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

Kyle Buettner¹, Sina Malakouti², Xiang Lorraine Li¹,², Adriana Kovashka¹,²
¹Intelligent Systems Program, ²Department 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.

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.

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Prompting Method</th>
<th>USA Acc</th>
<th>USA ∆</th>
<th>Asia Acc</th>
<th>Asia ∆</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/32</td>
<td>Zero-Shot CLIP [4]</td>
<td>50.3</td>
<td>-</td>
<td>46.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GeneralLLM [3]</td>
<td>49.9</td>
<td>-0.4</td>
<td>46.1</td>
<td>-0.5</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt</td>
<td>50.0</td>
<td>0.0</td>
<td>46.5</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>CountryLLM</td>
<td>50.3</td>
<td>0.0</td>
<td>47.7</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt+LLM</td>
<td>51.2</td>
<td>0.9</td>
<td>47.8</td>
<td>1.6</td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>Zero-Shot CLIP [4]</td>
<td>53.9</td>
<td>-</td>
<td>50.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GeneralLLM [3]</td>
<td>54.6</td>
<td>+0.7</td>
<td>52.2</td>
<td>+2.0</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt</td>
<td>54.7</td>
<td>+0.8</td>
<td>50.7</td>
<td>+2.3</td>
</tr>
<tr>
<td></td>
<td>CountryLLM</td>
<td>54.9</td>
<td>+1.0</td>
<td>51.4</td>
<td>+2.0</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt+LLM</td>
<td>54.9</td>
<td>+1.0</td>
<td>51.4</td>
<td>+2.0</td>
</tr>
<tr>
<td>RN50</td>
<td>Zero-Shot CLIP [4]</td>
<td>46.8</td>
<td>-</td>
<td>43.4</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GeneralLLM [3]</td>
<td>48.6</td>
<td>+1.8</td>
<td>45.4</td>
<td>+2.0</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt</td>
<td>47.5</td>
<td>+0.7</td>
<td>43.9</td>
<td>+0.5</td>
</tr>
<tr>
<td></td>
<td>CountryLLM</td>
<td>48.2</td>
<td>+1.4</td>
<td>45.7</td>
<td>+2.3</td>
</tr>
<tr>
<td></td>
<td>CountryInPrompt+LLM</td>
<td>49.2</td>
<td>+2.4</td>
<td>45.4</td>
<td>+2.0</td>
</tr>
</tbody>
</table>

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-
Table 2. Regularizing soft prompts with geography knowledge, top-1 bal. accuracy on DollarStreet, with target organized by income status. Gains w.r.t. 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.

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

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

**Ablation: regularization weight.** In Table 3, we show CountryInPrompt+LLM performance for various choices of \( \lambda \). 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 \( G_t \)): (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, 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 test 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 metal material
- white or brown color
- benches on side
- next to shower

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

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

Table 7. Varying geo. ensemble ($G_t$) for CIP+LLMReg method, on DollarStreet. Encoder=ViT-B/16. n=# countries in ensemble.

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
Table 9. Performance on DollarStreet classes with less than 1% recall in CoOp, respective to continents. with ViT/B16. Shown are gains/losses 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Africa Threshold t (# Classes)</th>
<th>Asia Threshold t (# Classes)</th>
<th>Americas Threshold t (# Classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;40% (30) ∆55% (82) ∆94% ∆</td>
<td>&lt;40% (18) ∆42% ∆78% ∆95% ∆</td>
<td>&lt;40% (30) ∆69% ∆92% ∆</td>
</tr>
<tr>
<td>CoOp [7]</td>
<td>27.4 - 37.7 - 48.5 - 53.9 -</td>
<td>36.0 - 42.3 - 55.6 - 61.5 -</td>
<td>33.1 - 42.0 - 45.6 - 51.9 - 61.5 -</td>
</tr>
<tr>
<td>CoCoOp [6]</td>
<td>27.4 0.0 37.5 - 0.2 48.7 - 0.2 54.3 - 0.4</td>
<td>32.6 ± 1.4 ± 42.3 ± 0.2 55.4 ± 0.2</td>
<td>31.1 ± 0.2 ± 42.0 ± 0.2 45.1 ± 0.2</td>
</tr>
<tr>
<td>KgCoOp</td>
<td>28.0 ± 0.6 ± 39.2 ± 0.4 ± 49.3 ± 0.8 ± 54.4 ± 0.5</td>
<td>34.7 ± 3.8 ± 44.4 ± 2.4 ± 57.3 ± 2.9</td>
<td>38.1 ± 0.4 ± 44.0 ± 1.0 ± 57.3 ± 1.0</td>
</tr>
<tr>
<td>CIPReg</td>
<td>31.9 ± 4.0 ± 41.5 ± 3.8 ± 51.7 ± 3.2 ± 56.8 ± 2.9</td>
<td>35.9 ± 5.0 ± 45.1 ± 3.0 ± 57.8 ± 2.9</td>
<td>38.1 ± 0.3 ± 42.0 ± 0.2 ± 56.8 ± 0.2</td>
</tr>
<tr>
<td>LLMReg</td>
<td>29.0 ± 1.4 ± 41.0 ± 2.4 ± 50.3 ± 1.3 ± 55.6 ± 1.7</td>
<td>35.4 ± 4.5 ± 44.2 ± 2.2 ± 57.3 ± 1.7</td>
<td>38.1 ± 0.4 ± 44.0 ± 1.0 ± 57.3 ± 1.0</td>
</tr>
<tr>
<td>CIP+LLMReg</td>
<td>32.0 ± 4.0 ± 42.0 ± 4.0 ± 51.9 ± 3.9 ± 57.2 ± 3.9</td>
<td>37.3 ± 6.0 ± 46.0 ± 3.9 ± 58.4 ± 2.9</td>
<td>40.6 ± 3.8 ± 48.1 ± 3.9 ± 58.4 ± 2.9</td>
</tr>
</tbody>
</table>

Table 10. Performance on GeoNet classes with less than 1% recall in CoOp, with ViT/B16. Gains w.r.t. CoOp of geography knowledge regularization are especially large for CoOp’s difficult classes (+5.6 in <5% scenario). CIP = CountryInPrompt, LLM = CountryLLM, CIP+LLM = CountryInPrompt+LLM.

<table>
<thead>
<tr>
<th>Method</th>
<th>Threshold t (# Classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;5% (36) ∆</td>
</tr>
<tr>
<td>CoOp [7]</td>
<td>1.4</td>
</tr>
<tr>
<td>CoCoOp [6]</td>
<td>4.4 ± 1.8</td>
</tr>
<tr>
<td>KgCoOp [5]</td>
<td>4.3 ± 1.6</td>
</tr>
<tr>
<td>CIPReg</td>
<td>6.7 ± 5.3</td>
</tr>
<tr>
<td>LLMReg</td>
<td>3.6 ± 2.2</td>
</tr>
<tr>
<td>CIP+LLMReg</td>
<td>7.0 ± 4.6 ± 24.2 ± 4.6</td>
</tr>
</tbody>
</table>

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
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.
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


