Can Vision-Language Models be a Good Guesser?
Exploring VLMs for Times and Location Reasoning
Supplementary Materials

Gengyuan Zhang\textsuperscript{1,2} Yurui Zhang\textsuperscript{3} Kerui Zhang\textsuperscript{1} Volker Tresp\textsuperscript{1,2}
\textsuperscript{1} LMU Munich, Munich, Germany
\textsuperscript{2} Munich Center for Machine Learning, Munich, Germany
\textsuperscript{3} Technical University of Munich
zhang@dbs.ifi.lmu.de

A. Dataset WikiTiLo
A.1. List of countries in WikiTiLo

The countries included in WikiTiLo and their regions are listed in Tab. 1. These countries are almost evenly distributed in 7 regions defined by their cultural and geographical affinity with reference of UNESCO\textsuperscript{1} and sorted alphabetically.

<table>
<thead>
<tr>
<th>Country</th>
<th>Region</th>
<th>Country</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Middle East</td>
<td>Argentina</td>
<td>Latin America</td>
</tr>
<tr>
<td>Australia</td>
<td>NA, EU, and OC</td>
<td>Brazil</td>
<td>Latin America</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>Southern Asia</td>
<td>China</td>
<td>Eastern Asia</td>
</tr>
<tr>
<td>Germany</td>
<td>NA, EU, and OC</td>
<td>India</td>
<td>Southern Asia</td>
</tr>
<tr>
<td>Indonesia</td>
<td>South-Eastern Asia</td>
<td>Iran</td>
<td>Middle East</td>
</tr>
<tr>
<td>Japan</td>
<td>Eastern Asia</td>
<td>Kazakhstan</td>
<td>Central Asia</td>
</tr>
<tr>
<td>Kenya</td>
<td>Sub-Saharan Africa</td>
<td>Kyrgyzstan</td>
<td>Central Asia</td>
</tr>
<tr>
<td>Malaysia</td>
<td>South-Eastern Asia</td>
<td>Mexico</td>
<td>Latin America</td>
</tr>
<tr>
<td>Nigeria</td>
<td>Sub-Saharan Africa</td>
<td>North Korea</td>
<td>Eastern Asia</td>
</tr>
<tr>
<td>Pakistan</td>
<td>Middle East</td>
<td>Rwanda</td>
<td>Sub-Saharan Africa</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>Middle East</td>
<td>South Africa</td>
<td>Sub-Saharan Africa</td>
</tr>
<tr>
<td>South Korea</td>
<td>Eastern Asia</td>
<td>Sri Lanka</td>
<td>Southern Asia</td>
</tr>
<tr>
<td>Tajikistan</td>
<td>Central Asia</td>
<td>Thailand</td>
<td>South-Eastern Asia</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>Central Asia</td>
<td>United States</td>
<td>NA, EU, and OC</td>
</tr>
<tr>
<td>Uzbekistan</td>
<td>Central Asia</td>
<td>Vietnam</td>
<td>South-Eastern Asia</td>
</tr>
</tbody>
</table>

Table 1. Countries with corresponding regions (NA, EU, and OC is the abbreviation of North America, Europa, and Oceania).

A.2. Data curation

In order to guarantee that WikiTiLo comprises images that are characteristics of socio-cultural visual hints, we conduct a manual image curation based on image visual cues on raw images in Wikimedia Commons, as in Fig. 1. We try to ensure that the space identity and time period of each image can be distinguished from the architectural patterns, costume styles, language types, movement postures, photo colors, and quality, or other fine-grained features.

A.3. Data distribution

The dataset distribution in location and times can be found in Fig. 2.

B. Visual encoders of discriminative VLMs

We compare all the visual encoders of discriminative Vision Language Models and Vision Models we used for references in the paper in the dimension of the dataset, visual encoder, and textual encoder in Tab. 2.

---

\textsuperscript{1}https://population.un.org/wpp/DefinitionOfRegions/
C. Impact of shot numbers

We studied the impact of shot number for OpenFlamingo as in Fig. 3. Especially for \textsc{ReasoningTime}s, we find the output prediction is more unstable when having more in-context shots and deteriorates the performance.

D. Visualization

We show the visualization of the transportation plan of word patch alignment on times classification as Fig. 7 in the main body. For Time-relevant questions, the attended patches seem less specific. Generally, visual tokens in the background instead of foreground objects have seemingly dominant contributions.

E. Prompts used for generative VLMs on \textsc{ReasoningTime}s

We also list all the prompts used for OpenFlamingo and LLaMA-Adapter V2 used for \textsc{ReasoningTime}s in Fig. 5 as in the paper for references.

F. Rationale examples for Chain-of-Thought

We annotate a subset of images with rationale in \textsc{Reasoning} tasks for OpenFlamingo Chain-of-Thought. Here, we showcase some examples of images and the rationale associated. We attempt to include visual details that are relevant for reasoning about times and locations for humans.

G. Case study

G.1. Failure on reasoning regions

We observe that generative VLMs perform worse in reasoning regions than reasoning countries, which is against intuition. We conduct a qualitative case study on the failure cases as in Fig. 8. It is shown that generative VLMs actually cannot really ground the reasoning process, especially in two-step reasoning. Even if the model gives correct reasoning that implies the country, it still fails to correlate to the corresponding countries. This again shows us the performance of models contained by the language models.

G.2. Dataset bias

We compare the model prediction of original images and transfer the images into three common image styles: low
Figure 4. Visualization of transportation plan of word patch alignment on times classification. Best viewed zoomed in. Rows from top to bottom: ViLT, CLIP, and BLIP. Columns from left to right: Afghanistan(Middle East) in 2000, Argentina(Latin America) in 1980, Japan(Eastern Asia) in 1940, and Germany(Europe) in 1880.

OpenFlamingo Cloze Test
<image> Output: This is a historical photo taken in the 19th Century.
Short answer: in the 19th Century

OpenFlamingo VQA
We divide time into 4 eras. These 4 eras are in the 19th Century, between 1900 and 1950, between 1950 and 2000, in the 21st Century.
<image> Question: When was this photo taken?
Short answer: in the 21st Century

OpenFlamingo VQA - CoT
We divide time into 4 eras. These 4 eras are in the 19th Century, between 1900 and 1950, between 1950 and 2000, in the 21st Century.
<image> Question: When was this photo taken?
Answer: Because the people in this photograph are dressed in attire typical of the Qing Dynasty in China. Therefore, it can be inferred that this photograph was taken during the Qing Dynasty. this photo was taken in the 19th Century.

OpenFlamingo VQA - CoT
The photograph was taken in one of the following eight areas. These 8 areas are "Central Asia," "Southern Asia," "Latin America," "Northern America, Europe and Oceania," "Middle East," "Eastern Asia," "South-Eastern Asia," "Sub-Saharan Africa."
<image> Question: In which area was this photograph taken?
Short answer: Southern Asia

OpenFlamingo VQA - CoT
The photograph was taken in one of the following eight areas. These 8 areas are "Central Asia," "Southern Asia," "Latin America," "Northern America, Europe and Oceania," "Middle East," "Eastern Asia," "South-Eastern Asia," "Sub-Saharan Africa."
<image> Question: In which area was this photograph taken?
Answer: Because in the photo, there is a man wearing a turban, and the photo includes a mosque, this photo was taken in the Middle East.

LLaMA-Adapter V2 Instruction
Instruction: This photograph was taken during one of the following 4 periods. We divide these 4 periods as in the 19th Century, between 1900 and 1950, between 1950 and 2000, in the 21st Century. In which period was this photo taken?

LLaMA-Adapter V2 Instruction
Instruction: The photograph was taken in one of the following eight regions. These eight regions are "Latin America," "Northern America, Europe and Oceania," ... "Eastern Asia," "South-Eastern Asia," "And Sub-Saharan Africa."

Figure 5. We list respectively the prompt templates we used in OpenFlamingo for each protocol for REASONING for times in (a), instructions for LLaMA-Adapter V2 for times in (b).

Figure 6. We list the prompt templates we used in location reasoning for OpenFlamingo in (a) and for LLaMA-Adapter V2 in (b).

on contexts, such as in-context demonstrations and instructions, and expose the hallucination problem [1]. Therefore, image details and style biases cannot help or influence the model reason.
Because this photo was taken in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this photo is in color, and there is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is an old photo, this photo was taken before 1900.

Because this photo is in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this photo was taken in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this photo is in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this photo was taken in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this photo was taken in color, and there is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.

Because this photo is in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this photo was taken in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this photo is in color, and the people in the photo are wearing modern clothing, so it is not a historical photo; this photo was taken in the 21st Century.

Because this is a black and white photograph, and it features a vintage car, this photo was taken between 1900 and 1950.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.

Because this is an old photo, this photo was taken before 1900.