Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning

Anonymous TACL submission

Abstract

The lack of reasoning capabilities in Vision-Language Models (VLMs) has remained at the forefront of research discourse. We posit that this behavior stems from a *reporting bias* in their training data. That is, how people communicate about visual content, by default, omits tacit information needed to supervise some types of reasoning; e.g., "at the game today!" is a more likely caption than "a photo of 37 people standing behind the field." We investigate the pragmatics of the data underlying popular VLMs like OpenCLIP, LLAVA-1.5 and Molmo, and find that four reasoning skills (spatial, temporal, negation, and counting) are not sufficiently represented. With a set of curated benchmarks, we demonstrate that: (i) frequency of skill-requiring instances predicts model performance; (ii) contrary to popular belief, scaling data size, model size, and to multiple languages does *not* result in emergence of these skills; but, promisingly, (iii) incorporating annotations specifically collected to obtain tacit information is effective. Our findings highlight the need for intentional, reasoning-aware data collection methods, rather than counting on scale for emergence of reasoning capabilities.

1 Introduction

Research in Vision-Language Models (VLMs) grapples with a paradox: despite their impressive performance on standardized benchmarks (Liu et al., 2024a; Deitke et al., 2024; OpenAI, 2024), these models often falter on tasks requiring counting (Paiss et al., 2023), spatial reasoning (Liu et al., 2023; Kamath et al., 2023b) and compositional reasoning (Ma et al., 2023; Zhao et al., 2022; Parcalabescu et al., 2022; Yuksekgonul et al., 2023; Kamath et al., 2024). We hypothesize that these gaps stem from a *reporting bias* in vision-language data. Put simply: when discussing images online, people systematically omit certain types of information, e.g., spatial prepositions. We leverage long-standing bodies of work in pragmatics, linguistics, and cognitive science to identify four types of tacit reasoning systematically omitted by people: *spatial reasoning, counting, negations,* and *temporal reasoning*.

We analyze popular pretraining corpora LAION (Schuhmann et al., 2022), LLAVA-1.5 (Liu et al., 2024a) and PixMo (Deitke et al., 2024), and validate that reporting bias occurs when people write alt-text (as in LAION) or are asked to annotate images with captions (as in LLAVA-1.5). To investigate potential correlations between training data and (a lack of) image-text reasoning skills, we curate evaluation questions that require these four types of reasoning. We evaluate a wide variety of contrastive and generative VLMs on these benchmarks and show that, in line with our hypothesis, existing models perform poorly (on average, open-source models fall 54 points behind human performance) unless they are explicitly trained with datasets that require such skills.

Crucially, data+model scaling alone is unlikely to lead to emergent reasoning[1] — *as the human behaviors underlying the reporting bias do not change with scale*. Extrapolating scaling performance on our evaluations suggests, e.g., that CLIP (Radford et al., 2021) would need to be trained with an intractable amount of data or number of model parameters to meet human performance on our benchmarks. Adding multilingual diversity to CLIP's training data by translating non-English captions in web-scale corpora to English, as in Nguyen et al. (2024), also does not improve model performance, showing that the reporting bias is not

[1] Unlike the success it has shown in perception and recognition tasks (Cherti et al., 2023), which are better represented naturally in training corpora (Udandarao et al., 2025).

specific to the English language.

Finally, we study whether such reasoning can be elicited from annotators when explicitly prompted to do so. We find that, for the same underlying images sourced from COCO, instructions from LLAVA and PixMo data collection elicit 2–3 times more instances of counting and spatial reasoning than instructions from COCO. Further, with carefully-written instructions we present, negation and temporal reasoning can also be successfully elicited. The prevalence of reasoning-related information in training data corresponds with improved reasoning capabilities of the corresponding models; thus, these results show promise to improve model reasoning via intentional, reasoning-aware data collection, rather than simply scaling.

Our contributions are: (1) revealing the reporting bias in vision-language, validated with three open-source image-text corpora; (2) re-purposing benchmarks for VLM reasoning and evaluating top-performing contrastive and generative VLMs; (3) revealing that scaling up data, parameters and multilingual diversity do not result in emergent reasoning; and (4) showing that reasoning-aware data collection is possible, and shows promise to improve model reasoning capabilities.

2 Related Work

Reporting bias is a well-studied phenomenon in the area of NLP, having presented itself as the "common sense problem", e.g., "people murder" is a more likely bigram than "people breathe" in text[2], leading models trained on this text to incorrectly believe that the former action is more likely to occur than the latter (Sap et al., 2019b; Shwartz et al., 2020). This was overcome with the introduction of large-scale commonsense corpora (Bosselut et al., 2019; Sap et al., 2019a) to provide models the lacking information. We study this phenomenon in vision-language data, tackling types of reasoning beyond common sense.

In the vision-language domain, Ye et al. (2024) show that people from different cultures describe different features of the same image when provided the same instructions. Nguyen et al. (2024) further show that by translating non-English captions to English, VLMs' zero-shot classification performance increases. We acknowledge the increased coverage of information by speakers from different languages, and ask the question: are there types of information omitted by *everyone*?

Several recent works have studied various failure cases of **VLM reasoning** (Ma et al., 2023; Diwan et al., 2022; Zhao et al., 2022; Kamath et al., 2023b). In response, other work focuses on improving the quality of the training data, by re-captioning the images (Nguyen et al., 2023; Lai et al., 2024; Betker et al.) and/or collecting proprietary data (OpenAI, 2024). We investigate a possible cause behind these failure cases, and study open-source datasets to determine whether annotators require specific instructions to include data otherwise omitted due to reporting bias.

Cherti et al. (2023) show that the performance of contrastive VLMs improves across several tasks with an increase in scale of model and training data size. However, this has shown to not be the case for reasoning tasks (Al-Tahan et al., 2024). In contrast, we investigate a reason *why* this behavior occurs. Further, our benchmarks target specific types of reasoning, and contain primarily real-world images. Additionally, we study both contrastive and generative VLMs.

3 Reporting Bias in Vision-Language Reasoning

While several recent vision-language benchmarks have exposed surprising failure modes of powerful VLMs in various reasoning tasks (Thrush et al., 2022; Kamath et al., 2023b; Zhao et al., 2022; Ma et al., 2023), each of these takes a different stance on the cause of the issue, and thus, its solution. Yuksekgonul et al. (2023); Hsieh et al. (2023); Doveh et al. (2023a,b) claim that the failure arises from over-reliance of the image-text pretraining task on batch size to learn detailed embeddings with the contrastive loss, and thus try to improve model compositional reasoning performance by introducing hard negatives to the batch. Zeng et al. (2021) states that image-level captions are insufficient to teach the model fine-grained understanding, and introduces hierarchical losses based on region captions. However, limited work to date studies the *data* these models are trained on.

No matter how large our corpora become, they are sourced from captions written largely by humans. As such, they exhibit the natural patterns and idiosyncracies of how humans understand and describe images. In this section, we leverage long-standing theories of the same — sourced from var-

[2]That people breathe is too obvious of a fact to be expressed in writing.

ious fields, including linguistics, pragmatics, and cognitive science — to arrive at hypotheses of under-represented information in web-scale corpora due to reporting bias. We then test their accuracy by investigating the training datasets of open-source contrastive and generative VLMs.

3.1 Theory-based Hypotheses of Omitted Types of Reasoning

When people communicate, they do not do so in a vacuum. Cognitive semantics points out people take cues from a variety of sources such as intents, perspectives, and topic of discussion to shape what we intend to say (Langacker, 2015; Talmy, 1972). We use specific words, e.g., adverbs, adjectives, and prepositions, to be as expressive as required by the context of the discussion. Moreover, Pragmatics tells us that we are highly organized in how we achieve this: we abide by a tacit set of cooperative principles that is expected in communication (Grice, 1975; Goodman and Frank, 2016). These topics have been investigated extensively by various efforts in linguistics, cognitive science, and child language acquisition, inter alia.

Here, we posit that such principles of communication can help explain the reporting bias we observe in the data. In writing captions, we produce text that best communicates what we observe. Thus, we expect captions to be subject to the same communicative principles that guide much of our utterances. At the same time, however, caption data is produced in a restricted setting that lacks communicative context that would produce the desired expressiveness. Without knowledge like the topic of discussion and limited understanding of who the caption consumers will be, the caption writers have only basic principles and common knowledge to guide their writing. As a result, the captions lack the expressive cues necessary to train vision-language models to count, to use negations, and to do spatial and temporal reasoning.

People tend to omit spatial and temporal language. Spatial language such as "left of", "above" or "below" and temporal prepositions such as "before" or "after" are central to enabling spatial and temporal reasoning respectively. However, unless explicitly directed, people may not naturally produce such language in captioning.

Pragmatic studies in conversational maxims, known as Gricean Maxims (Grice, 1975), suggest that what information is revealed and how much is revealed is counter-weighed by the expectation to be direct, to be concise and not to misdirect in communication. For example, maxims suggest that even if "a cat left of a dog" is a logically accurate description of an image, a person might opt for "a cat and a dog" because "left of" assigns undue importance to one over the other. Expressive as it may be, choosing the former caption when there is no explicit reason to do so would be deceitful (Maxim of Quality) or would impose a perspective that cannot be justified: whether it is the left of the viewer of the image, or of the subject in the image (Maxim of Manner).

In the same way, given an image of a boy throwing a ball, writing "and after, the ball will fall" would allow for temporal reasoning for a model, but such captions are likely to be avoided because they are too obvious (Maxim of Quantity) or due to insufficient knowledge or evidence about the described event (Maxim of Quality).

Even when spatial preposition use may be merited, studies in cognitive linguistics suggest that captioning may be limited by the existence of default relationships, which we only overlook when the situation calls for it (Talmy, 1972). For example, when grounding one object (a Figure) with respect to another (a Ground), humans will naturally choose the smaller and easier to move entity as the Figure (e.g., "a poster above a bed" is more likely than "a bed under a poster"). If they are equally sized and movable, we will disprefer the use of spatial language without ulterior reasons.

Such ulterior reasons or perspectives, as theories in linguistics suggest, are provided by discourse mechanisms like the Question under Discussion (QUD) — the implicit or explicitly stated question being addressed in a discourse (Von Stutterheim and Klein, 1989). Something as simple as knowing that the image being captioned is a shot of someone's newly adopted cat (given the picture of a dog and a cat) would provide perspectives on how to frame a caption: what to (de)emphasize, what to focus on, or simply, what to talk about. This is not a natural artifact of a restricted annotation setting. Specific prompting (discussed in Section 4) such as "focus on the cat" or "cat was just adopted" would be necessary to provide an actionable QUD to trigger the temporal or spatial language we want represented in the data.

People tend to omit counting. Why people may omit object counts in image captions is explained

by the expectation that a speaker should maximize the information conveyed while keeping the statement brief (Maxim of Quantity; Rational Speech Act (Frank and Goodman, 2012; Goodman and Frank, 2016)). The informational value added by "six cats" compared to "a group of cats" is negligible without further context, while requiring more effort on the speaker's part (counting the objects). Moreover, since there are very few contexts in which the listener cares whether there were exactly "six cats" compared to "a group of cats", i.e., it is rarely the QUD, there is no need for the writer to assume it (Maxim of Relevance).

People tend to omit negations. Intuitively, there is no rational reason for a person to write "there are no parrots" given a picture of a dog and a cat without further context. Much like counting, it would provide more information than necessary to describe the image (Maxim of Quantity; Rational Speech Act) and assign importance when none is merited (Maxim of Quality). Additionally, concepts of sentence processing related to psycholinguistics and child language acquisition (Tian and Breheny, 2016; Pea, 1978) suggest that negations are more costly and slower to process than positive statements, and are thus not preferred.

3.2 Testing Hypotheses in Open-Source Image-Text Corpora

In this section, we estimate the frequency of the aforementioned types of reasoning in popular open-source image-text corpora, to test our hypothesis that they occur rarely. We study the training data for OpenCLIP (Cherti et al., 2023), LLAVA-1.5 (Liu et al., 2024a) and Molmo (Deitke et al., 2024). Where OpenCLIP is only trained on LAION (Schuhmann et al., 2022), LLAVA-1.5 and Molmo are additionally trained on open-source academic datasets. We combine the text from all constituent datasets to run this study, taking sampling rates into account as well.

To perform this study, we list keywords corresponding to each type of reasoning, e.g., to study the prevalence of spatial language, we search for the keyword "right of" (among others, c.f. Appendix). While this includes false positives ("right of way"), it serves as a loose upper bound of the prevalence of the spatial relation in the dataset. For each of these keywords, we perform a string search in the listed corpora and list the percentage occurrence of the strings, shown in Table 1.

We then sample 100 data points corresponding to each type of reasoning in each corpus and manually calculate the number of data points in which the reasoning is truly represented and visible in the image, i.e., the true positive rate. We then calculate a rough estimate of the true number of occurrences of that type of reasoning in the corpus (Estimated True Occurrence in Table 1). Examples of data points contain keywords and do or do not operationalize reasoning are shown in Figure 1.

As seen in Table 1, the types of reasoning we study are indeed infrequent in the corpora, verifying our hypotheses from Section 3.1. To put these numbers in context, the word "blue" alone appears in 2% of LAION captions, and tends to be clearly visible in the image (e.g., "a pair of *blue* shoes").

4 The Effect of Annotator Instructions on Reporting Bias

4.1 Existing Datasets

To determine the effect of annotator instructions, we first study a dataset where there were no annotator instructions provided at all: LAION (Schuhmann et al., 2022), which was scraped from alt-text fields of images on the internet. We see from Table 1 that LAION has low representation across all four types of reasoning we study.

We next look at COCO's (Chen et al., 2015) crowdsourced captions, where annotators were given no specific prompting that would engage them in reasoning. In fact, they were explicitly instructed to "*not* describe things that might have happened in the future or past". Accordingly, we observe the effect of the instructions leading to an even lower occurrence of temporal reasoning in COCO as compared to the non-existent instructions of LAION. Interestingly, however, we see that the prevalence of spatial language and counting in COCO is higher than that of LAION. Having temporal reasoning restricted, annotators may have turned to focus more closely on describing the objects in the image.

For LLAVA-1.5 (Liu et al., 2024a) (pretraining and finetuning data, combined), the instructions required discussion of "object counts" and "relative positions between objects", among other non-reasoning-related instructions. This leads to higher occurrences of both counting and spatial reasoning than in COCO. However, it is worth pointing out that their estimated true occurrences are not higher than that of COCO. This may be

Data	Spatial		Counting		Negation		Temporal	
	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.	Occurr.	Est. True Occurr.
LAION-2B	0.3	0.1	8.8	1.7	0.8	0.1	0.9	0.2
COCO	3.7	3.7	10.8	10.4	0.2	0.1	0.2	0.1
LLAVA-1.5 (train)	5.8	4.7	12.4	6.0	5.2	1.4	1.7	0.6
Molmo (train)	3.3	2.2	28.8	16.8	6.0	3.2	2.9	0.3

Table 1: Percentage Occurrences and Estimated True Occurrences of reasoning-related keywords in popular open-source image-text corpora and training datasets of open-source VLMs.

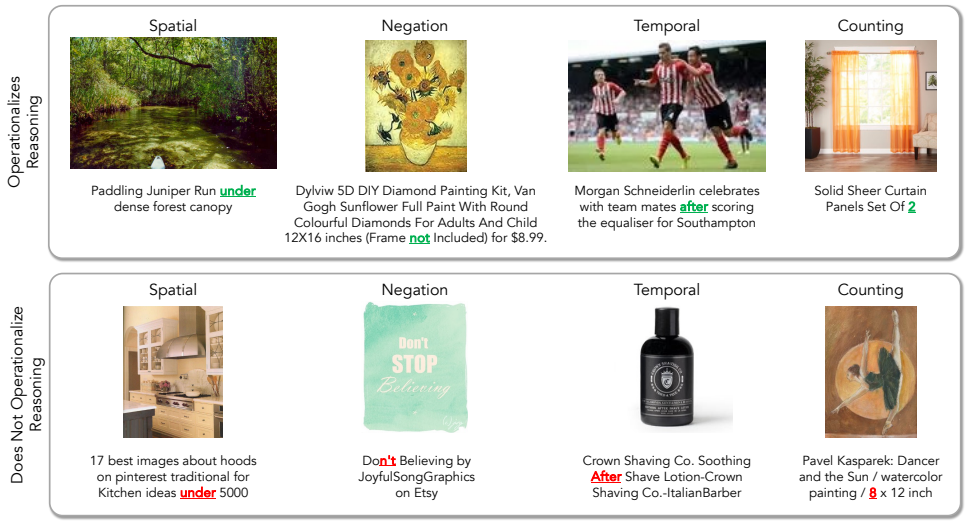

Figure 1: Examples from LAION-2B of data points that contain reasoning-related keywords that do and do not operationalize the reasoning capability itself.

explained by LLAVA's use of GPT4 as annotator for instruction tuning data. Our analysis shows that many of the false positives are in fact spurious descriptions that use counts and spatial language (e.g., a "left of" that is actually a "right of"), which is consistent with GPT4's weaknesses in reasoning. Had human annotators been employed, we expect to have observed higher true occurrences of counting and spatial language.

Finally, for Molmo's (Deitke et al., 2024) pre-training data, the annotators were instructed to discuss "objects and their counts" and "positions of the objects", among other non-reasoning-related instructions. Molmo's training data includes PixMo as well as other academic datasets, e.g. TallyQA (Acharya et al., 2019) and VQAv2 (Goyal et al., 2017). As seen in Table 1, specific instructions for counting and spatial leads to increased prevalence of spatial and counting reasoning. Without specific instructions, negations and temporal remaining remain low, as in LLAVA and LAION. It is important to note that there is additional data in PixMo to assist models with spatial reasoning and counting that is in the form of bounding box coordinates, and as such is not included in the above occurrence estimates.

4.2 Controlled Study

From these observations, we hypothesize that reporting bias occurs unless annotators are specifically prompted to include each type of otherwise-omitted information. To test this, we carry out a controlled study where annotators are given a fixed set of 100 images randomly sampled from COCO and requested to caption them. We provide them with one of four sets of annotator instructions: the original COCO captioning instruction, the LLAVA-1.5 captioning instruction, the PixMo captioning instruction, and instructions we write. We re-format the instructions slightly (e.g., PixMo captions were collected via audio, not text), but we retain the exact wording of what annotators were requested to include and not include in the captions. In our own instructions, we ask specifically for all four types of reasoning we study. All sets of instructions are provided in the Appendix.

We use Prolific[3] to collect participants for the study. They were asked to write a caption of at least 8 words (the minimum caption length in COCO), but were encouraged to make the captions as long as needed to include the requested information (which varied based on the instruction set). By not constraining the caption length, we mirror the tendency of people to communicate concisely (Maxim of Quantity). Annotators were paid $15 per hour of estimated work, with a bonus if they spent longer on the task. This allowed us to simulate the concise nature of communication (the annotators did not know they would be paid additionally) while paying annotators fairly.

We then check the 100 written captions for percentage occurrences as in Section 3.2, manually calculating the true positive rate. The results are shown in Table 2. When annotators are not asked to include anything specific, as in COCO, they do tend to use some spatial- and counting-related words, but no negation- or temporal-related words. Adding requests for spatial and counting reasoning, as in LLAVA-1.5 and PixMo, significantly increases the occurrence of words related to those types of reasoning, but not to temporal relations or negations. Thus, it is critical to be intentional with annotator instructions, if representation of various types of reasoning is desired. By specifically instructing all four phenomena, as in our instructions, we see that the prevalence of all four types of reasoning increases compared to COCO, showing that all types of reasoning *can* be elicited from annotators, *if* they are explicitly asked for the same. In terms of our aforementioned linguistics studies, by making the Question Under Discussion explicit, we are able to elicit the desired information.

We perform an additional study to determine whether forcing increased caption length (as in dense annotation schema, e.g., Deitke et al. (2024) requiring annotators to speak about the image for a full minute) yields reasoning-related information without the need for specific instructions. We find that it increases the occurrence of the types of reasoning people were already predisposed to in the original COCO study, but not of the other types of reasoning. Details are in the Appendix.

We will see in Section 6 that the occurrence of reasoning-related data in training predicts model performance on that type of reasoning. As such, our study makes it clear that annotator instructions are the key to overcoming reporting bias and improving model reasoning capabilities.

[3] https://www.prolific.com/ [accessed 2/2025]

Instructions	Spatial	Counting	Negation	Temporal
COCO	8	23	2	2
LLAVA-1.5	17	38	3	0
PixMo	21	43	12	1
Ours	14	39	52	44

Table 2: Percentage True Occurrences (manually calculated) of reasoning-related keywords in each set of 100 captions collected with different instructions for the controlled study.

5 Benchmarks

In this section we discuss the benchmarks we use to evaluate models on our four recognized phenomena. In several cases, we modify existing benchmarks to suit our needs. All benchmarks are multiple-choice caption options given an image, as shown in Figure 2. In the case of contrastive VLMs, e.g., CLIP, they are evaluated on image-text matching; in the case of generative VLMs, e.g., Molmo, they are evaluated in a multiple-choice QA setting (but for counting, as discussed below). These two are not comparable to each other, as only the latter has access to all options at once; however, they are both the most favorable settings for each type of model, allowing us to better estimate model capability.

Spatial reasoning. To evaluate spatial reasoning, we take the What'sUp benchmark from Kamath et al. (2023b). We use only Subset A of What'sUp, targeting four spatial relations: *on*, *under*, *left of* and *right of*. This data consists of an image of two basic household objects in a spatial relation to each other, with no distractors. It is perfectly balanced between the four possible prepositions and consists of 412 data points.

Counting. To evaluate counting, we take CountBench from Paiss et al. (2023). Originally consisting of an image with a caption directly from LAION containing some count of objects in the image (e.g., "background photo of three light bulbs"), we convert this dataset into the image-text matching format by manually reducing each caption to <count><objects> (e.g., "3 light bulbs") and adding alternate captions for each other count within 2–10. This has the added advantage of disallowing cheating by models who have seen the

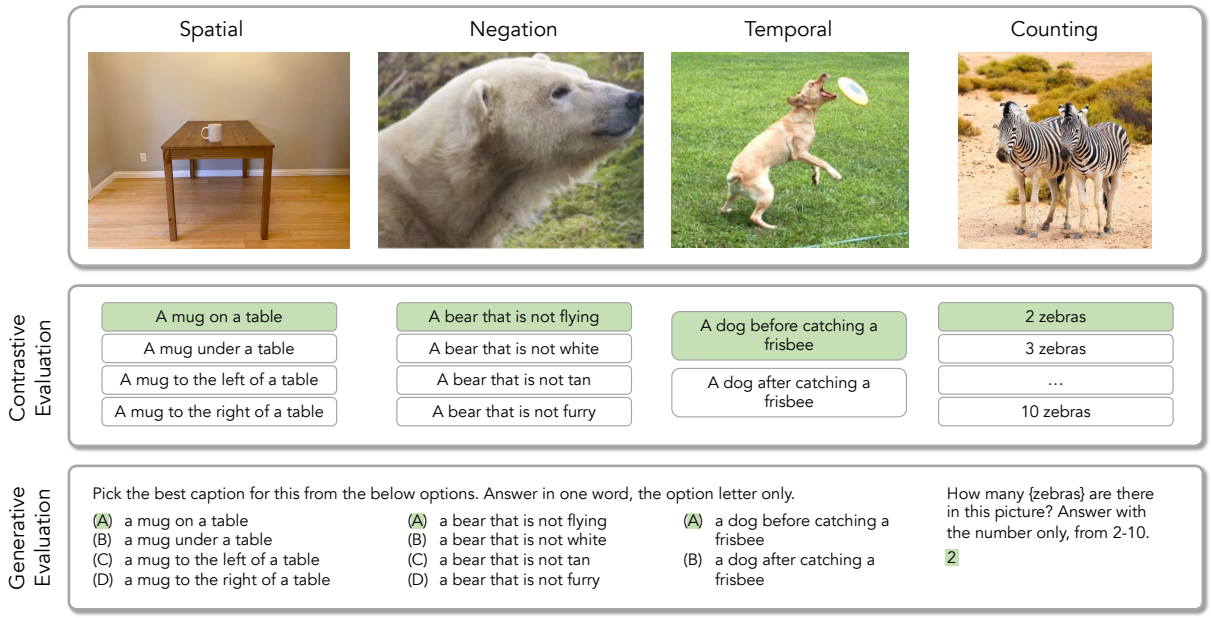

Figure 2: Examples from our four benchmarks for contrastive and generative evaluations. The generative evaluation is in MCQ format but for counting, for which a free form output with a given range yielded higher scores.

exact image-text pair from CountBench in LAION before. For generative VLMs, we find that all models we evaluate perform better when allowed to answer the question directly with a number compared to the MCQ format, and thus choose the former format for this task alone, as shown in Figure 2. There are 507 data points in this dataset, approximately balanced across the 9 counts.[4]

Negations. To evaluate negations, we repurpose the VAW benchmark (Pham et al., 2021). This dataset contains both positive and negative attributes for a given bounding box. We crop the bounding box (discarding those of insufficient size), then write the templated caption "a photo of a [object name] that is not [attribute]" with three positive attributes and one negative attribute, resulting in exactly one correct caption for the image. We obtain 800 such data points.

Temporal reasoning. To evaluate temporal reasoning, we begin with the ControlledImCaps benchmark (Kamath et al., 2023a), which contains *pairs* of images with corresponding captions. We select the temporal relation subset of this benchmark and split each data point into two data points with one image and two caption options each. We thus obtain 200 data points, perfectly balanced across the sets of two options.

[4]CountBench has 540 images and is perfectly balanced, but several images are no longer available [as of 02/2025].

6 Experiments and Results

Although we have shown that web-scale corpora do not tend to have significant representation of language related to the types of reasoning we study, it is not self-evident that models would necessarily be bad at these types of reasoning because of it. They may not require a significant amount of representation to learn the skill; or, alternately, they may be able to learn it from data that operationalized the skill but was not caught by our keyword-matching occurrence prediction.

In this section, we evaluate popular contrastive and generative VLMs on our benchmarks to ascertain their reasoning capabilities. We then study the effect on contrastive model performance of scaling both the model parameter size and the training data size, as well as the effect of adding multilingual diversity to the training data. Finally, we discuss the performance of popular closed-source VLMs on our benchmarks.

6.1 Models

Contrastive VLMs. We evaluate OpenCLIP (Cherti et al., 2023) models of different sizes: ViT-B/32, ViT-B/16, ViT-L/14, ViT-g/14, and ViT-H/14, as well as OpenCLIP ViT-B/32 trained with multilingual diversity, i.e, with non-English captions translated to English added to the data (Nguyen et al., 2024). We evaluate these models with an image-text matching task.

	Model	Spatial	Negation	Counting	Temporal
(a)	CLIP ViT-B/32	30.6	11.5	43.4	58.5
	+ ML Div.	27.4	15.5	23.3	51.5
	CLIP ViT-B/16	27.7	12.7	48.1	55.0
	CLIP ViT-L/14	28.4	12.3	64.1	52.0
	CLIP ViT-g/14	28.4	12.7	59.0	52.0
	CLIP ViT-H/14	26.0	13.2	60.0	59.0
(b)	LLAVA-1.5-7B	37.6	33.4	47.3	72.5
	LLAVA-1.5-13B	61.7	28.4	48.9	74.5
	Molmo 7B-O	75.5	38.4	77.5	78.0
	Molmo 7B-D	87.6	41.3	83.8	80.5
(c)	LLAVA-1.6-m7B	60.0	40.6	52.9	70.0
	QwenVL 7B-Chat	47.1	24.2	84.6	67.5
	Qwen2VL 7B-Inst.	98.3	56.1	85.8	84.0
	GPT4o	91.5	22.2	90.9	95.0
	GPT o1	97.6	64.7	88.2	97.0
	Gemini 1.5-Flash	98.5	46.4	84.6	81.5
	Gemini 1.5-Pro	92.0	49.0	87.8	85.0
	Claude-3 Haiku	65.5	28.9	83.4	70.0
	Claude-3.5 Sonnet	95.4	42.0	92.3	83.5
	Random Chance	25.0	25.0	11.1	50.0
	Human Estimate	100	100	100	100

Table 3: Results on our benchmarks of: (a) Contrastive VLMs, (b) Open-Source Generative VLMs, (c) Closed-Source Generative VLMs. All models fall far behind human performance on multiple types of reasoning.

Generative VLMs. We evaluate two generative VLMs that have completely open-source training data: LLAVA-1.5 (Liu et al., 2024a) and Molmo (Deitke et al., 2024), corresponding to our studies in previous sections. We further evaluate several generative VLMs with mixed- or closed-source training data: Qwen-VL (Bai et al., 2023), Qwen2-VL (Wang et al., 2024), LLAVA-1.6-Mistral (Liu et al., 2024b), GPT4o and o1 (OpenAI, 2024), Gemini-1.5 Flash and -1.5 Pro (Team et al., 2024), and Claude-3 Haiku and -3.5 Sonnet (Anthropic, 2024). For these models, we evaluate the model in a multiple-choice QA format (except for counting, which is free-form with a given range, as discussed in Section 5).

6.2 Results

Contrastive VLMs. Table 3(a) shows the performance of OpenCLIP models on our benchmarks. The contrastive VLMs score slightly above random chance on spatial reasoning and temporal reasoning, but score far less than random chance on negations. We find that CLIP tends to ignore negations, scoring the inverse of their attribute detection performance (c.f. Appendix). The models perform fairly well on counting, although it is worth noting that the counting benchmark was initially sourced from OpenCLIP training data.

Generative VLMs. Table 3(b) shows the performance of open-source generative VLMs. The generative models outperform the contrastive models on average, but fall far behind human performance across all tasks, especially negation. Scaling up LLAVA-1.5 significantly improves spatial reasoning performance, but no other type of reasoning.

6.3 Scaling Laws

In this section, we evaluate the aforementioned OpenCLIP models with different training data sizes (LAION-80M, LAION-400M, LAION-2B) and number of data points seen during training (3B, 13B, 34B), obtaining 32 models in total. Each of these is evaluated on our benchmarks to obtain scaling laws, as in Cherti et al. (2023). The resulting graphs are shown in Figure 3. In contrast to CLIP behavior on pure perception tasks such as ImageNet (Deng et al., 2009), where the loss drops steeply with an increase in data and/or parameter scale (Cherti et al., 2023), on our benchmarks we see different patterns: on spatial reasoning, the scaling law struggles to fit the data points, but it is clear that the loss does not drop with an increase in compute; on counting, increasing compute does seem to help, but noting the log scale, the amount of compute would need to be several orders of magnitude higher to reach human performance at 0 loss; on negation, increasing compute helps very slightly, but the loss remains very high ($\sim$87%), and an intractable amount of compute would be needed to reach human performance (at loss 0%); and on temporal reasoning, increasing compute does not improve performance.

Note that the prevalence of counting data far surpasses that of negations, temporal or spatial relations (c.f. Section 3.2), explaining its relatively high performance — although the frequency is still low on average compared to popular attributes, and the model performance is far behind human performance, which is 100%.

When we disentangle model scale from data scale, we see very similar trends. From this, we infer that neither scaling up the model size, nor the training data size, improves model performance beyond what is seen in Figure 3 — proving that the underlying problem of reporting bias cannot be mitigated with scale alone, as an intractable amount of compute in the form of training data and/or model parameters would be needed to reach human performance on these benchmarks.

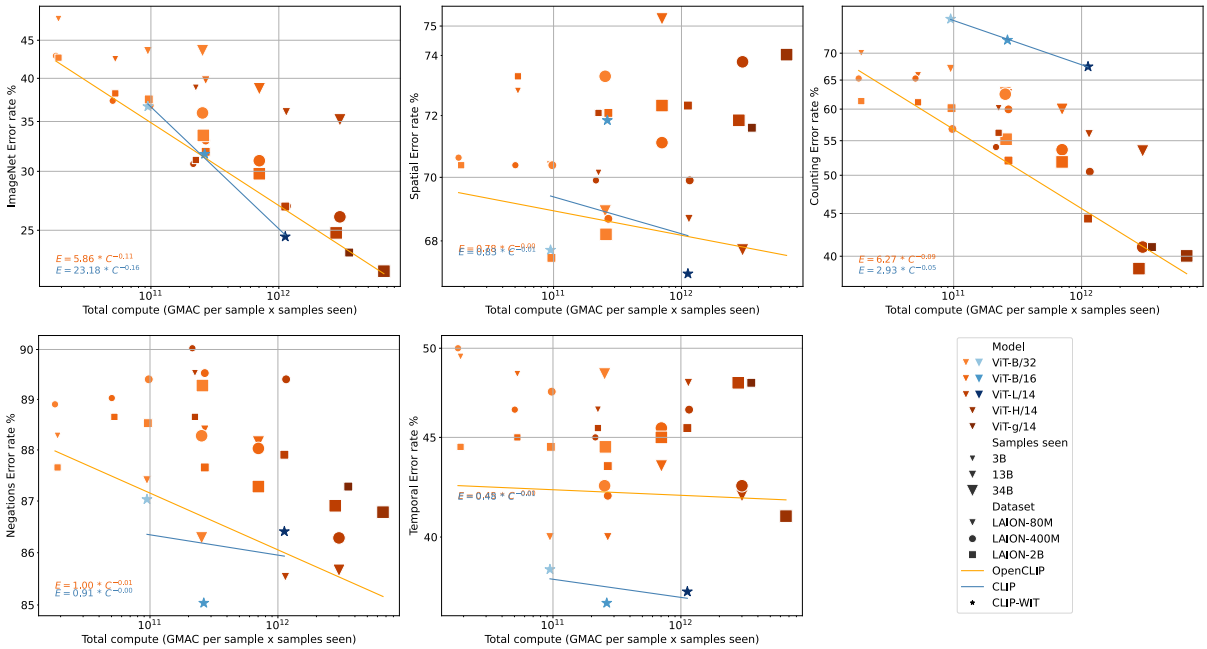

Figure 3: Scaling laws for OpenCLIP models on ImageNet (top left) compared to our benchmarks on spatial, counting, negation and temporal tasks respectively. Note the log-log plots and differing y axes across graphs.

6.4 Adding Captions from Other Languages

Nguyen et al. (2024) showed that adding multilingual diversity to the training data of contrastive VLMs by translating non-English web-scraped alt-text to English can significantly improve their performance on classification tasks; this work was rooted in Ye et al. (2024), which highlighted the difference in semantic content in images discussed when people using different languages captioned the same image. We ask: is leveraging this multilingual diversity sufficient to circumvent the reporting bias seen in image-text corpora? To study this, we evaluate the OpenCLIP ViT-B/32 model from Nguyen et al. (2024) on our benchmarks. As seen in Table 3, this model actually underperforms the OpenCLIP ViT-B/32 model trained on LAION English captions alone — showing that these types of reasoning are omitted by *all* speakers.

6.5 Closed-Source Generative Models

Top-performing closed-source models perform well on our benchmarks, although they still fall behind human performance, especially on negation and temporal reasoning. As the details behind the data collection and training are not public, it is difficult to draw inferences from these results; however, the importance of data quality in addition to scale is clear from the efforts invested in data collection (OpenAI, 2024).

7 Conclusion and Future Work

We study the *reporting bias* in vision-language: specifically, the systematic omission of types of information by people captioning images, which then form the image-text corpora popular VLMs are trained on. By identifying human behaviors rooted in linguistics, pragmatics, and cognitive science, we predict the types of information omitted, verify their lack in public image-text corpora, and show that contrastive and generative VLMs trained on this data perform poorly on the types of reasoning corresponding to the missing information. Further, we reveal the importance of the instructions provided to annotators during data collection, showing that intentional collection shows promise in improving representation of reasoning-related data in training corpora, which could in turn improve reasoning capabilities of VLMs.

Future research directions include: (1) automating the identification of significant gaps in text and images in training corpora; (2) synthesizing data to fill those gaps; (3) finetuning models on augmented data using different methods, e.g., hard negative finetuning of contrastive models, or instruction finetuning of generative models; and (4) eliciting captions that avoid the reporting bias in a more natural way than programmatic augmentation, e.g., by identifying communicative intents that naturally call for this information.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084.

Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. 2024. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *Advances in Neural Information Processing Systems*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023a. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023a. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The hard positive truth about vision-language compositionality. *European Conference on Computer Vision (ECCV)*.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, and Meng Cao. 2024. Veclip: Improving clip training via visual-enriched captions.

Ronald W Langacker. 2015. Construal. *Handbook of cognitive linguistics*, 39:120–142.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving multimodal datasets with image captioning.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. In *Advances in Neural Information Processing Systems*, volume 37, pages 91430–91459.

OpenAI. 2024. Gpt-4o system card.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Roy Pea. 1978. *The development of negation in early child language*. Ph.D. thesis, University of Oxford.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13018–13028.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Leonard Talmy. 1972. Semantic structures in english and atsugewi.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Ye Tian and Richard Breheny. 2016. Dynamic pragmatic view of negation processing. *Negation and polarity: Experimental perspectives*, pages 21–43.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2025. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *Advances in Neural Information Processing Systems*, 37:61735–61792.

Christiane Von Stutterheim and Wolfgang Klein. 1989. Referential movement in descriptive and narrative discourse. In *North-Holland Linguistic Series: Linguistic Variations*, volume 54, pages 39–76. Elsevier.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2024. Computer vision datasets and models exhibit cultural and linguistic diversity in perception.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

A Appendix

A.1 Details about Occurrence of Reasoning-Related Keywords

Keywords. The keywords we search for are: (1) "on top of", "under", "left of" and "right of" for spatial reasoning; (2) "before" and "after" for temporal reasoning; (3) "two"–"ten" and "2"–"10" for counting; and (4) "not" and "n't" for negations.

Estimating True Positive Rate. We discard keyword occurrences that do not operationalize the types of reasoning we study, e.g., "jeans under $25" does not encourage spatial reasoning.

A.2 Details about the Controlled Study

Instructions provided. The instructions provided to annotators are kept as close as possible to the original papers, with the reasoning-related words kept verbatim. Instructions are visible to the crowdworkers as they scroll through the images they annotate, as shown in Figure 4.

Figure 4: Instructions provided for the COCO (top left), LLAVA-1.5 (top right), PixMo (bottom left) and our (bottom right) sets of instructions.

Length experiment. We study whether asking annotators to write longer captions increases the types of reasoning represented. We collect an additional 50 captions of the first 50 COCO images from our study, with the same instructions as COCO captions. However, we require here that the captions are all at least 50 words. In these 50 captions, 10 have spatial reasoning, 25 have counting, and none have negations/temporal reasoning. The prevalence of spatial and counting is about double that of the study with an 8-word minimum. It is clear that increasing the caption length does encourage some types of reasoning, but it does not serve as a solution to increasing representation of all types of reasoning.

Counting. We see that the majority of object counts are the number 2, which is easy for annotators to count. However, upon closer inspection of the data, we also see that there are simply fewer images with >2 instances of any given object. This highlights the need to study reporting bias in the image space as well, rather than the text space alone, as discussed in Section 7.

A.3 Qualitative Observations

CLIP ignores negations. When evaluating negations, we observe that CLIP's performance on negated attributes ≈ 100 – attribute recognition performance. To investigate, we evaluate object negation, and find that CLIP's performance on negated objects $\approx$ 100 – object recognition performance: the data points on which CLIP gets the negated attribute/object correct are those on which it gets the attribute/object incorrect; showing that the model completely ignores the negation.

Models can count to smaller numbers better. When evaluating counting, we observe that contrastive and generative VLMs both perform better when counting small numbers than when counting large ones; which also correlates with the numbers' appearance in the training data: annotators are more likely to count smaller numbers of objects — as the number increases, they default to approximations such as "group of" and "several".

"Left" and "right" are the most difficult spatial relations for VLMs. Both contrastive and generative models struggle more with "left" and "right" than with "on" and "under". This also correlates with the relations' appearance in the training data, and validates our earlier hypotheses: due to the inherent ambiguity in these two relations ("left" from which perspective?), symmetric relations like "next to" are preferred over asymmetric types of grounding by annotators.

Contrastive VLMs can ignore keywords even when they do occur in the training data. We show that the phenomena we study are included rarely in captions. When they *are* included, though, it tends to be after the most salient information of the image is already captured by the caption, i.e., they are included as a "least significant bit" of information. As such, the contrastive loss allows the model to ignore these parts of the caption completely, as the salient image features are sufficient to retrieve the image in the batch.