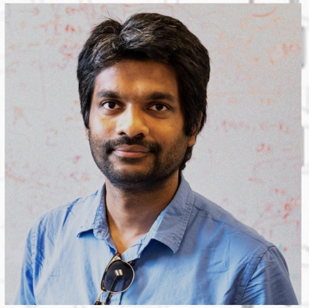
Feb 2023
AAAI Tutorials
Knowledge-Driven Vision-Language Pretraining

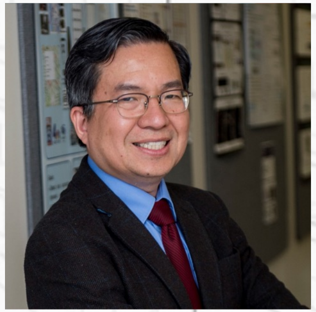
Knowledge-Driven Vision-Language Pretraining


Manling Li
UIUC


Xudong Lin
Columbia


Jie Lei
Meta AI


Mohit Bansal
UNC


Shih-Fu Chang
Columbia


Heng Ji
UIUC

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN COLUMBIA UNIVERSITY UNC Meta AI

Feb 2023
AAAI Tutorials
Knowledge-Driven Vision-Language
Pretraining

Commonsense Knowledge in V+L Pretraining

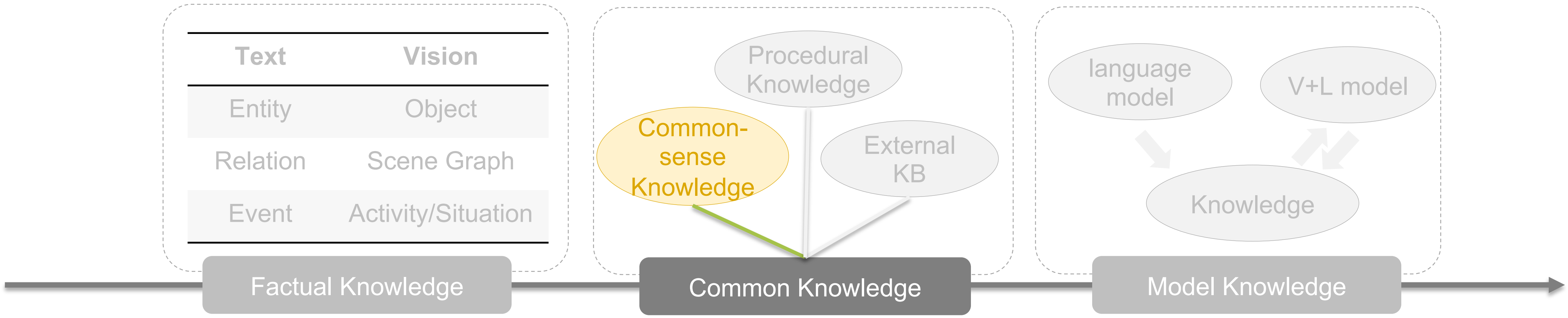
Knowledge-Driven Vision-Language Pretraining (Part III)

Manling Li
UIUC
manling2@illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
COLUMBIA UNIVERSITY
UNC
Meta AI

Commonsense Knowledge

Commonsense Knowledge is the basic facts and behaviors of the everyday world.



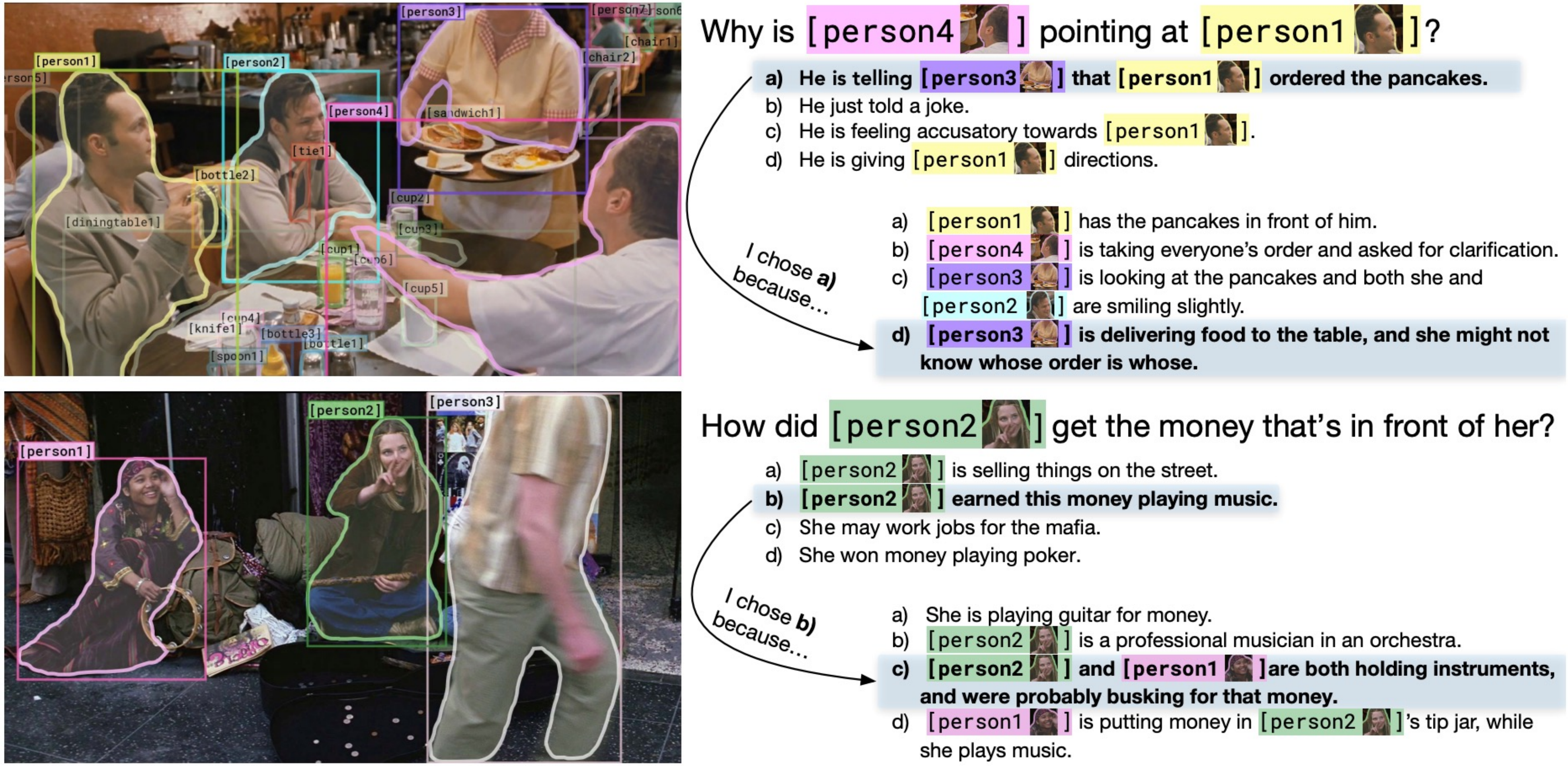
Outline

- What is Commonsense Knowledge
  - VCR
  - VisualCOMET
  - Physical Knowledge
- Knowledge → VLM
  - Knowledge Graph Riddles
  - Knowledge Graph Embedding
- VLM → Knowledge
  - Physical Knowledge
  - Unimodal vs Multimodal

Part 1: What is Visual Commonsense Knowledge?

Visual Commonsense Knowledge

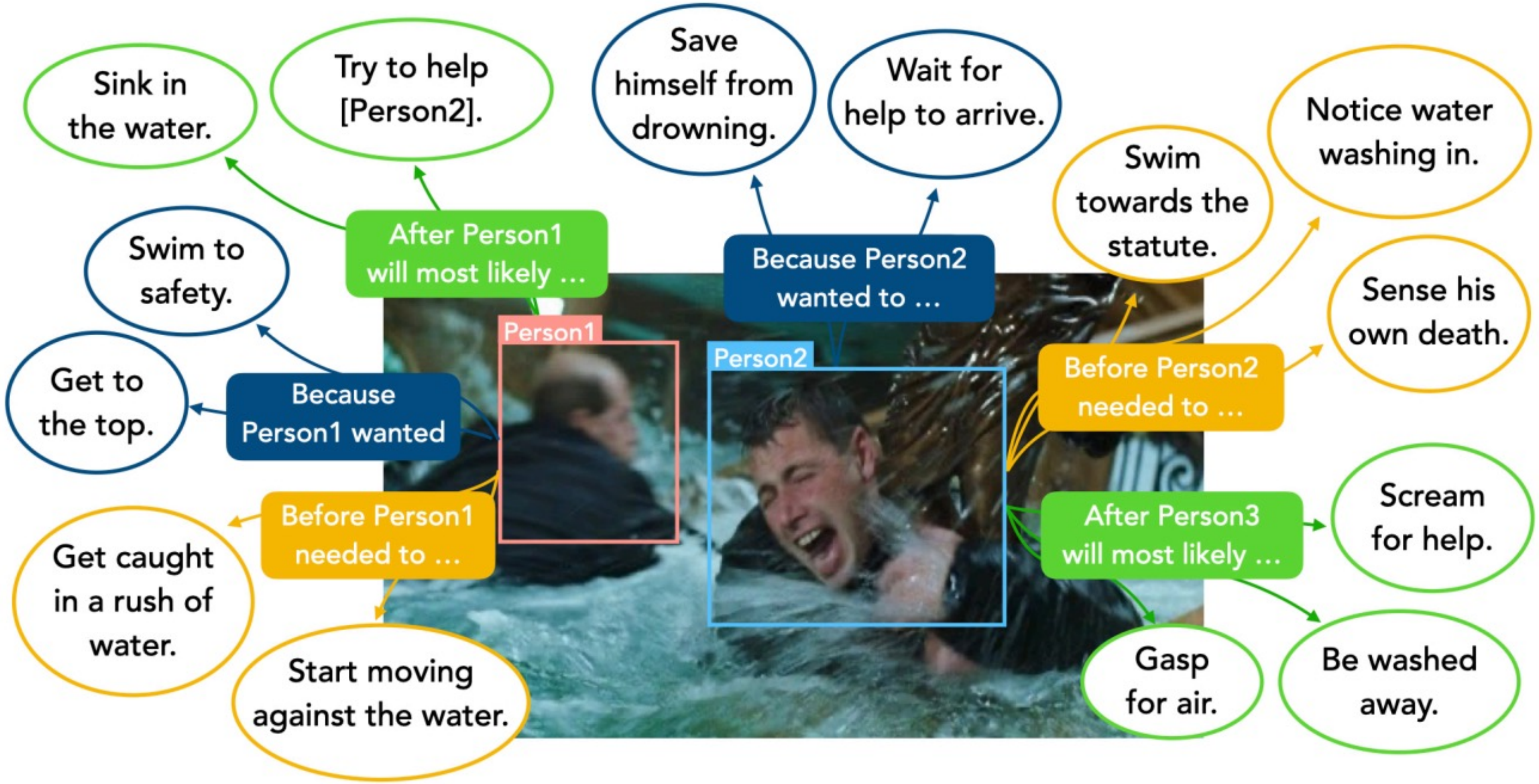
Visual Commonsense Reasoning (VCR): From Recognition to Cognition



From Recognition to Cognition: Visual Commonsense Reasoning. 2019

Visual Commonsense Knowledge

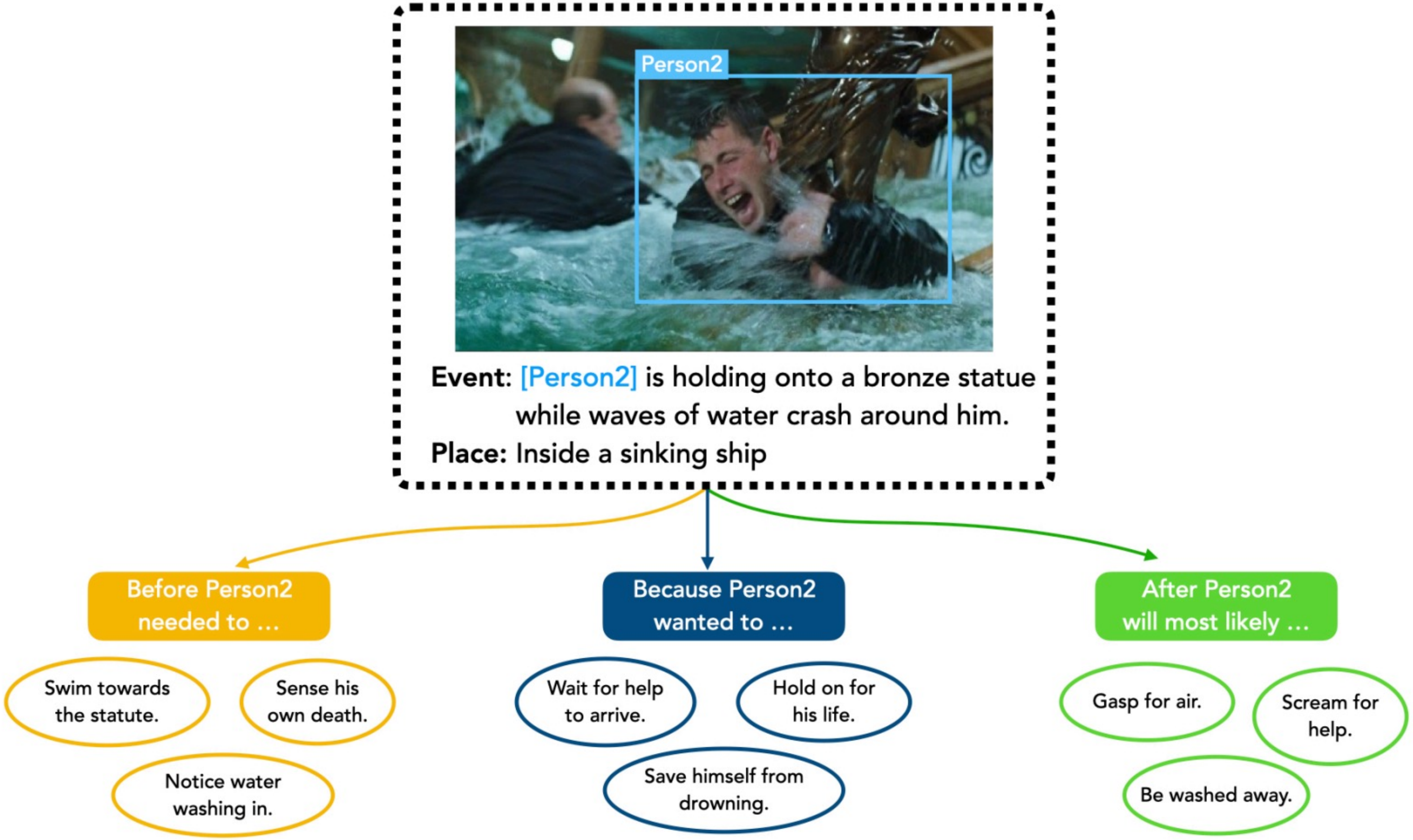
VisualCOMET: Cognitive Image Understanding via Visual Commonsense Graphs



VisualCOMET: Reasoning About the Dynamic Context of a Still Image https://mosaickg.apps.allenai.org/visual_comet

Visual Commonsense Knowledge

VisualCOMET Task Formulation: Generate the entire visual commonsense graph



Visual Commonsense Knowledge

Large Dataset Collection: There are in total 139,377 distinct Visual Commonsense Graphs over 59,356 images involving 1,465,704 commonsense inferences.

	Train	Dev	Test	Total
# Images/Places	47,595	5,973	5,968	**59,356**
# Events at Present	111,796	13,768	13,813	**139,377**
# Inferences on Events Before	467,025	58,773	58,413	584,211
# Inferences on Events After	469,430	58,665	58,323	586,418
# Inferences on Intents at Present	237,608	28,904	28,568	295,080
# Total Inferences	1,174,063	146,332	145,309	**1,465,704**

Physical Commonsense Knowledge

Physical Commonsense Knowledge can be learned via natural language.

a. Shape, Material, and Purpose		
[Goal]	Make an outdoor pillow	
[Sol1]	Blow into a tin can and tie with rubber band	✘
[Sol2]	Blow into a trash bag and tie with rubber band	✔
[Goal]	To make a hard shelled taco,	
[Sol1]	put seasoned beef, cheese, and lettuce onto the hard shell.	✘
[Sol2]	put seasoned beef, cheese, and lettuce into the hard shell.	✔
[Goal]	How do I find something I lost on the carpet?	
[Sol1]	Put a solid seal on the end of your vacuum and turn it on.	✘
[Sol2]	Put a hair net on the end of your vacuum and turn it on.	✔

b. Commonsense Convenience		
[Goal]	How to make sure all the clocks in the house are set accurately?	
[Sol1]	Get a solar clock for a reference and place it just outside a window that gets lots of sun. Use a system of call and response once a month, having one person stationed at the solar clock who yells out the correct time and have another person move to each of the indoor clocks to check if they are showing the right time. Adjust as necessary.	✘
[Sol2]	Replace all wind-ups with digital clocks. That way, you set them once, and that's it. Check the batteries once a year or if you notice anything looks a little off.	✔

PIQA: Reasoning about Physical Commonsense in Natural Language. AAAI 2020

The “Something Something” Dataset

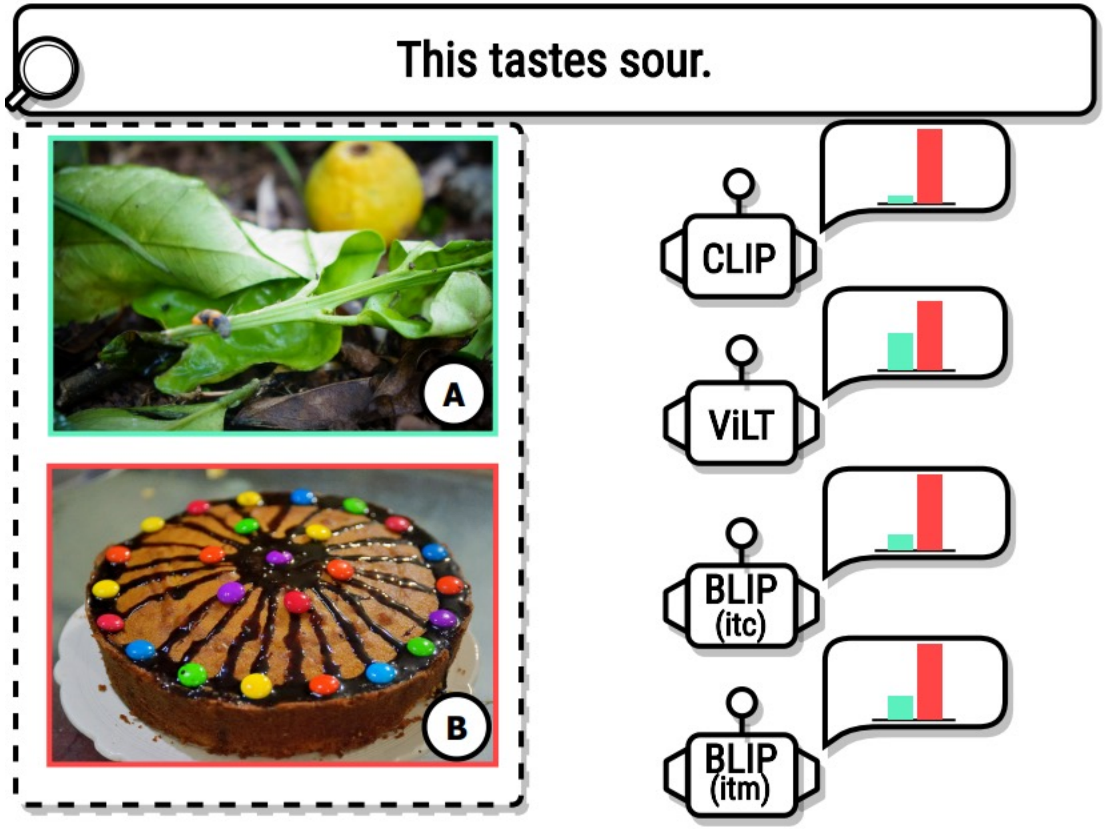
10 selected classes
Dropping [something]
Moving [something] from right to left
Moving [something] from left to right
Picking [something] up
Putting [something]
Poking [something]
Tearing [something]
Pouring [something]
Holding [something]
Showing [something] (almost no hand)

The “something something” video database for learning and evaluating visual common sense

Part 2: How can commonsense knowledge be learned via V+L pretraining?

Failures

Current V+L models lack abilities to capture commonsense knowledge:



Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles. 2022

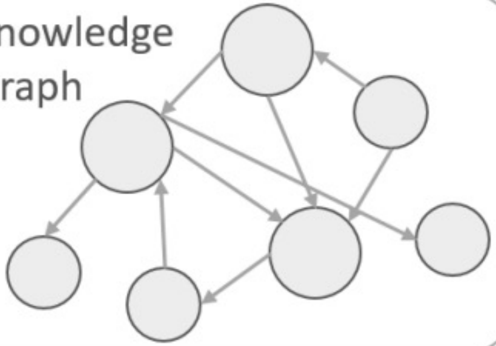
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability

Original image-text pair

A ***cat*** with a ***box*** in an ***office***.

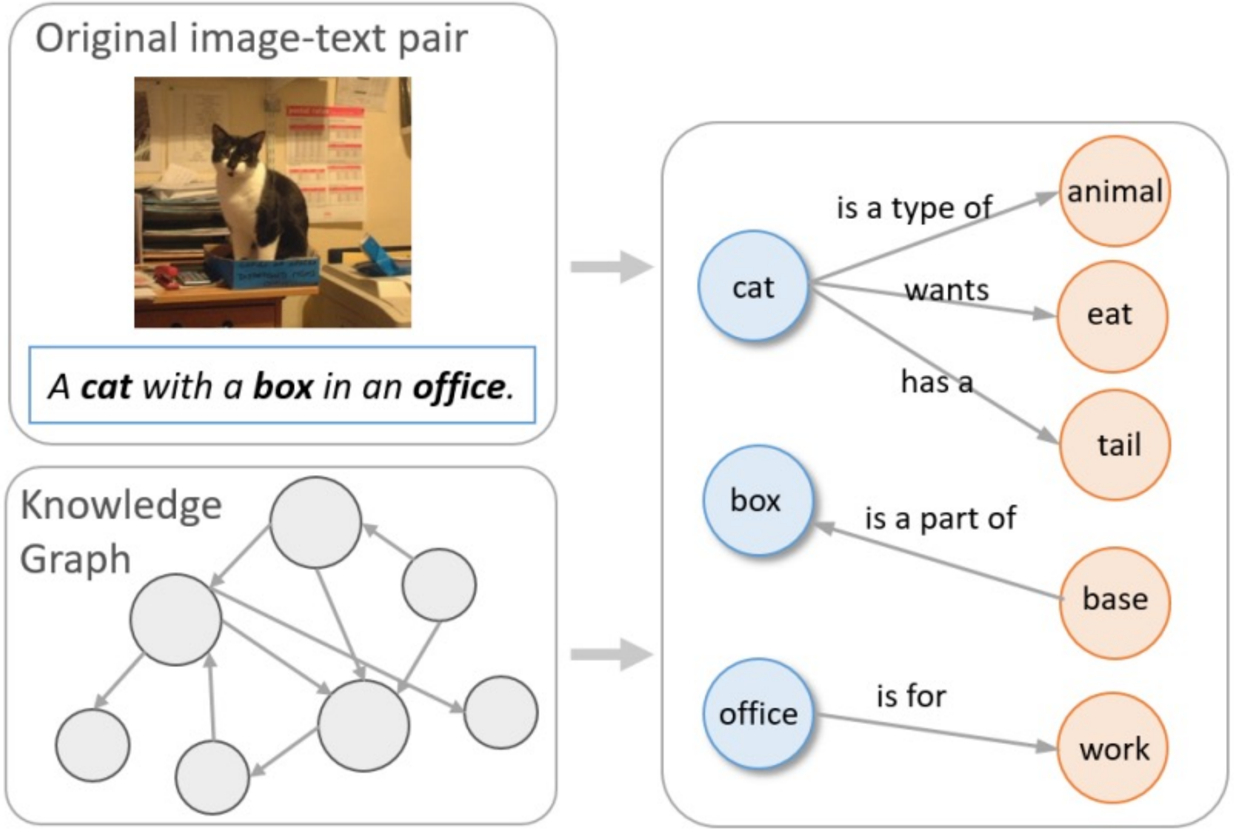
Knowledge Graph



Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles. 2022

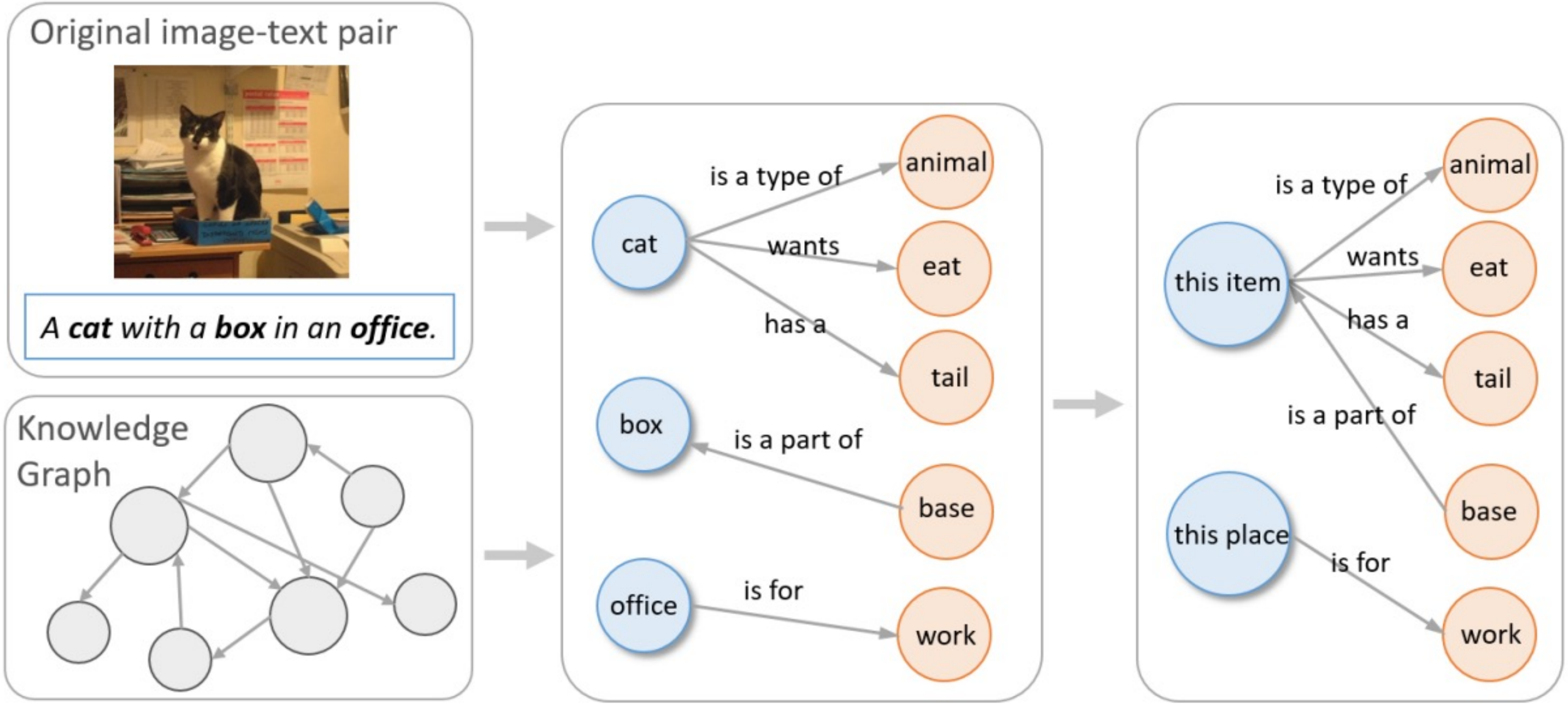
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



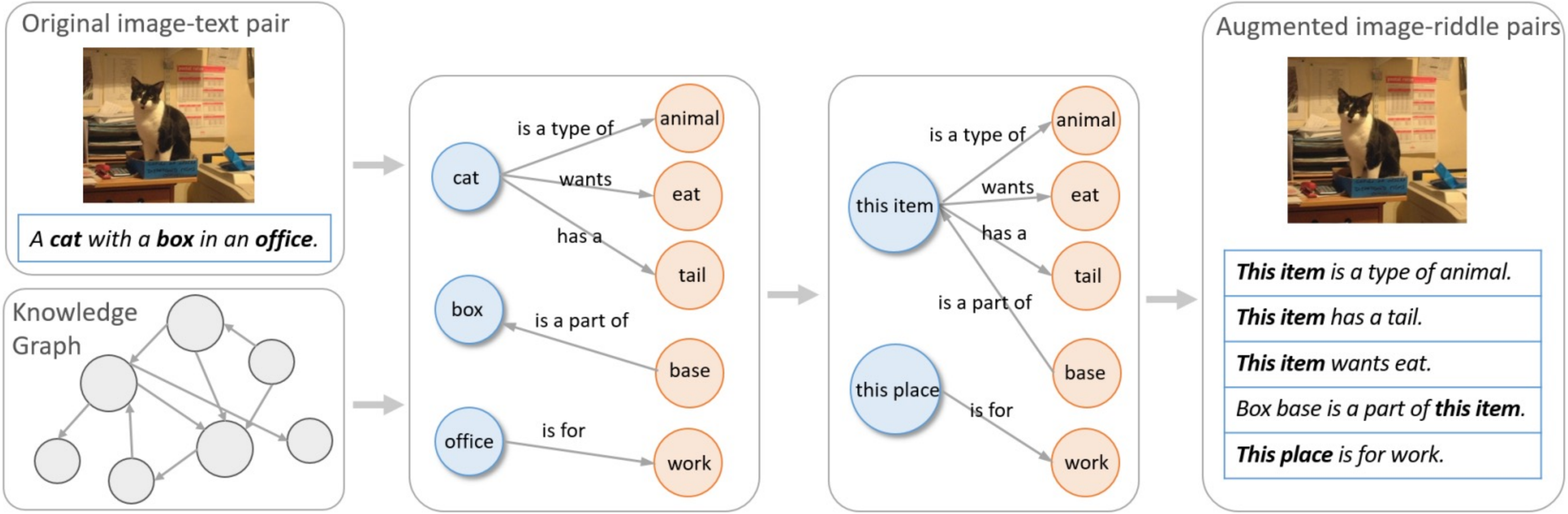
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



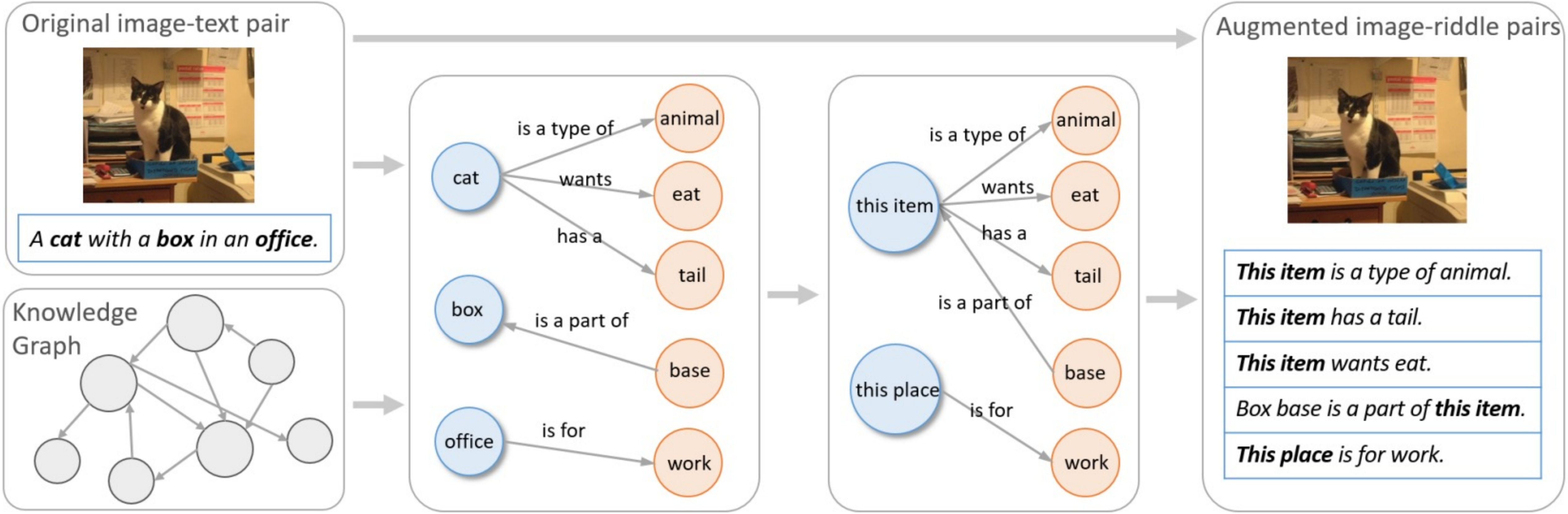
DANCE: Improving Commonsense in Vision-Language Models

DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability

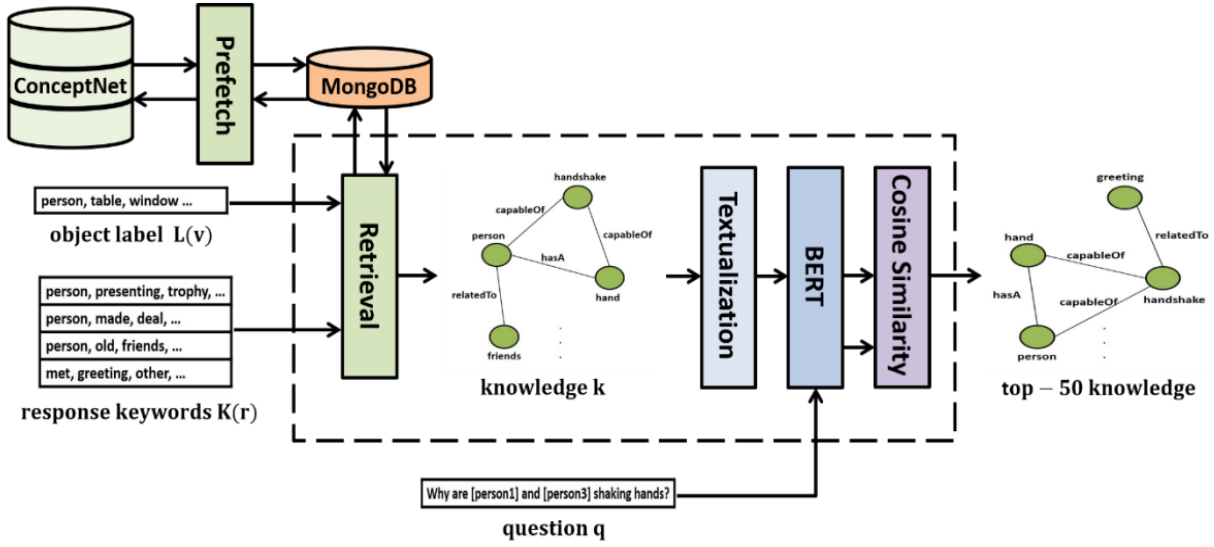


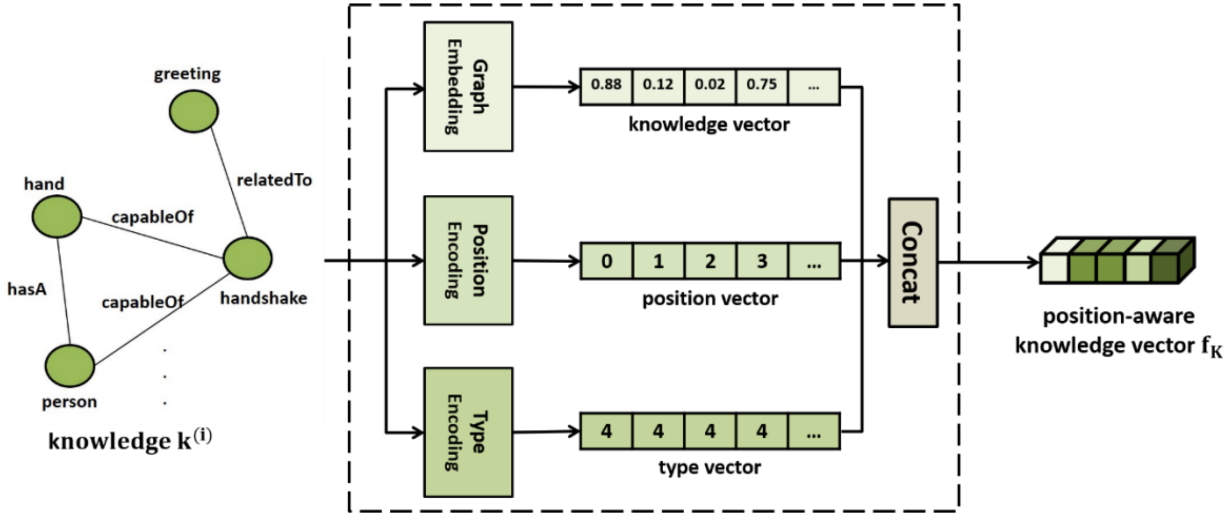
DANCE: Improving Commonsense in Vision-Language Models

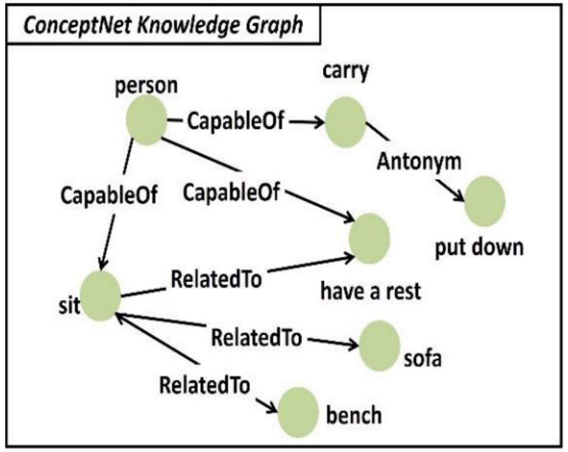
DANCE: Data Augmentation with kNowledge graph linearization for CommonsensE capability



Vision–Language Knowledge Co-Embedding







What is [person1] doing now?

ⓐ He is riding on a horse.	0.1%
ⓑ [person1] is being handed a fork to eat his food.	0.2%
ⓒ He is sitting and resting after carrying a heavy item into the room.	**98%**
ⓓ He is in the process of turning to look at [person1].	1.7%

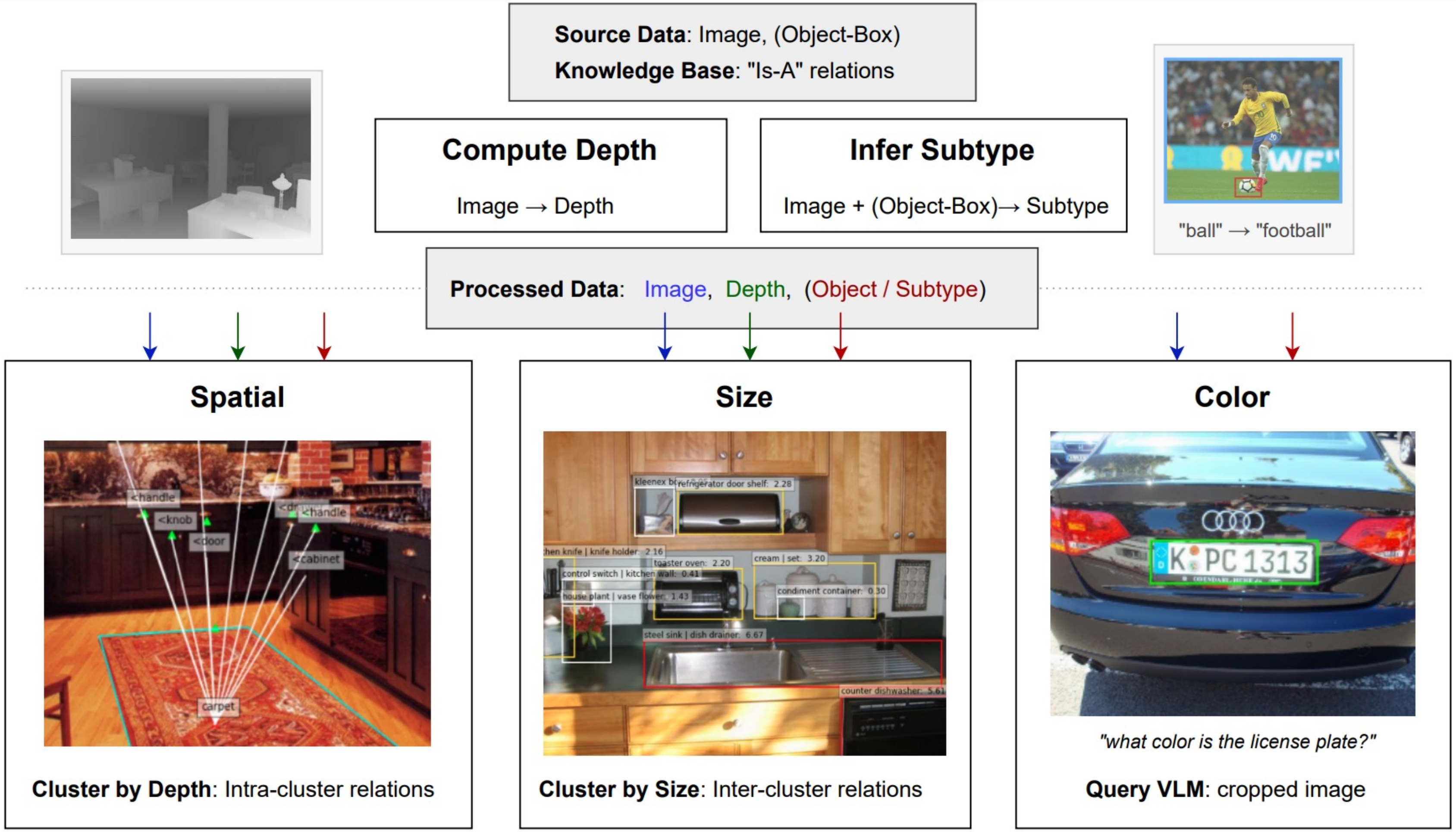
Because…

ⓐ He is sitting next to a cart with a large object on it. He is tired from pulling the kart because it is heavy.	0.2%
ⓑ He is using a brace, used to carry things.	14%
ⓒ He has brought many items into the fort which make him feel at ease.	0.8%
ⓓ He had a bench and has put it down.	**85%**

Part 3: Are VLMs commonsense KBs?

Probing “Visible” Physical Commonsense Knowledge

Visually accessible knowledge representing color, size and space



VIPHY: Probing “Visible” Physical Commonsense Knowledge

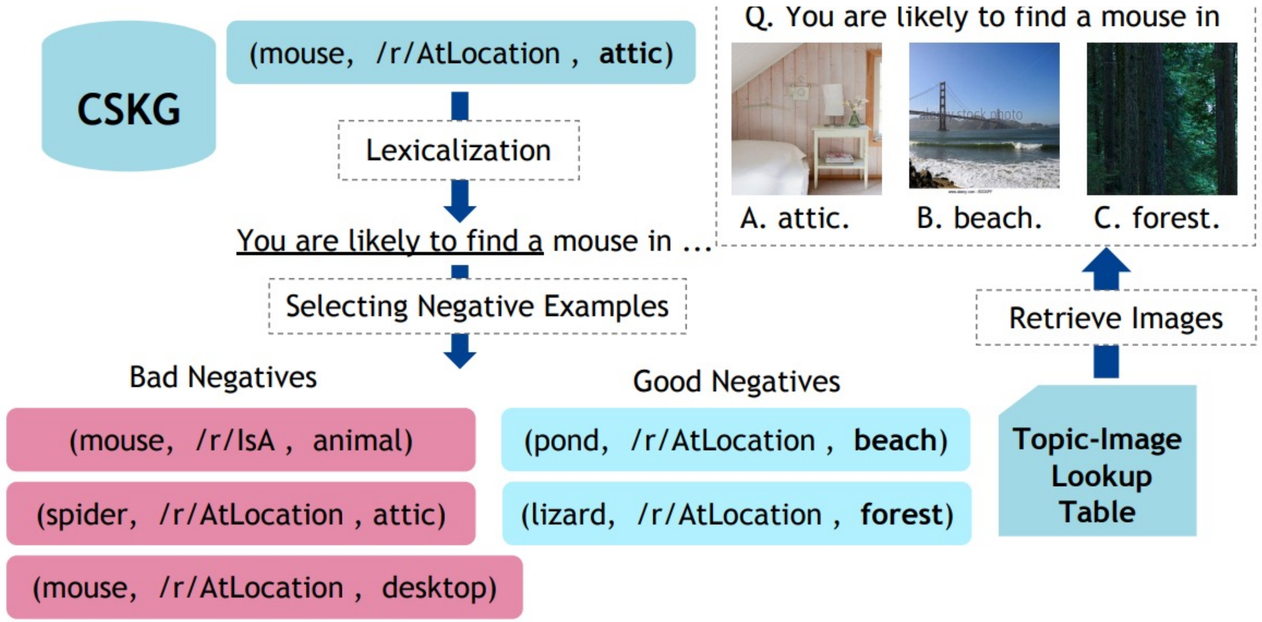
Probing “Visible” Physical Commonsense Knowledge

Visually accessible knowledge representing color, size and space

Task	Setting	Prompt
Color	ZS	O is of `[MASK]` color
	FT	`[CLS]` color of O
	QA	What is the color of O? (a) .. (b) ..
Size	ZS	O_1 is `[MASK]` than O_2 in size
	FT	`[CLS]` size of O_1 in comparison to O_2
	QA	what is the size of O_1 in comparison to O_2? (a) .. (b) ..
Spatial	ZS	in a S, the O_1 is located `[MASK]` the O_2
	FT	`[CLS]` in a S, the O_1 is located in comparison to O_2
	QA	in a S, where is O_1 is located in comparison to O_2? (a) .. (b) ..

Are Visual-Linguistic Models Commonsense KBs?

CS dimension	Starting prompt	Answer candidates	# Instances
part-whole	Furry animals have	A_1: effect of chilling innovation. **A_2: millions of hair.** A_3: hole in.	1,165
taxonomic	Recruit is a way to	A_1: rate. **A_2: enlist.** A_3: slope.	1,323
distinctness	Shade is not	A_1: flat. A_2: postal worker. **A_3: sunny.**	828
similarity	Throw up is a synonym of	A_1: rutinic acid. A_2: random. **A_3: vomit.**	644
quality	A wet floor is	**A_1: slippery.** A_2: light brown. A_3: abbreviated to unido.	1,840
utility	A fork is used for	A_1: speed of transit. A_2: confuse voters. **A_3: picking up food.**	2,090
creation	Music is created by	A_1: olive oil mill. A_2: mapping process. **A_3: instruments.**	100
temporal	Going for a haircut requires	**A_1: finding barber.** A_2: hard examinations. A_3: write persuasively.	1,889
spatial	You are likely to find a document folder in	**A_1: file drawer.** A_2: madagascar jungle. A_3: minerals.	1,599
desire	You would thank someone because you want to	A_1: accomplish mutual goal. **A_2: feel good.** A_3: cool off.	1,781



Are Visual-Linguistic Models Commonsense KBs?

CS dimension	Starting prompt	Answer candidates	# Instances
part-whole	Furry animals have	A_1: effect of chilling innovation. **A_2: millions of hair.** A_3: hole in.	1,165
taxonomic	Recruit is a way to	A_1: rate. **A_2: enlist.** A_3: slope.	1,323
distinctness	Shade is not	A_1: flat. A_2: postal worker. **A_3: sunny.**	828
similarity	Throw up is a synonym of	A_1: rutinic acid. A_2: random. **A_3: vomit.**	644
quality	A wet floor is	**A_1: slippery.** A_2: light brown. A_3: abbreviated to unido.	1,840
utility	A fork is used for	A_1: speed of transit. A_2: confuse voters. **A_3: picking up food.**	2,090
creation	Music is created by	A_1: olive oil mill. A_2: mapping process. **A_3: instruments.**	100
temporal	Going for a haircut requires	**A_1: finding barber.** A_2: hard examinations. A_3: write persuasively.	1,889
spatial	You are likely to find a document folder in	**A_1: file drawer.** A_2: madagascar jungle. A_3: minerals.	1,599
desire	You would thank someone because you want to	A_1: accomplish mutual goal. **A_2: feel good.** A_3: cool off.	1,781

dim.: spatial

You are likely to find vegetables in:
A. workplace.
B. stationary shop.
C. *garden.*

dim.: part-whole

A boat has:

A. reached legal age.
B. *sails*
C. different rules.

dim.: quality

A hill can be:

A. *steep.*
B. about to change.
C. important for normal living.

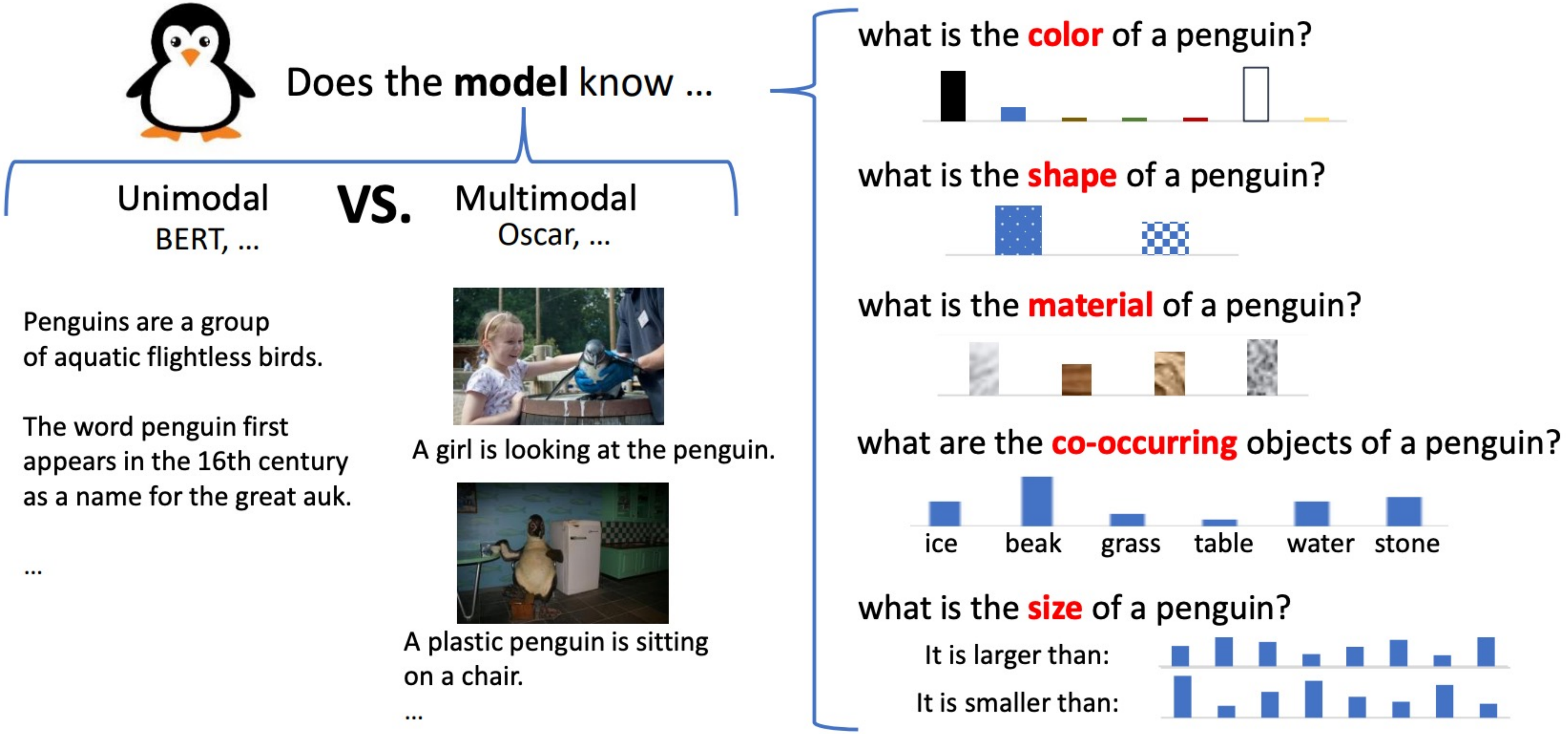
Are Visual-Linguistic Models Commonsense KBs?

Visual Commonsense Knowledge is more difficult than textual knowledge.

row	Images	part-whole 1,165	taxonomic 1,323	distinctness 828	similarity 644	quality 1,840	utility 2,090	creation 100	temporal 1,189	spatial 1,599	desire 1,781	All 13,259
1 RoBERTa	–	68.5	61.8	80.2	**67.4**	69.7	**74.2**	72.0	**60.9**	54.8	**65.9**	**67.5**
2 BERT	–	62.8	71.2	80.1	54.8	68.1	72.4	74.0	53.7	52.4	60.4	65.0
3 BERT_{CC}	–	68.4	62.0	66.6	51.1	66.0	65.4	62.0	53.6	**63.7**	58.3	61.9
4 UNITER_BERT_{T}	–	70.1	**74.5**	**81.4**	62.4	**72.0**	73.8	**79.0**	54.5	53.9	61.5	66.5
5 UNITER_{T}	–	**70.9**	59.8	71.3	51.2	69.9	71.5	71.0	52.7	61.5	62.5	64.0
6 VILBERT_{T}	–	63.9	60.3	64.9	46.7	66.1	71.2	58.0	52.2	61.0	62.8	60.7
7 UNITER_{TV}	retrieved	63.0	54.0	65.9	46.4	62.4	65.4	62.0	49.2	57.4	58.5	58.4
8 VILBERT_{TV}	retrieved	55.0	49.9	55.9	42.2	57.4	60.5	52.0	47.2	52.9	56.6	53.0
9 $\text{UNITER}_{T\tilde{V}}$	dummy	61.5	51.6	63.4	42.2	63.6	66.4	55.0	49.4	58.2	59.7	57.1
10 $\text{VILBERT}_{T\tilde{V}}$	dummy	60.4	58.9	64.9	43.9	63.4	65.5	55.0	48.4	56.8	62.0	57.9
11 UNITER_{V}	retrieved	36.4	36.6	40.1	38.5	34.2	36.6	32.0	34.8	36.2	34.3	36.0
12 VILBERT_{V}	retrieved	37.8	35.1	37.7	39.8	36.8	35.7	41.0	33.0	37.6	34.0	36.8
13 $\text{UNITER}_{\tilde{V}}$	dummy	30.8	26.3	45.7	28.6	29.2	28.7	19.0	28.7	29.6	30.7	29.7
14 $\text{VILBERT}_{\tilde{V}}$	dummy	34.8	35.8	50.5	40.4	30.4	31.1	30.0	29.4	33.5	30.1	34.6

Unimodal vs Multimodal models?

Unimodal and multimodal models' abilities to capture visual commonsense knowledge



Unimodal vs Multimodal models?

ViComTe dataset on five relation types: color, shape, material, size, and visual co-occurrence

Relation	**# Classes**	**# (subj, obj) Pairs**	**Ex Template**	**Ex (subj, obj) Pair**
color	12	2877	[subj] *can be of color* [obj]	(*sky*, *blue*)
shape	12	706	[subj] *has shape* [obj] .	(*egg*, *oval*)
material	18	1423	[subj] *is made of* [obj] .	(*sofa*, *cloth*)
size (smaller)	107	2000	[subj] *is smaller than* [obj] .	(*book*, *elephant*)
size (larger)	107	2000	[subj] *is larger than* [obj] .	(*face*, *spoon*)
co-occurrence	5939	2108	[subj] *co-occurs with* [obj] .	(*fence*, *horse*)

Unimodal vs Multimodal models?

Unimodal and multimodal models' abilities to capture visual commonsense knowledge

Source	Group	Spearman ρ	# Subjs	Avg # Occ	Top5 # Occ	Btm5 # Occ	Acc@1
VG	All	64.3 ± 23.9	355	1252.6	64.6	308.6	
	SINGLE	62.2 ± 24.0	131	494.9	64.6	1181.6	80.2
	MULTI	69.3 ± 20.7	136	1156.1	2062.2	347.0	
	ANY	58.4 ± 27.1	88	2529.6	8452.4	1213.4	
Wikipedia	All	33.4 ± 30.6	302	543.6	1758.0	49.8	
	SINGLE	29.6 ± 29.9	110	352.2	345.8	35.0	35.5
	MULTI	33.9 ± 30.9	119	500.8	1242.0	27.6	
	ANY	38.2 ± 30.4	73	902.0	3000.2	161.2	

Visual Commonsense in Pretrained Unimodal and Multimodal Models. NAACL 2022

Future Direction: Adding commonsense knowledge to pretraining

Knowledge for Model

?

- In-context prompt
- data augmentation
- data selection

Knowledge for data

Knowledge → VLM

Knowledge Graph Riddles

Knowledge Graph Embedding

VLM → Knowledge

Physical knowledge

Unimodal vs Multimodal

Future Direction: Physical Knowledge Enhanced LM/VLM

Humans learn a huge amount of knowledge about the external world via **multisensory experience and interactions**, however, current **LLM/VLM** are trained with **static datasets**, thus **lacks understanding of the physical world**.

Spatial Relation

Z: Put object A to the left of object B. Then, put object B in front of object A. Then, put object C to the left of object A. Which object is directly behind object B?

Object C is directly behind object B.

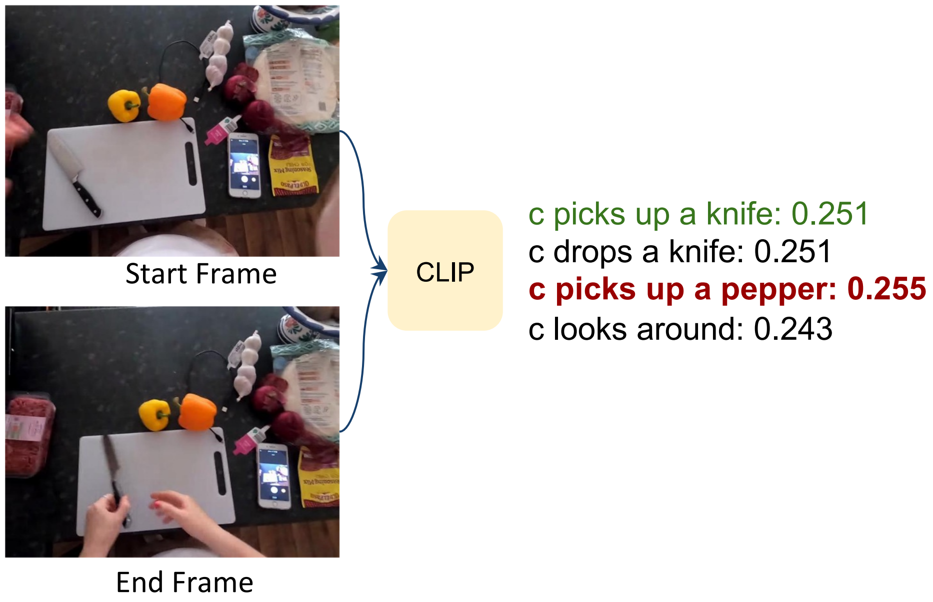
Knowledge requiring embodiment

Z: Imagine you are a human being. Put your left hand on the back of your head. Can you still see your left hand?

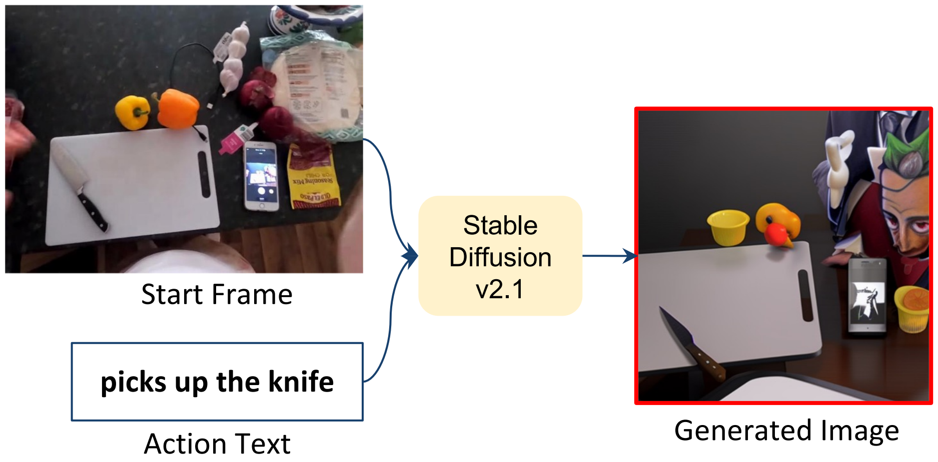
Yes, I can still see my left hand as it is positioned on the back of my head.

Future Direction:
Physical Knowledge Enhanced LM/VLM

Humans learn a huge amount of knowledge about the external world via **multisensory experience and interactions**, however, current **LLM/VLM** are trained with **static datasets**, thus **lacks understanding of the physical world**.



Action classification



Causal Effect of Actions

Physical Interactions involving actions and objects

Future Direction: Physical Knowledge Enhanced LM/VLM

From Reading/Seeing to Doing: From passive perception to interaction with the world.