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

Gengyuan Zhang [1,2] **Yurui Zhang** [3] **Kerui Zhang** [1] **Volker Tresp** [1,2]
[1] LMU Munich, Munich, Germany
[2] Munich Center for Machine Learning, Munich, Germany
[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[1] and sorted alphabetically.

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

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.

[1] https://population.un.org/wpp/DefinitionOfRegions/

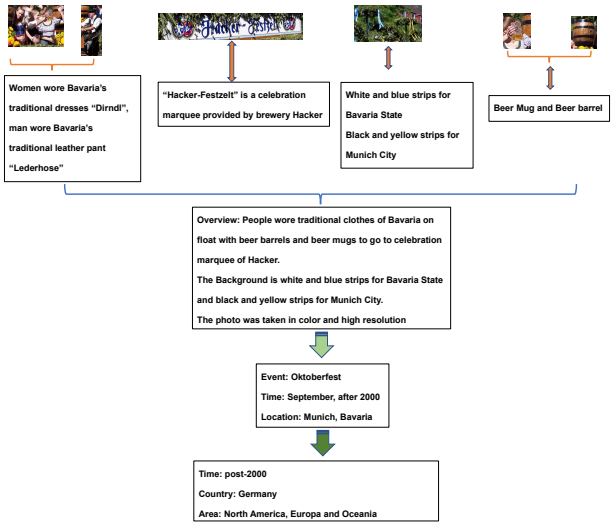


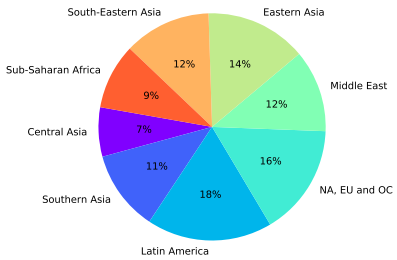
Figure 1. An example of manual image curating to determine its time and location. By doing this, the images of WikiTiLo are grounded by multiple scene text, faces, object segments from the image, and colors and resolution of the image.

A.3. Data distribution

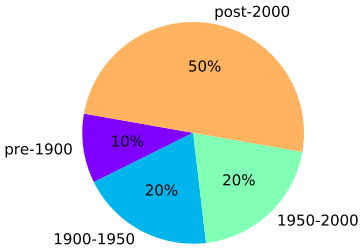
The dataset distribution in location and times can be found in Fig. 2.

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.



a Location Distribution



b Times Distribution

Figure 2. The WikiTiLo dataset exhibits a diverse distribution of times and locations. In terms of time, the images range from 1827 to the post-2000 era. Regarding location, we selected 30 countries whose images met our filtering criteria, representing eight regions. Due to the development of Internet media, images taken after 2000 constitute a significant portion of the dataset. Conversely, images taken before 1900 only account for 10% of the dataset, primarily due to limited data availability and poor quality.

Model	Visual Encoder	Text Encoder	Dataset	Multimodal
ResNet-50	ResNet	-	ImageNet [4]	✘
ViLT	Patch Emb.	BERT	MSCOCO [3], GCC [5]	✔
CLIP-ViT	ViT	Transf.	Online data	✔
CLIP-RN	ResNet	Transf.	Online data	✔
BLIP	ViT	BERT	CapFilt [2],14M [2], 129M [2]	✔

Table 2. Comparison of disminitative VLMs. Variants of different datasets and encoders will be denoted by suffixes.

C. Impact of shot numbers

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

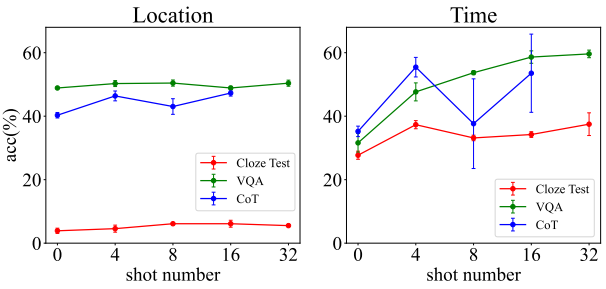


Figure 3. Impact of different shot numbers for OpenFlamingo. More in-context shots do not substantially increase the performance on REASONING$_{\text{LOCATION}}$, but achieve a higher accuracy on REASONING$_{\text{TIMES}}$.

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 Times-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 REASONING$_{\text{TIMES}}$

We also list all the prompts used for OpenFlamingo and LLaMA-Adapter V2 used for REASONING$_{\text{TIMES}}$ 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 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

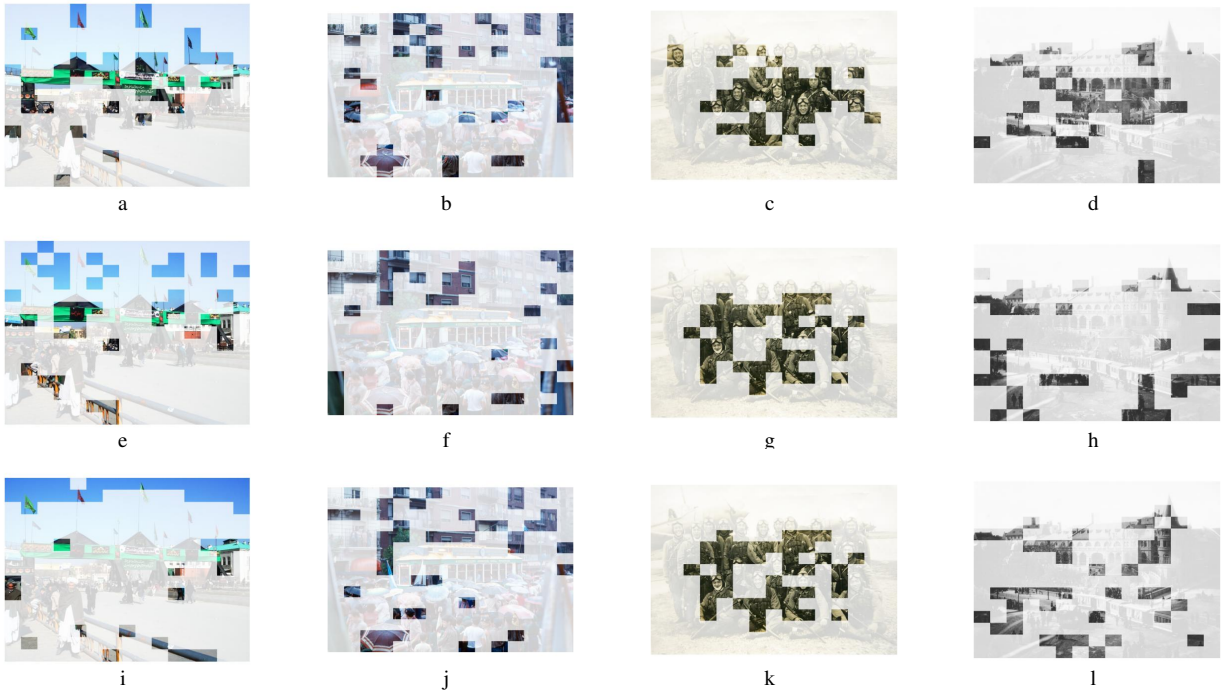


Figure 4. Visualization of transportation plan of word patch alignment on times classification. Best viewed zoomed in. Rows from top to bottom: ViLT, CLIP, and BLIP. Columns from left to right: Afghanistan(Middle East) in 2000, Argentina(Latin America) in 1980, Japan(Eastern Asia) in 1940, and Germany(Europe) in 1880.

OpenFlamingo Cloze Test
<image> *Output*: This is a historical photo taken in the 19th Century. <|endofchunk|>
Open Flamingo 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 <|endofchunk|>
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.<|endofchunk|>

a

LLaMA-Adapter V2 Instruction[a]
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[b]
Instruction: In which period was this photo taken?

b

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).

OpenFlamingo Cloze Test
<image> *Output*: This is a local photo taken in area Latin America. <|endofchunk|>
Open Flamingo VQA
The photograph was taken in one of the following eight areas. These eight 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 <|endofchunk|>
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. <|endofchunk|>

a

LLaMA-Adapter V2 Instruction[a]
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."
LLaMA-Adapter V2 Instruction[b]
Instruction: In which geopolitical region was this photo taken?

b

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

quality, grayscale, and sketch. We select several example photos and show how generative VLMs fail in these cases. We find predictions of generative VLMs are not really grounded by visual cues of images. Answers depend 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.

Image	Location	Time	Image	Location	Time
Eastern Asia South Korea 1940	the Asian woman in the photo is wearing traditional Korean dress, and there are porcelain wares displayed on the cabinet next to her	this is a black and white photograph, and the women in it are wearing old-fashioned traditional attire	Southern Asia India 2000	in the photo there is a Ganesha, which is one of the best-known and most worshipped deities in the Hindu pantheon	this is a color photograph with a high resolution, and it also features modern buildings
Eastern Asia China 1880	the photo features an interior in Chinese style, adorned with a plaque bearing Chinese characters and furnished with ancient Chinese furniture	this is a black and white photograph with slight damage, and it depicts ancient Chinese houses. Therefore, it can be considered a historical photograph	Middle East Afghanistan 1970	in the photo, there is an elderly person wearing a turban standing in a barren land, with the houses built on a hill and a snow-capped mountain in the background, which are consistent with the attire and landscape of Afghan	although this is a color photograph, the pixel resolution is low, indicating that it was not taken recently
Central Asia Kazakhstan 2000	in the crowd of people in the photo, there are some women wearing headscarves, and the photo includes snow-capped mountains and Kazakh yurts	this is a color photograph, and it has a high resolution	Northern America, Europe and Oceania Germany 1980	the white young men in the photo are climbing over the Berlin Wall	this photograph captures the Berlin Wall, which was dismantled after 1989. Therefore, it can be inferred that this photo was taken before 1989

Figure 7. Rationale examples we used as few shot samples in protocol OpenFlamingo VQA with CoT.

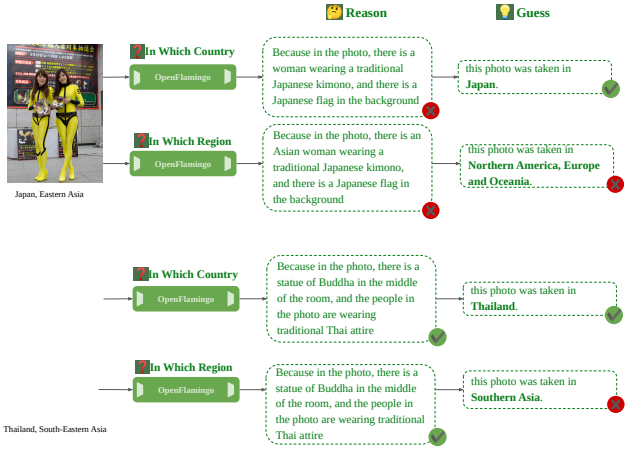


Figure 8. In some cases, the model can correctly predict the country level but fails in the region reasoning. We speculate that the visual cues could not be mapped exactly to the region definitions in our experiments.

Figure 9. Case study of generative VLMs under different image biases. We compare the model prediction of original images and transfer the images into three common image styles that prevail in the dataset: low quality, grayscale, and sketch. The factual generation is marked in green, and the wrong generation is in red.

References

[1] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 3

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[5] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018. 2