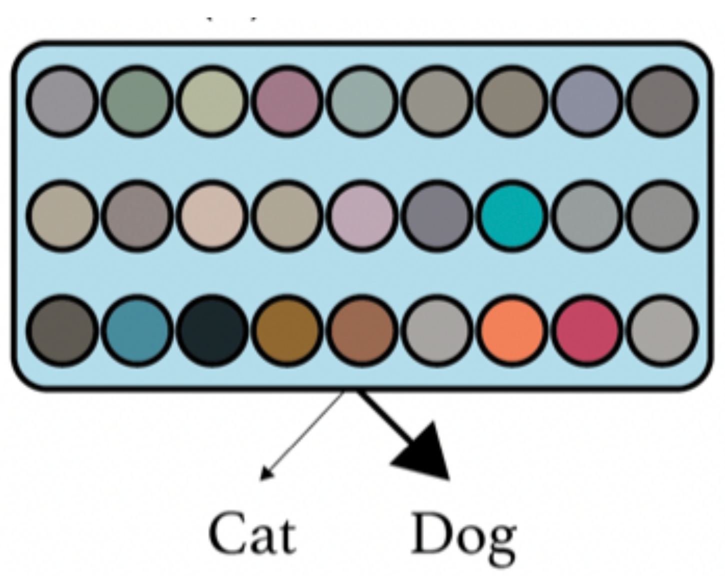


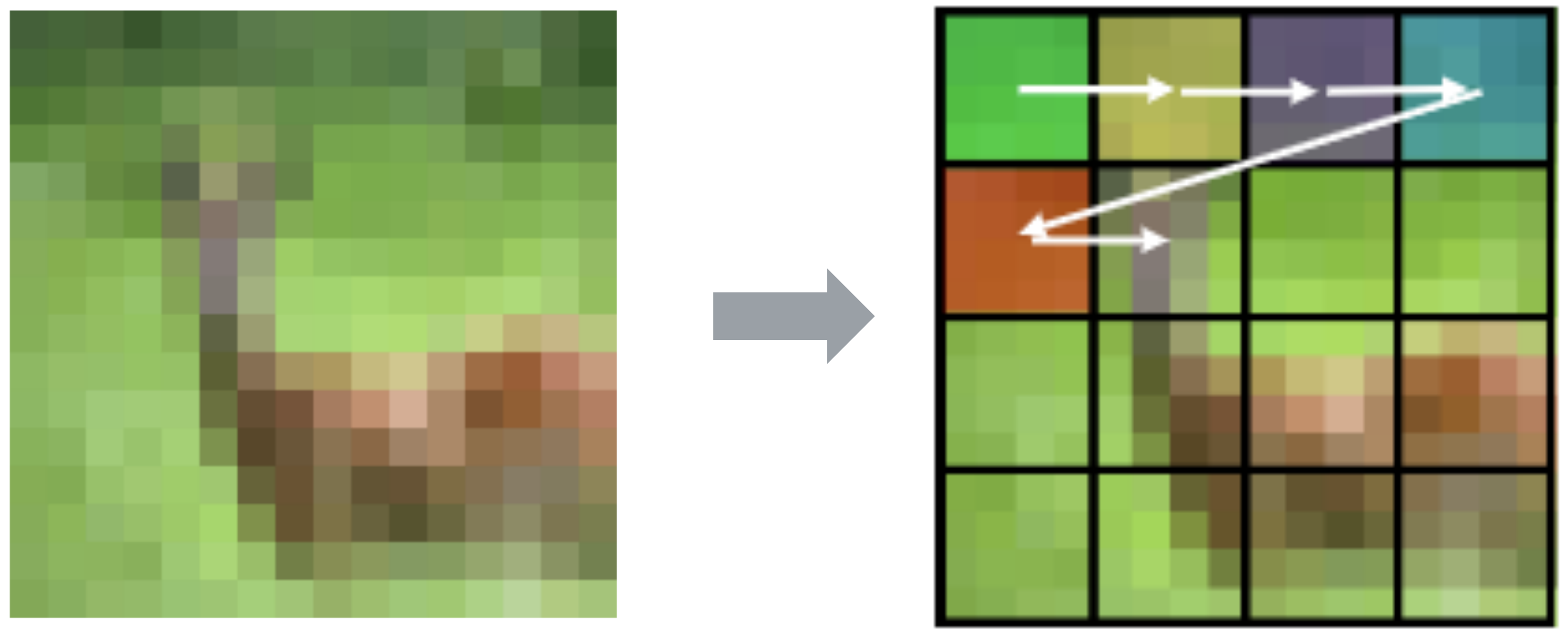
	Linear Probing ↑	Finetuning ↑
Patch-**raster**	55.5	78.7
Patch-**random**	75.5	87.5

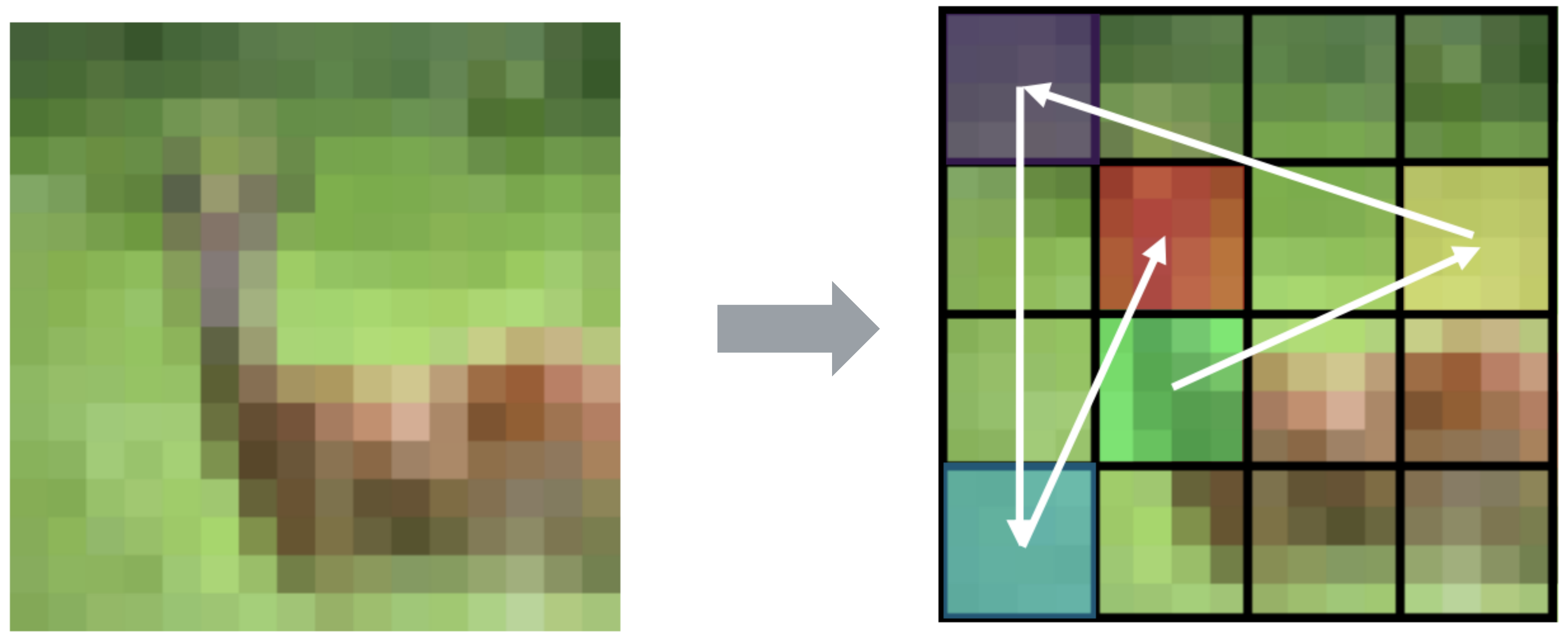
Image **Serialization**

ImageNet100





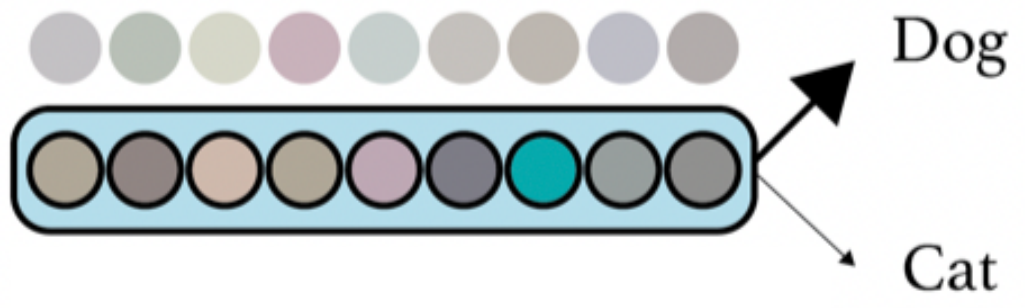


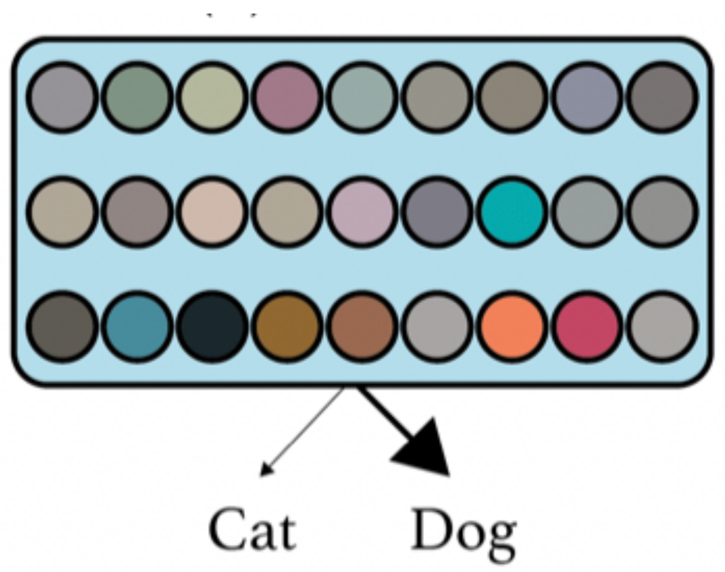


	Linear Probing ↑	Finetuning ↑
Patch-**raster**	49.4	82.1
Patch-**random**	**53.0**	**84.2**

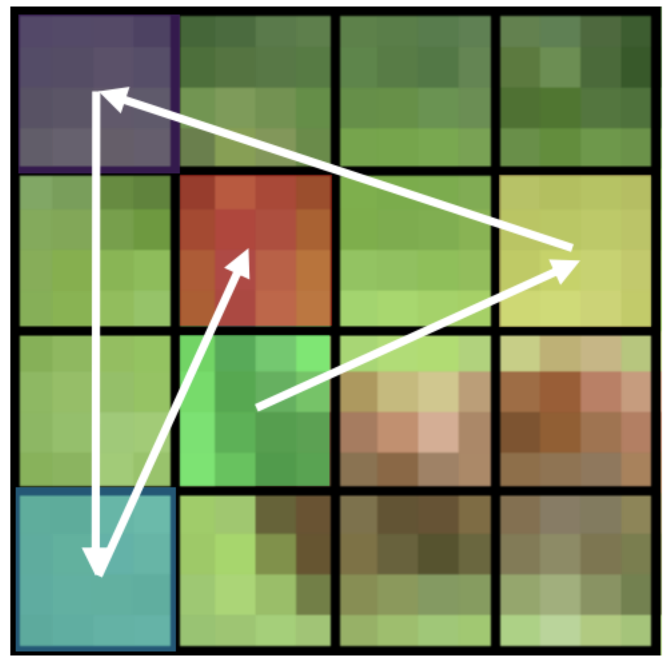
Token **Grouping** (into segments)

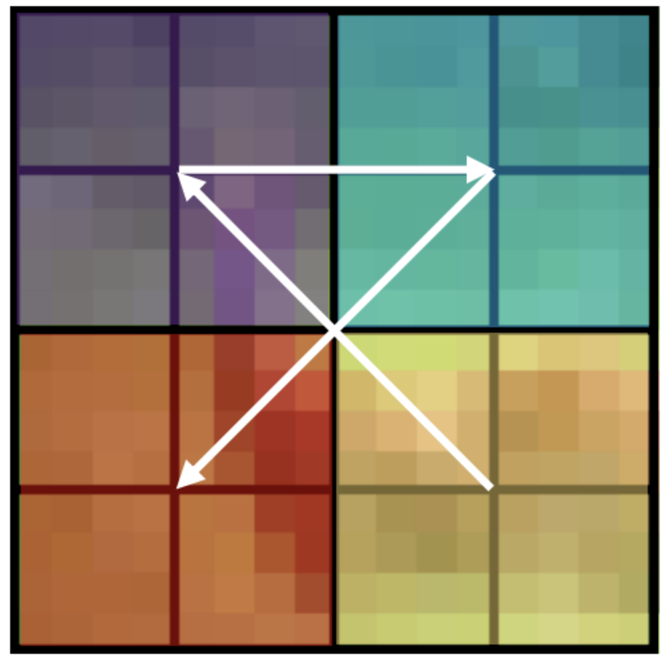
ImageNet100





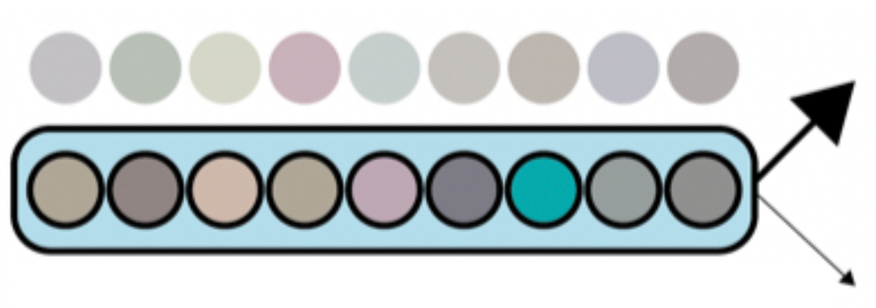
	Linear Probing ↑	Finetuning ↑
Patch-random	53.0	84.2
Segment-random	**64.8**	**86.2**

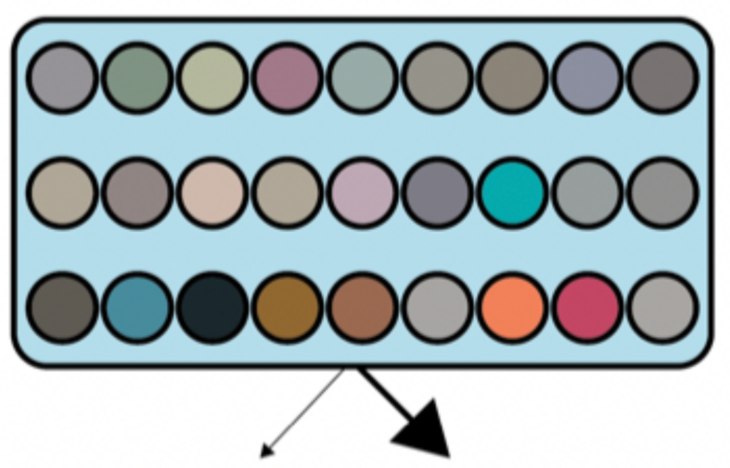


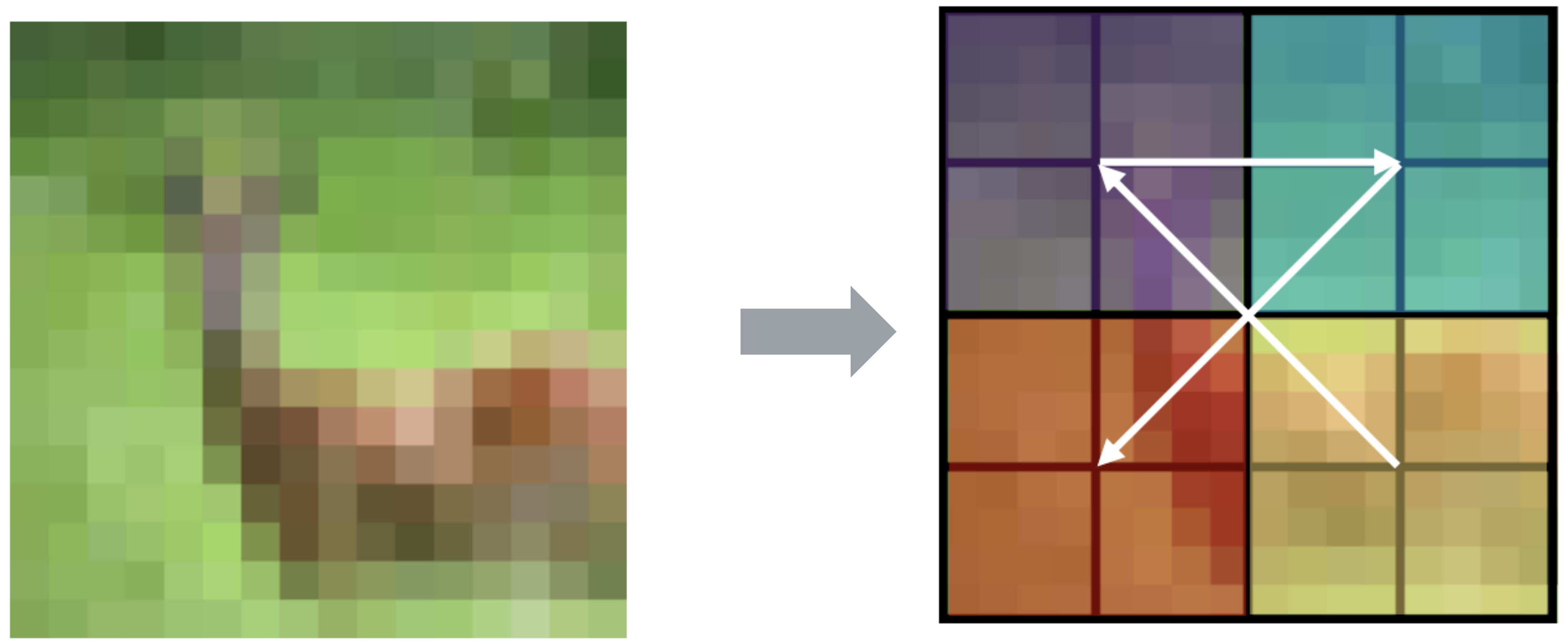


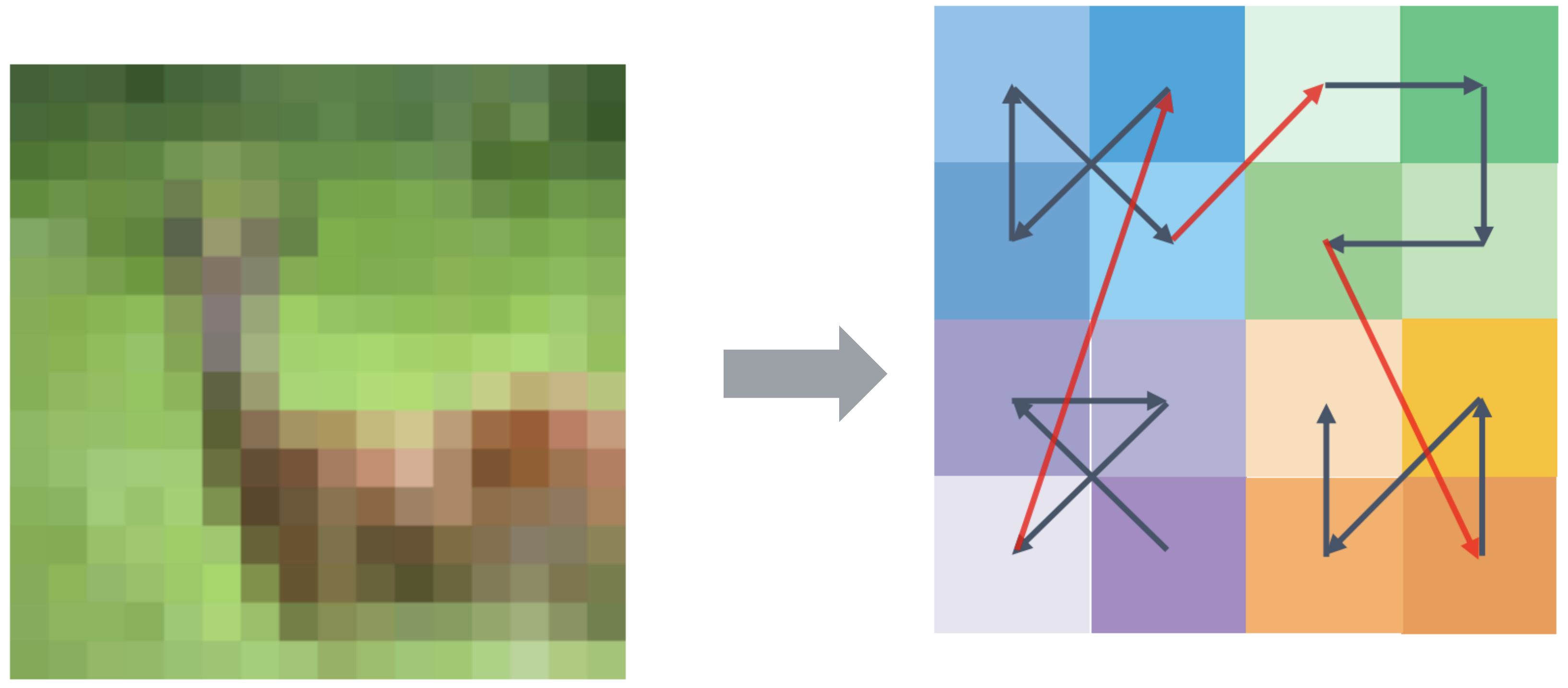
Token **Grouping** (into segments)

ImageNet100









	Linear Probing ↑	Finetuning ↑
Segment-random	64.9	85.3
Hierarchical-random	<u>**65.8**</u>	<u>**85.6**</u>

Visualizations

Low-data **Image Classification** (on CIFAR10/100 Datasets)

Model	CIFAR10		CIFAR100	
	LIN	FT	LIN	FT
Supervised		91.3		64.13
DINO (Caron et al., 2021)	89.0	94.4	65.78	76.3
MAE (He et al., 2021)	87.3	95.9	54.0	81.1
RandSAC-Square	92.1	96.7	**69.7**	**81.5**
RandSAC-Blob	**93.9**	**96.9**	67.9	79.6

Image-level
Classification

Man, Woman, Horse

Image Classification (on ImageNet Dataset)

	Model	Backbone	Parameter	Linear	Fine-tune
Supervised	DeiT (Touvron et al., 2021)	ViT-B	86M	N/A	81.2
Clustering	DINO (Caron et al., 2021)	ViT-B	86M	78.2	82.8
Contrastive Learning	MoCo v3 (Chen et al., 2021b)	ViT-B	86M	76.7	83.2
Masked Image Modeling	BEIT (Bao et al., 2022)	ViT-B	86M	N/A	83.2
	MAE (He et al., 2021)	ViT-B	86M	68.0	83.6
Autoregressive Image Modeling	iGPT (Chen et al., 2020a)	iGPT-S	76M	41.9	N/A
	iGPT (Chen et al., 2020a)	iGPT-M	455M	54.5	N/A
	iGPT (Chen et al., 2020a)	iGPT-L	1362M	65.2	N/A
	RandSAC-Square (K=9)	ViT-B	86M	72.3	**83.7**
	RandSAC-Square (K=16→4)	ViT-B	86M	68.9	**83.9**

Image-level
Classification

Man, Woman, Horse

Object Detection (on COCO Dataset)

Method	Pre-Epochs	AP^{bbox}	AP^{mask}
DeiT (Touvron et al., 2021)	300	47.9	42.9
MoCo-v3 (Chen et al., 2021b)	300	47.9	42.7
DINO (Caron et al., 2021)	300	46.8	41.5
BEiT (Bao et al., 2022)	800	49.8	44.4
MAE (He et al., 2021)	1600	50.3	44.9
RandSAC-Square (K=16→4)	1600	**50.9**	**45.0**

Instance-level
Detection

Man, **Woman**,
Horse, **Horse**

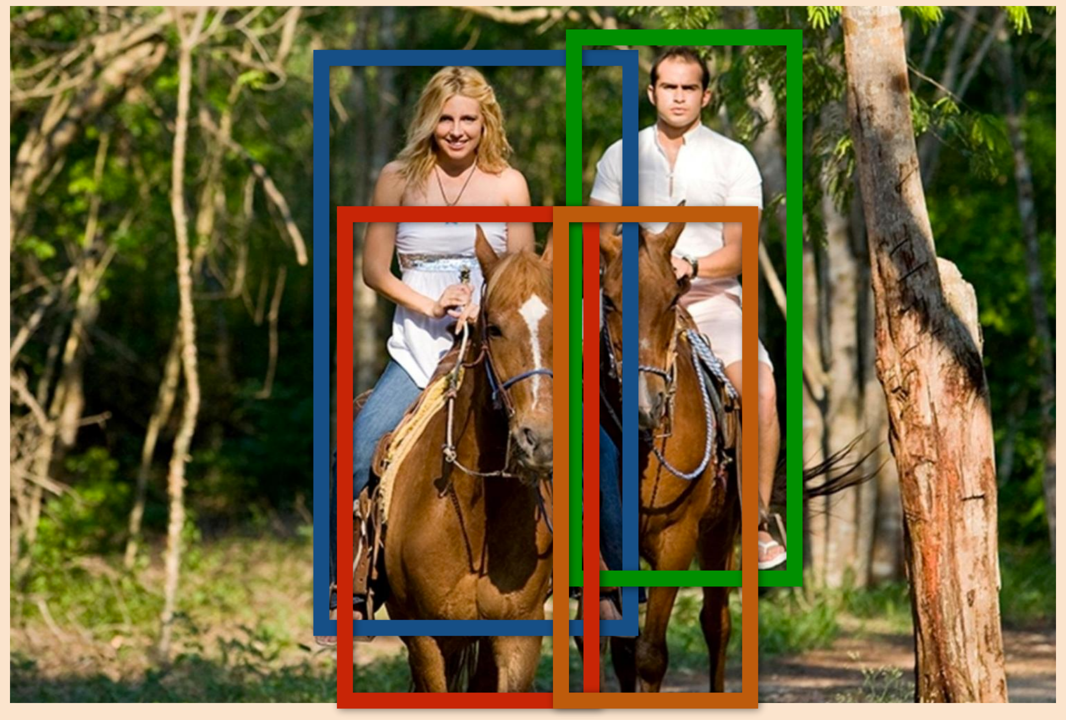


Image Segmentation (on ADE20K Dataset)

Method	Crops	Super.	Self-super.	mIoU
DeiT (Touvron et al., 2021)	1	✓	✗	47.0
MoCo v3 (Chen et al., 2021b)	2	✗	✓	47.2
DINO (Caron et al., 2021)	2+10	✗	✓	47.2
BEiT (Bao et al., 2022)	1	✗	✓	46.5
MAE	1	✗	✓	48.1
RandSAC-Square (K=9)	1	✗	✓	48.3
RandSAC-Square (K=16→4)	1	✗	✓	48.5

Instance-level
Segmentation

Man, **Woman**,
Horse, **Horse**

Compute Efficiency, Strategy 1:

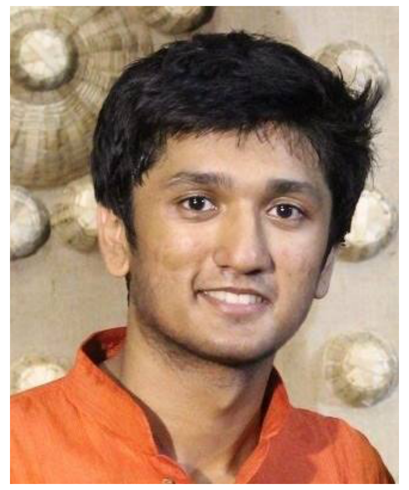
Iterative Refinement

Chapter 2:

Computational Efficiency and Data Bias

Scene Graphs

Siddhesh Khandelwal
(PhD, UBC)



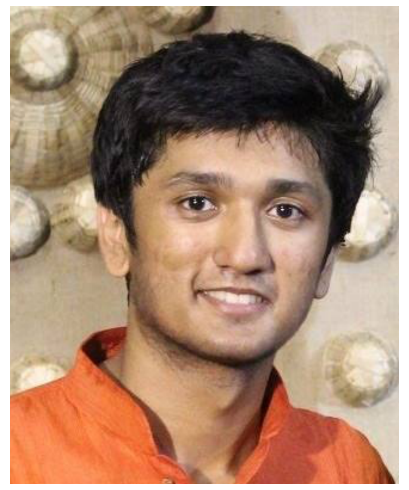
Scene Graphs are graph based representation of images that encode the **objects** in an image along with their **relationships**.

NEURAL INFORMATION PROCESSING SYSTEMS

Iterative Scene Graph Generation, Khandelwal, **Sigal,** *(2022)*

Scene Graphs

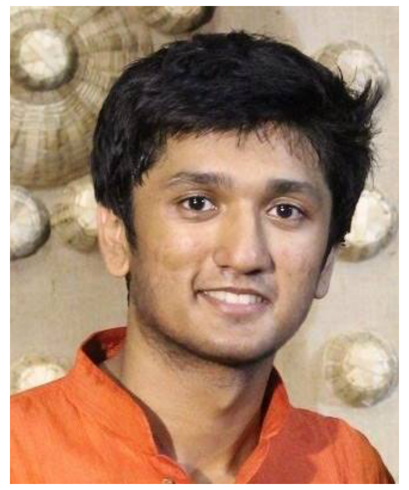
Siddhesh Khandelwal
(PhD, UBC)



Scene Graphs are graph based representation of images that encode the **objects** in an image along with their **relationships**.

Scene Graphs

Siddhesh Khandelwal
(PhD, UBC)



Scene Graphs are graph based representation of images that encode the **objects** in an image along with their **relationships**.

→ **Hat**

→ **Umbrella**

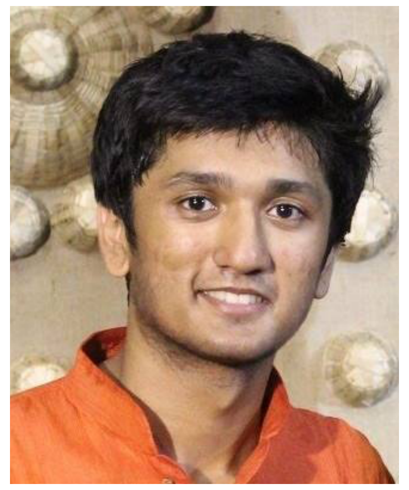
→ **Lamp post**

→ **Person**

Iterative Scene Graph Generation, Khandelwal, **Sigal,** *(2022)*

Scene Graphs

Siddhesh Khandelwal
(PhD, UBC)



Scene Graphs are graph based representation of images that encode the **objects** in an image along with their **relationships**.

Hat

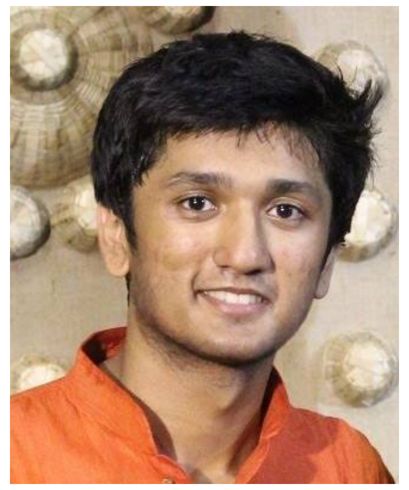
Person

Umbrella

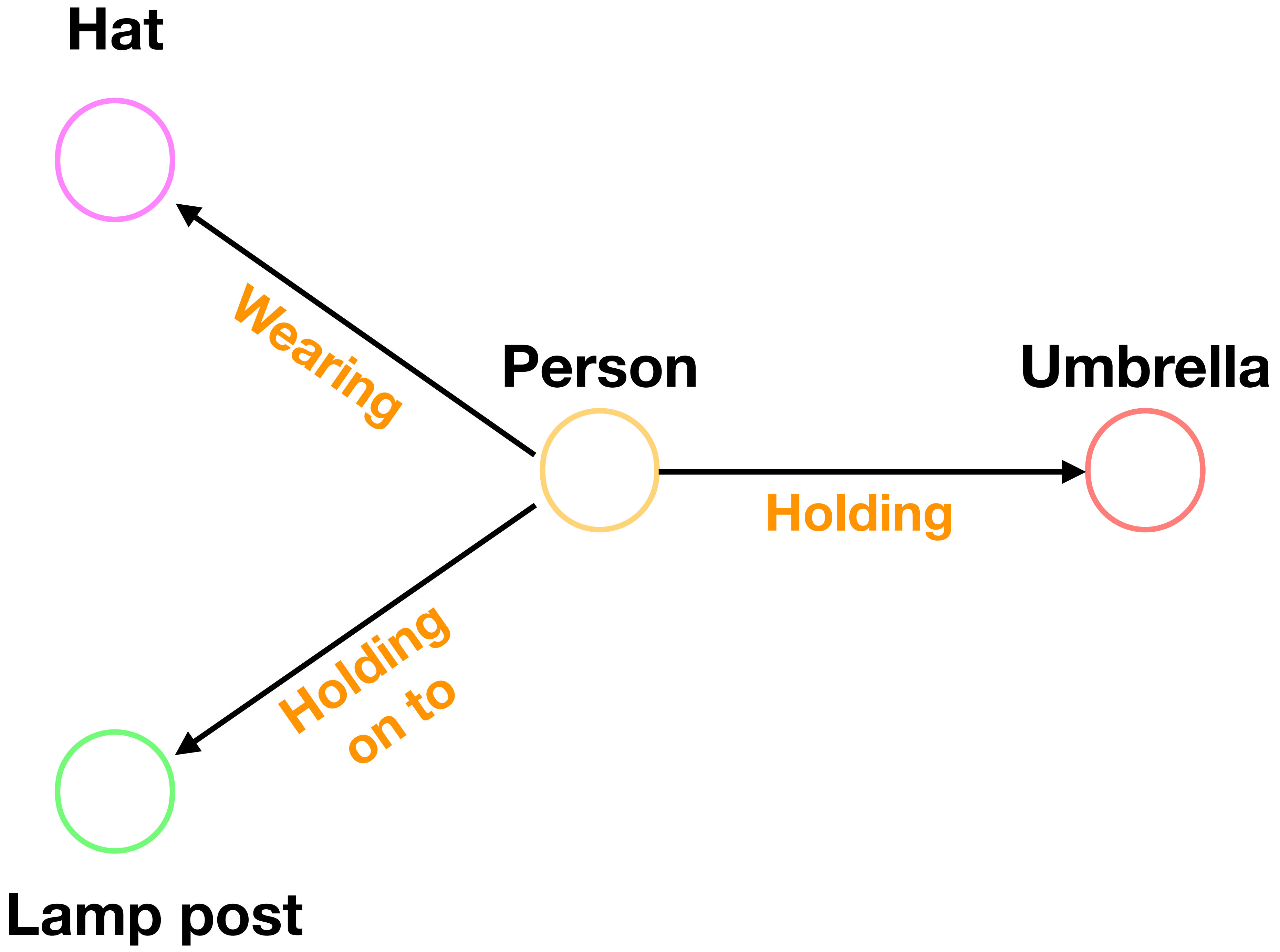
Lamp post

Scene Graphs

Siddhesh Khandelwal
(PhD, UBC)



Scene Graphs are graph based representation of images that encode the **objects** in an image along with their **relationships**.

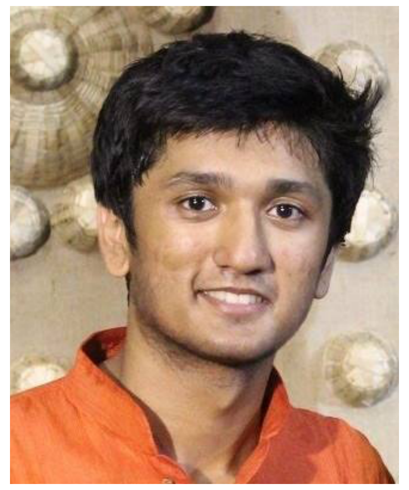


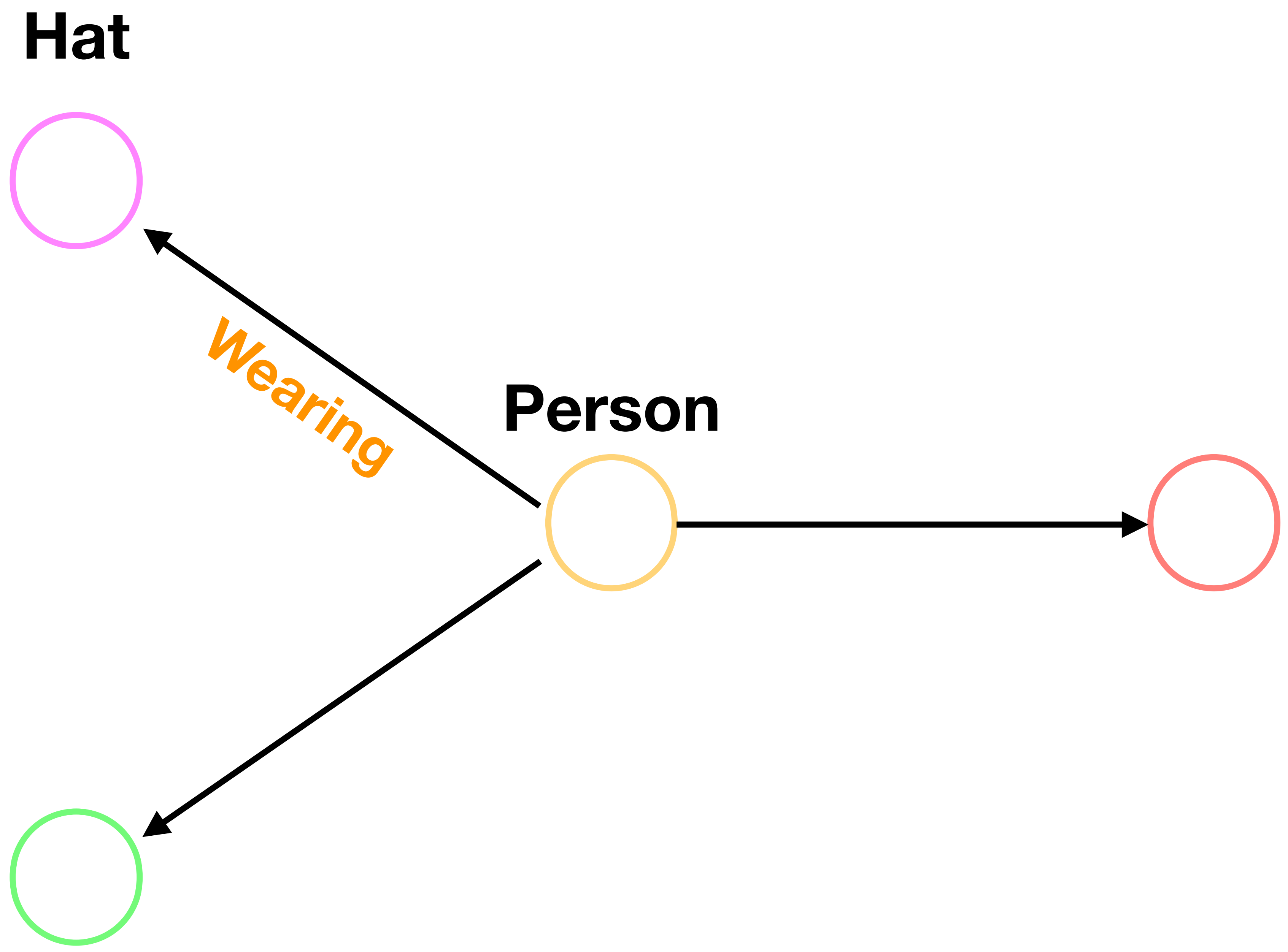
Iterative Scene Graph Generation, Khandelwal, **Sigal,** *(2022)*

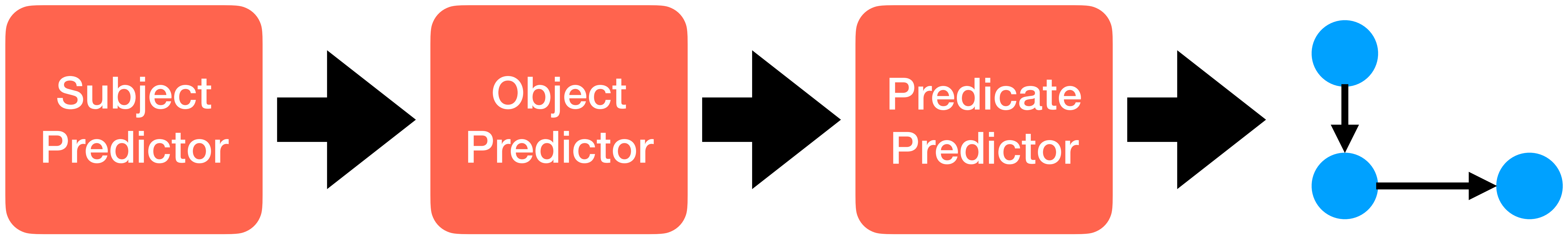
NEURAL INFORMATION PROCESSING SYSTEMS

Scene Graphs

Siddhesh Khandelwal
(PhD, UBC)

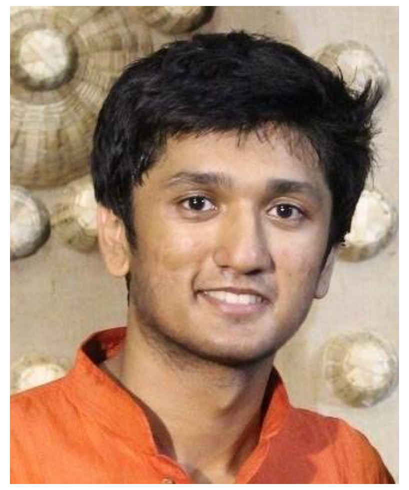


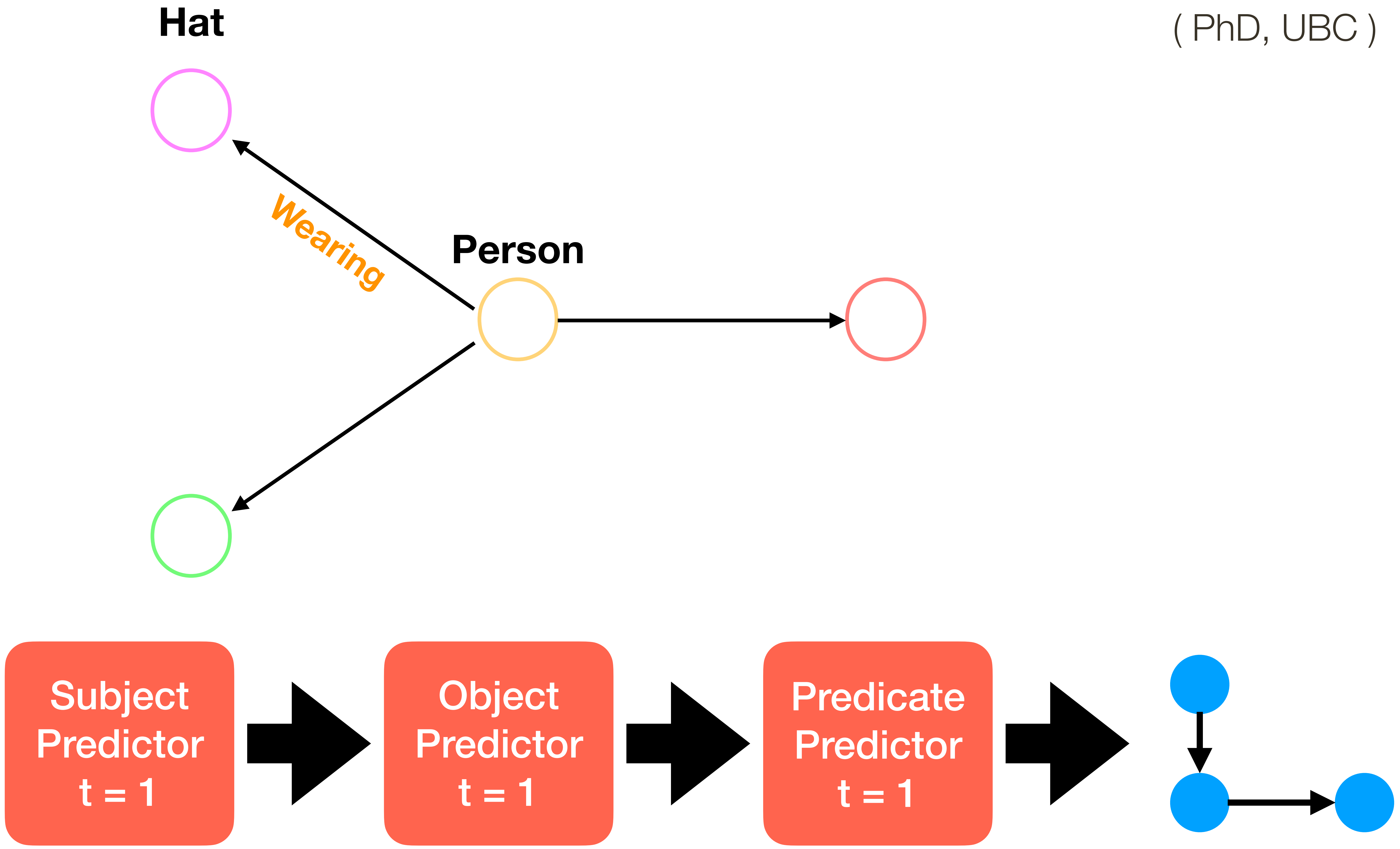




Scene Graphs

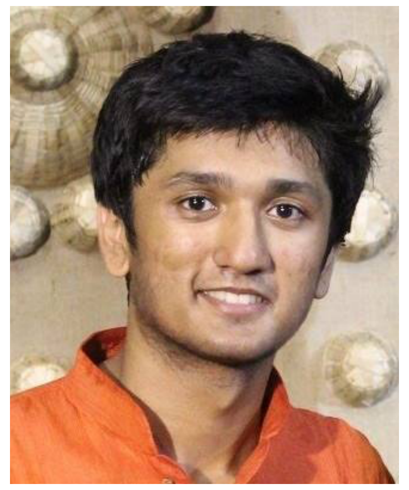
Siddhesh Khandelwal
(PhD, UBC)



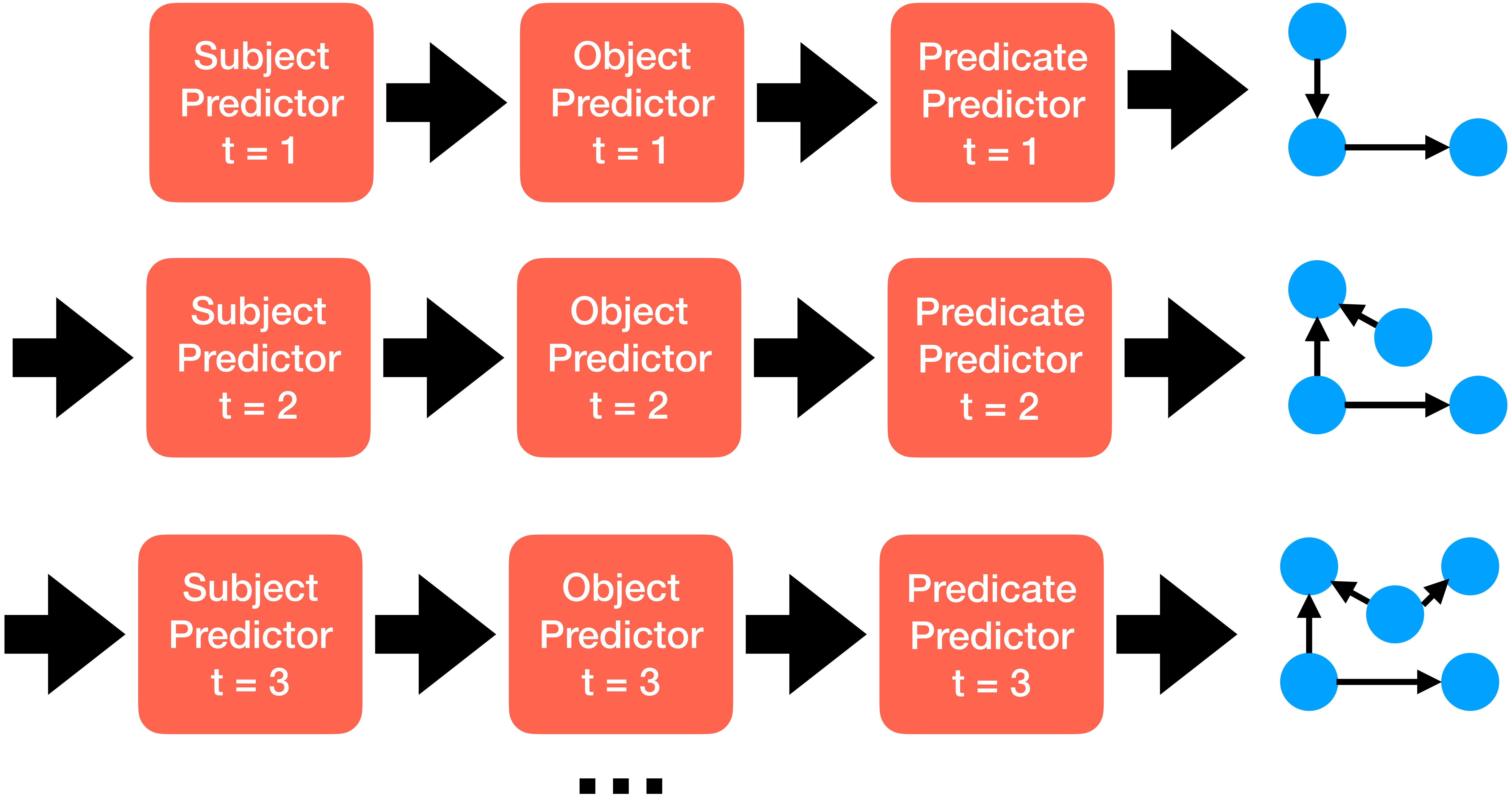


Iterative Scene Graph Generation

Siddhesh Khandelwal
(PhD, UBC)

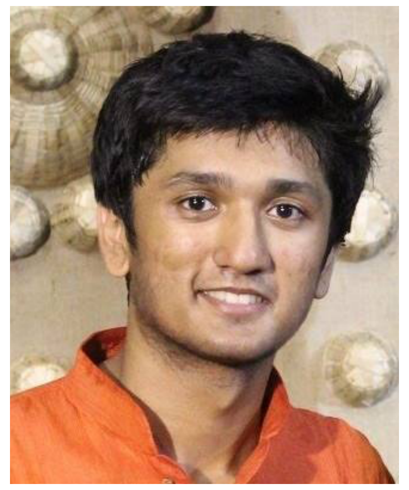


Key Insight: Formulate the problem of Scene Graph estimation as one of **iterative refinement**

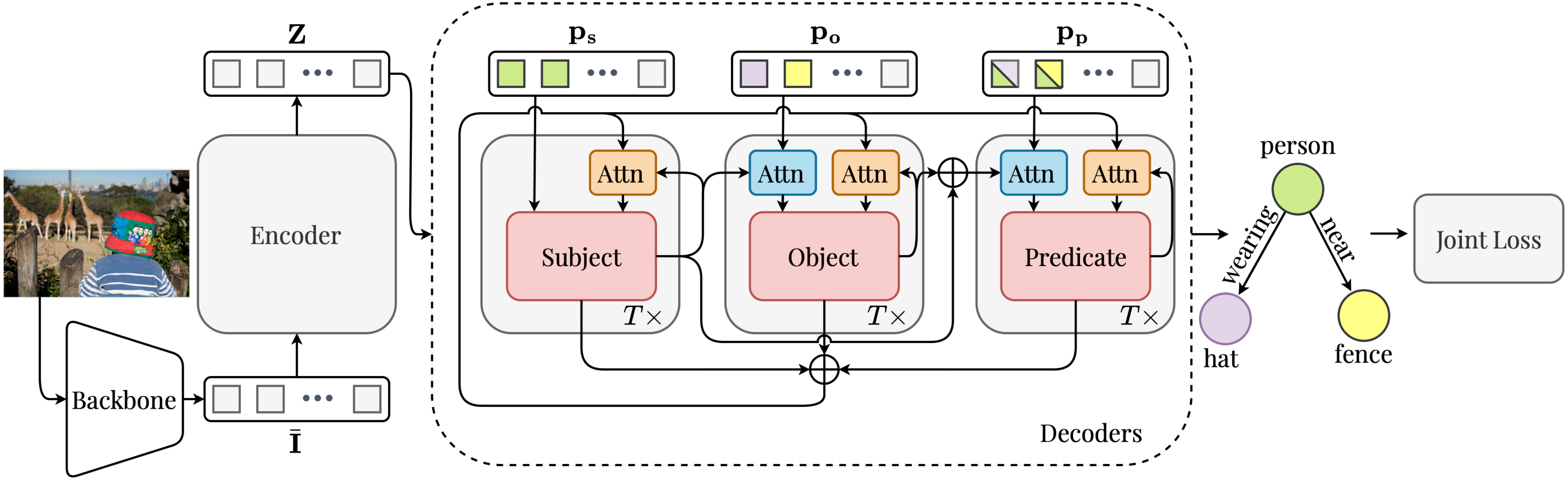


Transformer Based **Iterative Generation**

Siddhesh Khandelwal
(PhD, UBC)



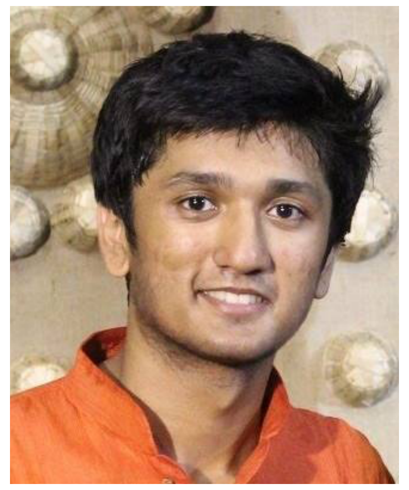
The iterative framework is realized using a novel transformer-based architecture



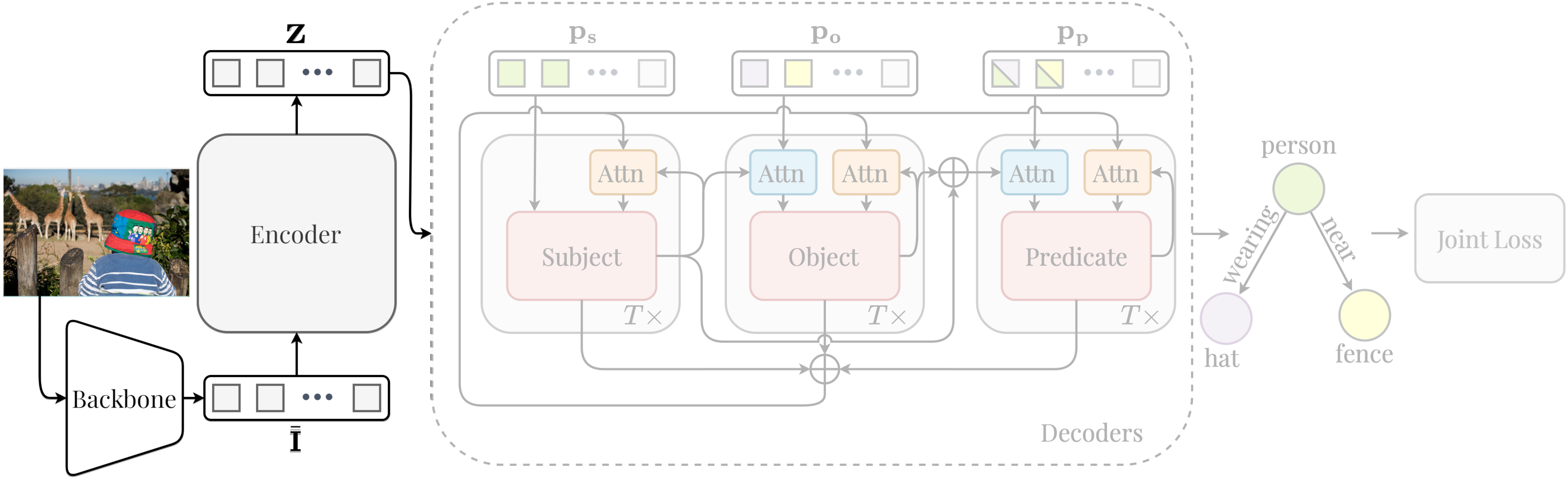
Iterative Scene Graph Generation, Khandelwal, **Sigal,** *(2022)*

Image **Encoder**

Siddhesh Khandelwal
(PhD, UBC)



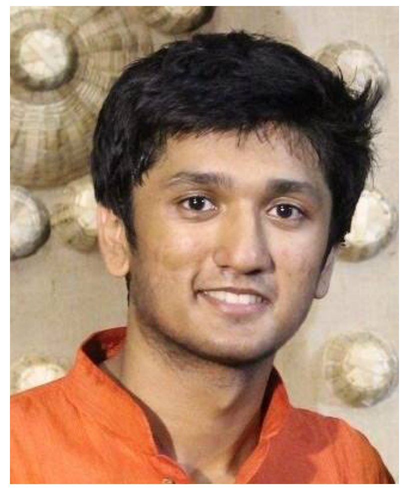
Similar to DETR[1], the encoder is a multi-layer transformer network that encodes image into a feature representation



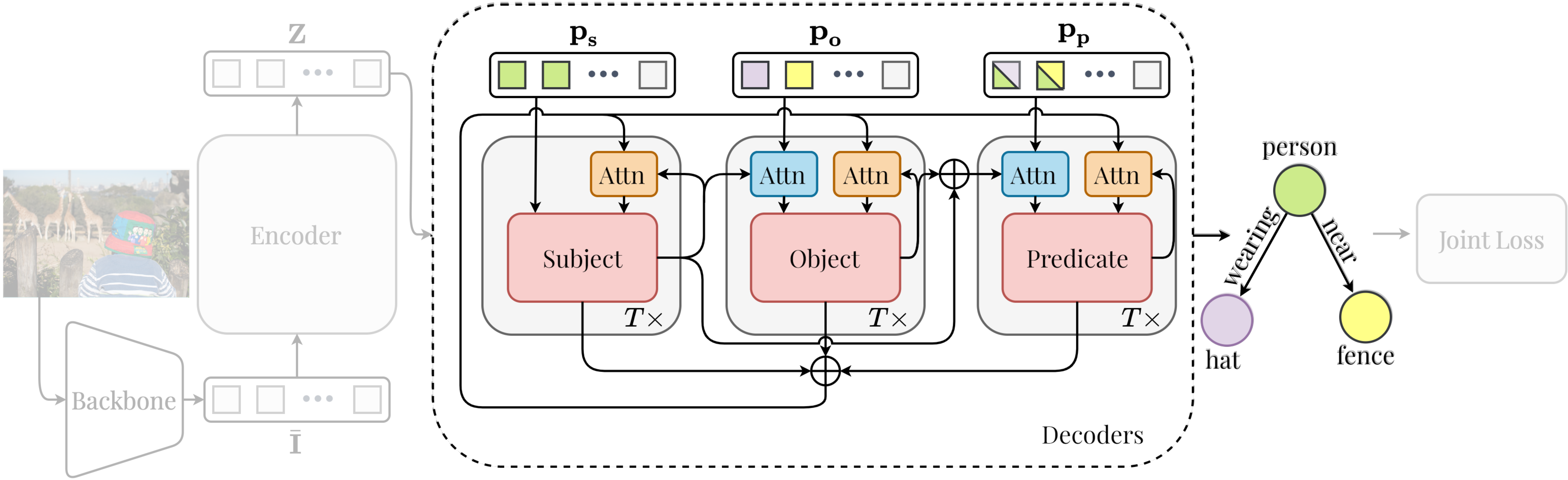
[1] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Springer, Cham, 2020.

Triplet **Decoder**

Siddhesh Khandelwal
(PhD, UBC)



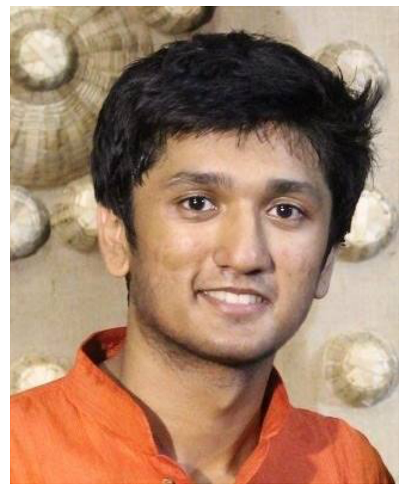
Each of the subject, object, and predicate predictors is a multi-layer transformer



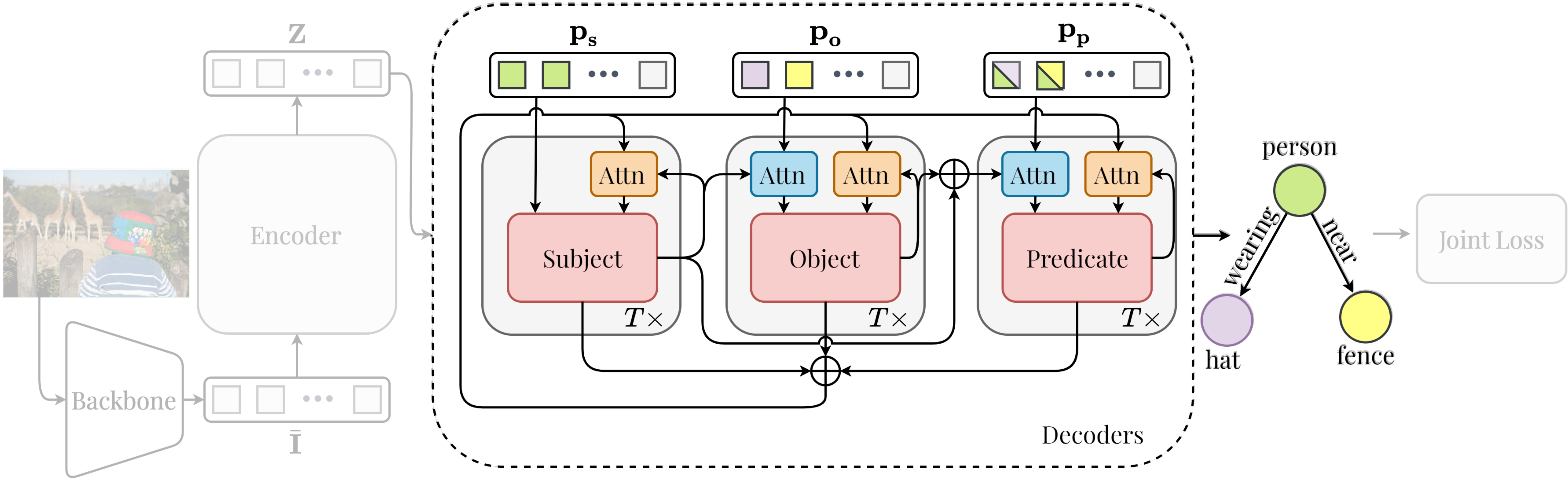
Iterative Scene Graph Generation, Khandelwal, **Sigal,** *(2022)*

Triplet **Decoder**

Siddhesh Khandelwal
(PhD, UBC)

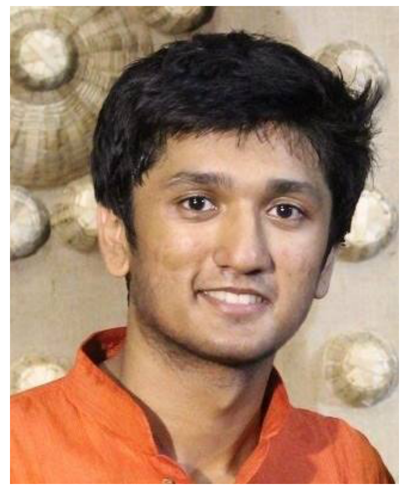


The iterative framework is modelled explicitly by using **two kinds of conditioning** and implicitly by a **joint loss**

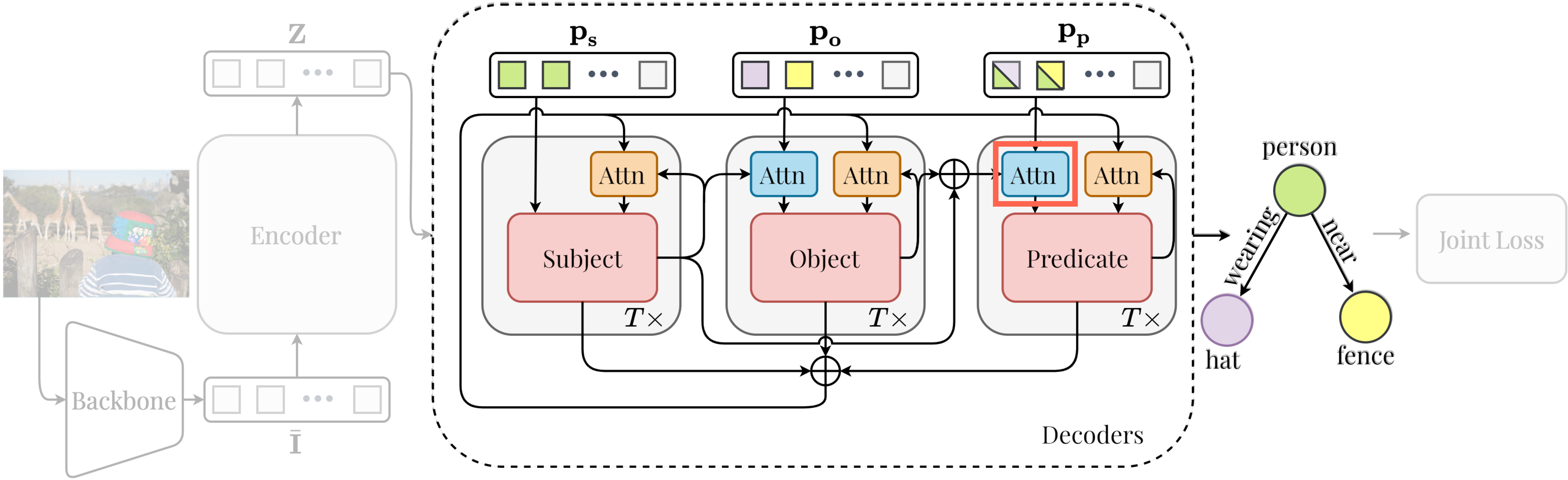


Conditioning **Within Step**

Siddhesh Khandelwal
(PhD, UBC)

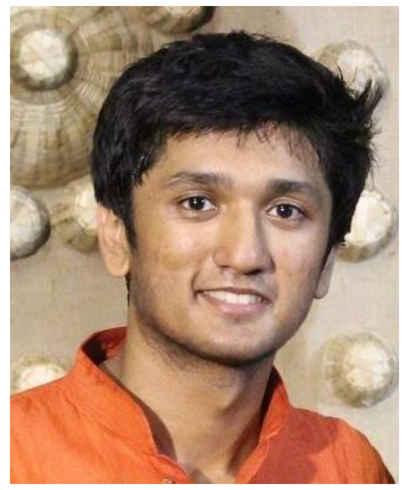


The predicate predictor within a particular step t is conditioned on the subject and object decoder outputs at step t

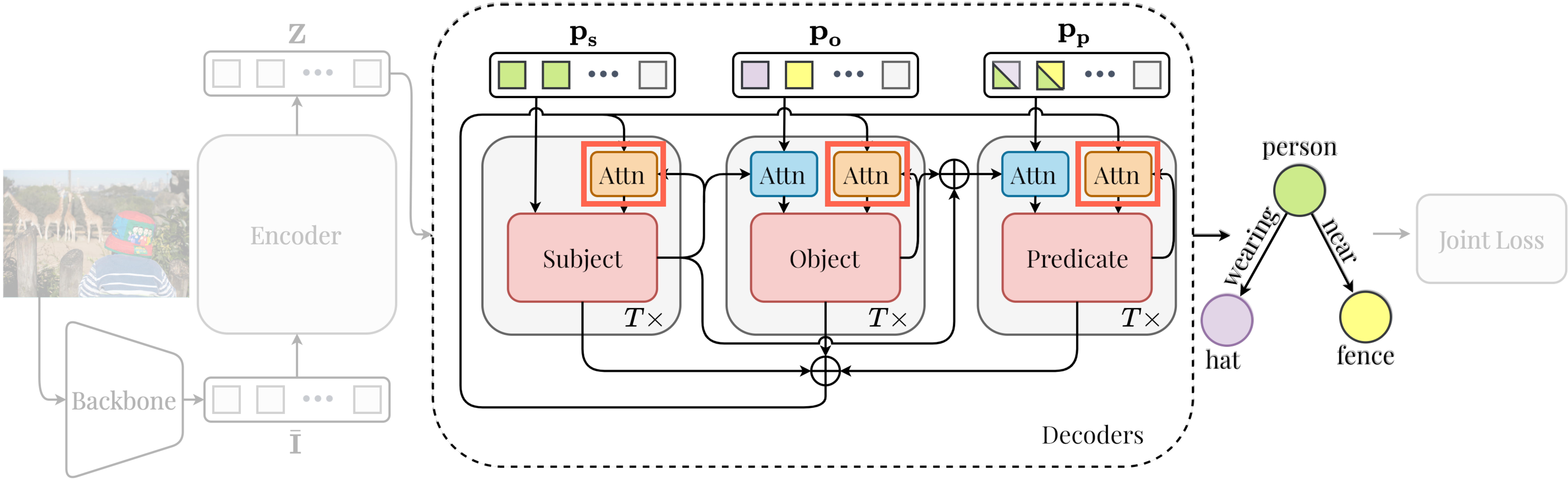


Conditioning **Across Steps**

Siddhesh Khandelwal
(PhD, UBC)

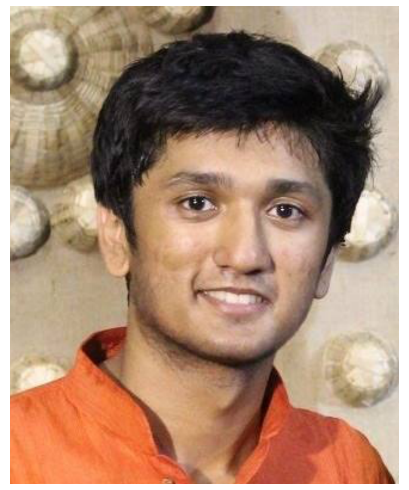


The predicate decoder within a particular step t is conditioned on the previous graph estimate from step $t - 1$

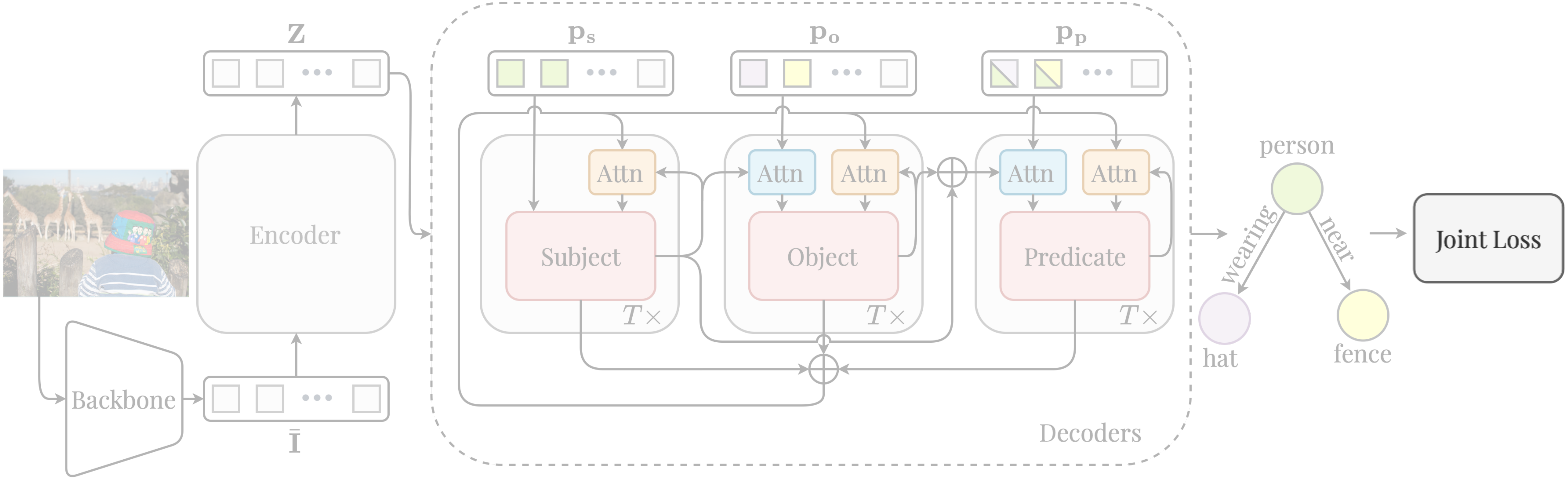


Joint **Loss**

Siddhesh Khandelwal
(PhD, UBC)

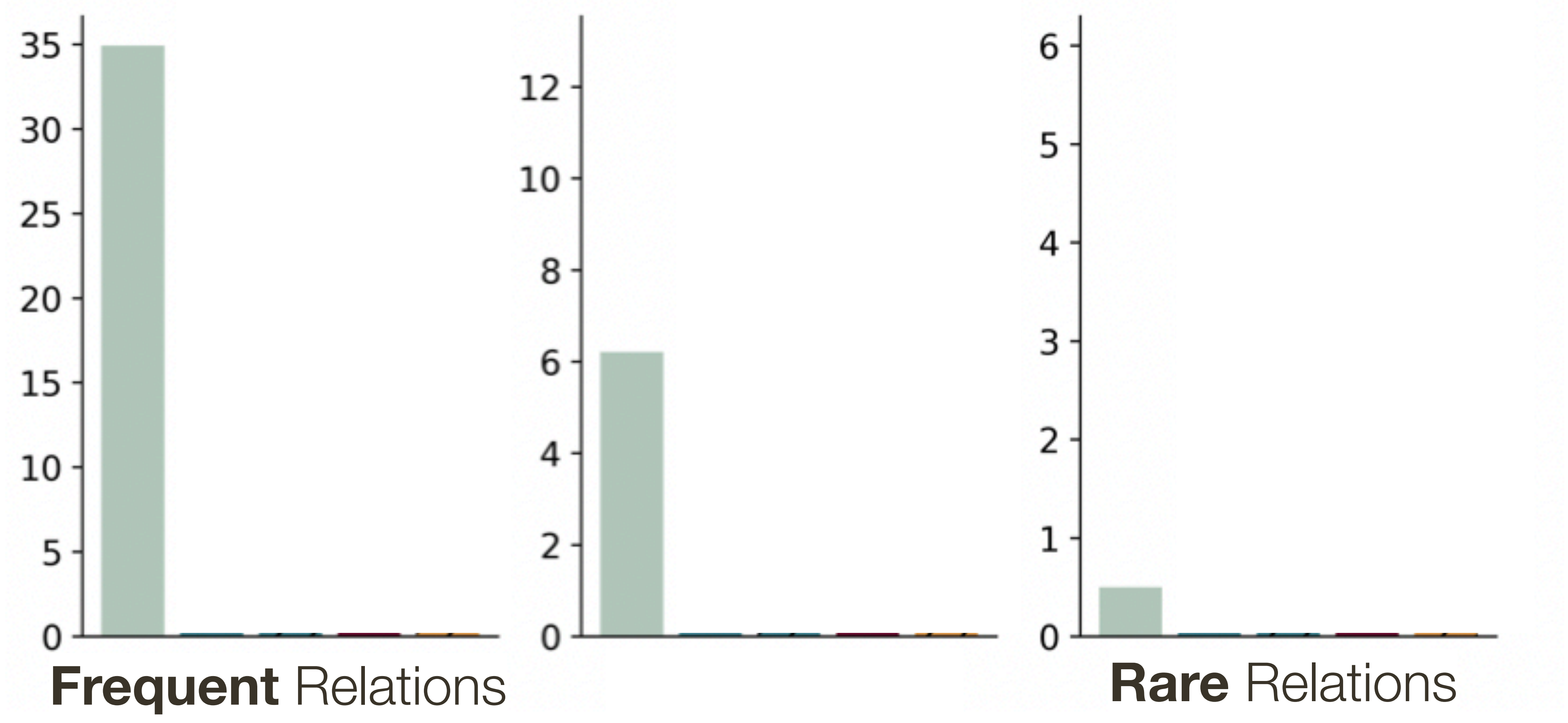


We additionally use a novel joint loss to ensure a valid scene graph is generated at each step. This loss **implicitly** enables refinement.



De-**Biasing**, **Strategy 1**:

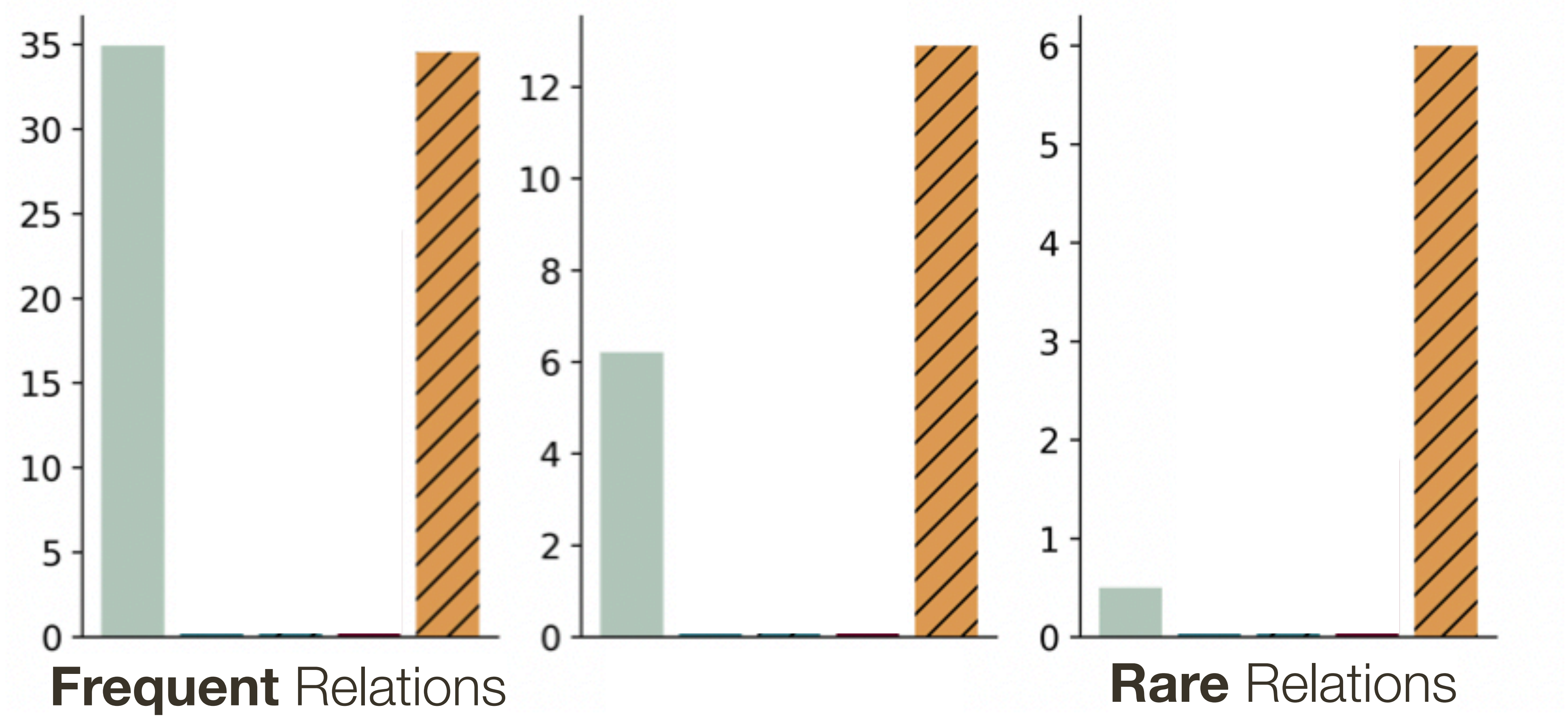
Data Re-sampling



[“Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation”, Li, Zhang, Wan, He, CVPR, 2021]

De-**Biasing**, **Strategy 1**:

Data Re-sampling



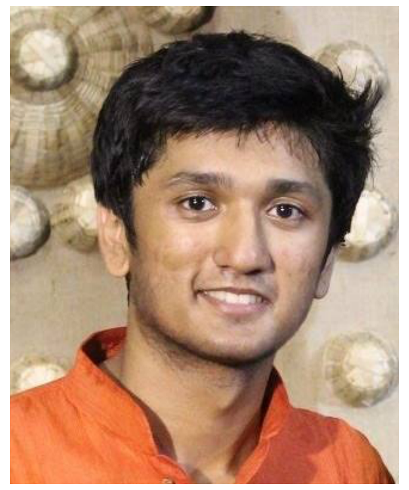
[“Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation”, Li, Zhang, Wan, He, CVPR, 2021]

De-**Biasing**, **Strategy 2**:

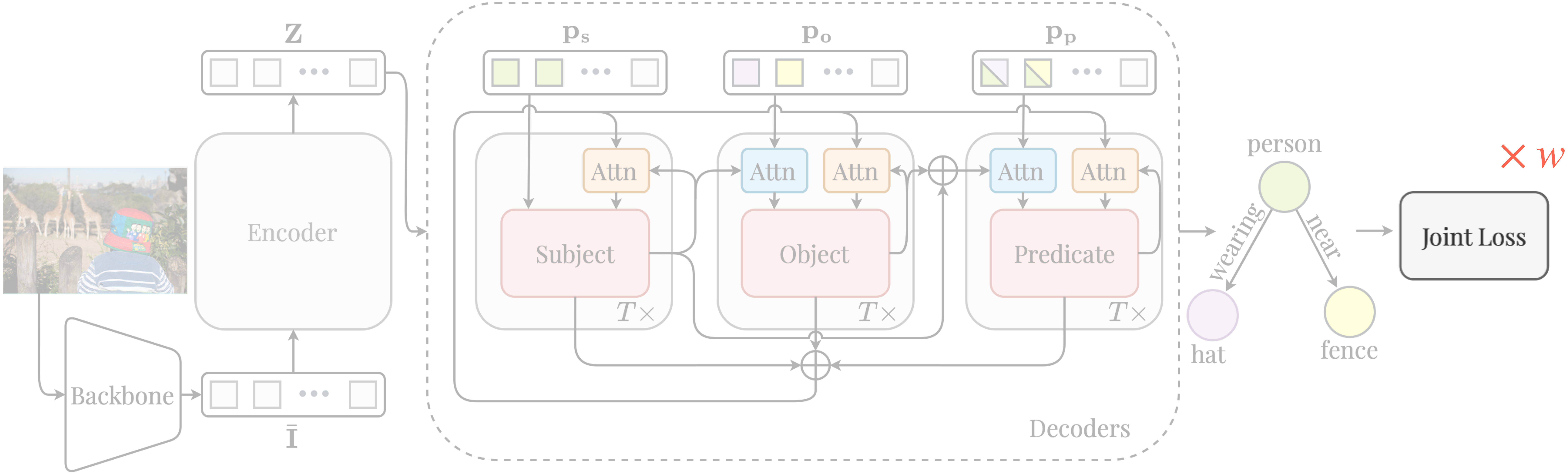
Loss Re-scaling

Loss **Re-weighting**

Siddhesh Khandelwal
(PhD, UBC)



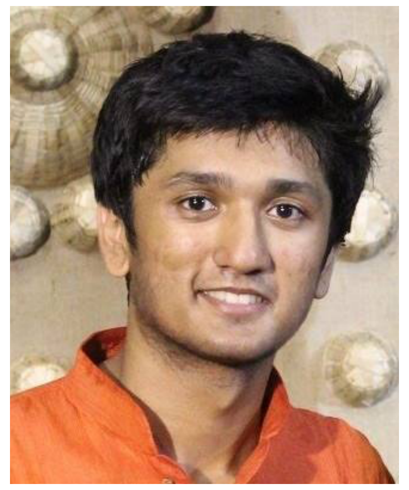
A loss re-weighting strategy is used to address the inherent long-tail nature of the task, giving our model flexibility to **trade-off** dominant for underrepresented classes



$$w_c = \max\left\{\left(\frac{\alpha}{\text{class frequency in training set}}\right)^{\beta}, 1.0\right\}$$

Experiments

Siddhesh Khandelwal
(PhD, UBC)



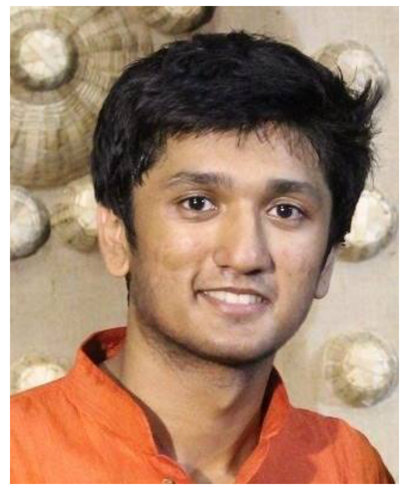
Our proposed transformer based approach outperforms existing baselines, while simultaneously operating on a wide spectrum of performance metrics.

Method	mR@50/100	R@50/100	hR@50/100	Head	Body	Tail
BGNN [30, 29]	8.6 / 10.3	28.2 / 33.8	13.2 / 15.8	29.1	12.6	2.2
RelDN [53, 29]	4.4 / 5.4	30.3 / 34.8	7.7 / 9.3	31.3	2.3	0.0
AS-Net [4]	6.1 / 7.2	18.7 / 21.1	9.2 / 10.7	19.6	7.7	2.7
HOTR [27]	9.4 / 12.0	23.5 / 27.7	13.4 / 16.7	26.1	16.2	3.4
Concurrent Work						
SGTR$_{M=1}$ [29]	12.0 / 14.6	25.1 / 26.6	16.2 / 18.8	27.1	17.2	6.9
SGTR$_{M=3}$ [29]	12.0 / 15.2	24.6 / 28.4	16.1 / 19.8	28.2	18.6	7.1
SGTR$_{M=3,\text{BGNN [30]}}$ [29]	15.8 / 20.1	20.6 / 25.0	17.9 / 22.3	21.7	21.6	17.1
Ours$_{(\alpha=0.0,\beta=*)}$	8.0 / 8.8	29.7 / 32.1	12.6 / 13.8	31.7	9.0	1.4
Ours$_{(\alpha=0.14,\beta=0.5)}$	14.4 / 16.4	27.9 / 30.4	19.0 / 21.3	30.0	17.3	11.2
Ours$_{(\alpha=0.07,\beta=0.75)}$	15.7 / 17.8	27.2 / 29.8	19.9 / 22.3	28.5	18.8	13.3
Ours$_{(\alpha=0.14,\beta=0.75)}$	15.8 / 18.2	26.1 / 28.7	19.7 / 22.3	28.2	19.4	13.8
Ours$_{(\alpha=0.14,\beta=0.75),\text{BGNN [30]}}$	17.1 / 19.2	22.9 / 25.7	19.6 / 22.0	24.4	20.2	16.4
Ours$_{(\alpha=0.14,\beta=0.75),M=3}$	19.5 / 23.4	30.8 / 35.6	23.9 / 28.2	32.9	28.1	15.8

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

Siddhesh Khandelwal
(PhD, UBC)

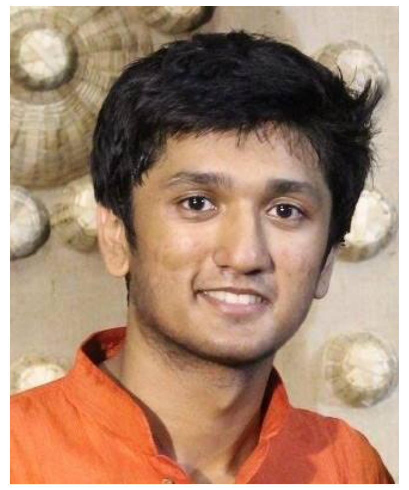


Our proposed transformer based approach outperforms existing baselines, while simultaneously operating on a wide spectrum of performance metrics.

Method	mR@50/100	R@50/100	hR@50/100	Head	Body	Tail
BGNN [30, 29]	8.6 / 10.3	28.2 / 33.8	13.2 / 15.8	29.1	12.6	2.2
RelDN [53, 29]	4.4 / 5.4	30.3 / 34.8	7.7 / 9.3	31.3	2.3	0.0
AS-Net [4]	6.1 / 7.2	18.7 / 21.1	9.2 / 10.7	19.6	7.7	2.7
HOTR [27]	9.4 / 12.0	23.5 / 27.7	13.4 / 16.7	26.1	16.2	3.4
Concurrent Work						
SGTR$_{M=1}$ [29]	12.0 / 14.6	25.1 / 26.6	16.2 / 18.8	27.1	17.2	6.9
SGTR$_{M=3}$ [29]	12.0 / 15.2	24.6 / 28.4	16.1 / 19.8	28.2	18.6	7.1
SGTR$_{M=3,\text{BGNN [30]}}$ [29]	15.8 / 20.1	20.6 / 25.0	17.9 / 22.3	21.7	21.6	17.1
Ours$_{(\alpha=0.0,\beta=*)}$	8.0 / 8.8	29.7 / 32.1	12.6 / 13.8	31.7	9.0	1.4
Ours$_{(\alpha=0.14,\beta=0.5)}$	14.4 / 16.4	27.9 / 30.4	19.0 / 21.3	30.0	17.3	11.2
Ours$_{(\alpha=0.07,\beta=0.75)}$	15.7 / 17.8	27.2 / 29.8	19.9 / 22.3	28.5	18.8	13.3
Ours$_{(\alpha=0.14,\beta=0.75)}$	15.8 / 18.2	26.1 / 28.7	19.7 / 22.3	28.2	19.4	13.8
Ours$_{(\alpha=0.14,\beta=0.75),\text{BGNN [30]}}$	17.1 / 19.2	22.9 / 25.7	19.6 / 22.0	24.4	20.2	16.4
Ours$_{(\alpha=0.14,\beta=0.75),M=3}$	19.5 / 23.4	30.8 / 35.6	23.9 / 28.2	32.9	28.1	15.8

Experiments

Siddhesh Khandelwal
(PhD, UBC)

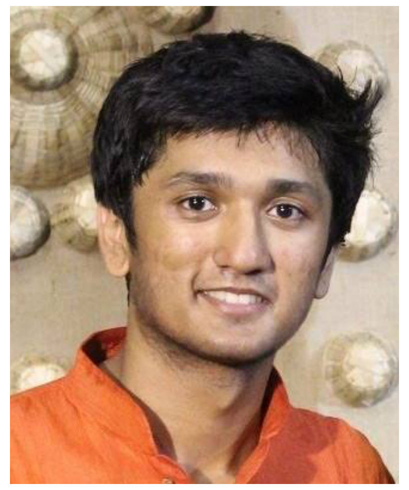


Our proposed transformer based approach outperforms existing baselines, while simultaneously operating on a wide spectrum of performance metrics.

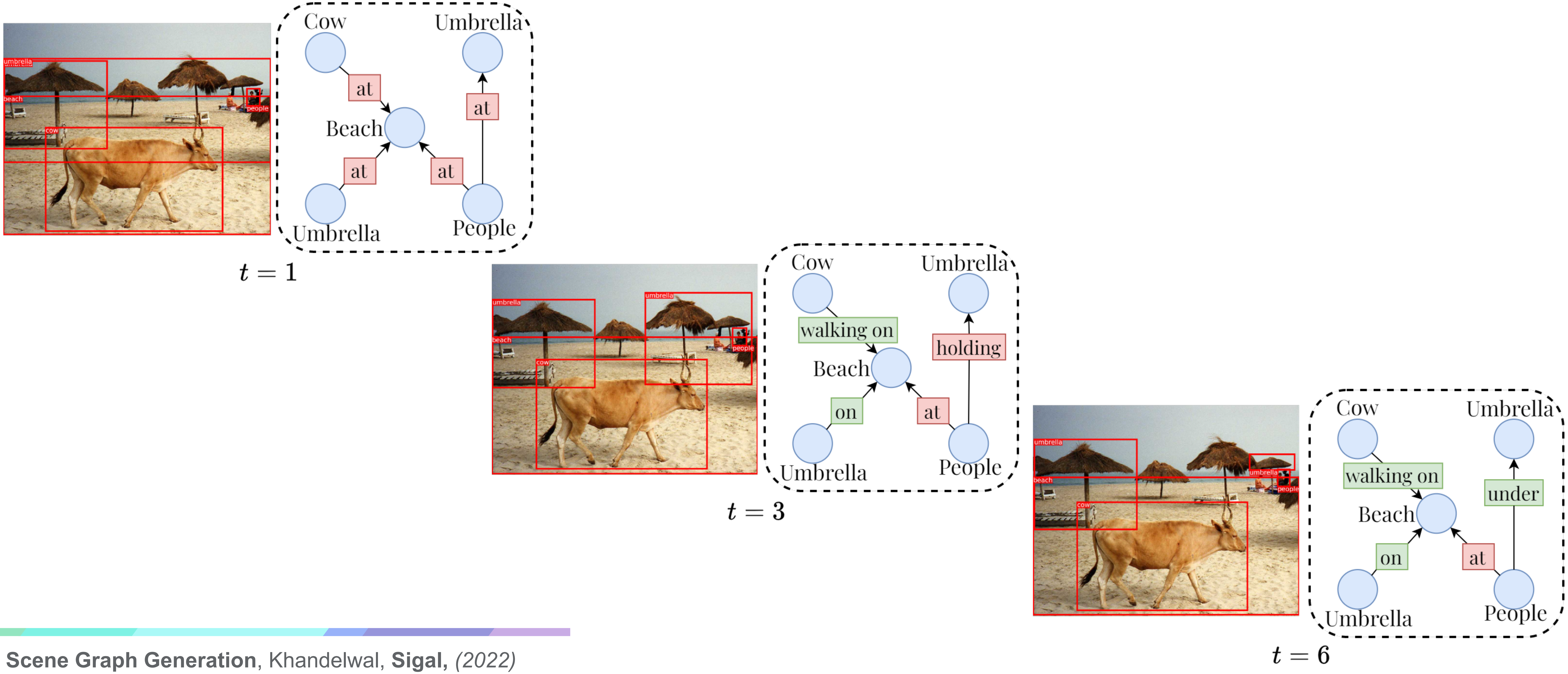
Method	mR@50/100	R@50/100	hR@50/100	Head	Body	Tail
BGNN [30, 29]	8.6 / 10.3	28.2 / 33.8	13.2 / 15.8	29.1	12.6	2.2
RelDN [53, 29]	4.4 / 5.4	30.3 / 34.8	7.7 / 9.3	31.3	2.3	0.0
AS-Net [4]	6.1 / 7.2	18.7 / 21.1	9.2 / 10.7	19.6	7.7	2.7
HOTR [27]	9.4 / 12.0	23.5 / 27.7	13.4 / 16.7	26.1	16.2	3.4
Concurrent Work						
$\text{SGTR}_{M=1}$ [29]	12.0 / 14.6	25.1 / 26.6	16.2 / 18.8	27.1	17.2	6.9
$\text{SGTR}_{M=3}$ [29]	12.0 / 15.2	24.6 / 28.4	16.1 / 19.8	28.2	18.6	7.1
$\text{SGTR}_{M=3,\text{BGNN [30]}}$ [29]	15.8 / 20.1	20.6 / 25.0	17.9 / 22.3	21.7	21.6	17.1
$\text{Ours}_{(\alpha=0.0,\beta=*)}$	8.0 / 8.8	29.7 / 32.1	12.6 / 13.8	31.7	9.0	1.4
$\text{Ours}_{(\alpha=0.14,\beta=0.5)}$	14.4 / 16.4	27.9 / 30.4	19.0 / 21.3	30.0	17.3	11.2
$\text{Ours}_{(\alpha=0.07,\beta=0.75)}$	15.7 / 17.8	27.2 / 29.8	19.9 / 22.3	28.5	18.8	13.3
$\text{Ours}_{(\alpha=0.14,\beta=0.75)}$	15.8 / 18.2	26.1 / 28.7	19.7 / 22.3	28.2	19.4	13.8
$\text{Ours}_{(\alpha=0.14,\beta=0.75),\text{BGNN [30]}}$	17.1 / 19.2	22.9 / 25.7	19.6 / 22.0	24.4	20.2	16.4
$\text{Ours}_{(\alpha=0.14,\beta=0.75),M=3}$	19.5 / 23.4	30.8 / 35.6	23.9 / 28.2	32.9	28.1	15.8

Visualization

Siddhesh Khandelwal
(PhD, UBC)

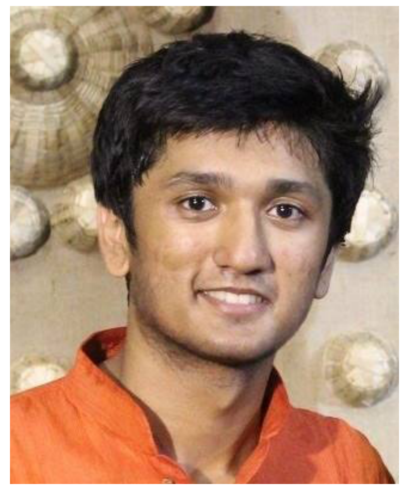


The graph quality **improves** over multiple refinement steps

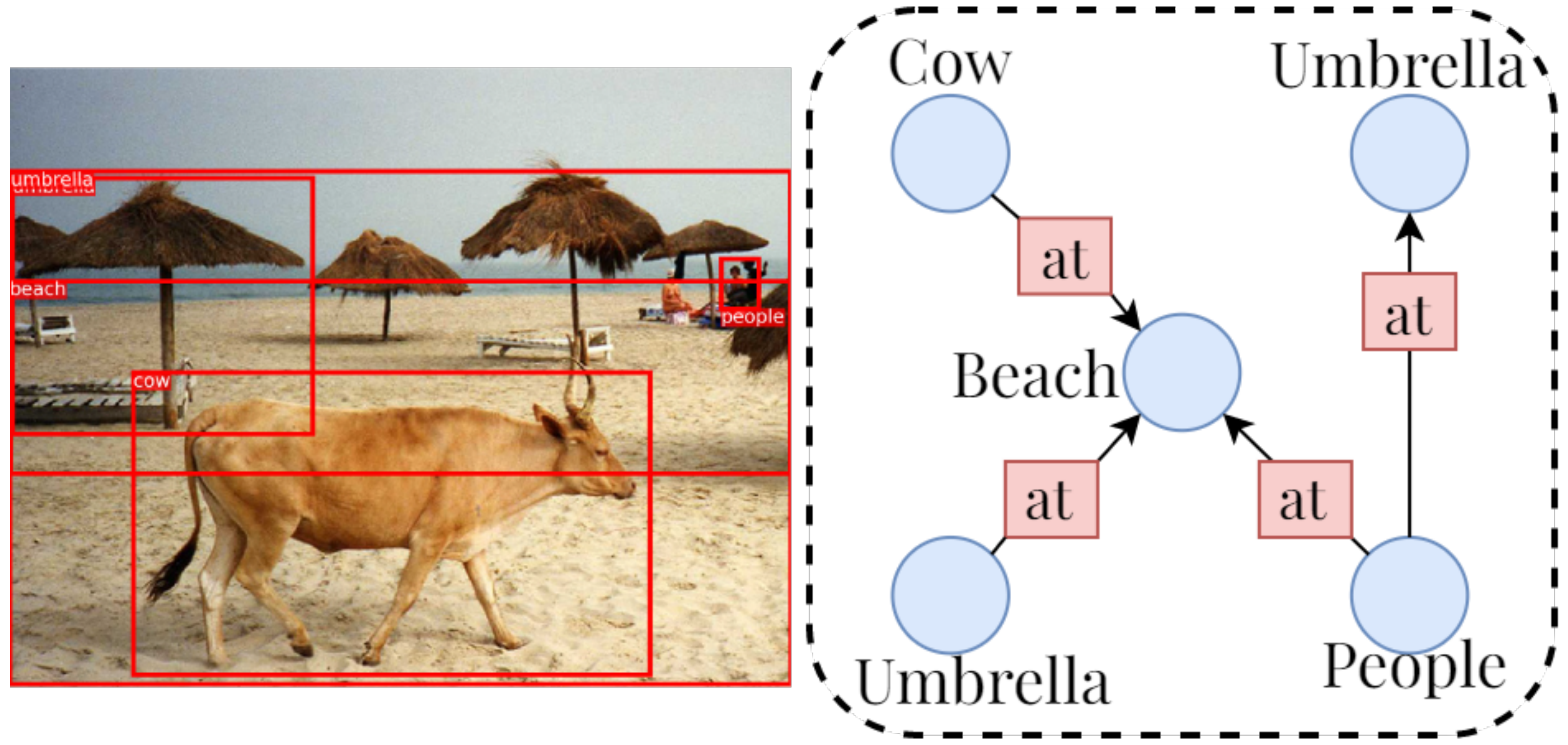


Visualization

Siddhesh Khandelwal
(PhD, UBC)

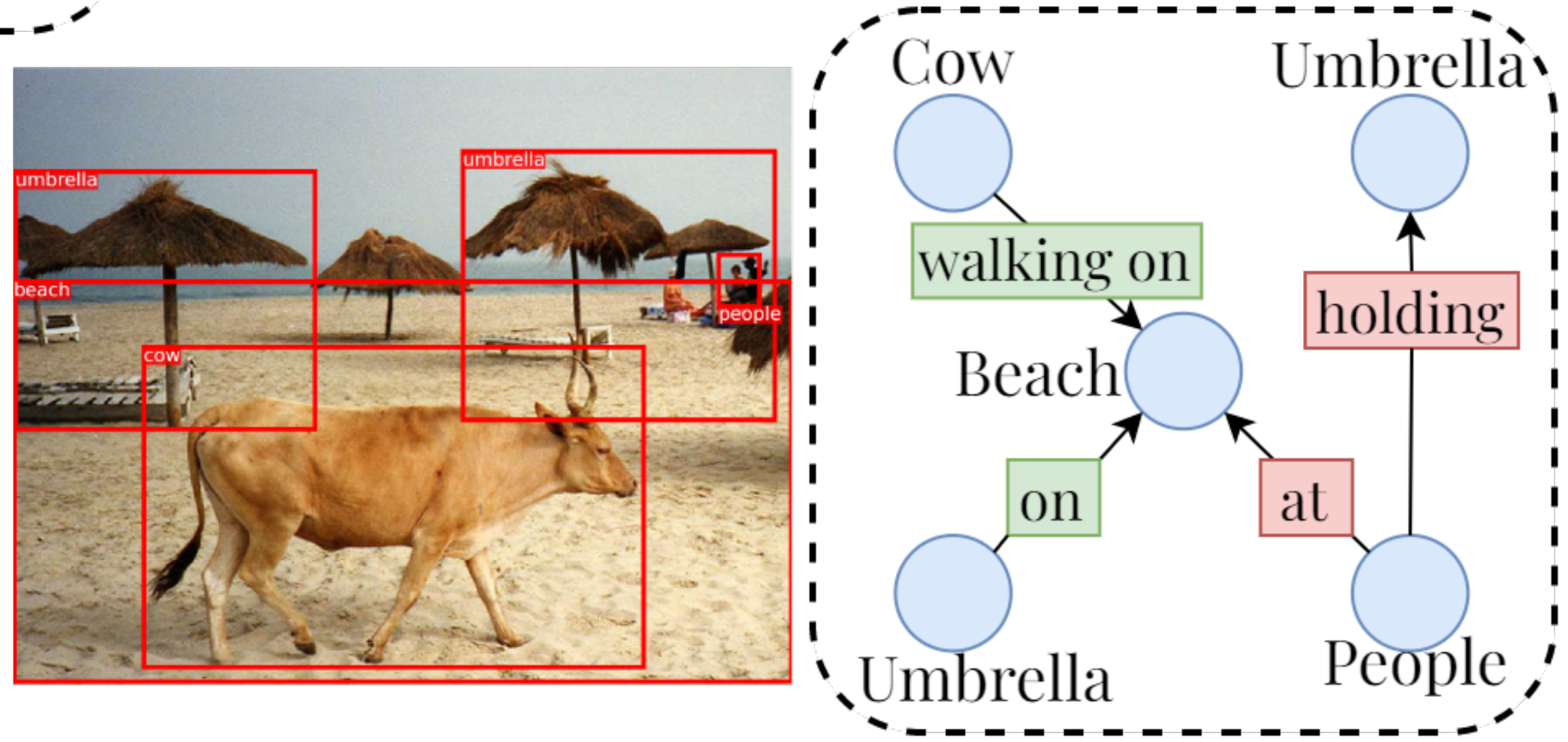


The graph quality **improves** over multiple refinement steps

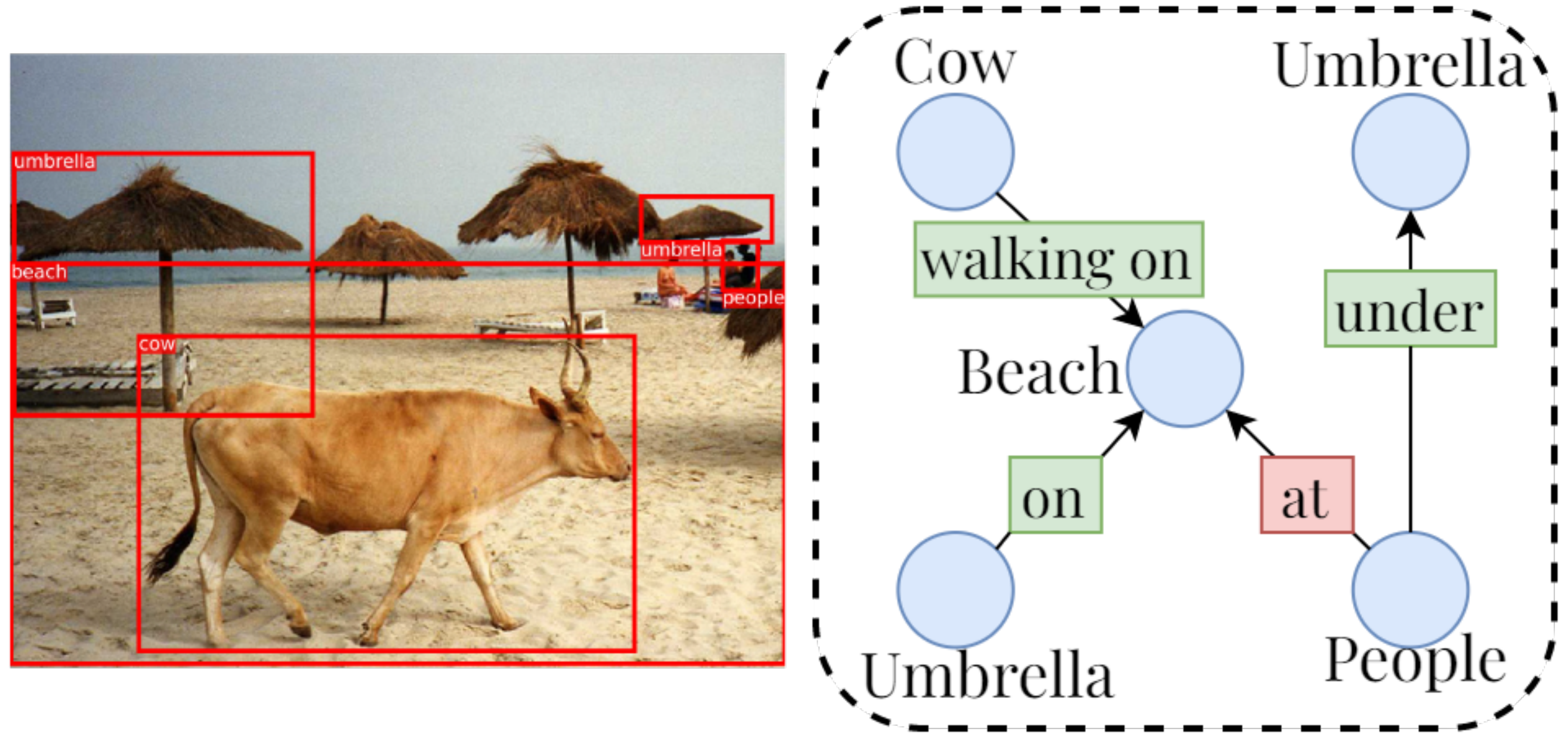


$t = 1$

This also means we can trade off **quality** for **computation**



$t = 3$

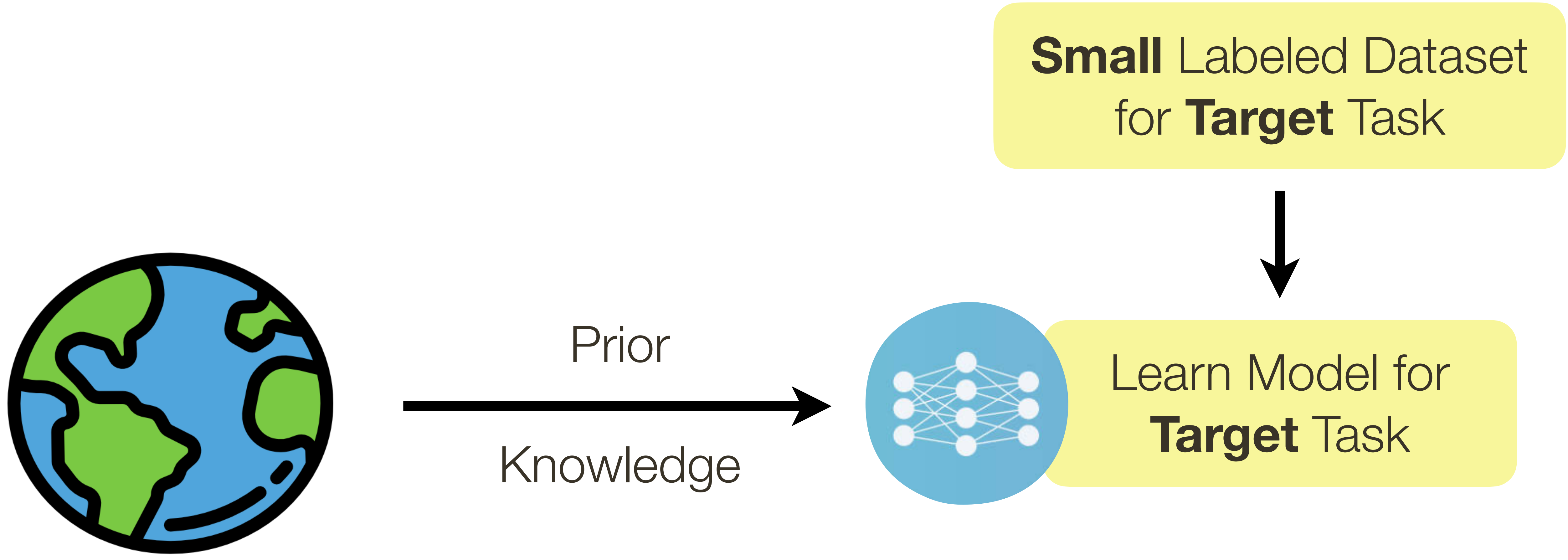


$t = 6$

Data Efficiency, Strategy 4:

Adding Prior Knowledge

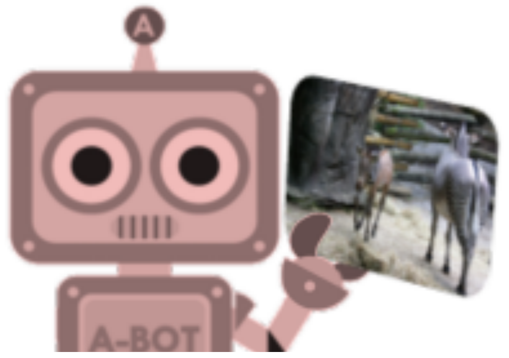
(Case-study in Common Sense)



Knowledge-based **Visual Question Answering**

Aditya Chinchure
(MSc, UBC)

Question:
This animal is known for many acute senses including what?



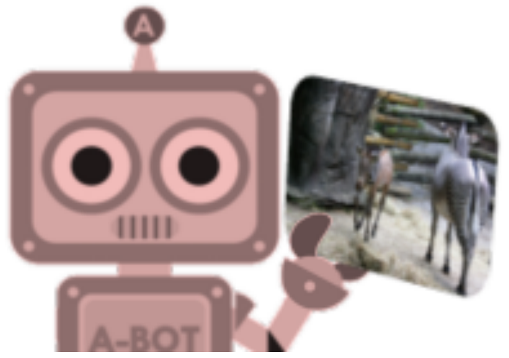
hearing and smell

VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge, Ravi*, Chinchure*, **Sigal**, Liao and Shwartz, *(2022)*

Knowledge-based **Visual Question Answering**

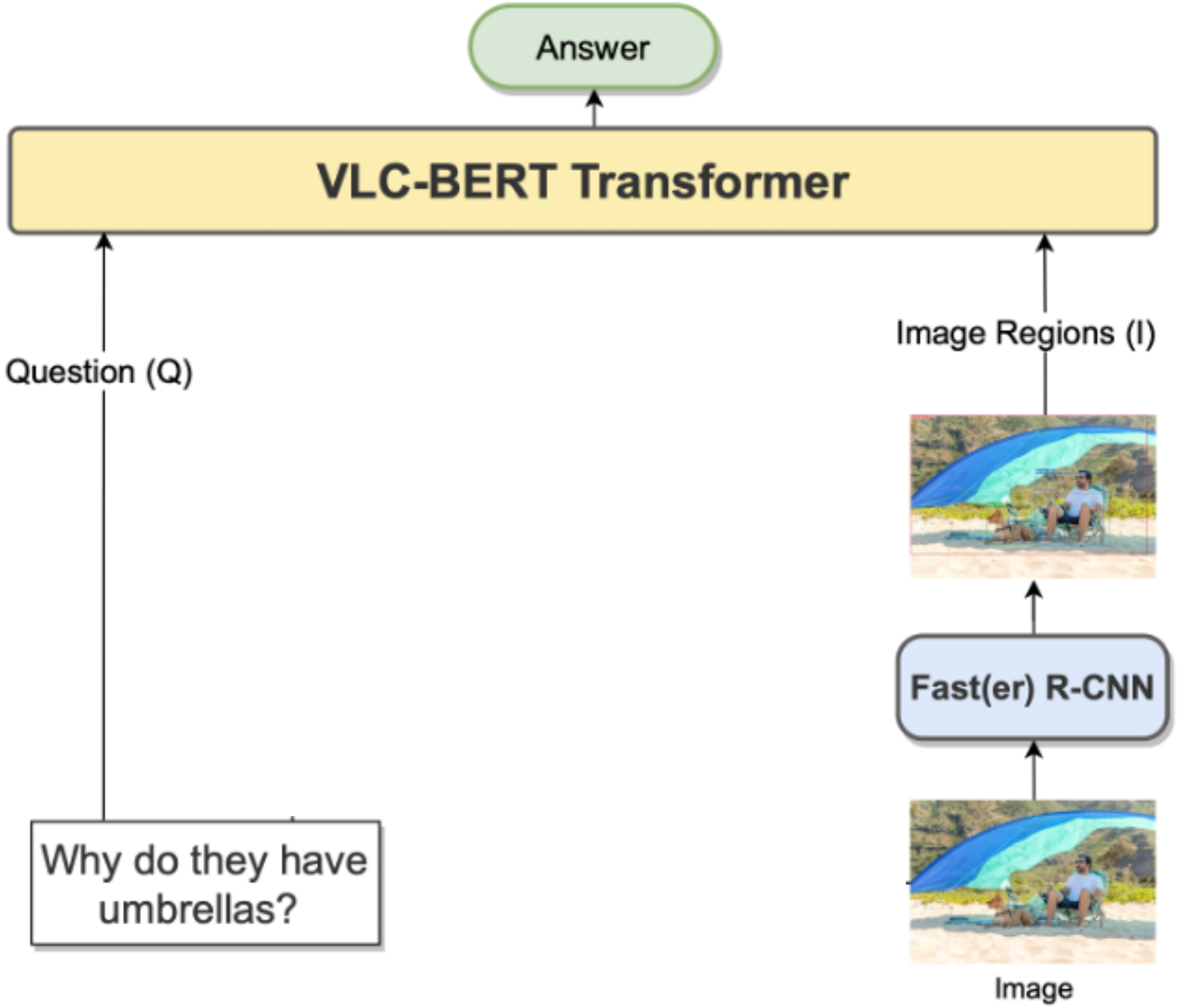
Aditya Chinchure
(MSc, UBC)

Question:
This animal is known for many acute senses including what?

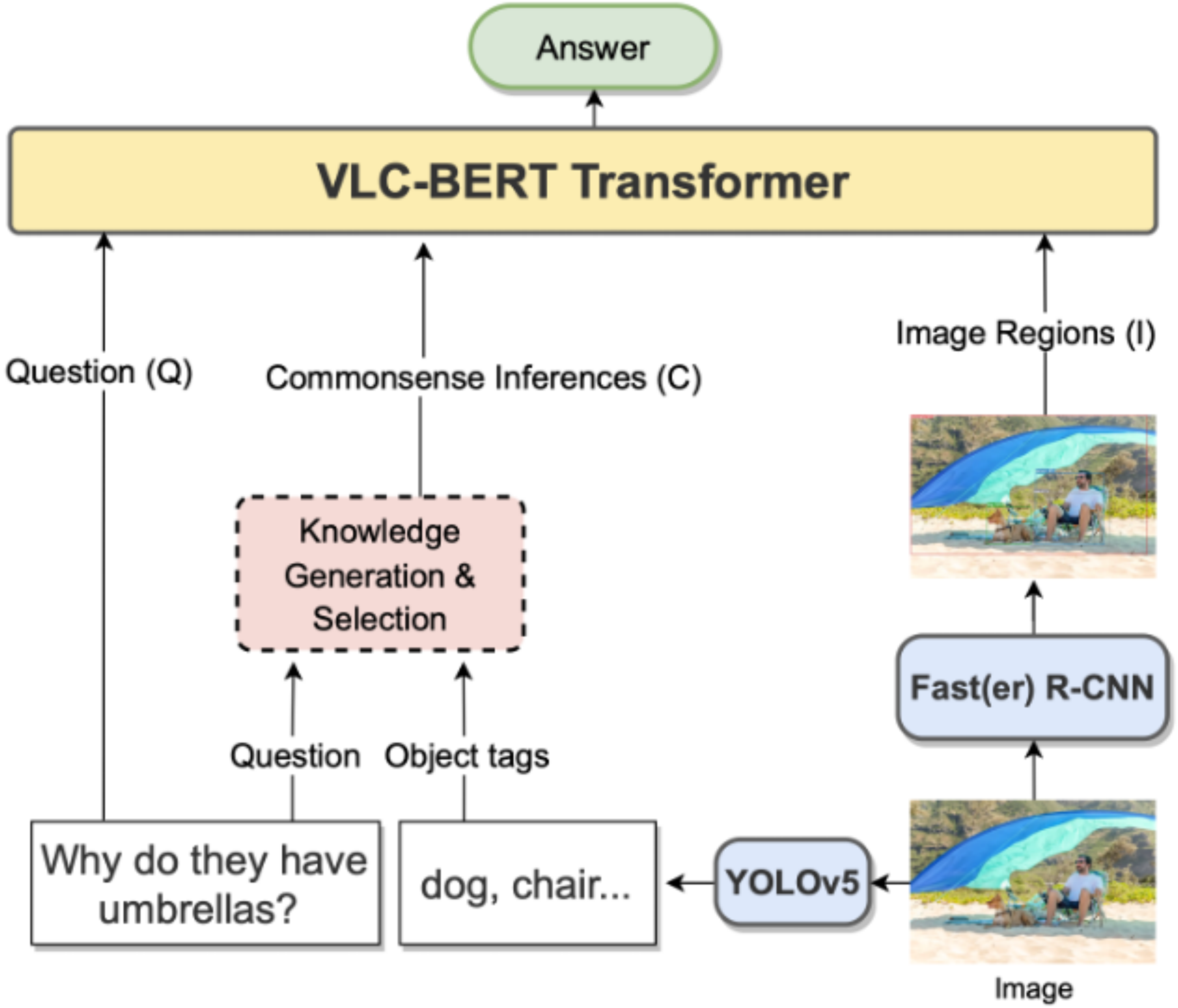


Requires **visual knowledge** that the cat is present, but also **common sense** semantic knowledge about cats as specie

Knowledge-based **Visual Question Answering**

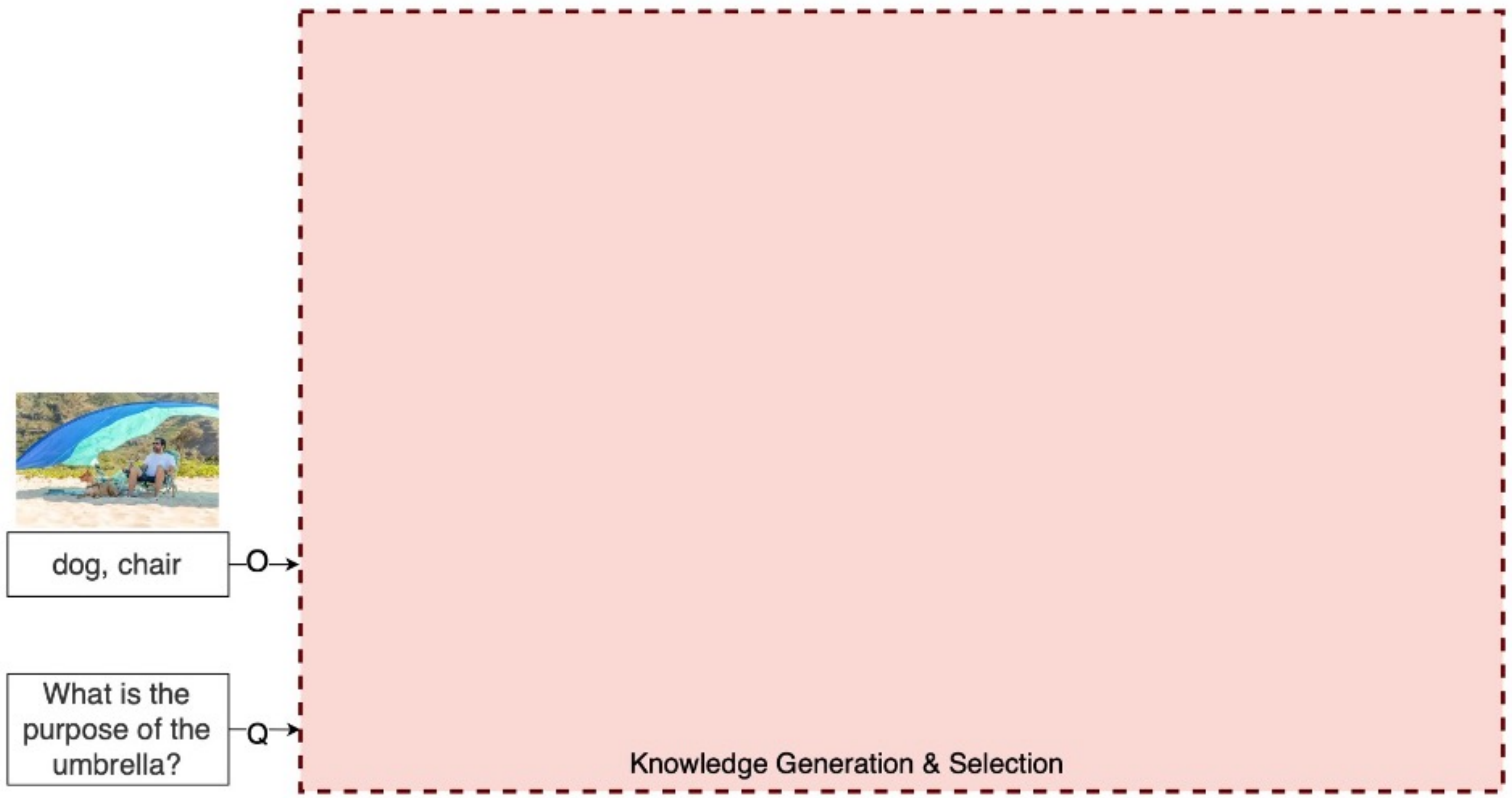


Knowledge-based **Visual Question Answering**



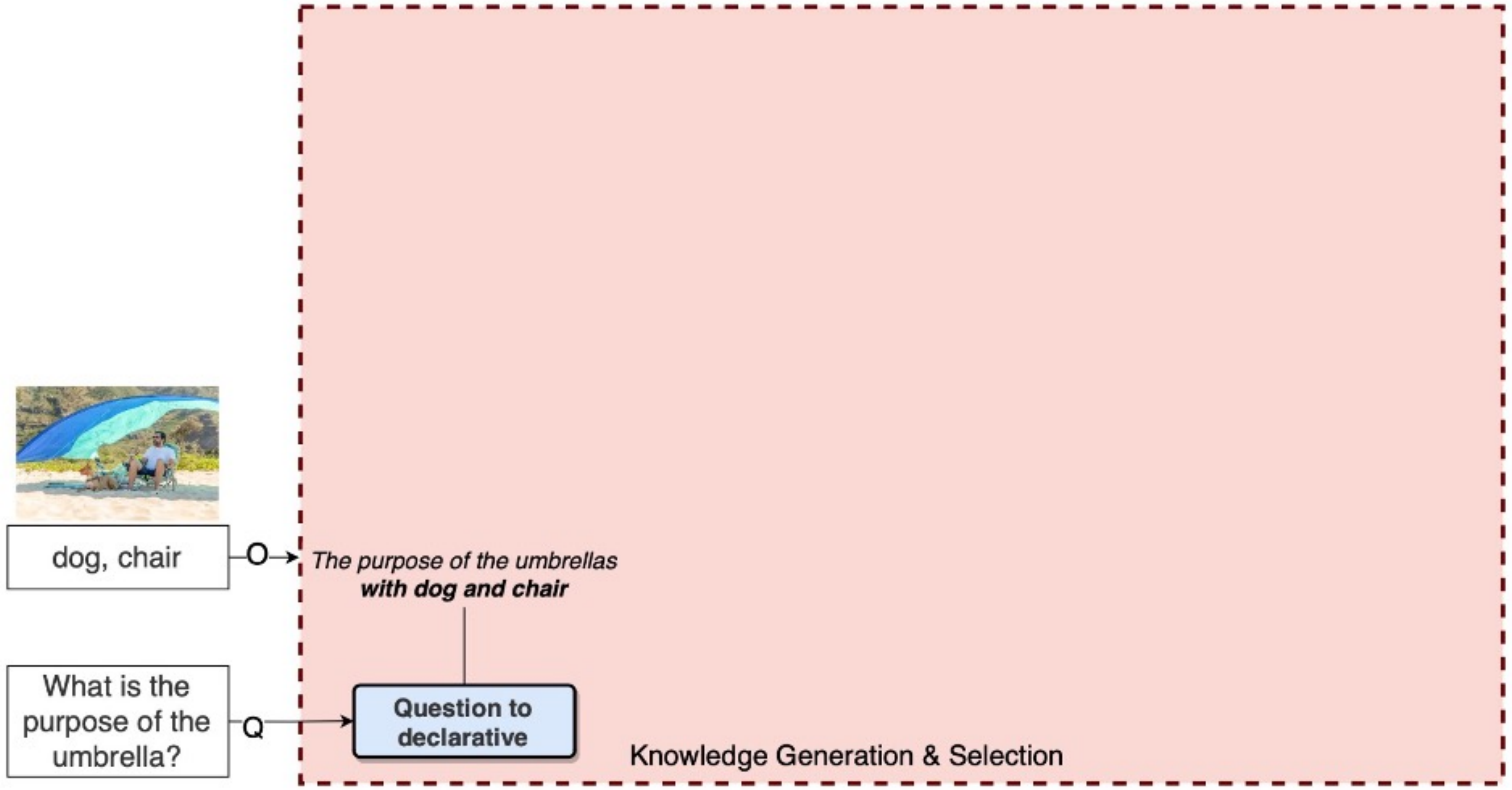
VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge, Ravi*, Chinchure*, **Sigal**, Liao and Shwartz, *(2022)*

Knowledge **Generation** & **Selection**



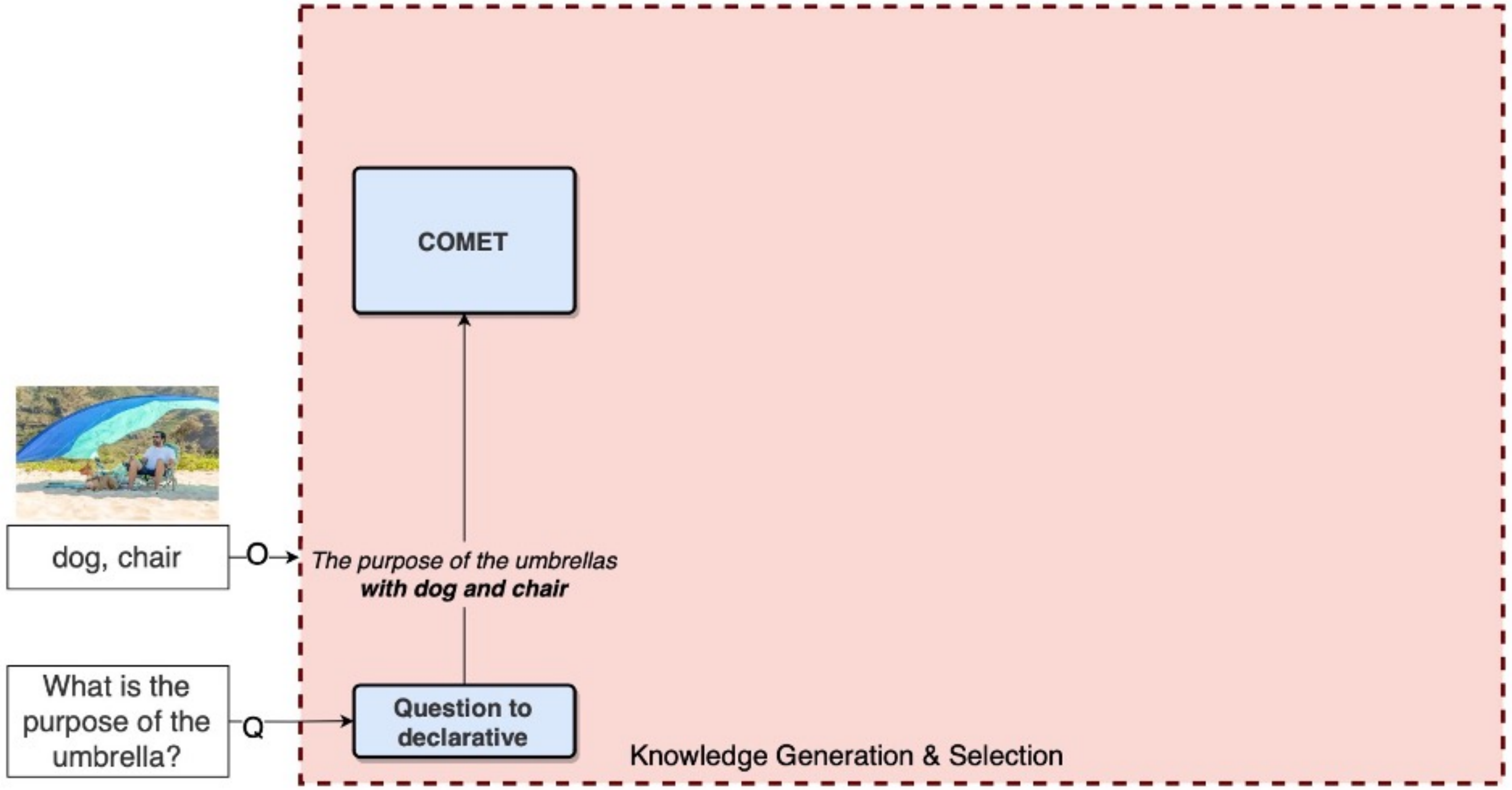
Knowledge **Generation** & **Selection**

Convert question into declarative statement and concatenate detected objects



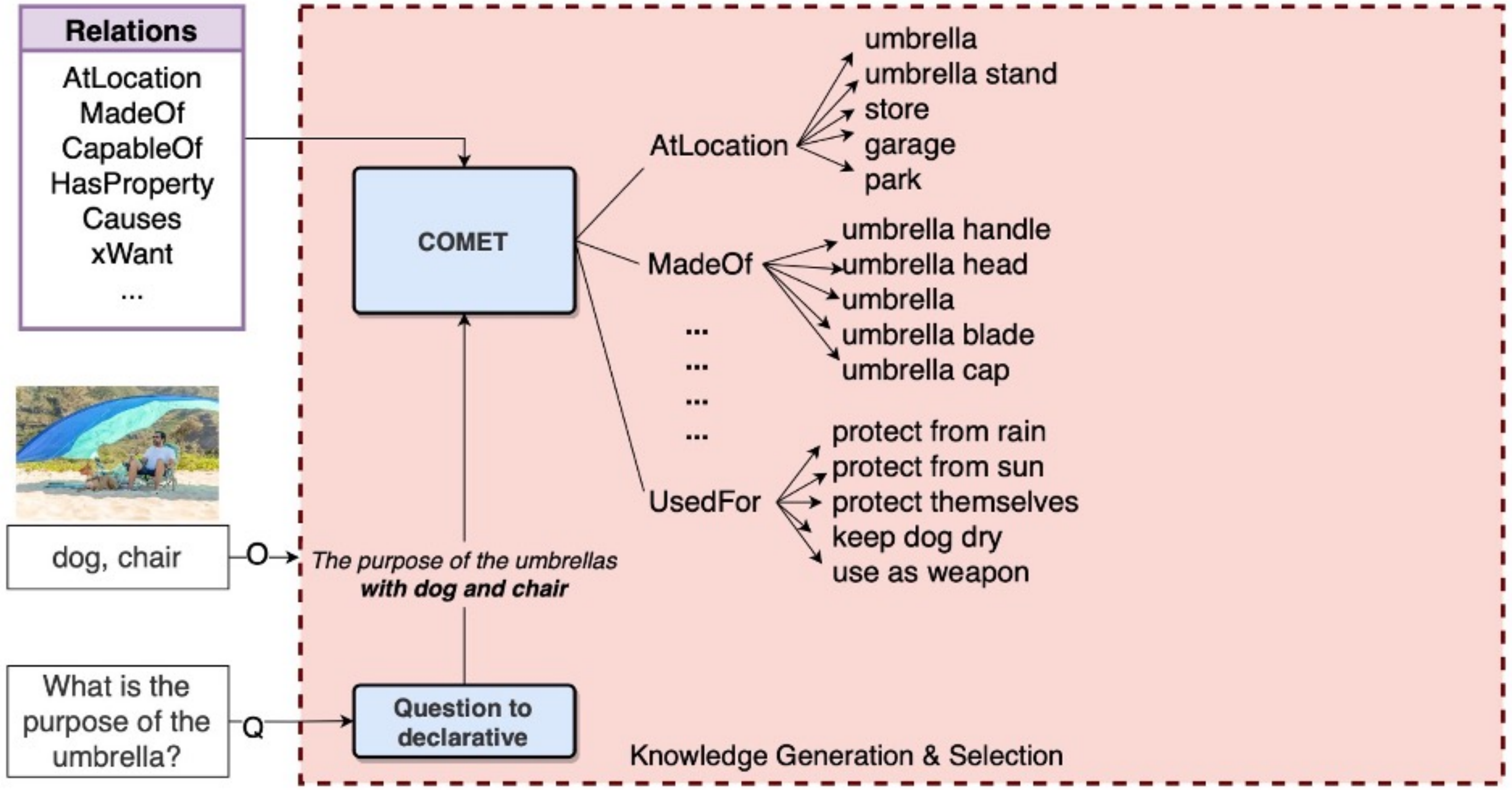
Knowledge **Generation** & **Selection**

Query a neural knowledge-based model to extract common sense inferences



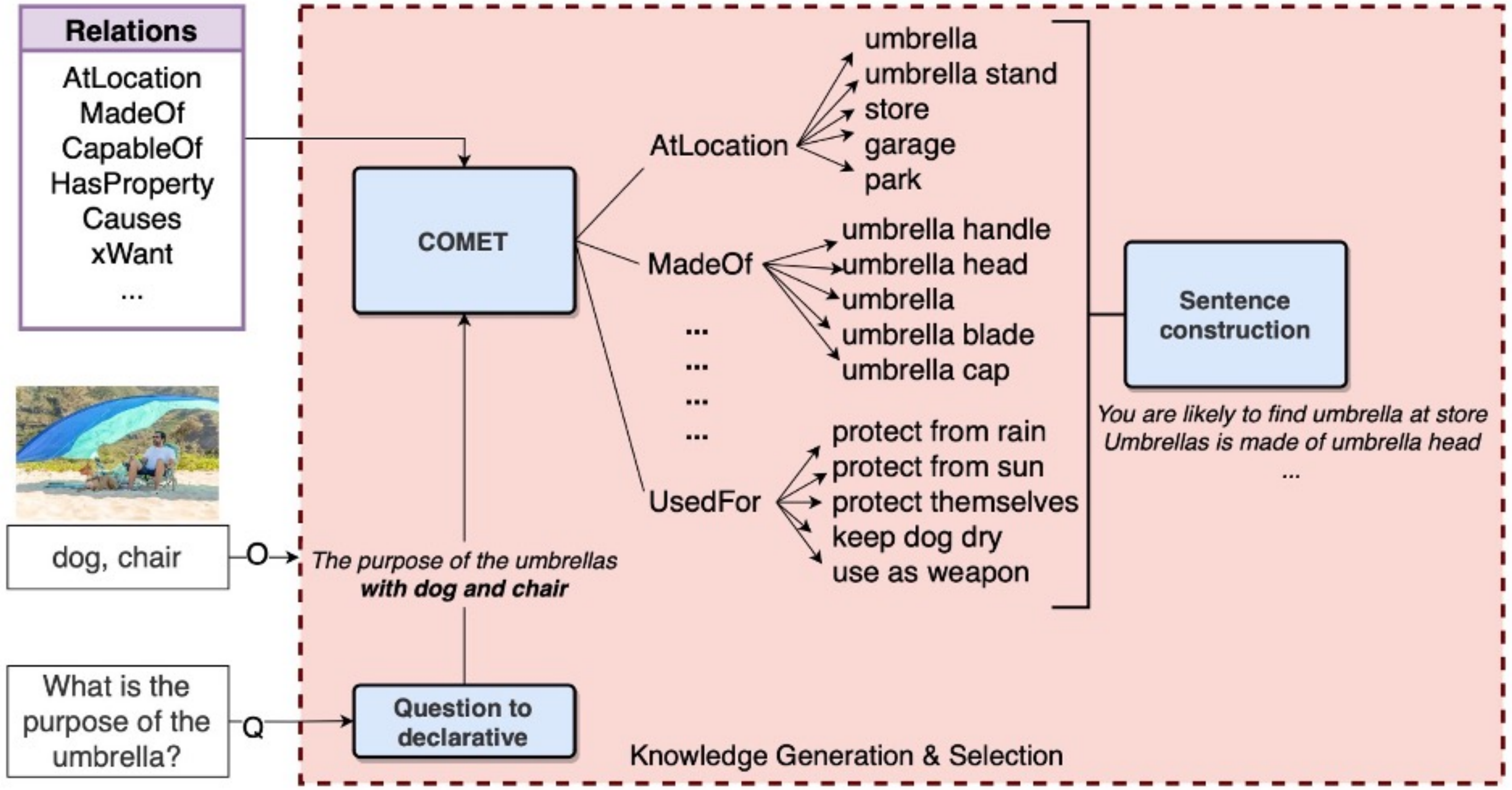
Knowledge **Generation** & **Selection**

Query a neural knowledge-based model to extract common sense inferences



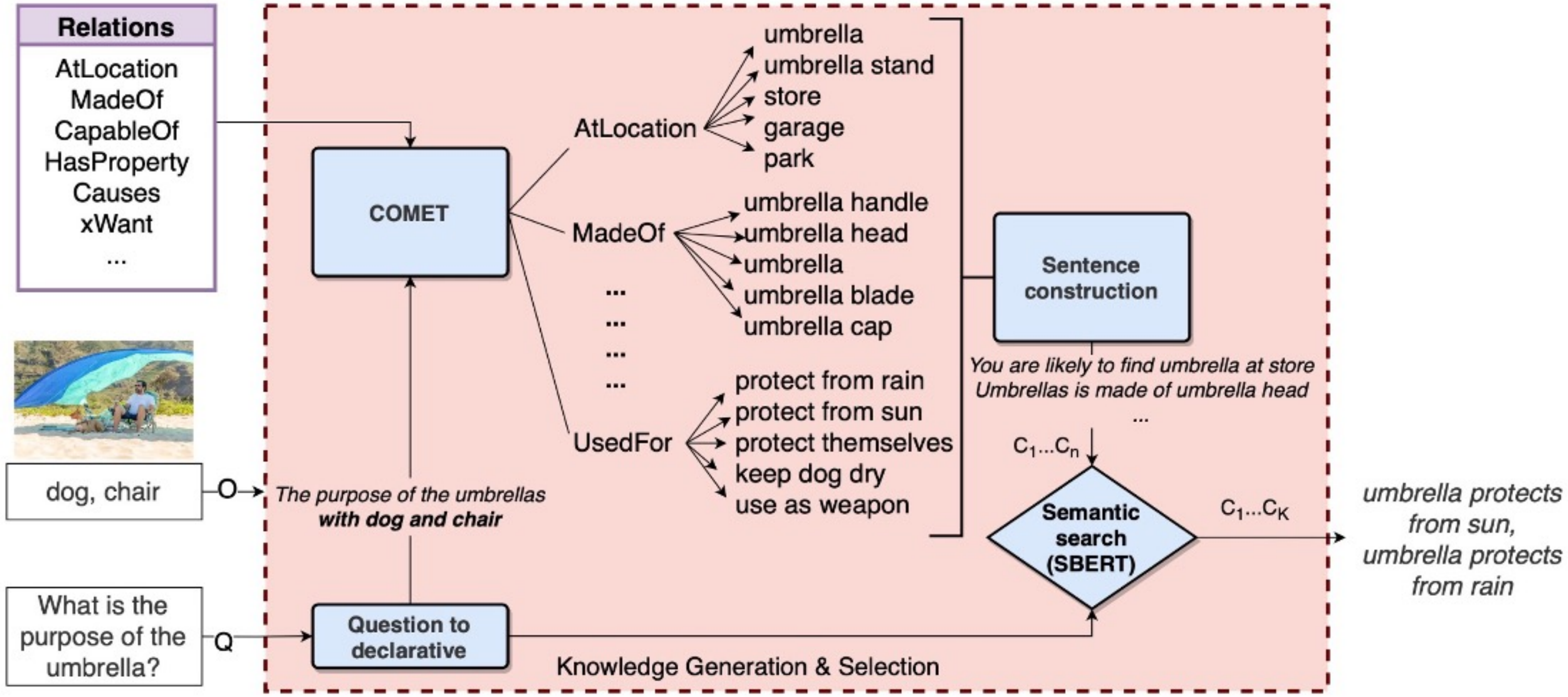
Knowledge **Generation** & **Selection**

Convert inferences into sentences using lingual templates

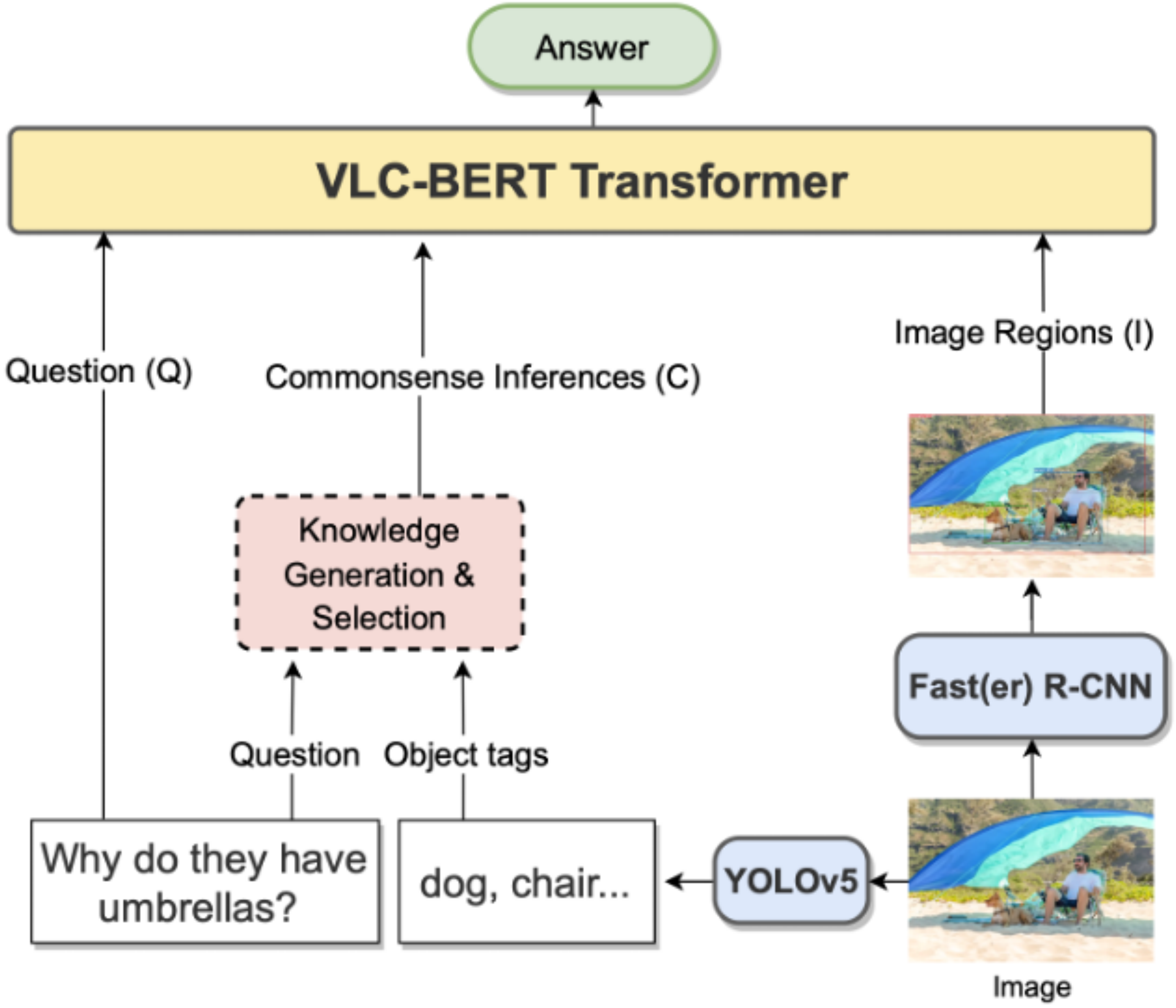


Knowledge **Generation** & **Selection**

Select inferences that are most relevant for the current question



Knowledge-based **Visual Question Answering**



VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge, Ravi*, Chinchure*, **Sigal**, Liao and Shwartz, *(2022)*

Results

Method	Knowledge Sources	OK-VQA	A-OKVQA	Approx. Params
ViLBERT [36]	-	-	25.85	116M
LXMERT [36]	-	-	25.89	-
BAN + AN [29]	Wikipedia	25.61	-	-
BAN + KG-AUG [20]	Wikipedia + ConceptNet	26.71	-	-
MUTAN + AN [29]	Wikipedia	27.84	-	-
ConceptBert [9]	ConceptNet	33.66	-	118M
KRISP [28]	Wikipedia + ConceptNet	32.31	27.1	116M
KRISP [28]	Wikipedia + ConceptNet + VQA P.T.	38.9	-	116M
Visual Retriever-Reader [26]	Google Search	39.2	-	-
MAVEx [47]	Wikipedia + ConceptNet + Google Images	41.37	-	-
GPV2 [18,36]	Web Search (Web10k) + COCO P.T.	-	40.7	220M
PICa-Base [48]	GPT-3	43.3	-	175B
PICa-Full [48]	GPT-3	48.0	-	175B
KAT [14]	Wikidata + GPT-3	54.41	-	175B
VLC-BERT (Ours)	VQA P.T. + COMET	43.14	38.05	118M

Question Answering

Q: What are people doing?
Q: What time of the year is it?
Q: Are the people married?

VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge, Ravi*, Chinchure*, **Sigal**, Liao and Shwartz, *(2022)*

Qualitative Results

Q: What is the object the man is on made from? *Tags: skateboard, bench*
VLC-BERT base: Metal
VLC-BERT COMET: Wood

Commonsense Inferences (C):
The object is made of made from wood (0.52)
Before, the skateboard is made from wood happens (0.4)
The object is used for to skate on it (0.03)
You are likely to find the object in skate park (0.02)
Sometimes, the object causes the object is made from (0.01)

Q: This was used to keep the house warm before central air? *Tags: potted plant, couch*
VLC-BERT base: Heat
VLC-BERT COMET: Fire

Commonsense Inferences (C):
This can make a fire (0.27)
This is used for use as a blanket (0.2)
Sometimes, this causes hot (0.17)
Sometimes, this causes cold (0.15)
This is made up of heating (0.1)

Q: What is the person doing? *Tags: kite, skateboard*
VLC-BERT base: Skateboard
VLC-BERT COMET: Fly kite

Commonsense Inferences (C):
The person can ride the kite (0.25)
The person can fly kite (0.22)
The person is made of the kite to be flying (0.19)
Before, the person needed to have a kite (0.18)
After, the person rides the kite happens (0.14)

Qualitative Results

Q: What is the person doing? *Tags: kite, skateboard*
VLC-BERT base: Skateboard
VLC-BERT COMET: Fly kite

Commonsense Inferences (C):
The person can ride the kite (0.25)
The person can fly kite (0.22)
The person is made of the kite to be flying (0.19)
Before, the person needed to have a kite (0.18)
After, the person rides the kite happens (0.14)

To **conclude …**

- **Data-efficient Learning**
  - **Large-model + Transfer-learning**
  - **Multi-task learning + Fine-tuning**
  - **Foundational Model + Fine-tuning**
  - **Prior-knowledge Integration**
  - In-context Learning, Prompting
  - *Many other techniques* …

- **Compute-efficient Inference**
  - **Iterative refinement with early stopping (a.k.a. cascades)**
  - *Many other techniques* …

- **Data-bias Mitigation**
  - **Data re-sampling**
  - **Loss re-weighting**
  - *Many other techniques* …


thank you


ICICS/CS BUILDING
eat. sleep. code.
UBC