Weak Multimodal Supervision for Robust Visual Recognition

Adriana Kovashka

Associate Professor

Department of Computer Science

University of Pittsburgh

(Original) Research Motivation

- Media content affects public opinion and societal outcomes (e.g. via elections)
- We want to understand agenda/intent in the media
  - Images in political articles, visual advertisements
- Data is limited, supervision is noisy
  - Images with intent not as abundant as general images
  - Expensive to carefully annotate all knowledge required
  - Text in ads/articles helps understand images, but relation is indirect, abstract
- Can we learn useful models from noisy data?
- How to transfer knowledge from other domains?

Outline

1. Visual reasoning for advertisements
   - Visual/language reasoning about ad strategies (ICCV 2021, ACL 2023)
2. Learning concepts from language supervision
   - Modeling complementary image-text relationships (BMVC 2018, ECCV 2020, TPAMI 2022)
   - Modeling variance and noise in object naming (EACL 2024, NAACL Findings 2024)
   - Investigating the role of language context (WACV 2024)
3. Recognizing objects with unexpected appearance
   - Across modalities (ICMR 2023)
   - Across geographies (CVPR 2024)

Decoding image advertisements

Recognized Concepts

people	commerce	street
business	stock	city

What Does It Mean?

Food at Burger King must taste really good since even competitor's employee secretly buys it.

Image Captioning

A man standing in front of a display of food.

A man standing in front of a display of a store.

- State-of-the-art vision systems (were) inadequate to describe the messages hidden behind purposefully designed advertisements

Predictions from *Clarifai* and *Vinyals et al.*

Challenges: Human associations (symbols)

Challenges: Complex relation of image/text

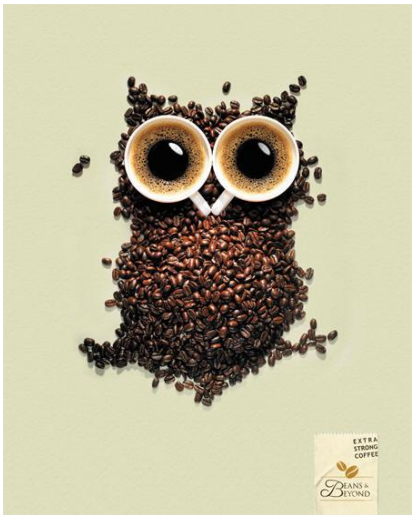
- Is it an advertisement for …

← Public Service Announcement
Against Human Trafficking

"Winter Collection"

Commercial Advertisement →
Premium Women Clothing

Challenges: Atypical objects



Challenges: Implied physical processes

Advertisement dataset

We collected **64,832** advertisement images with:

- 204,340 **topic** annotations
- 102,340 **sentiment** annotations
- 202,090 **action/reason (what/why)** annotations
- 64,131 **symbolism** annotations
- 20,000 **strategy** annotations
- 11,130 **slogan** annotations

http://cs.pitt.edu/~kovashka/ads

Hussain, Zhang, Zhang, Ye, Thomas, Agha, Ong and Kovashka, CVPR 2017

Decoding image advertisements

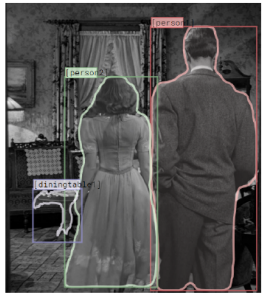
- What message does the ad convey (*action*), and what arguments does it provide for taking the suggested action (*reason*)?
- Multiple-choice task: Given *k* options for action-reason statements, pick one that matches the image

Correct action-reason statements:
- I should drink evian because it helps you recover
- I should buy Evian because it keeps us young
- I should drink Evian because it will keep me like a baby

Hussain, Zhang, Zhang, Ye, Thomas, Agha, Ong and Kovashka, CVPR 2017

Shortcuts in Visual Commonsense Reasoning

Shortcut effects: Example and definition. Consider the following example from the VCR dataset (Zellers et al. 2019). In the figure, [person1] (male) is on the right and [person2] (female) is on the left.



*Question: What does **[person1]** think of **[person2]**'s **dress**?*
*Correct answer: **[person1]** thinks **[person2]** looks stunning in her **dress**.*
Incorrect #1: She does not approve.
Incorrect #2: [person2] is a girl and girls like to wear makeup.
Incorrect #3: [person1] is confused and annoyed by [person2] following her in the store.

Table 5: **Our method enables the most robust training.** All results show Q→A except for the bottom two which show QA→R. The best method per group on Q→A is **bolded**, and the best method per task is underlined.

Method	Std Val	Rule-Singular	Rule-Plural	AdvTop-1
Baseline (B2T2)	68.5	63.3	65.3	37.0
Masking 0.05	**69.3**	**63.9**	**66.0**	48.8
Masking 0.10	68.7	62.8	64.7	50.1
Masking 0.15	68.2	62.0	63.3	**50.6**
Masking 0.30	64.1	56.6	56.8	47.5
Masking 0.05 + MLM	68.5	62.9	64.8	47.3
Masking 0.10 + MLM	69.1	63.8	65.0	**50.6**
Ours-CL Init0.30 Decay1e-4	69.6	64.5	64.7	51.7
Ours-CL Init0.30 Decay5e-5	**69.9**	**65.9**	**66.8**	54.5
Ours-CL Init0.50 Decay1e-4	69.4	65.0	65.0	53.0
Ours-CL Init0.50 Decay5e-5	69.8	65.4	66.3	**54.9**
Baseline (B2T2)	68.5	64.9	67.9	34.7
Ours-CL Init0.30 Decay5e-5	70.6	66.6	70.4	47.9

[val-54]

Q: Where is [2] going ?

Original Val data
A0 [2] is going into the store .
A1 [2] is getting into a carriage .
A2 [1] is going to the bathroom .
A3 [1] is going outside to play after the conversation with [2] is over .

Modified by rule (A single person)
A0 He is going into the store .
A1 [2] is getting into a carriage .
✗ **A2 [2] is going to the bathroom .**
A3 [1] is going outside to play after the conversation with [2] is over .

Modified by an adversarial model
A0 [MASK] is going into the store .
A1 [2] is getting into a [MASK] .
✗ **A2 [MASK] is going to the bathroom**
A3 [1] is [MASK] outside to play after the conversation with [2] is over .

[val-270]

Q: What are [1, 2] feeling ?

Original Val data
A0 [1, 2] do not like the restaurant .
A1 They are apprehensive .
A2 They are both feeling happy .
A3 [1, 2] are feeling drunk .

Modified by rule (A group of people)
A0 [1, 2] do not like the restaurant .
A1 [1, 2] are apprehensive .
A2 They are both feeling happy .
A3 [1, 2] are feeling drunk .

Modified by an adversarial model
✗ **A0 [1, 2] do not like the [MASK] .**
A1 They are apprehensive [MASK]
A2 They are [MASK] feeling happy .
A3 [1, 2] are feeling [MASK] .

Ye and Kovashka, AAAI 2021

Domain-Robust Visual Question Answering

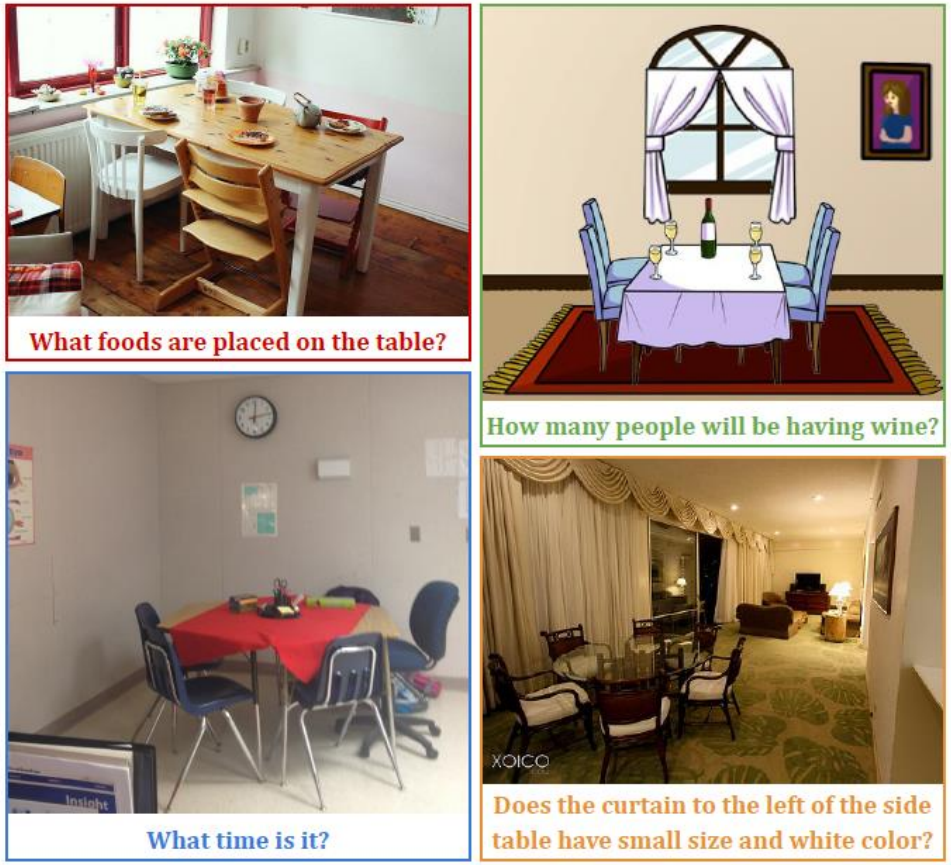


Figure 1. The same visual setting can be captured in different ways in VQA datasets, and paired with different information needs (questions). They may require deduction using visual contents, reading from a specific region of the image, or reasoning about complex spatial relationships. All examples are selected from real VQA datasets, *i.e.* VQA v2, VQA Abstract, VizWiz and GQA.

Zhang, Maidment, Diab, Kovashka, Hwa, CVPR 2021

Atypical images with persuasive intent

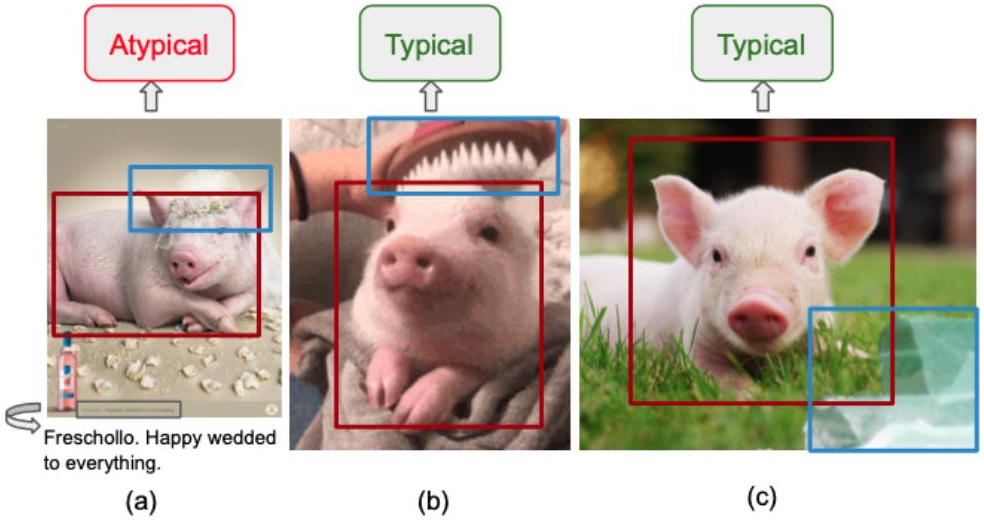


Figure 1. These images illustrate the importance of object interactions and their spatial relative position for atypicality detection. (a) Pig wearing a bridal veil is atypical; (b) If a handled brush instead of a veil is on top of the pig's head, then the image is typical; (c) If the veil's location is different, the image may also be typical.

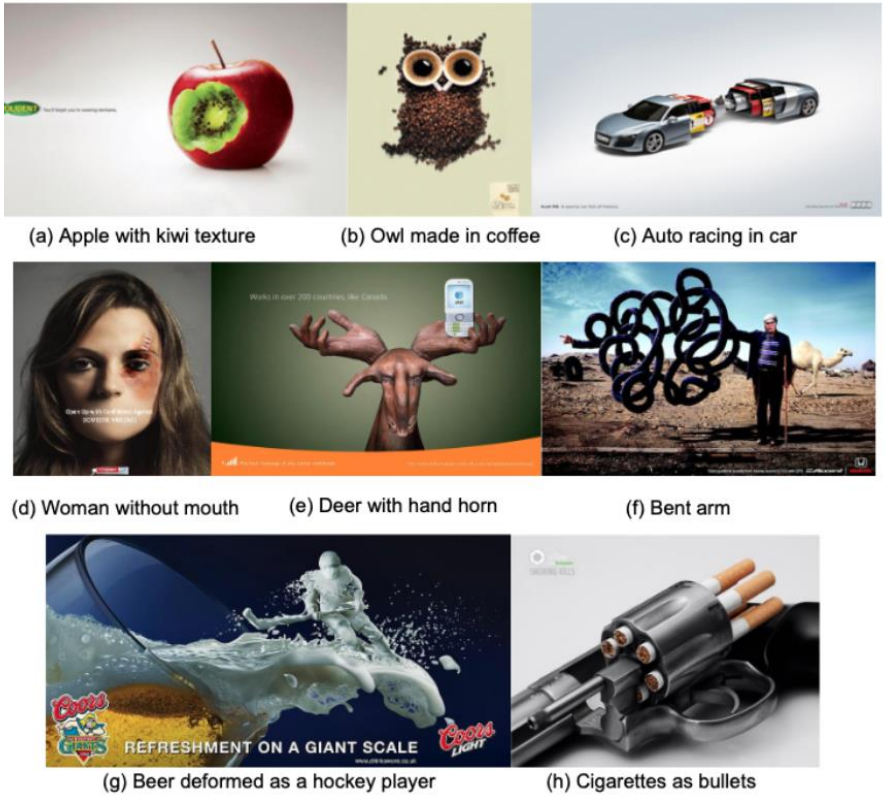


Figure 2. Atypical object transformations in ads; [48]'s dataset.

Guo, Hwa and Kovashka, ICCV 2021

Decoding Symbolism in Language Models

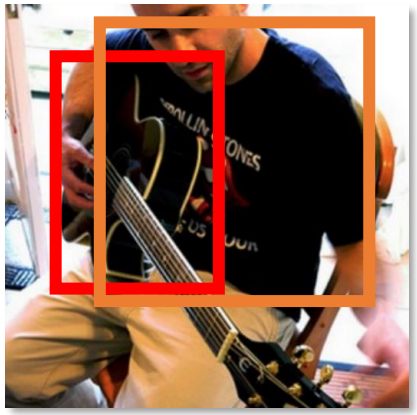
Relationship Type	Count	Example (signifier - signified)	Example (situated signifier - signified)
UsedFor	52	makeup - beauty	cartoon candy running on a treadmill - health
HasProperty	46	child - youth	workers sitting closely in a sofa - comfort
RelatedTo	47	mountain - adventure	cigarette smoke in the shape of mushroom cloud - danger
Others	94	chocolate - love	foot stepping on tombstone - death
Indirect	116	giraffe - love	shoes made out of red bull cans - strong

Table 2: Relationship types of *signifier-signified* in the set of advertising symbolism.

Guo, Hwa and Kovashka, ACL 2023

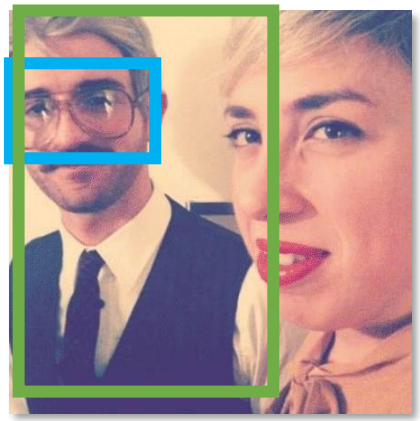
Understanding the relation of image/text

Image Captioning



man in black shirt is playing guitar.

Visual Question Answering



Who is wearing glasses?
man

Persuasive Ads using Visual Rhetoric

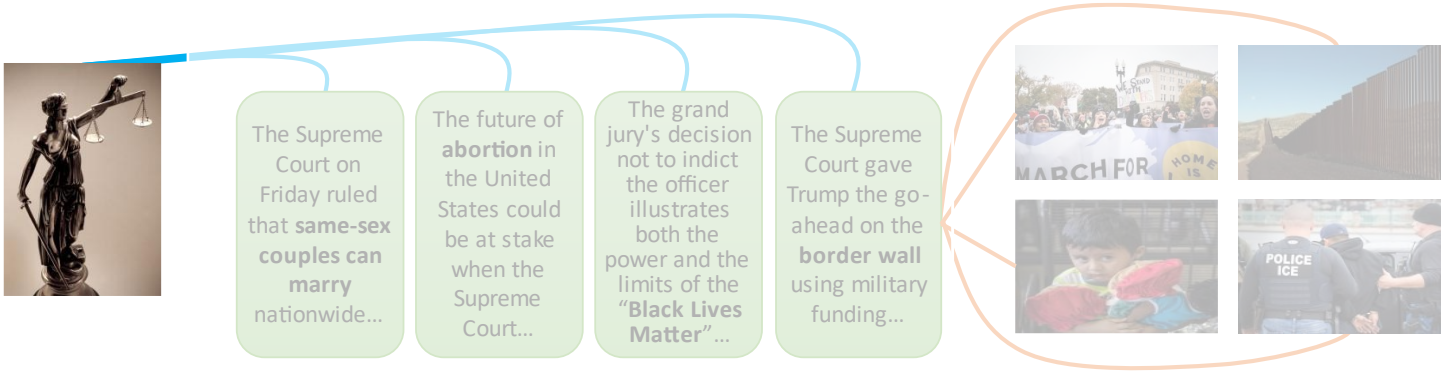
They can't afford to wait for evolution.
Dolphin? Earphone?

Redundant messages are transmitted independently across channels.

Elements across channels do not align or are connected in non-literal way.

Zhang, Hwa and Kovashka, BMVC 2018

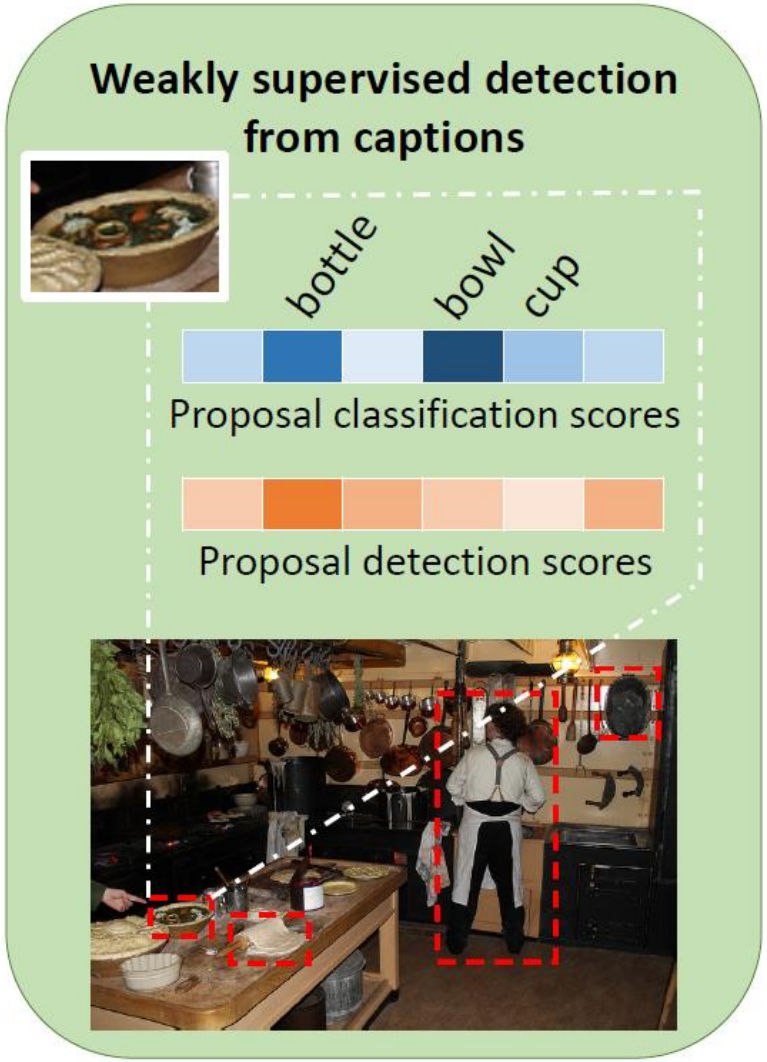
Multimodal complementarity



- Modalities complementary; each contributes to overall message the multimedia document seeks to convey
  - Figurative / symbolic usages (e.g. Themis, justice)
  - Abstract semantic concepts (e.g. immigration) can have diverse visual expressions

Thomas and Kovashka, ECCV 2020

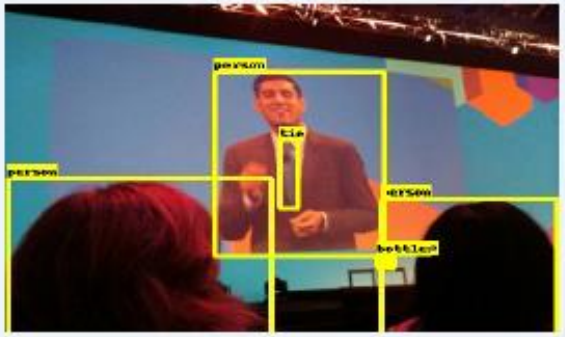
Learning object detectors from captions



Ye et al., ICCV 2019

Captions don’t mention all objects present

- This causes a challenge for standard weakly-supervised detection methods

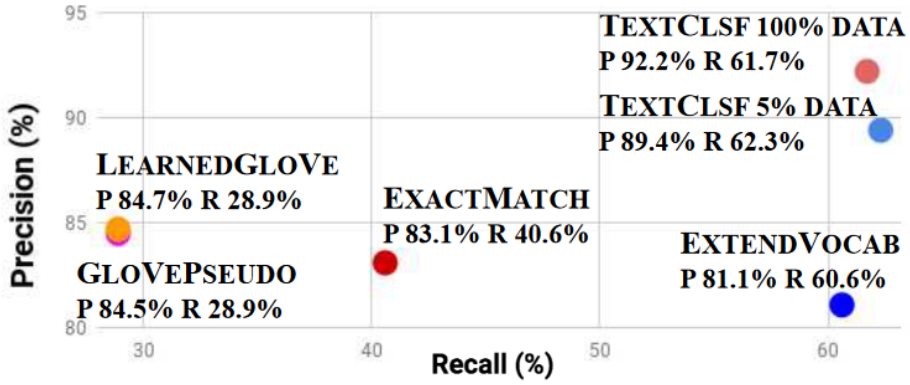


- *People watch a man **delivering a lecture** on a screen*
- *A large screen showing a person **wearing a suit***
- *An audience is looking at a film of a man talking that is projected onto a wall*

GROUNDTRUTH objects: person, tie, bottle

In this example, the object “tie” cannot be extracted using a lexical matching method, but it can be inferred through reasoning (ties are worn at formal events)

- Can learn to map from captions to labels efficiently



Ye et al., ICCV 2019

Dealing with noise in extracted labels

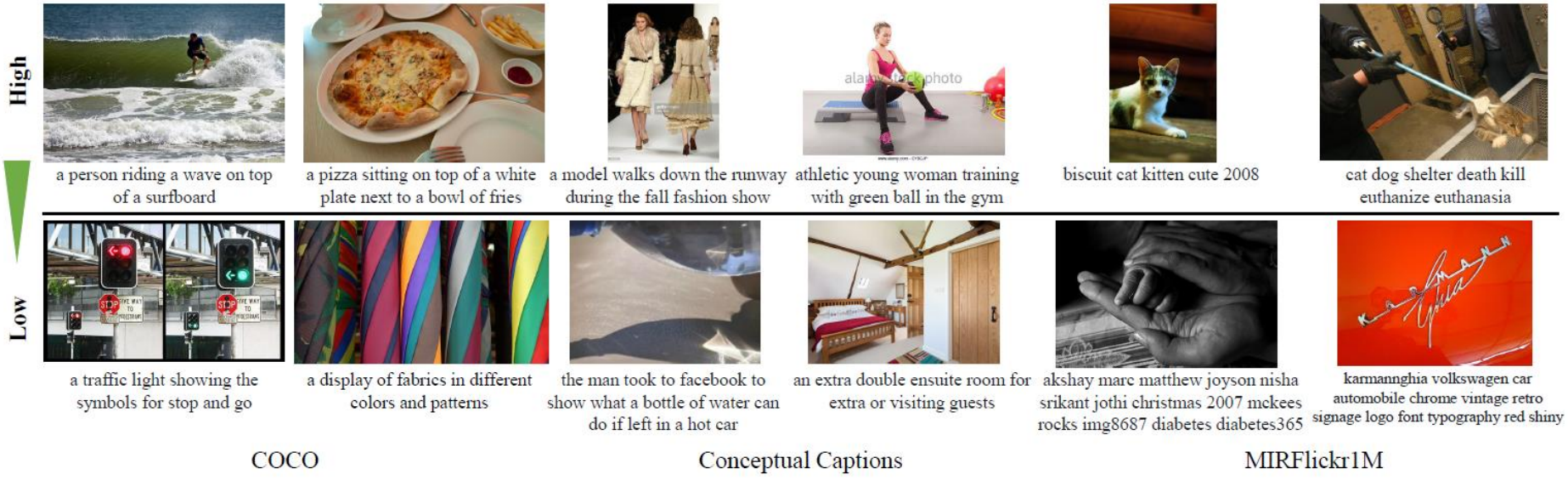


Fig. 11: Image-caption pairs with high homogeneity scores on the top, and low scores on the bottom.

Unal et al., TPAMI 2022

VEIL: Vetting Extracted Image Labels from In-the-Wild Captions

Figure 1: Examples of noisy extracted labels (underlined) from our **C**aption **La**bel **N**oise dataset. We categorize types of similar context present instead of the underlined object, as well as types of visual defects and linguistic indicators that are useful for detecting noise.

Rai and Kovashka, EACL 2024

VEIL: Vetting Extracted Image Labels from In-the-Wild Captions

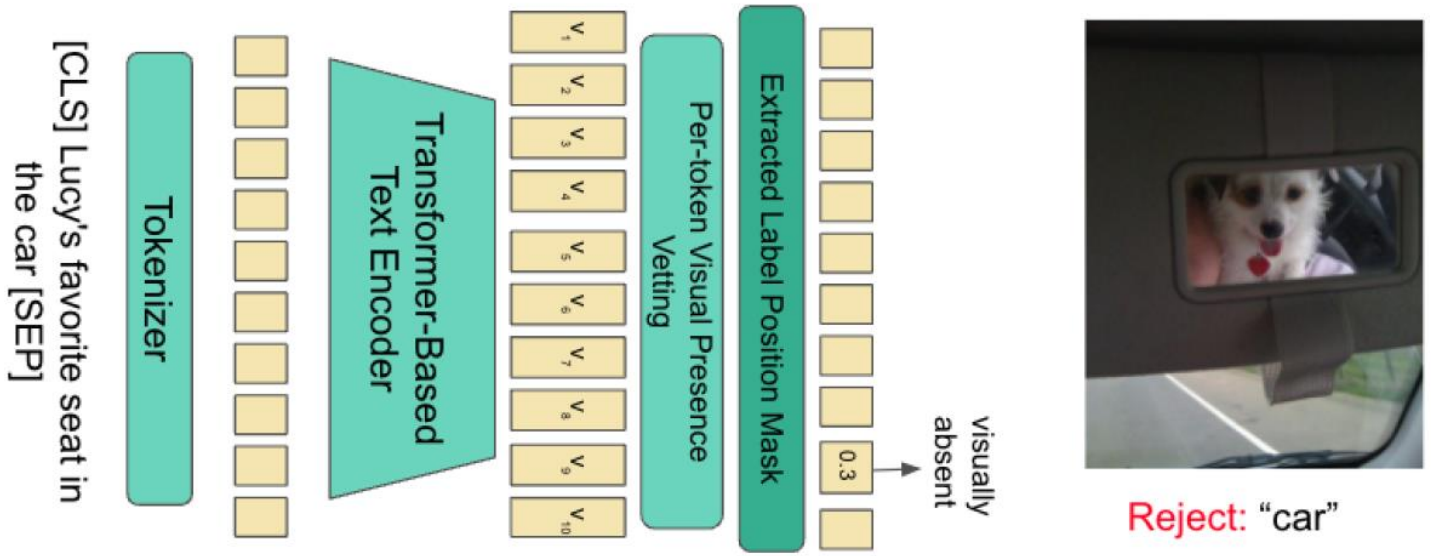


Figure 2: VEIL architecture. In this example, only "dog" is an extracted label and it fails the vetting process. The masking layer masks visual presence predictions for text tokens not corresponding to an extracted label.

Rai and Kovashka, EACL 2024

Synonym relations affect object detection learned on vision-language data

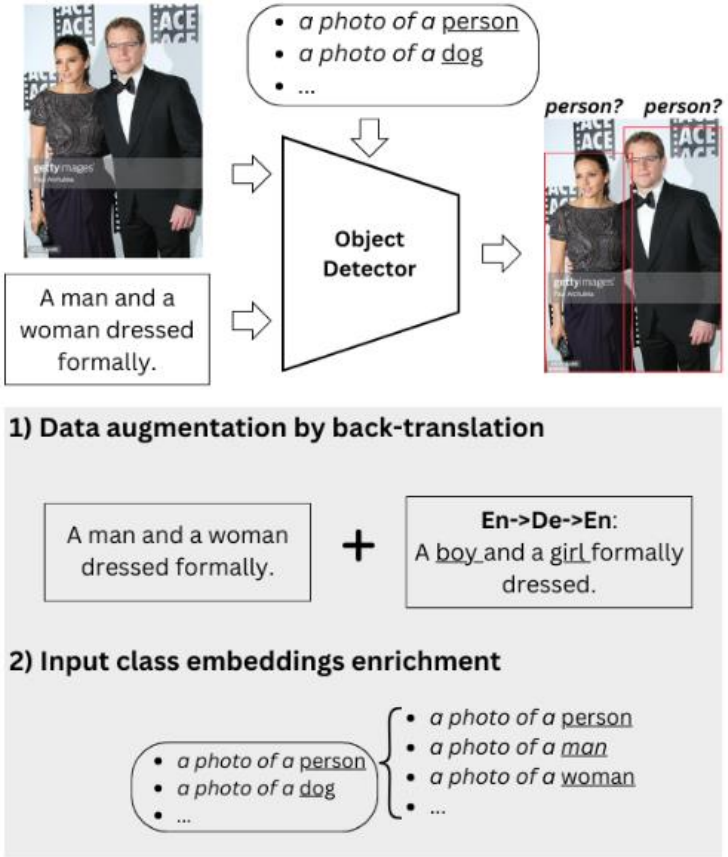


Figure 1: Top: input to an open-vocabulary object detector: images, class embeddings, and captions; and its output: bounding boxes with associated labels. Bottom: our approaches. 1) Data augmentation by back-translation: add captions back-translated from a foreign language; 2) Class embeddings enrichment: consider synonyms when extracting class embeddings.

Nebbia and Kovashka, NAACL 2024 Findings

Synonym relations affect object detection learned on vision-language data

Original	German	Russian
A **skate board rider** does a trick in front of a building.	A **skateboarder** does a trick in front of a building.	A **skater** does a trick in front of the building.
Three adults help a **youngster** follow a sheet of instructions.	Three adults help a **teenager** follow a sheet of instructions.	Three adults help the **teenager** follow instructions.

Table 1: Examples of (left) original COCO captions, (middle) captions back-translated from German, and (right) captions back-translated from Russian.

Captions	COCO names	Synonyms mean (std)	Avg.
Class embeddings: COCO names			
Original	**44.45**	33.87 (5.94)	35.63
BT: German	*44.23*	34.25 (5.32)	35.91
Class embeddings: enriched			
Original	43.58	*37.25* (4.56)	*38.31*
BT: German	37.48	36.75 (4.56)	36.87
Curriculum	43.49	**37.93** (3.22)	**38.85**

Table 3: Class embedding enrichment: mAP@0.5 (as %) evaluated on COCO class embeddings ("COCO names") and on synonyms embeddings ("Synonyms").

Nebbia and Kovashka, NAACL 2024 Findings

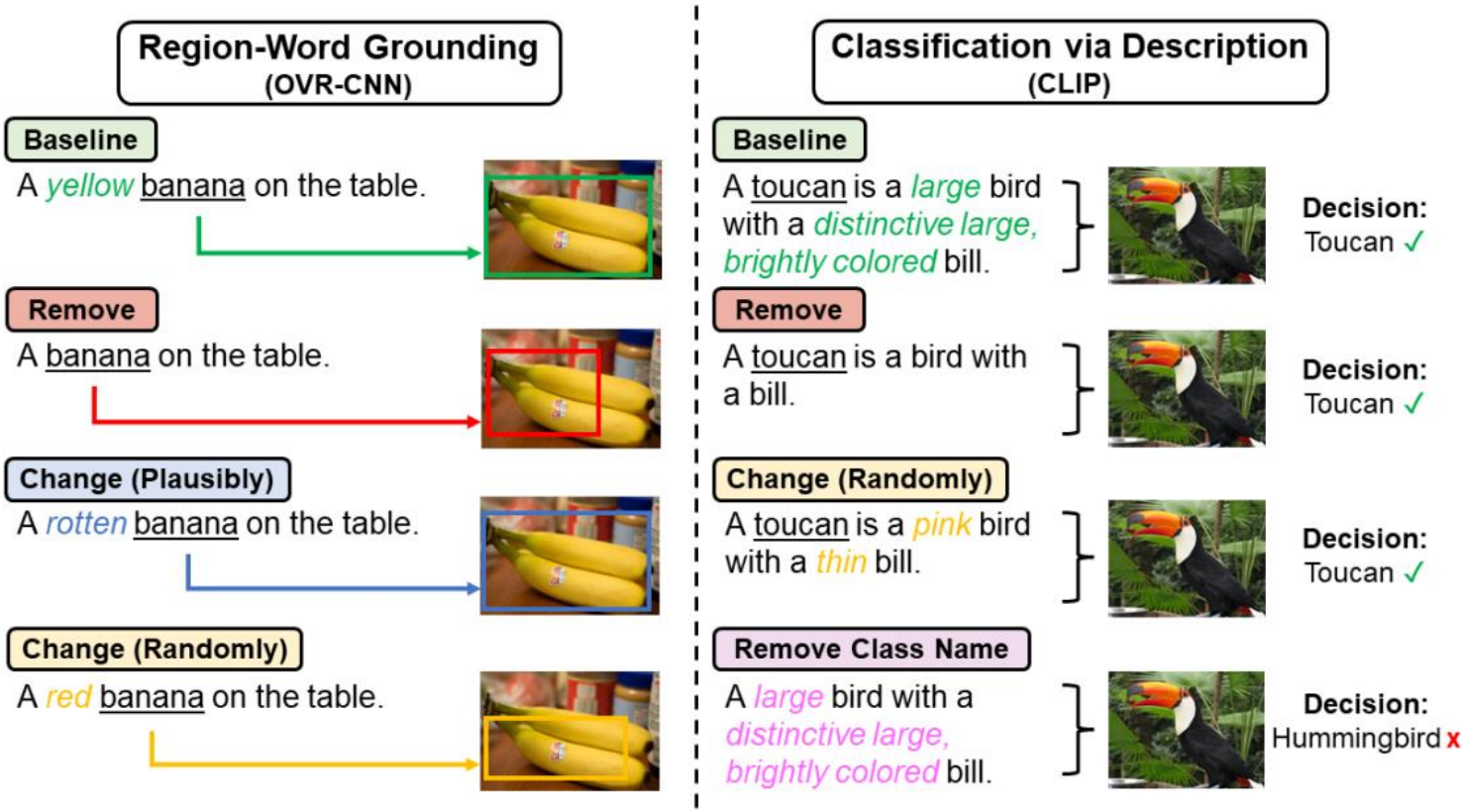
Investigating the Role of Attribute Context in Vision-Language Models for Object Detection

- **RQ1:** Does attribute context play a role in VL pretraining for **object detection**?
  - Not significantly by default
- **RQ2:** Does learning to ground objects to **contextualized word embeddings** utilize attribute meaning?
  - Not significantly, unless adjective-perturbed caption negatives are used
- **RQ3:** Do VL models perform well at tasks where objects are described with/in terms of attributes?
  - No, there is room for improvement
  - Using class descriptions without class names and with only attributes is notably not effective
- **RQ4:** Can **contrastive negative sampling of captions** increase a model's ability to use attribute context?
  - Yes, in both region-word alignment pretraining and CLIP finetuning
- **RQ5:** What sampling mechanisms are most effective?
  - In OVR-CNN pretraining, both plausible and random sampling are effective
  - In CLIP finetuning, random sampling is most effective

Buettner and Kovashka, WACV 2024

Measuring attribute sensitivity

- **Our method:** Measure performance under various text changes
  - **Baseline** – original caption/description
  - **Removing attributes –** losing helpful information – *does performance drop?*
  - **Changing attributes (plausibly/randomly)** – making text incorrect – *does performance drop?*
  - **Removing class names -** Is describing an object by attributes alone effective?

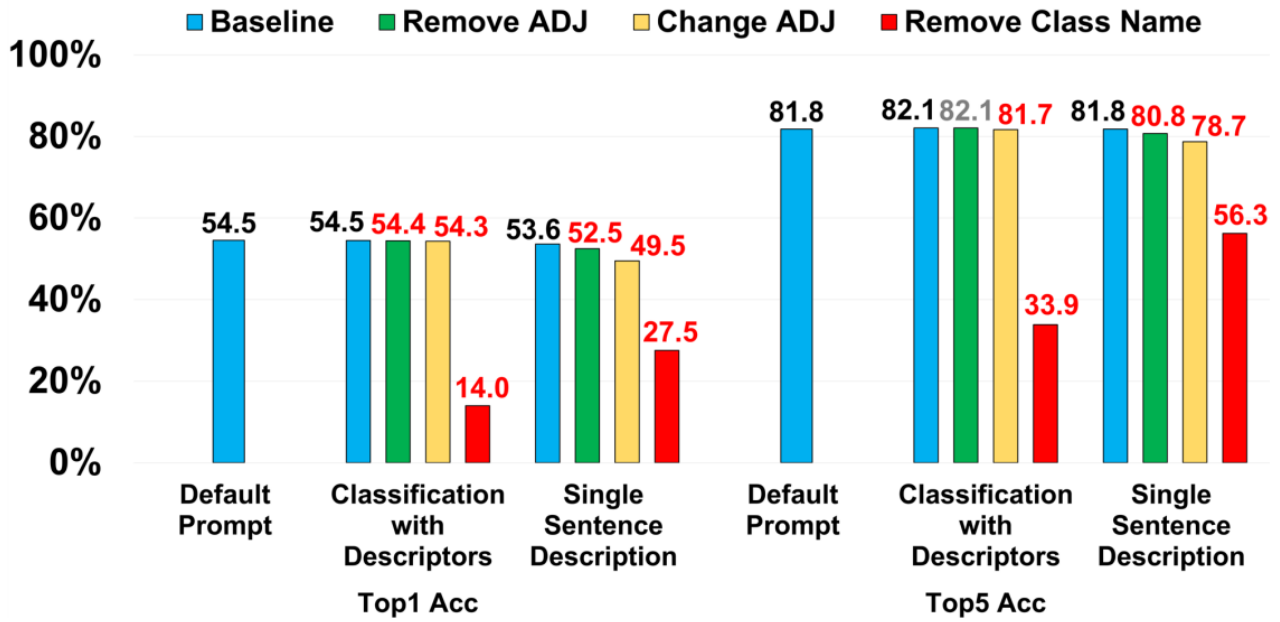


For this study, we consider attributes as adjective modifiers detected with a dependency parser.

Buettner and Kovashka, WACV 2024

To what extent can CLIP effectively use attribute information?

- We test classification via description[1,2] with ImageNet-V2[3]
  - We find limited sensitivity in removing/changing attribute settings
    - More sensitivity observed in single-sentence case
  - We find **removing class names** to significantly reduce performance
    - Attribute-only descriptions are not sufficient for class differentiation



University of Pittsburgh

[1] Menon and Vondrick, Visual classification via description from large language models, ICLR 2023.
[2] Pratt et al., What does a platypus look like? Generating customized prompts for zero-shot image classification, ICCV 2023.
[3] Recht et. al, Do ImageNet classifiers generalize to ImageNet?, ICML 2019.

Towards Shape-regularized Learning for Mitigating Texture Bias in CNNs

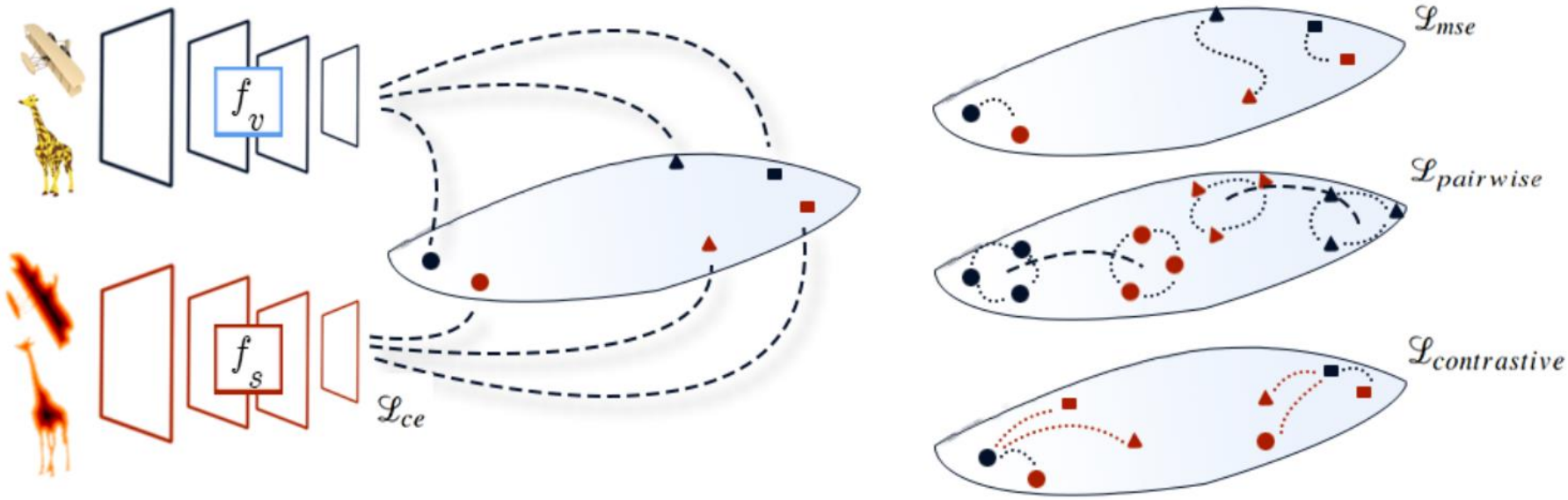


Figure 3: (Left) ResNet18 (f_v) is trained to learn image features using the shape-bias loss. A pretrained ResNet18 (f_s) is used as feature extractor for 2D distance transforms. Different categories are depicted by distinct shapes (▲, ●, ■) in the latent space. (Red denotes shape features. Blue denotes image features. (Right) The shape-bias loss is composed of 3 losses, (a) mean squared error (b) pairwise loss and (c) contrastive loss. In addition, we use cross-entropy loss (for supervised learning). Blue dotted lines denote distance minimization. Red dotted lines denote distance maximization.

Sinha and Kovashka, ICMR 2023

Recognizing objects across geography

- Objects (e.g. cleaning equipment, plates, beds) look different in different countries; can knowledge from LLMs help us bridge the gaps in visual appearance?

cleaning equipment: ['bucket shape', 'handles for carrying', 'bright colors such as blue or green', 'mop or broom attachments', 'cleaning solution bottles', 'sponges or scrub brushes', 'dustpans or trash bags']

bed: ['rectangular shape', 'wooden frame', 'raised platform', 'woven mats or fabrics on top', 'often decorated with shells or feathers', 'may have mosquito nets or curtains']

roof: ['pointed or sloping shape', 'made of thatch or palm leaves', 'different colors and textures', 'often raised on stilts or posts', 'may have decorative elements such as carvings or patterns']

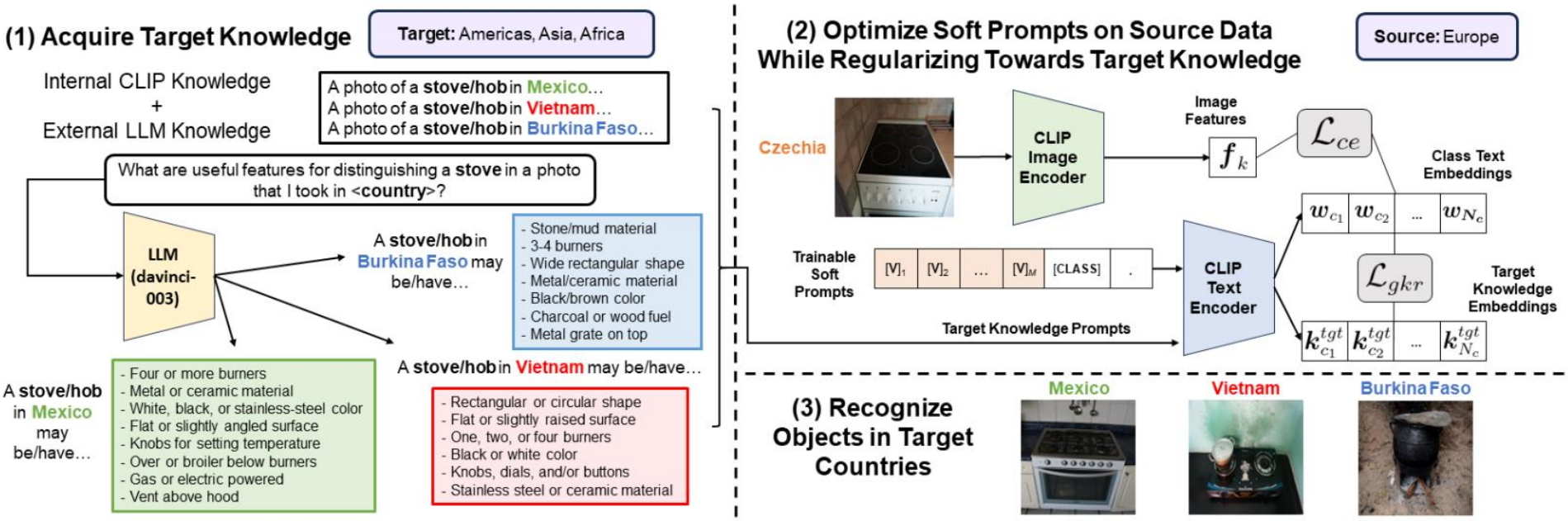
plate: ['round or oval shape', 'ceramic or clay material', 'intricate designs or patterns', 'earthy colors such as brown or beige', 'varying sizes and depths', 'may be used for serving food or as decorative objects']

pet: ['furry or feathered bodies', 'wagging tails or flapping wings', 'domesticated breeds such as cats or dogs', 'collars or leashes', 'obedience to human commands', 'friendly or playful behavior', 'common pets such as parrots or pigs.']

salt: ['white color', 'crystalline texture', 'irregular shapes', 'small granules', 'may appear in piles or mounds']

Figure 1. Examples of images from Papua New Guinea from DollarStreet [62] for which our proposed method allows us to confidently recognize the right category, while the baselines produce incorrect answers. We include information about the object's appearance in this country from a large language model, shown on the right of each image.

Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition



Buettner et al., CVPR 2024

Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition

Encoder	Prompting Method	*Top-1 Accuracy*										*Top-3 Accuracy*									
		Europe		**Africa**		**Asia**		**Americas**		**Total**		**Europe**		**Africa**		**Asia**		**Americas**		**Total**	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
ViT-B/32	Zero-Shot CLIP [36]	59.1	-	43.7	-	50.8	-	**55.3**	-	51.7	-	81.1	-	64.8	-	72.3	-	**77.4**	-	73.7	-
	GeneralLLM [30]	57.3	-1.8	44.3	+0.6	50.9	+0.1	54.6	-0.7	51.4	-0.3	78.8	-2.3	64.5	-0.3	72.1	-0.2	75.7	-1.7	73.0	-0.7
	CountryInPrompt	57.5	-1.6	45.2	+1.5	51.9	+1.1	55.0	-0.3	52.1	+0.4	80.2	-0.9	65.5	+0.7	73.3	+1.0	76.9	-0.5	73.9	+0.2
	CountryLLM	59.4	+0.3	45.2	+1.5	52.1	+1.3	**55.3**	0.0	52.6	+0.9	80.9	-0.2	66.4	+1.6	**73.6**	+1.3	**77.4**	0.0	74.6	+0.9
	CountryInPrompt+LLM	**60.8**	+1.7	**45.3**	+1.6	**52.2**	+1.4	55.0	-0.3	**52.8**	+1.1	**81.5**	+0.4	**67.4**	+2.6	**73.6**	+1.3	76.7	-0.7	**74.7**	+1.0
ViT-B/16	Zero-Shot CLIP [36]	64.3	-	46.9	-	53.9	-	**60.1**	-	55.5	-	84.3	-	69.3	-	75.9	-	81.1	-	77.2	-
	GeneralLLM [30]	64.2	-0.1	48.8	+1.9	**56.0**	+2.1	58.5	-1.6	56.8	+1.3	83.9	-0.4	71.1	+1.8	76.3	+0.4	80.4	-0.7	77.9	+0.7
	CountryInPrompt	63.9	-0.4	49.6	+2.7	55.7	+1.8	59.3	-0.8	56.6	+1.1	84.0	-0.3	71.3	+2.0	76.5	+0.6	80.0	-1.1	77.7	+0.5
	CountryLLM	65.2	+0.9	49.6	+2.7	55.6	+1.7	59.7	-0.4	57.0	+1.5	84.3	0.0	71.8	+2.5	**77.5**	+1.6	**81.5**	+0.4	**78.8**	+1.6
	CountryInPrompt+LLM	**65.5**	+1.2	**50.8**	+3.9	**56.0**	+2.1	59.7	-0.4	**57.4**	+1.9	**85.5**	+1.2	**72.5**	+3.2	77.0	+1.1	80.9	-0.2	78.7	+1.5
RN50	Zero-Shot CLIP [36]	53.0	-	38.0	-	44.4	-	49.8	-	45.7	-	76.5	-	60.2	-	66.4	-	**72.7**	-	68.1	-
	GeneralLLM [30]	55.5	+2.5	40.9	+2.9	46.9	+2.5	50.3	+0.5	47.9	+2.2	76.0	-0.5	61.2	+1.0	67.7	+1.3	71.1	-1.6	68.6	+0.5
	CountryInPrompt	54.5	+1.5	**43.4**	+5.4	47.0	+2.6	50.8	+1.0	48.4	+2.7	76.0	-0.5	**64.0**	+3.8	68.7	+2.3	**72.7**	0.0	**70.0**	+1.9
	CountryLLM	56.2	+3.2	41.1	+3.1	47.3	+2.9	50.4	+0.6	48.3	+2.6	**77.2**	+0.7	62.5	+2.3	**68.8**	+2.4	72.4	-0.3	**70.0**	+1.9
	CountryInPrompt+LLM	**56.4**	+3.4	43.0	+5.0	**48.0**	+3.6	**50.9**	+1.1	**49.1**	+3.4	76.7	+0.2	63.1	+2.9	68.3	+1.9	71.1	-1.6	69.4	+1.3

Table 1. **Zero-shot CLIP inference with descriptive knowledge prompts, top-1/3 balanced accuracy (Acc) on DollarStreet.** Strategies to capture CLIP's internal country knowledge (CountryInPrompt), external LLM country knowledge (CountryLLM), and their combination (CountryInPrompt+LLM), often improve vs. the zero-shot CLIP baseline (prompt "a photo of a/an <object>"), especially on Africa and Asia; gains in green, drops in red. CountryLLM notably outperforms the GeneralLLM [30] baseline.

Buettner et al., CVPR 2024

Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition

Encoder	Prompting Method	*Source* Europe		*Target* Africa		Asia		Americas		Total	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
ViT-B/16	CoOp [52]	72.2	-	53.9	-	61.5	-	68.6	-	61.7	-
	CoCoOp [51]	73.2	-	54.3	-	61.2	-	68.3	-	61.4	-
	KgCoOp [47]	73.1	-	54.4	-	62.6	-	68.7	-	62.4	-
	CountryInPrompt Reg	71.8	*-1.4*	56.8	+2.4	63.0	+0.4	69.8	+1.1	63.5	+1.1
	CountryLLM Reg	73.2	*0.0*	55.6	+1.2	63.0	+0.4	70.0	+1.3	63.2	+0.8
	CountryInPrompt+LLM Reg	**73.6**	*+0.4*	**57.2**	+2.8	**63.8**	+1.2	**70.3**	+1.6	**64.0**	+1.6
RN50	CoOp [52]	64.6	-	45.2	-	51.6	-	59.5	-	52.2	-
	CoCoOp [51]	62.9	-	44.5	-	51.0	-	58.3	-	51.4	-
	KgCoOp [47]	63.5	-	46.3	-	53.9	-	**60.5**	-	53.9	-
	CountryInPrompt Reg	63.5	*-1.1*	48.0	+1.7	53.9	0.0	60.3	-0.2	54.3	+0.4
	CountryLLM Reg	64.5	*-0.1*	47.4	+1.1	54.2	+0.3	59.9	-0.6	54.3	+0.4
	CountryInPrompt+LLM Reg	**65.5**	*+0.9*	**48.1**	+1.8	**54.5**	+0.6	60.4	-0.1	**54.8**	+0.9

Table 2. **Regularizing soft prompts with geographical knowledge, top-1 bal. acc. on DollarStreet.** We emphasize that our regularization aims to improve **target** performance, rather than source (gray, *italicized*). Gains/drops are shown vs. the *best* of soft prompt baselines (shaded). CountryInPrompt+LLM Reg achieves notable gains in target, especially on Africa. Methods use 16 shots per class.

Buettner et al., CVPR 2024

Recognizing objects across geography

- Language supervision has different structure across languages; not all languages mention all objects

English: A man and lady posing together
German: Young and happy appearing couple is sitting on an **office table** standing in the corner and **hugging**
Ukrainian: A man **hugs** a girl, they are sitting on a **white office table**

Czech: A young girl coming out of the **metro** in Prague in white clothes with dark glasses
Quechua: A woman in white **carries** her white robe in her hand
Swahili: A woman **carrying** her merchandise

Arabic: An African man stands in front of a fiber mill
Ukrainian: African man at the rope loom at the exhibition
Swahili: An African man showing a **cardamom tree** and some ropes from the tree in front of a motor machine

English: An aerial view of the city with red roof **buildings**
Arabic: Aerial view of a city

English: A macro shot view of spring **snowdrop** flowers
Quechua: **White** flowers with green leaves drying in the ground

English: **A pasta noodle cutter** on the wooden table
Farsi: A **metal** cooking utensil with a red wooden **handle**

Figure 3. Representative examples of language differences in captions from the XM3600 dataset [68]. Noteworthy differences in mentioned objects and actions are highlighted.

	eu-mean	eu-stdev	ar-mean	ar-stdev	hi-mean	hi-stdev	id-mean	id-stdev	asia-mean	asia-stdev	sw
tree(s)	270.5^{---}	92.9	349	19.799	**581.5**$^{+++}$	214.3	286	49.5	274.7	63.52	**383**
mountain(s)	171.1^{---}	47.78	183	24.0416	185.5	31.82	173	42.43	**218**$^{+++}$	16.52	**208**
street	**100.9**	30.22	**124**$^{+++}$	50.9117	61	7.071	38.5^{---}	10.61	76.67	19.01	82
car(s)	207.3	19.96	235	24.0416	**239**	50.91	204—	11.31	220	17.78	**270+++**
building(s)	244.8^{---}	69.3	281.5	40.3051	329	108.9	**383.5**	84.15	253.3	49.9	**502**$^{+++}$
restaurant	45.82	13.71	**54**	7.07107	19^{---}	5.657	**50.5**$^{+++}$	13.44	42.67	6.11	21
table	156.7	52.82	162.5	58.6899	**240**$^{+++}$	12.73	**228**	93.34	185.3	43.66	121^{---}
plate	112.5	25.93	90^{---}	12.7279	105.5	10.61	109.5	33.23	**119.3**$^{+++}$	5.132	**113**
box	18.14	4.454	15.5^{---}	0.70711	15.5	2.121	**28**$^{+++}$	4.243	**24**	2.646	18
bottle	10.23^{---}	2.654	12	0	10.5	2.121	11	4.243	**14.67**$^{+++}$	0.577	**18**
dog	26.23	5.108	28	1.41421	**31.5**$^{+++}$	4.95	29.5	0.707	20.67^{---}	5.508	**34**
woman	135.5	23.72	127	5.65685	114^{---}	31.11	**164.5**$^{+++}$	20.51	133.3	27.65	**160**

Table 1. Language shifts in terms of concept mentions in different languages. We grouped European languages (eu), Arabic and Farsi (ar), Hindi and Bengali (hi), Indonesian and Thai (id), Asian languages (asia), and report Swahili on its own (sw). The largest two numbers per row are bolded. Observe the differences between the language with highest (+++) and lowest (−−−) counts, which are significantly larger than the within-group standard deviations.

Recognizing objects across geography

- Need to recognize rare, culture-specific objects (with limited data)

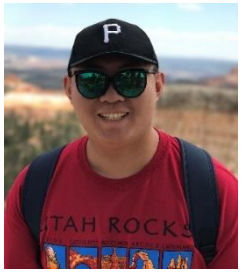
Figure 4. Rare objects unique to specific cultures, with their most distinctive visual attribute (and a related cateogory).

Acknowledgements

- Funding

- Students involved in highlighted projects

Keren Ye



Mingda Zhang

Chris Thomas

Erhan Unal

Kyle Buettner

Arushi Rai

Giacomo Nebbia