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

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

Abstract

Vision-Language Models (VLMs) are expected to be capable of reasoning with commonsense knowledge as human beings. One example is that humans can reason where and when an image is taken based on their knowledge. This makes us wonder if, based on visual cues, Vision-Language Models that are pre-trained with large-scale image-text resources can achieve and even surpass human capability in reasoning times and location. To address this question, we propose a two-stage RECOGNITION & REASONING *probing task applied to discriminative and generative VLMs to uncover whether VLMs can recognize times and location-relevant features and further reason about it. To facilitate the studies, we introduce WikiTiLo, a well-curated image dataset compromising images with rich socio-cultural cues. In extensive evaluation experiments, we find that although VLMs can effectively retain times and location-relevant features in visual encoders, they still fail to make perfect reasoning with context-conditioned visual features. The dataset is available at* `https://github.com/gengyuanmax/WikiTiLo`.

Figure 1. An example image of times and location reasoning in WikiTiLo: Can you tell where and when is this picture taken?

1. Introduction

Vision-Language Models (VLMs) have exhibited remarkable advancements in enhancing multi-modal learning. On the one hand, discriminative VLMs such as CLIP [25], BLIP [19], and ImageBind [11] offer powerful encoders that possess excellent representation and demonstrate significant transferability on recognition and understanding tasks. On the other hand, generative VLMs like LLaMA Adapter [10, 39], BLIP2 [18], Flamingo [2], and LLaVA [22] combine visual encoders from discriminative VLMs with Large Language Models (LLMs) such as LLaMA [33] and GPT3 [5] to further bridge visual features and leverage the exceptional reasoning abilities of LLMs.

With VLMs trained on extensive knowledge corpora comprising multi-modal data, an intriguing question arises: can VLMs reason about the implicit sociocultural backgrounds associated with an image, such as its times and location? Just as in the popular game 'GeoGuesser,' where players are tasked with locating an image based on visual cues, we contemplate whether VLMs can also exhibit a similar aptitude for being a 'good guesser.' This kind of commonsense understanding necessitates that the model can infer the times and location of an image by comprehending not only the local visual evidence but also the underlying concepts. For example, from Fig. 1, humans with related experience can easily infer that the image was most likely taken in Germany and during the famous Oktoberfest based on the cues such as people wearing traditional Bavarian clothes and celebrating with beers and barrels.

Under this background, our work aims to investigate the times and location reasoning capabilities of VLMs. Consequently, we put forward two research questions:
RQ1: Can discriminative VLMs recognize times and location-relevant features from visual input?

RQ2: Can generative VLMs reason about times and locations associated with images based on visual cues?

To address these questions step by step, we devise a two-stage probing task: RECOGNITION and REASONING, applied to discriminative and generative Vision-Language Models (VLMs) as illustrated in Fig. 2. Discriminative and generative VLMs are not just parallel, but generative VLMs are also built upon the encoders of discriminative models.

In the first stage RECOGNITION, we assess whether the visual encoders within discriminative VLMs can successfully identify distinctive features for location and times reasoning using a classification task. Visual features of discriminative VLMs are context-agnostic, which means they are not dependent on tasks and questions. In the second stage REASONING, assuming that the visual encoder can identify salient features within its pre-trained representations, we evaluate the ability of generative VLMs, with the visual encoder intact, to conduct times and location reasoning through an open-ended question-answering task. The generative VLMs are now based on context-conditioned visual features and the powerful reasoning ability of large language models.

We also construct a new dataset for times and location reasoning, WikiTiLo (**Wiki**Common **Ti**mes and **Lo**cation), which comprises images captured over a broad time range and is geographically balanced to mitigate cultural bias. The dataset has been carefully curated to ensure that each image contains distinct visual cues that align with human expert knowledge.

This is one of the first works to investigate whether state-of-the-art large pre-trained Vision-Language Models, enhanced with Large Language Model techniques, are capable of times and location reasoning. Our contributions to this work include:

1. we construct a new dataset, WikiTiLo, for times and location reasoning, with a focus on the socio-cultural background behind images;

2. we propose a two-stage probing strategy, RECOGNITION and REASONING, to evaluate the ability of VLMs to recognize times and location-relevant visual cues and subsequently reason about times and location in a generative setting;

3. we evaluate three discriminative VLMs and two generative VLMs on this same benchmark. Experiments show that visual encoders in discriminative VLMs can generate context-agnostic visual features that help identify times/locations, but generative VLMs fail to reason based on the visual cues.

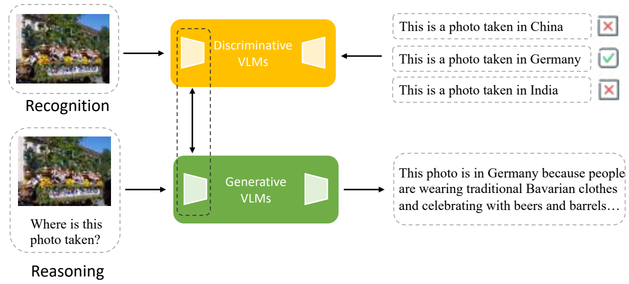


Figure 2. We apply a two-stage probing task RECOGNITION and REASONING to both discriminative and generative VLMs. We evaluate RECOGNITION on discriminative VLMs and REASONINGon generative VLMs with the visual encoder intact.

2. Related Work

Vision-Language Model Vision-language models bridge the gap between flourishing studies in computer vision and natural language studies. Recent works, including [6, 8, 11, 16, 17, 19, 20, 23–25, 36–38] aim to learn a representation of images and texts in a joint feature space and unify vision and language tasks like and visual question answering [16, 17, 31]. Along with the boost of pre-training generative Large Language Models (LLM) like GPT-3 [5], LLaMA [33] manifest their excellent commonsense reasoning ability in broad tasks and are fastly adapted to multi-modal scenarios as in works including Flamingo [2], LLaMA-adpter [10], BLIP-2 [18].

Times and Location Reasoning Recent studies on image times and geo-location prediction [15, 29, 35] show an interesting capacity of deep learning models. CLIP [25] also reports a similar Geolocalization classification. Some geolocation datasets such as Cross-View Time Dataset (CVT) [27], GT-CrossView [34], and Cross-View Image Geolocalization (CVIG) [21] have been provided for the downstream task as Geolocalization. But unlike geolocation estimation, our work is more focused on commonsense knowledge-based reasoning and pays more attention to the sociocultural perspective of images when constructing the dataset. TARA [9] is a preceding work on exploring CLIP's commonsense knowledge in this regard; compared to TARA, we construct a dataset consisting of images with a **wider** timespan and **unbiased** location distribution and evaluate on a wider range of VLMs.

Model Probing Model probing is first proposed in Natural Language Processing to investigate whether representations of Language Models have already learned specific linguistic properties [13, 30]. [4] proposes a framework to probe visual superpixels as an equivalence of word tokens. However, probing multimodal models is a relatively new area. [7, 25, 26, 28] attempt to investigate multimodal probing tasks such as object counting and position identification to learn whether VLMs are capable of understanding multimodal concepts. Our work aims to utilize a simple linear probing [1] method to probe whether VLMs can reason commonsense knowledge like times and location.

3. Dataset: WikiTiLo

We construct the WikiTiLo (**Wiki**Common **Ti**mes and **Lo**cation) dataset, which consists of 6296 images with annotation of the specific time and country where images are taken. The dataset covers 30 countries in Europe, Asia, America, and Africa and the years from 1826 to 2021. We also consider regions as a higher level of geological concept other than countries since the country border does not equate with cultural borders [32]. We partition these countries into 8 regions according to UNESCO based on geography and cultural proximity. The designation can be found in Supplementary Material.

3.1. Data Curation

Wikimedia Commons[1] which is a project of the Wikimedia Foundation, contains a media repository of open images, sounds, videos, and other media. The images of Wikimedia Commons are accumulated from different countries and regions around the world, from different historical periods, and in various categories.

We select the images manually based on such a criterion that the location identity and time period features of each image can be distinguished from architectural patterns, costume styles, languages, social events, photo colors, and quality or other fine-grained features by humans. In particular, to avoid image distribution bias, we attempt to control image origins and balance the ratio of countries that range from more visible areas like Europe and America to less attended areas like Africa and Central Asia, according to socio-cultural characteristics and geographical regions. A detailed selection paradigm is in Supplementary Material.

In total, we have included 30 countries from which images can meet our criteria to represent different regions. Fig. 3 presents some example images in WikiTiLo.

3.2. Statistics

The times and location proportions of WikiTiLo are demonstrated in Supplementary Material. All images are nearly evenly distributed in 8 regions. Half of the images are taken after 2000 due to the multimedia era and 10% are before 1900 due to limited resources and bad quality. For training tasks in the linear probing setting, we use 80% of the entire dataset as the training set, 10% as the validation set, and 10% for the evaluation reported.

[1]https://commons.wikimedia.org/wiki/Main_Page

4. Methodology

Given an image i, the model estimates the posterior probability of answer space a. As in Fig. 4, in RECOGNITION stage, the posterior probability $\mathbb{P}(a)$ is conditioned on context-agnostic visual feature v like visual features of CLIP. In REASONING stage, $\mathbb{P}(a|v_c)$ is conditioned on context-aware visual feature v_c as in MLLMs, contexts include instructions, interleaved image-text demonstrations, and questions. v_c is conditioned not just on raw images, but also on task-specific and question-related contexts.

We probe whether context-agnostic visual features of discriminative VLMs contain prominent times/location-relevant cues in the RECOGNITION step and probe whether generative VLMs can reason about times and location on context-conditioned visual features in the REASONING . For both stages, we evaluate models on three tasks, **Times**, **Location(Region)**, and **Location(Country)**, to determine in which era/region/country the photo is taken.

4.1. RECOGNITION

In the RECOGNITION probing stage, we test whether discriminative VLMs can recognize prominent times/location-relevant visual features via probing the features of visual encoders on multi-class classification tasks. $\text{RECOGNITION}_{\text{TIMES}}$ is a fine-grained image classification task where models predict the occurrence time of an image. This task evaluates whether visual representations of a model contain commonsense knowledge of historical times. $\text{RECOGNITION}_{\text{LOCATION}}$ is a fine-grained image classification task where models predict the location where an image is taken. We test on both country level and region level that aligns with the dataset information. We attempt to evaluate the visual encoders of discriminative VLMs in two settings: zero-shot with a prediction calculated as in Eq. 1 and linear probing as in Eq. 2.

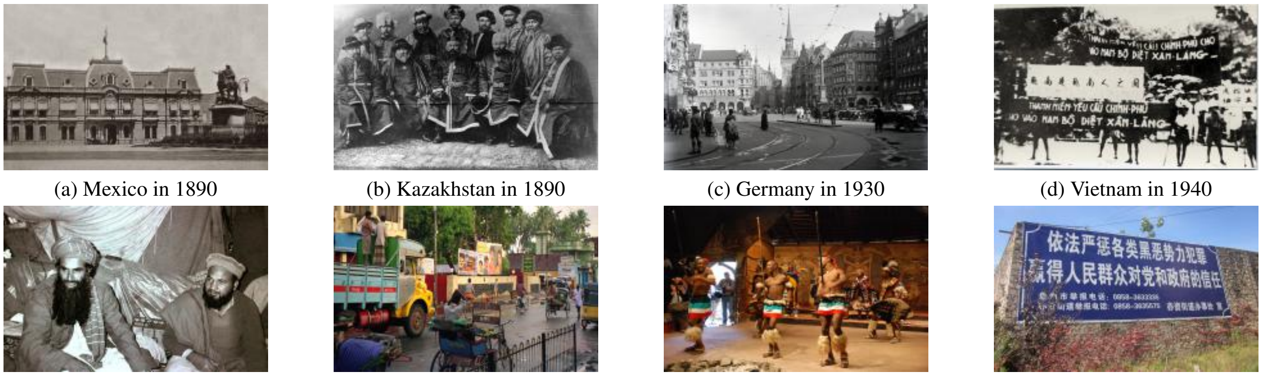
$$\mathbb{P}_{zs}(a_i|v) = \frac{\exp(\text{sim}(\text{VE}(v), \text{TE}(p \circ c_i)))}{\sum_j^C \exp(\text{sim}(\text{VE}(v), \text{TE}(p \circ c_j)))} \tag{1}$$

$$\mathbb{P}_{lp}(a_i|v) = \text{softmax}(\text{linear}(\text{VE}(v)))_i \tag{2}$$

where p is hard prompts we use, VE and TE are visual and textual encoders in VLMs, and c is the class label.

4.2. REASONING

Upon the hypothesis that visual encoders extract recognizable visual features for times and location cues, we further evaluate generative VLMs that can give us good reasoning about times and location based on the cues. We propose using open-ended visual question-answering tasks $\text{REASONING}_{\text{TIMES}}$ and $\text{REASONING}_{\text{LOCATION}}$ to reason times and location.



(a) Mexico in 1890 (b) Kazakhstan in 1890 (c) Germany in 1930 (d) Vietnam in 1940

(e) Afghanistan in 1980 (f) India in 1980 (g) South Africa in 2000 (h) China in 2010

Figure 3. Some example images in the WikiTiLo exhibit abundant visual cues with a sociocultural background, such as stylish buildings, text on images, and traditional clothing. These visual cues enable humans to reason and draw conclusions based on the provided evidence.

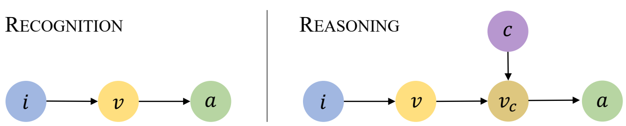


Figure 4. In RECOGNITION, we calculate discriminative VLMs' posterior $\mathbb{P}(a|v)$ of the context-agnostic visual features v of an image i. In REASONING, the context-conditioned visual features v_c are dependent on intact visual features v and question context c, including instructions, prompts, and demonstrations; the posterior probability of possible answers is denoted as $\mathbb{P}(a|v_c)$.

As shown in Fig. 4, the posterior probability of the answer space is conditioned on context-conditioned visual features. A generative VLM generates answers that are conditioned on visual features V_c, as in Eq. 3.

$$\mathbb{P}_{gen}(a_i|v_c) = \prod_j \mathbb{P}(a_{i,j}|\phi(v,c), a_{i,1:j-1}) \tag{3}$$

ϕ is a bridge function that projects visual features v and contexts c into the language space, as perceiver resampler in Flamingo [2], adapters in LLaMA-adpter [10], and Q-Former in BLIP-2 [18]. Generative VLMs can excellently leverage the commonsense reasoning abilities of large language models when predicting the answers.

5. Experiments

Model selection For RECOGNITION stage, we compare three Vision-Language models: ViLT [16], CLIP [25], and BLIP [19]. In the Linear Probing setting, we also choose one pre-trained ResNet-50 [12] as a pure vision baseline. For REASONING stage, we evaluate two generative VLMs, OpenFlamingo [3] and LLaMA-adapter V2 [10], both with CLIP-ViT-L/14, which is one of the best-performing encoders in the RECOGNITION task.

OpenFlamingo Cloze Test
`<image>` *Output*: This is a local photo taken in country China. `<|endofchunk|>`
Open Flamingo VQA
The photograph was taken in one of the following 30 countries. 30 countries are ...
`<image>` *Question*: In which country was this photograph taken?
Short answer: China. `<|endofchunk|>`
OpenFlamingo VQA - CoT
The photograph was taken in one of the following 30 countries. 30 countries are ...
`<image>` *Question*: In which country was this photograph taken?
Answer: Because in the photo, there is an elderly person wearing a turban standing in a barren land, with houses built on a hill and a snow-capped mountain in the background, which are consistent with the attire and landscape of Afghan, this photo was taken in Afghanistan.`<|endofchunk|>`

Figure 5. We list templates for each protocol for REASONING$_{\text{LOCATION}}$. For OpenFlamingo, few-shot examples are used to facilitate in-context learning. In the case of the zero-shot setting, two text samples are retained but without image tokens `<image>`.

Evaluation We evaluate models on three tasks as mentioned above: Times, Location(Region), and Location(Country). Answer space for Times includes 4 time periods: pre-1900, 1900-1950, 1950-2000, and post-2000, and includes 8 different regions. For Location(Country), the model should predict in which of the 30 countries the photo is taken.

For all experimental settings, we evaluate the performance using accuracy, precision, and F1 score averaged across classes. For RECOGNITION, we retrieve the top-1 rank answer as the prediction for all models with the prompts mentioned above. For REASONING, considering the free-form responses generated by language models, we post-process all the predictions and filter the keywords that are relevant to the questions. We require an Exact Match for evaluating generative VLMs. The only exception is that for RECOGNITION on Times, we remap the prediction of LLaMA-Adapter V2-Instruction^b to the time intervals in our settings since strict instruction following is hard to guarantee. Any generated answer that does not directly address the question is considered a mismatch.

Human Performance serves as an important baseline. Hu-

LLaMA-Adapter V2 Instruction[a]
Instruction: The photograph was taken in one of the following 30 countries. These 30 countries are ... In which country was this photo taken?
LLaMA-Adapter V2 Instruction[b]
Instruction: In which country was this photo taken?

Figure 6. For the LLaMA-Adapter V2, we use a simple question as the instruction. For the case with labels, the label sentence was added in the instruction before the question.

man common sense knowledge about times and location reasoning varies substantially between individuals depending on their background and education. Hence we conducted tests with 12 participants, each assigned 60 randomly sampled images from the test set. All participants have varied backgrounds from six countries across four regions and do not receive any pre-training prior to the test. We report the average accuracy as the human-level baseline.

5.1. Experimental Setups

Zero-shot Setting To evaluate the zero-shot performance of discriminative VLMs, we adopt a prompt "*A historical/recent/contemporary image with black-and-white/color taken in [times period]*" for RECOGNITION$_{\text{TIMES}}$ classification and "*a local photograph from [country] in [region]*" for RECOGNITION$_{\text{LOCATION}}$ classification. We calculate the image-text similarities across all candidates and retrieve the top rank as predicted classes.

Linear Probing Setting We used a single linear layer with vision encoders' embedding size as input size. The class size is 4 for predicting times, 8 for regions, and 30 for countries. The visual encoders of the model were frozen during training. Each linear probe was trained for 18 epochs. We search the hyperparameter, including learning rate in the range from 0.05 to 0.0001 and learning rate decay among $[0.1, 0.2, 0.5]$ with ray.tune[2]. For each task, we use CE loss.

Generation Setting We employ two evaluation protocols for OpenFlamingo: VQA and Cloze Test. In both protocols, we utilize either in-context demonstrations or explicit prompts to constrain the generated answer to be within the set of candidate answers. If the generated answer fails to directly address the given question, it is considered incorrect. The results were averaged on 4 trials with different random seeds. We also conduct tests with varying numbers of in-context shots in the range of $[0, 4, 8, 16, 32]$. In the zero-shot setting, we retain two text samples while excluding image tokens `<image>` and image samples. To solicit rationale, we also attempt to output Chain-of-Thought(CoT) for prediction by annotating a subset of samples with rationales, which is showcased in Supplementary Material. For LLaMA-Adapter V2, we use two different instructions for question answering on times and location reasoning. Details of all the prompts used can be found in Fig. 5 to Fig. 6.

[2]https://www.ray.io/ray-tune

6. Results

6.1. RECOGNITION

We compare the model performance on zero-shot and linear probing in RECOGNITION stage.

Zero-shot Performance Tab. 1 shows the results of the zero-shot performance of VLMs on RECOGNITION$_{\text{TIMES}}$ and RECOGNITION$_{\text{LOCATION}}$ tasks. We notice that CLIP variants achieve impressive performance on both tasks and outperform other models by a large margin, which corresponds with the findings of excellent zero-shot performance in [25]. Amongst all variants, model performance on times classification does not vary much from model architecture. By contrast, BLIP performs poorly on both tasks.

We also find that among each VLM family, models with a larger visual encoder perform slightly better than others on times classification but have a bigger advantage on location classification. Datasets also are critical for model performance. CLIP is trained on large-scale image-text data and manifests the best capabilities of reasoning times and location. However, we still cannot ground the failure of ViLT and BLIP fully since the domain shift between pre-training datasets and WikiTiLo might be huge.

Compared to human performance, all CLIP models outperform our testee on both tasks. Noteworthily, human performance varies individually according to their background, with a variance of 10.41% in times classification and 8.69% in location classification.

Linear Probing Performance Tab. 2-3 shows the results of model performance on the Times and Location(Region) task in the linear probing setting. We report the models' best performance after the hyperparameter search. We use pre-trained ResNet-50 as a pure-vision baseline. ResNet-50 shows non-inferior performance on the Times task but cannot compare to most Vision-Language models on the Location(Region) except ViLT.

All models enjoy a prominent performance improvement compared to their zero-shot baselines. CLIP models still achieve much better performance than other VLMs. In particular, BLIP models also enjoy a large improvement in both tasks. Thus, we deduce that both CLIP and BLIP have learned comparably informative and discriminative features for these two tasks in pre-training already, which profits from their pre-training data scale.

We conclude that reasoning location is a harder and more interesting task than reasoning times. This could be explained by the fact that location classification is a more fine-grained task, which requires the model to distinguish more detailed visual cues at a commonsense level, like un-

	Model	Times			Country			Region		
		Accuracy	Precision	F1-score	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score
RECOGNITION	CLIP-ViT-B/32	78.57%	70.66%	70.66%	44.28%	43.11%	40.19%	63.65%	67.34%	64.42%
	CLIP-ViT-B/16	75.71%	67.94%	70.08%	55.23%	54.16%	53.71%	77.62%	77.23%	76.86%
	CLIP-ViT-L/14	76.03%	69.73&	69.62%	68.25%	60.77%	61.87%	85.56%	86.47%	85.72%
	CLIP-ViT-L/14@336px	**79.05%**	**73.66%**	**73.75%**	**72.85%**	**64.52%**	**65.79%**	**88.25%**	**88.92%**	**88.43%**
	BLIP-129M	30.95%	46.81%	46.14%	35.23%	35.30%	30.07%	46.51%	57.02%	49.05%
	BLIP-129M-Coco	32.22%	50.13%	43.49%	35.23%	34.25%	28.74%	47.78%	52.89%	48.14%
	BLIP-129M-Flickr	58.57%	54.33%	56.19%	33.49%	29.84%	27.32%	48.89%	50.44%	47.46%
	BLIP-ViT-L	44.44%	54.97%	52.86%	*42.22%*	*43.30%*	*40.10%*	54.60%	56.73%	54.57%
	BLIP-ViT-L-Coco	42.06%	51.75%	49.40%	39.20%	38.93%	34.61%	47.78%	52.89%	48.14%
	BLIP-ViT-L-Flickr	*66.51%*	*58.37%*	*60.14%*	40.47%	41.24%	37.54%	*60.00%*	*61.08%*	*59.38%*
	ViLT-Coco	*70.16%*	*58.51%*	*57.78%*	3.65%	3.60%	3.10%	16.98%	17.39%	17.32%
	ViLT-Flickr30K	51.90%	48.84%	49.21%	*7.93%*	*4.00%*	*4.80%*	*20.32%*	*20.37%*	*19.81%*
REASONING	OpenFlamingo-Cloze Test	27.70%	26.36%	11.49%	3.89%	3.69%	2.18%	4.72%	8.62%	4.72%
	OpenFlamingo-VQA	31.59%	*30.36%*	*28.60%*	*48.88%*	*53.78%*	*41.19%*	22.49%	30.49%	18.64%
	OpenFlamingo-VQA CoT	*35.21%*	29.36%	28.42%	40.3%	45.24%	33.17%	*24.04%*	*39.39%*	*19.27%*
	LLaMA-Adapter V2-Instruction[a]	*58.02%*	28.04%	32.88%	23.05%	*52.64%*	18.66%	*19.07%*	*26.59%*	*13.01%*
	LLaMA-Adapter V2-Instruction[b]	34.34%	*58.59%*	*37.70%*	*45.62%*	51.57%	*35.50%*	11.12%	10.05%	5.99%
	Frequency baseline	25.07%	25.29%	23.27%	3.33%	2.95%	2.88%	12.53%	12.59%	12.24%
	Human baseline(average)	67.42%	-	-	48.30%	-	-	62.42%	-	-

Table 1. Results of performance without training for RECOGNITION and REASONING. The highest performance among all models is highlighted in **bold**, and the best performance of each VLM is marked in *italic*.

Model	Accuracy	Precision	F1-score
ResNet-50	80.63%	73.71%	71.89%
CLIP-RN50	85.87%	81.26%	80.73%
CLIP-RN101	86.67%	80.71%	80.71%
CLIP-RN50x16	90.32%	86.24%	86.39%
CLIP-ViT-B/32	89.37%	85.67%	85.04%
CLIP-ViT-B/16	88.25%	83.04%	82.95%
CLIP-ViT-L/14	90.63%	86.20%	86.14%
CLIP-ViT-L/14@336px	**92.70%**	**89.40%**	**89.64%**
BLIP-129M	86.51%	81.36%	80.51%
BLIP-129M-Coco	85.87%	81.16%	80.77%
BLIP-129M-Flickr	86.67%	82.42%	80.91%
BLIP-ViT-L	*88.41%*	*84.72%*	*84.02%*
BLIP-ViT-L-Coco	86.67%	82.45%	81.47%
BLIP-ViT-L-Flickr	87.94%	83.73%	83.23%
ViLT-Coco	80.00%	75.05%	72.72%
ViLT-Flickr30K	*80.79%*	*75.86%*	*73.19%*

Table 2. Results of Linear Probing for Times

Model	Accuracy	Precision	F1-score
ResNet-50	54.13%	54.85%	54.45%
CLIP-RN50	72.06%	71.98%	72.04%
CLIP-RN101	74.44%	74.44%	74.41%
CLIP-RN50x16	86.67%	87.31%	86.93%
CLIP-ViT-B/32	79.37%	79.36%	79.11%
CLIP-ViT-B/16	84.60%	85.18%	84.88%
CLIP-ViT-L/14	90.95%	91.36%	90.92%
CLIP-ViT-L/14@336px	**93.33%**	**93.90%**	**93.37%**
BLIP-129M	75.08%	75.10%	75.26%
BLIP-129M-Coco	77.14%	76.77%	76.99%
BLIP-129M-Flickr	76.98%	77.01%	76.81%
BLIP-ViT-L	76.51%	77.37%	76.77 %
BLIP-ViT-L-Coco	78.10%	77.92%	78.04%
BLIP-ViT-L-Flickr	*78.73%*	*78.64%*	*78.53%*
ViLT-Coco	42.70%	*45.61%*	42.12%
ViLT-Flickr30K	*45.08%*	42.91%	*43.01%*

Table 3. Results of Linear Probing for Location(Region)

derstanding distinct geological and cultural elements.

We also find that visual encoders have an impact on the linear probing results. ViLT still shows a poor performance and is even worse than ResNet-50. Since CLIP-RN50 with the same encoder structure and BLIP trained with the same datasets still shows a good performance, we can only attribute it to ViLT with a simple linear patch embedding cannot produce informative visual features for such tasks.

Visualization Inspired by [16], we adopt a cross-modal alignment of times/location tokens and image patches on linear probing models for visualization, to show that visual encoders can recognize relevant discriminative features. We use the prompts in the zero-shot setting. We select the first 25% of visual patches that have more mass transported from word tokens of labels in prompts to visual patches to emphasize. Fig. 7 and Supplementary Material. show the visualization of location classification. For Times-relevant questions, the attended patches seem less specific. Generally, visual tokens in the background instead of foreground objects have seemingly dominant contributions.

Interestingly, we find that for RECOGNITION$_{\text{LOCATION}}$, models tend to attend to specific regions like scene texts and human clothing. CLIP can locate the Chinese characters on the board, as shown in Fig. 7(i) and banner texts reading "Malaysian" in Fig. 7(f). When the model performs poorly, as shown in Fig. 7(b), attended patches are less meaningful. This shows us that VLMs can recognize distinct and discriminative visual cues on an image that help reason commonsense knowledge about the location.

Figure 7. Visualization of transportation plan of word patch alignment on location classification. Best viewed zoomed in. Rows from top to bottom: ViLT, CLIP, and BLIP. Columns from left to right: China (Eastern Asia) in 2010, Malaysia (South-Eastern Asia) in 1990, Tajikistan (Central Asia) in 2000, Bangladesh (Southern Asia) in 1970.

6.2. REASONING

Performance of In-Context Learning-based Reasoning From Tab. 1, we also find that OpenFlamingo Cloze Test only achieves frequency baseline results. This reveals that in-context demos can teach the generative LLMs the mapping to output label space but do not leverage any semantic information from visual features. In the VQA protocol, OpenFlamingo has higher accuracy compared to the Cloze Test protocol. However, enforcing the model to output Chain-of-Thought does not improve the performance. Regarding the impact of the number of shots in in-context learning, we show that more shots do not improve performance on location reasoning substantially, as in Supplementary Material. Especially for REASONING$_{\text{TIMES}}$, we find the output prediction is more unstable when having more in-context shots and deteriorates the performance. *Interestingly*, we find that models perform even worse on reasoning regions than countries and the language models fail to relate the countries to their affiliated regions in some cases, of which we include a case study in Supplementary Material.

Performance of Instruction-based Reasoning LLaMA-Adapter V2 exhibits unsatisfying performance compared to other models in general. With Instruction^a, the model's performance declines when provided with label instructions. With Instruction^b, the model achieves performance that is comparable to OpenFlamingo VQA as well as the averaged human baselines. We also find different instructions have a big impact on the model performance on REASONING.

6.3. Analysis

We notice the performance gap between RECOGNITION and REASONING. Generative models also can hardly achieve averaged human baselines, which cannot live up to the expectations since large-scale VLMs should benefit more from a massive knowledge base in our assumption.

Based on the fact that in RECOGNITION stage, good performance of discriminative VLMs shows visual encoders can extract good visual features even without any task contexts, we reasonably speculate that the performance gap lies in two aspects. (1) Context-conditioned visual features v_c cannot retain answer-relevant information, and (2) large language models fail to reason based on the visual cues.

Since for most generative VLMs, the projection modules are deeply coupled with the language models, like adapters used in LLaMA-adapter, it is hard to probe the context-conditioned visual features v_c via linear prober. We conduct a qualitative study on failure cases. To validate our hypotheses, we conducted a case study on the failure cases.

Failure Cases We investigate the failure cases in the REASONING phase of generative VLMs. In Fig. 8, we analyze different types of failure cases for OpenFlamingo and LLaMA-Adapter V2. By requiring OpenFlamingo to output Chain-of-Thought and providing clear instructions to LLaMA-Adapter V2, both models can make grounded guesses. However, failure guesses can occur in the following scenarios: 1) when OpenFlamingo fails to generate a grounded reason based on an image, and 2) when Open-

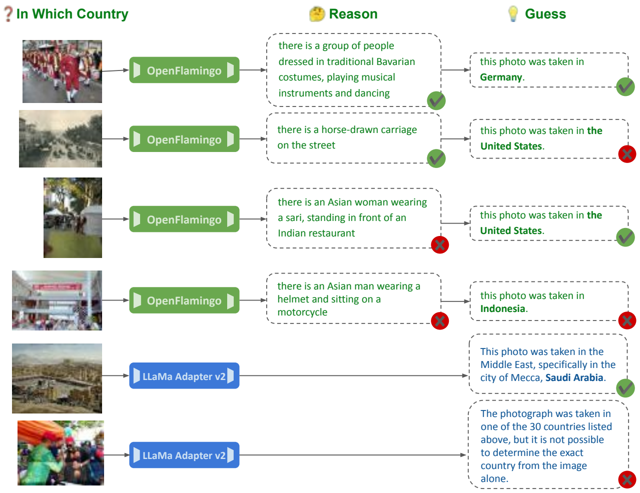


Figure 8. Example cases of interest from OpenFlamingo VQA with Chain of Thought, and LLaMA Adapter V2. We find that the answers generated are not grounded on the visual cues and the reasoning and predictions are not coherent, which both fail.

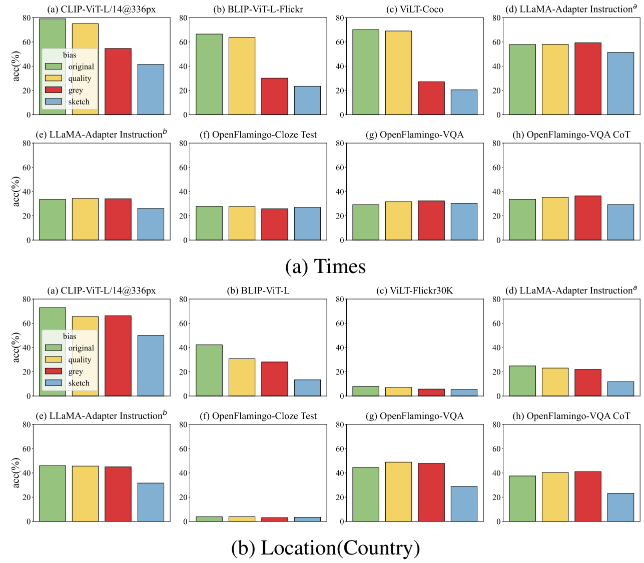


Figure 9. We investigate how different image biases influence model performance on tasks related to Times and Location (Country). We compare the test set under four settings: original, lower quality, grayscale, and sketch. Our findings reveal that for discriminative VLMs, there is a significant drop in model performance when we change image styles. However, generative VLMs are less affected by these changes. Please zoom in for a better view.

Flamingo cannot generate coherent answers based on the reason. We find that these failures are primarily caused by OpenFlamingo being heavily influenced by the few-shot demos, which limits its ability to extract relevant and useful information to a significant extent. On the other hand, for LLaMA-Adapter V2, predictions are rejected when the model fails to locate relevant visual cues in an image, despite humans being able to reason effectively. This highlights that generative VLMs still struggle to fully leverage the visual cues in images for times and location reasoning.

Dataset Bias Humans sometimes rely on extraneous information such as image quality, color, or style to determine times of photo-taking. Considering the images in WikiTiLo come from different media resources, it is crucial to investigate whether the performance of VLMs will also be affected by these factors. We compare 8 models, including 3 discriminative VLMs and 5 generative VLMs, on the original test set, and 3 style-transferred test sets: images in lower quality, in gray scale, and in sketch-version. We assume that if the reasoning is based on image style bias instead of image details, the model results will deteriorate after style transfer.

As shown in Fig. 9, lower quality does not really influence the model performance since visual encoders always reshape the input images, and images in sketches are always unrecognizable and significantly deteriorate model performance. However, grayscale images can evidently decrease the performance of discriminative VLMs.

Nevertheless, generative VLMs are almost unaffected by biases. Again, we conducted a failure case study as in Supplementary Material. 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 [14]. Therefore, both image details and styles cannot help reasoning.

7. Conclusion

In this work, we propose a two-stage probing task RECOGNITION and REASONING and introduce a new dataset WikiTiLo, to explore VLMs' times and location reasoning ability Based on the findings of RECOGNITION and REASONING tasks on the WikiTiLo dataset, we have probed the capabilities of VLMs in times and location reasoning. Experiments indicate that discriminative VLMs, such as CLIP, effectively extract context-agnostic visual features for identifying times and locations. However, generative VLMs struggle to reason correctly due to the inability of context-conditioned visual features to retain relevant information or the language models' failure to reason based on visual cues, according to the experiments and case studies. This motivates further exploration into how large language models can better utilize visual features in future studies.

Acknowledgements

This work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2, 4

[3] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, Mar. 2023. 4

[4] Dominika Basaj, Witold Oleszkiewicz, Igor Sieradzki, Michał Górszczak, Barbara Rychalska, Tomasz Trzciński, and Bartosz Zieliński. Explaining self-supervised image representations with visual probing. Freiburg, Germany: International Joint Conferences on Artificial Intelligence, 2021. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 2

[7] Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 3

[8] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[9] Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, and Dan Roth. There'sa time and place for reasoning beyond the image. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1138–1149, 2022. 2

[10] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1, 2, 4

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 2

[14] 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. 8

[15] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Predicting good features for image geo-localization using per-bundle vlad. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1170–1178, 2015. 2

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2, 4, 6

[17] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 2, 4

[19] 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. 1, 2, 4

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[21] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[24] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022. 2

[25] 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. 1, 2, 3, 4, 5

[26] Philipp J. Rösch and Jindřich Libovický. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1031–1041, Seattle, United States, July 2022. Association for Computational Linguistics. 3

[27] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a dynamic map of visual appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12444, 2020. 2

[28] Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257, 2022. 3

[29] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018. 2

[30] Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534, 2016. 2

[31] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[32] Vas Taras, Piers Steel, and Bradley L Kirkman. Does country equate with culture? beyond geography in the search for cultural boundaries. *Management International Review*, 56:455–487, 2016. 3

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[34] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 494–509. Springer, 2016. 2

[35] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 37–55. Springer, 2016. 2

[36] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2

[37] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[38] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[39] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 1