Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition (Supplementary Material)

Kyle Buettner[1], Sina Malakouti[2], Xiang Lorraine Li[1,2], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA
{buettnerk, sem238}@pitt.edu, {xianglli, kovashka}@cs.pitt.edu
https://krbuettner.github.io/GeoKnowledgePrompting

The supplementary material is organized as extensions to Sec 5.1 through 5.3 in the main paper (zero-shot inference, soft prompting, and further analysis). We provide in-depth experiments and ablations to support our main findings and method design. We also provide additional examples.

1. Zero-Shot Inference

Qualitative examples. In Fig. 1, we provide additional examples of zero-shot CLIP inference on DollarStreet using geography-specific descriptors (CountryLLM). It is demonstrated that geography knowledge is successfully able to capture diverse object forms and designs. The top activating descriptors in these examples highlight materials ("thatch" for *roof* in Myanmar, "glass" for *stove/hob* in Spain) and colors ("yellow"/"orange" for *spices* in India). LLM context enables CLIP to be probed for its own cultural knowledge (*e.g.* "traditional Chinese musical instrument" for *instrument* and "Chinese characters such as Kanji, Hanzi, or Pinyin" for *wall decoration*). Cultural conventions are better captured, exhibited by "squatting style" activating for *toilet* in Nepal (such form is not common in Western geographies). In error cases, some classes with related descriptors may be confused (*cooking pots* and *stove/hob*). Such errors suggest improvement is needed in CLIP's understanding of natural language concepts. Error cases may also result because of ambiguity with close categories, as shown by *home* vs. *roof* in Colombia (where the *home* descriptors seem fairly accurate). Nonetheless, it is interesting to observe that a successful prediction can occur even when descriptors from other categories strongly activate.

DollarStreet performance by country. The zero-shot, continent-level DollarStreet results in Table 1 of the main paper can be further broken down into country-level performance. We particularly show CountryInPrompt+LLM vs. GeneralLLM (i.e. full geography knowledge vs. general knowledge) with ViT-B/16 in Fig. 2. This figure notably exhibits per-country *overall* accuracy instead of balanced ac-

Figure 1. **Qualitative examples of success/failure cases (CountryLLM).** We show the prediction (green if correct, red if not) as well as the prediction confidence and the top 5 descriptors (with CLIP similarity scores) for each image. Encoder = ViT-B/16.

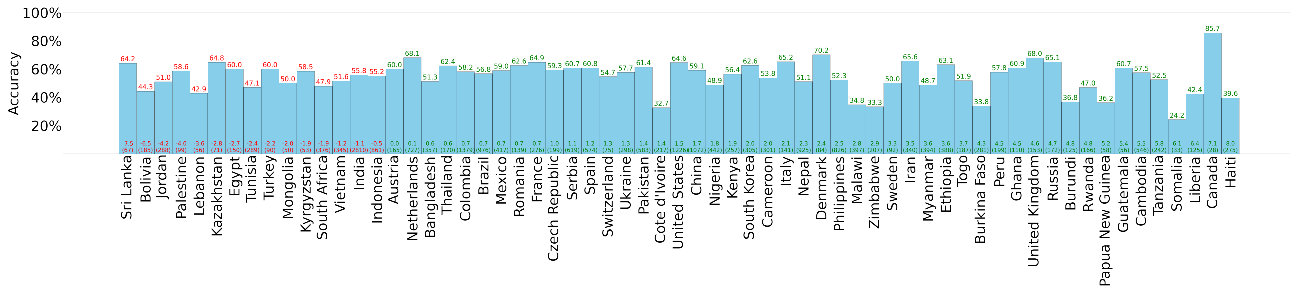


Figure 2. **Country-level overall accuracy in zero-shot inference with CountryInPrompt+LLM, gains/drops shown vs. GeneralLLM (descriptors not specific to geographies)**, for ViT-B/16. Note that geography knowledge integration is generally effective across countries, demonstrated by performance improvements in 48/63 countries. The overall accuracy over all countries is 55.8% for CountryInPrompt+LLM and 54.6% for GeneralLLM.

Encoder	Prompting Method	*Top-1 Accuracy* USA Acc	Δ	Asia Acc	Δ	*Top-3 Accuracy* USA Acc	Δ	Asia Acc	Δ
ViT-B/32	Zero-Shot CLIP [4]	50.3	-	46.2	-	69.6	-	66.1	-
	GeneralLLM [3]	49.9	-0.4	46.6	+0.4	69.1	-0.5	67.0	+0.9
	CountryInPrompt	50.0	-0.3	46.5	+0.3	69.1	-0.5	66.3	+0.2
	CountryLLM	50.3	0.0	47.7	+1.5	69.7	+0.1	67.1	+1.0
	CountryInPrompt+LLM	**51.2**	+0.9	**47.8**	+1.6	**69.8**	+0.2	**67.7**	+1.6
ViT-B/16	Zero-Shot CLIP [4]	53.9	-	50.2	-	72.8	-	69.2	-
	GeneralLLM [3]	54.6	+0.7	52.2	+2.0	73.4	+0.6	70.9	+1.7
	CountryInPrompt	54.6	+0.7	50.7	+0.5	73.3	+0.5	70.5	+1.3
	CountryLLM	54.7	+0.8	**52.5**	+2.3	73.5	+0.7	**71.1**	+1.9
	CountryInPrompt+LLM	**54.9**	+1.0	51.4	+1.2	**74.1**	+1.3	70.9	+1.7
RN50	Zero-Shot CLIP [4]	46.8	-	43.4	-	65.6	-	63.2	-
	GeneralLLM [3]	48.6	+1.8	45.4	+2.0	67.3	+1.7	65.5	+2.3
	CountryInPrompt	47.5	+0.7	43.9	+0.5	66.8	+1.2	63.7	+0.5
	CountryLLM	48.2	+1.4	**45.7**	+2.3	67.3	+1.7	65.2	+2.0
	CountryInPrompt+LLM	**49.2**	+2.4	45.4	+2.0	**67.8**	+2.2	**65.7**	+2.5

Table 1. **Zero-shot CLIP with descriptive knowledge prompts, top-1/3 balanced accuracy (Acc) on GeoNet.** Strategies to capture CLIP's internal country knowledge (CountryInPrompt), external LLM country knowledge (CountryLLM), and their combination (CountryInPrompt+LLM), improve the zero-shot CLIP baseline (prompt "a photo of a"). Gains in green, drops in red.

curacy due to limited examples per class per country. Here we see that compared to GeneralLLM, 48/63 countries have performance improvements with geography knowledge. 15 countries show drops, for which we posit a couple of reasons. For one, CLIP may have inadequate understanding of some detailed LLM descriptors due to imperfect vision-language alignment. Secondly, it is possible that the geography knowledge may be insufficient for some countries. Ensuring proper alignment between visual and language features and more adequate quality of LLM+VLM knowledge are important areas for future work.

GeoNet. We further show zero-shot inference with CLIP on GeoNet (test sets) in Table 1. The 10 most common countries in the "Asia" test set are used for our "Asia" evaluation to match the geography-specific LLM descriptors we acquire (i.e. China, India, Indonesia, Japan, Kenya, Malaysia, Singapore, Taiwan, Tanzania, and Thailand - GeoNet considers these as all "Asia"). Observe that the highest USA/Asia performance for each encoder is provided by one of the geography-specific prompting strategies. With ViT-B/32, CountryLLM has notably higher performance on Asia vs. GeneralLLM. With ViT-B/16 and RN50, CountryLLM improves vs. GeneralLLM on Asia, though top target differences are smaller compared to DollarStreet. We hypothesize that the general concept representations probed by LLMs and the ones respective to the top countries in Asia are relatively similar within GeoNet (*e.g.* 43% of images are from Japan/China, which are high-resource). In general, the default CLIP gaps between USA and Asia are not extremely large, indicating that CLIP has a notable degree of robustness on GeoNet. Also, unlike DollarStreet, some classes within GeoNet are mostly unique to geographies (*e.g. shoji* in traditional Japanese architecture), so cross-geography knowledge may not be as helpful.

2. Soft Prompting

Performance by income. Instead of reporting target performance by continent, in Table 2 we show results break-

Method	*Source* Europe		*Target* (Income Status) Low		Medium		High	
	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
CoOp [7]	72.2	-	44.3	-	61.6	-	71.1	-
CoCoOp [6]	73.2	-	44.4	-	61.4	-	70.5	-
KgCoOp [5]	73.1	-	43.4	-	62.5	-	72.6	-
CIP Reg	71.8	-1.4	46.0	+1.6	63.6	+1.1	72.4	-0.2
LLM Reg	73.2	0.0	45.2	+0.8	63.1	+0.6	**73.3**	+0.7
CIP+LLM Reg	**73.6**	+0.4	**46.8**	+2.4	**64.0**	+1.5	**73.3**	+0.7

Table 2. **Regularizing soft prompts with geography knowledge, top-1 bal. accuracy on DollarStreet, with target organized by income status.** Gains w.r.t. best baseline. Note that geo-diverse prompts help especially in the low-income scenario. CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM= CountryInPrompt+LLM. Encoder = ViT-B/16.

λ	*Source* Europe	*Target* Africa	Asia	Americas	Total
	Acc	Acc	Acc	Acc	Acc
2	73.1	55.9	63.0	**70.3**	63.4
4	**73.6**	57.2	**63.8**	**70.3**	**64.0**
6	72.8	**57.3**	63.5	**70.3**	63.8
8	72.7	56.7	63.1	68.9	63.1
10	70.7	55.4	61.6	68.1	62.0

Table 3. **Regularizing soft prompts with geography knowledge, top-1 bal. accuracy on DollarStreet, at varying values of λ.** Our method uses λ=4. Encoder = ViT-B/16.

ing down target performance by income (using the delineation of low, medium, and high-income buckets from [1]). While our method helps across all income levels, the gains are most significant in low-income scenarios.

Ablation: regularization weight. In Table 3, we show CountryInPrompt+LLM performance for various choices of λ. The highest total performance is achieved at $\lambda = 4$.

Experiment: source of knowledge. With CountryInPrompt+LLM, we test three different ways to select countries for knowledge aggregation (i.e. how to choose $\mathcal{G}_t$): (1) using unseen countries of interest (named *target*, with country count 49), (2) using countries seen during training (named *source*, with country count 14), and (3) using all countries in the dataset (named *all*, with country count 63). Shown in Table 4, we find that both *target* and *all* methods perform well on target geographies in comparison to source, and the target-only ensemble does best overall. This result indicates that including diverse knowledge of target countries best ensures geographical robustness across the world. With RN50, using a source-only ensemble performs poorly on Africa and Asia, but best on Americas (presumably due to some greater similarities, *e.g.* between the US, Canada, and Europe). Interestingly, we find that target regularization is best on the source test set, which we attribute to more domain-generalizable class representations achieved overall, given the use of diverse knowledge.

Comparison to *gpt-3.5-turbo*. In addition to *davinci-003*,

Encoder	$\mathcal{G}_t$	*Source* Europe	*Target* Africa	Asia	Americas	Total
		Acc	Acc	Acc	Acc	Acc
ViT-B/16	Target	**73.6**	**57.2**	**63.8**	**70.3**	**64.0**
	All	72.6	56.6	63.6	70.1	63.8
	Src	73.5	55.8	62.9	70.1	63.1
RN50	Target	**65.5**	**48.1**	**54.5**	60.4	**54.8**
	All	**65.5**	**48.1**	**54.5**	60.3	**54.8**
	Src	64.9	46.8	54.4	**60.6**	54.5

Table 4. **Regularizing soft prompts with geography knowledge, top-1 bal. accuracy on DollarStreet, with different countries in $\mathcal{G}_t$.** Comparisons are shown for CountryInPrompt+LLM at λ=4. Using a target country-only ensemble (49 countries) performs slightly better than using all countries (63 countries).

we test *gpt-3.5-turbo* (ChatGPT) as an LLM knowledge source. *gpt-3.5-turbo* notably needed significant prompt engineering to produce adequate descriptors. We use the following prompt for *gpt-3.5-turbo*:

> Task: For an object/concept name and country name provided, very concisely provide up to 10 visual features that can distinguish that object in a photo taken in that specific country. The key is to make sure the descriptions capture an object's key visual attributes and properties across the country. Examples include colors, textures, shapes, materials used, parts/components, common context/background, size, and possible designs/forms across the country. Consider common attributes specific to objects in that country and ensure descriptor diversity to represent regions with low socioeconomic status.
> These are strict output requirements:
> - Each description should be simple and interpretable by a child
> - Use only a few words per descriptor
> - Start directly in the form of a bulleted list
> - The output should complete this sentence: "A/an <object> which (is/has/etc.)"
> - Be specific, qualifying with visual adjectives, and do not be vague or general at all
> - Adjectives like "unique"/"diverse"/"distinctive" are not specific enough to help distinguish an object in a photo, so do not use them
> - Specific EX: "red color"/"small size"/"wooden hand"
> Use this example as a reference...
> To tell there is a bathtub in a photo in Japan, the following visual features are helpful:
> - short in length and deep
> - square shape
> - wooden, plastic, or steel material
> - white or brown color
> - benches on side
> - next to shower
> Now complete:
> To tell there is <object> in a photo in <country>, the following visual features are helpful:

Encoder	LLM	*Source* Europe Acc	*Target* Africa Acc	Asia Acc	Amer. Acc	Total Acc
ViT-B/16	*davinci003*	**73.6**	**57.2**	**63.8**	**70.3**	**64.0**
	gpt-3.5-turbo	71.9	**57.2**	63.1	69.9	63.6
RN50	*davinci003*	**65.5**	48.1	**54.5**	60.4	**54.8**
	gpt-3.5-turbo	64.8	**48.3**	54.4	**60.6**	**54.8**

Table 5. **Regularizing soft prompts with geography knowledge, top-1 bal. accuracy on DollarStreet, *davinci003* vs. *gpt-3.5-turbo*.** Comparisons are shown for CountryInPrompt+LLM at λ=4. While *davinci003* is observably more performant with ViT-B/16, our strategy also works well with *gpt-3.5-turbo*.

Method	*Source* Americas Acc	*Target* Africa Acc	Asia Acc	Europe Acc	Total Acc
CoOp	71.1	56.4	62.5	72.1	64.2
CoCoOp	70.7	55.8	62.4	72.2	64.2
KgCoOp	72.7	56.2	63.0	72.9	64.0
CIPReg	71.4 -1.3	57.4 +1.0	63.5 +0.5	72.6 -0.3	64.6 +0.4
LLMReg	71.7 -1.0	**57.6** +1.2	63.8 +0.8	73.3 +0.4	64.8 +0.6
CIP+LLMReg	**73.0** +0.3	57.4 +1.0	**64.0** +1.0	**73.4** +0.5	**65.1** +0.9

Table 6. **Geo knowledge regularization on DollarStreet, src = Americas, tgt = Africa,Asia,Europe.** Encoder = ViT-B/16.

Table 5 shows our findings comparing *gpt-3.5-turbo* vs. *davinci-003* with geography knowledge regularization. *davinci-003* is clearly the top LLM with respect to ViT-B/16, but results are more comparable with RN50. We leave a more rigid comparison study of LLM world knowledge for future work.

Experiment: America as source. We test America as a different source than Europe and show results in Table 6. The results show improvements across target continents, exhibiting that our approach generalizes to other sources.

Experiment: Choice of ensemble. Table 7 shows results when varying the ensemble to be only from certain continents. Notably, having diverse continents (Am/Af/As) leads to top performance overall. A diverse ensemble may allow for a more domain-generalizable representation overall. Regarding single-continent performance, it is observed that the Am/As ensembles do not result in the top Am/As performance respectively. We believe that countries in other continents can be informative for learning (*e.g.* if multiple countries use *adobe* for houses). Future fine-grained analysis can be done to identify optimal ensemble properties.

Experiment: Breaking down Americas. Having observed drops for Americas/RN50 in Table 2 of the main paper, we breakdown overall accuracy for North America (USA, Canada, Mexico) and Central/South America (Haiti, Bolivia, Brazil, Colombia, Peru, Guatemala) in Table 8. Our method improves over KgCoOp in Central/South America for both encoders, but not in North America for RN50,

$\mathcal{G}_t$	n	*Source* Eu Acc	*Target* Af Acc	As Acc	Am Acc	Total Acc
Am/Af/As	49	**73.6**	**57.2**	**63.8**	70.3	**64.0**
Am	9	73.4	56.9	63.6	70.0	63.7
Af	18	73.2	57.0	63.6	**70.4**	63.9
As	22	73.4	57.0	63.4	70.3	63.6

Table 7. **Varying geo. ensemble ($\mathcal{G}_t$) for CIP+LLMReg method, on DollarStreet.** Encoder=ViT-B/16. n=# countries in ensemble.

Region	n	ViT-B/16 KgCoOp Acc	ViT-B/16 CIP+LLMReg Acc	RN50 KgCoOp Acc	RN50 CIP+LLMReg Acc
SA/CA	2,795	66.6	**68.4** +1.8	57.1	**57.9** +0.8
NA	1,946	69.4	**70.0** +0.6	**60.9**	60.0 -0.9

Table 8. **Geo knowledge regularization on DollarStreet, perf. on North vs. South/Central Am. Src = Eu.** n = # of test images.

which explains the overall drops. We reason that North America does not benefit from knowledge constraints due to CLIP already being well-aligned to images in countries like the USA.

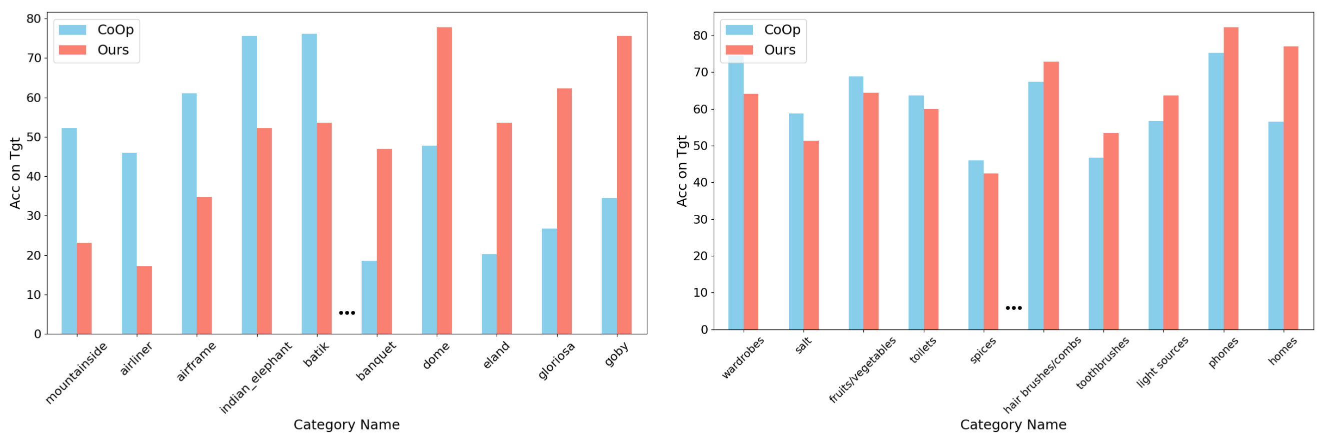
Classes by difficulty by continent: DollarStreet. In the main paper (Table 4), we show performance on “difficult” classes for CoOp overall. We provide further results analyzing performance on difficult classes with respect to each continent in DollarStreet, shown in Table 9. Similarly, our top method provides the top gains on the difficult classes, i.e. the <40% scenario, across every continent.

Classes with the most impact: GeoNet & DollarStreet. In Figure 3, we show classes where our regularization method has the most impact, in both the positive direction (gains) and negative direction (drops). On DollarStreet, it is notably effective for *homes*, which vary in appearance and construction materials across regions. On GeoNet, it benefits categories like *goby* (a type of fish), *gloriosa* (a type of flower), and *dome*, with domes differing in color between the USA (typically gray) and Asian countries (often yellow and orange). While *goby* and *gloriosa* generally look consistent worldwide, their images may experience context shifts due to environmental differences. On the other hand, general categories such as *airliner*, *mountainside*, and *salt* are adversely affected by geographical knowledge regularization. Ensuring good performance across all classes, perhaps through considering the adaptation of class representations at a finer-grained level, is needed in future work.

Classes by difficulty: GeoNet. We show a per-class breakdown of our knowledge regularization on GeoNet in Table 10. Like with DollarStreet in Table 4 of main, we show the thresholds t=40,60,80,100; however, since GeoNet has a large number of classes, we also show t=5 for a more

Method	Africa Threshold t (# Classes)								Asia Threshold t (# Classes)								Americas Threshold t (# Classes)							
	<40% (30)	Δ	**<60%** (55)	Δ	**<80%** (82)	Δ	**≤100%** (94)	Δ	**<40%** (18)	Δ	**<60%** (42)	Δ	**<80%** (78)	Δ	**≤100%** (95)	Δ	**<40%** (3)	Δ	**<60%** (30)	Δ	**<80%** (69)	Δ	**≤100%** (92)	Δ
CoOp [7]	27.4	-	37.7	-	48.5	-	53.9	-	30.9	-	42.1	-	55.6	-	61.5	-	24.9	-	50.8	-	61.9	-	68.6	-
CoCoOp [6]	27.4	0.0	37.5	-0.2	48.7	+0.2	54.3	+0.4	32.6	+1.7	42.3	+0.2	55.4	-0.2	61.2	-0.3	33.1	+8.2	51.9	+1.1	61.5	-0.4	68.3	-0.3
KgCoOp	28.0	+0.6	39.2	+1.5	49.3	+0.8	54.4	+0.5	34.7	+3.8	44.5	+2.4	57.3	+1.7	62.6	+1.1	38.1	+13.2	54.2	+3.4	62.6	+0.7	68.7	+0.1
CIPReg	31.9	+4.5	41.5	+3.8	51.7	+3.2	56.8	+2.9	35.9	+5.0	45.1	+3.0	57.8	+2.2	63.0	+1.5	40.7	+15.8	55.6	+4.8	63.5	+1.6	69.8	+1.2
LLMReg	29.0	+1.6	40.1	+2.4	50.3	+1.8	55.6	+1.7	35.4	+4.5	44.2	+2.1	57.3	+1.7	63.0	+1.5	39.6	+14.7	55.7	+4.9	64.1	+2.2	70.0	+1.4
CIP+LLMReg	**32.0**	+4.6	**42.0**	+4.3	**51.9**	+3.4	**57.2**	+3.3	**37.3**	+6.4	**46.0**	+3.9	**58.4**	+2.8	**63.8**	+2.3	**46.5**	+21.6	**56.8**	+6.0	**64.4**	+2.5	**70.3**	+1.7

Table 9. **Performance on DollarStreet classes with less than t% recall in CoOp, respective to continents**, with ViT-B/16. Shown are gains/drops w.r.t. CoOp. Our top method improves greatest in the <40% scenario, in *every* continent scenario. CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM= CountryInPrompt+LLM.



(a) **GeoNet** (b) **DollarStreet**

Figure 3. **Classwise comparison of Ours (CountryInPrompt+LLM) vs. CoOp on GeoNet and Dollarstreet.** We sort the top 25 most frequent categories based on $\Delta = Acc(Ours) - Acc(CoOp)$ and show the bottom 5 and top 5 categories. Leveraging geographical knowledge can be most useful on objects that may have geography-specific characteristics, while it may hurt generic category performance.

Method	Threshold t (# Classes)									
	<5% (36)	Δ	**<40%** (212)	Δ	**<60%** (345)	Δ	**<80%** (483)	Δ	**≤100%** (600)	Δ
CoOp [7]	1.4	-	19.6	-	30.7	-	41.9	-	51.2	-
CoCoOp [6]	4.4	+3.0	22.2	+2.6	33.5	+2.8	43.7	+1.8	52.6	+1.4
KgCoOp [5]	4.3	+2.9	22.4	+2.8	34.3	+3.6	43.9	+2.0	52.6	+1.4
CIPReg	6.7	+5.3	**24.2**	+4.6	35.5	+4.8	45.0	+3.1	53.5	+2.3
LLMReg	3.6	+2.2	23.2	+3.6	34.8	+4.1	44.5	+2.6	53.1	+1.9
CIP+LLMReg	**7.0**	+5.6	**24.2**	+4.6	**35.8**	+5.1	**45.4**	+3.5	**53.9**	+2.7

Table 10. **Performance on GeoNet classes with less than t% recall in CoOp**, with ViT-B/16. Gains w.r.t. CoOp of geography knowledge regularization are especially large for CoOp's difficult classes (+5.6 in <5%). CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM= CountryInPrompt+LLM.

aggressive threshold. We observe a similar observation that our method performs well on the most challenging classes at t=5 (+5.6% vs. CoOp baseline). CountryInPrompt regularization appears to drive performance of the hard classes in this case; CountryLLM regularization provides more evenly distributed improvements across thresholds.

3. Further Analysis

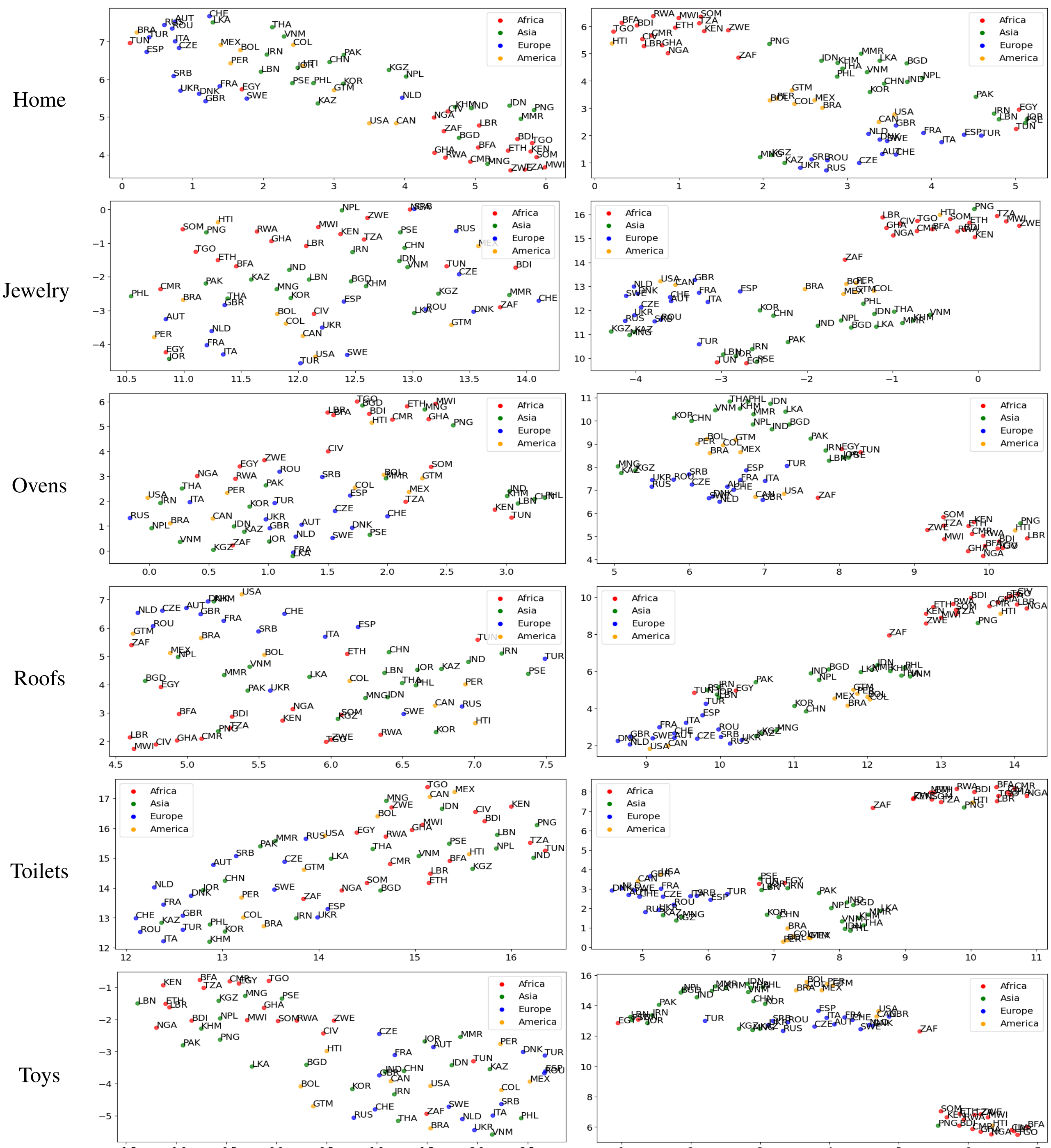
Descriptor topics. In Table 11, we show examples of words that appear amongst the geography-specific LLM descriptors for various DollarStreet categories. There exist some significant differences. For instance, for *toilet*, "squat" appears multiple times in Africa and Asia descriptors, but not in European descriptors. Similarly, for *roof*, "thatch" is more common in Africa and Asia than Europe and Americas. There are also some notable concepts that are common across regions, such as a *toilet* being "white" and *roof* being "metal". We advocate for future work that ensures factuality and proper representativeness of such concepts to extend utility to various regions.

UMAP for further categories. In Figure 4, UMAP [2] visualization is used to compare the class text embeddings of CountryLLM and CountryInPrompt+LLM across various DollarStreet classes. With CountryLLM, we note that LLM descriptors are often more alike among countries within the same continent than between different continents. For instance, due to cultural differences, people may use different

Class	Descriptor	Eu		Af		As		Am	
		Count	Rel	Count	Rel	Count	Rel	Count	Rel
home	stone	7	0.47	2	0.11	5	0.24	2	0.22
	mud	0	0.00	16	0.89	6	0.29	0	0.00
	bright colors	2	0.13	12	0.67	8	0.38	2	0.22
	balcony	14	0.93	2	0.11	14	0.67	8	0.89
	flower	7	0.47	0	0.00	0	0.00	3	0.33
toilet	squat	0	0.00	6	0.33	4	0.19	1	0.11
	white	15	1.00	14	0.78	20	0.95	9	1.00
	bidet	7	0.47	1	0.06	9	0.43	3	0.33
	button	6	0.40	2	0.11	5	0.24	1	0.11
	ceramic	8	0.53	12	0.67	14	0.67	5	0.56
roof	thatch	1	0.07	9	0.50	10	0.48	3	0.33
	straw	0	0.00	12	0.67	4	0.19	0	0.00
	terracotta	4	0.27	1	0.06	4	0.19	3	0.33
	metal	4	0.27	8	0.44	9	0.43	6	0.67
	clay	3	0.20	8	0.44	8	0.38	5	0.56

Table 11. **Example descriptor topics.** For various DollarStreet classes, we show examples of common words in the LLM descriptor sets across countries (grouped by continent). Count is the overall frequency within a continent, while rel. is the relative count (normalized by amount of countries in continent in DollarStreet). Country counts: Eu 15, Af 18, As 21, Am 9.

kinds of toilets in Africa compared to European countries. CountryInPrompt+LLM on the other hand shows much tighter clusters of countries, especially intra-continent, due to the addition of CLIP's internal knowledge.



(a) CountryLLM

(b) CountryInPrompt + LLM

Figure 4. **UMAP plots for various DollarStreet categories, CountryLLM vs. CountryInPrompt+LLM**. Country-specific class text embeddings often are close to those of neighboring countries. When CLIP's internal knowledge is added from (a) to (b) with the addition of country names, the clusters tighten.

References

[1] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 70–88, 2022.

[2] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[3] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations, ICLR*, 2023.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[5] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16816–16825, 2022.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.