Explaining CLIP's performance disparities on data from blind/low vision users

Daniela Massiceti[†] Camilla Longden[†] Agnieszka Słowik[†] Samuel Wills[◇]
Martin Grayson[†] Cecily Morrison[†]

[†]Microsoft Research [◇]The World Bank

Abstract

Large multi-modal models (LMMs) hold the potential to usher in a new era of automated visual assistance for people who are blind or low vision (BLV). Yet, these models have not been systematically evaluated on data captured by BLV users. We address this by empirically assessing CLIP, a widely-used LMM likely to underpin many assistive technologies. Testing 25 CLIP variants in a zero-shot classification task, we find that their accuracy is 15 percentage points lower on average for images captured by BLV users than web-crawled images. This disparity stems from CLIP's sensitivities to 1) image content (e.g. not recognizing disability objects as well as other objects); 2) image quality (e.g. not being robust to lighting variation); and 3) text content (e.g. not recognizing objects described by tactile adjectives as well as visual ones). We delve deeper with a textual analysis of three common pre-training datasets: LAION-400M, LAION-2B and DataComp-1B, showing that disability content is rarely mentioned. We then provide three examples that illustrate how the performance disparities extend to three downstream models underpinned by CLIP: OWL-ViT, CLIPSeg and DALL-E2. We find that few-shot learning with as few as 5 images can mitigate CLIP's quality-of-service disparities for BLV users in some scenarios, which we discuss alongside a set of other possible mitigations.

1. Introduction

AI-based applications hold the potential to help people who are blind and low vision (BLV) with everyday visual tasks [3, 5]. However, the popularity of video-calling services like Be My Eyes [1] suggest that human assistance is still often required due to the wide set of assistance tasks [44] and varying quality of BLV images [8, 17]. Recent advances in large multi-modal models (LMMs) [19, 49, 52] could potentially address these challenges, enabling a new era of automated visual assistance as highlighted by the early partnership between Open AI and Be My Eyes [2].

Despite the opportunity, little work has evaluated how well LMMs perform on data from BLV users. Performance disparities have been identified for other user groups [6, 36, 45, 52, 55, 66] but the evidence for BLV users is either anecdotal [49] or not specific to *large* multi-modal models [8]. Since BLV users are likely to be one of the biggest beneficiaries of LMMs, often in productivity- and safety-critical situations, it is important to extend studies to this group.

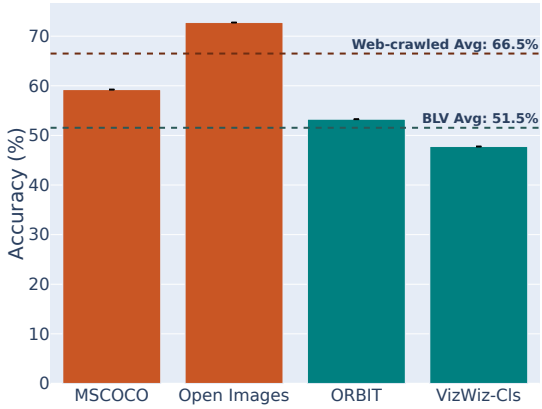

Figure 1. **CLIP's zero-shot object recognition accuracy is 15 percentage points lower in images from BLV users (ORBIT, VizWiz-Classification) versus web-crawled images (MSCOCO, Open Images).** Average accuracy (with 95% c.i.) in a standardized zero-shot image classification task is reported over 80-100K images per dataset for 25 CLIP variants.

To address this, we systematically evaluate CLIP, a widely used LMM with 8700+ citations and 24M+ downloads[1], on data from BLV users. CLIP's rich embeddings and strong zero-shot capabilities have led to it underpinning a wide range of downstream tasks including image classification [52], object detection [41, 42], semantic segmentation [37], image captioning [61, 63] and video recognition [35]. It has also been used to create large-scale datasets [23, 34, 57, 58] and evaluation metrics [26, 50]. As CLIP's pre-trained parameters are often used directly, poor performance can have wide-ranging implications for downstream assistive applications that use them.

We investigate CLIP's performance on BLV data along three dimensions: image content, image quality, and textual

[1]Statistics taken from Google Scholar and OpenAI's Hugging Face Hub (for CLIP ViT-L/14, ViT-B/32 and ViT-B/16) on 23 October 2023.

content. Visual content considers how well CLIP can recognize BLV-specific objects, such as guide canes. Visual quality assesses robustness to quality variations that characterize BLV images, such as blur and atypical framing [17]. Textual content examines performance on tactile descriptive words used by BLV users in contrast to visual ones, for example "plastic" versus "yellow". We study each dimension in the context of a zero-shot image classification task, providing a worst-case estimate on how well CLIP will serve downstream assistive applications if used out-of-the box.

Overall, we find that CLIP's zero-shot classification accuracy is 15 percentage points lower on BLV images compared to web-crawled images across 25 CLIP variants. These variants span architecture size (ViT-B/16 to ViT-g/14), pre-training dataset (WIT [52], LAION [57, 58], DataComp/CommonPool [23]) and pre-training dataset size (80M to 3.8B). On deeper inspection, underperformance stems from CLIP: 1) recognizing disability objects less well than non-disability ones, with 25 percentage points lower accuracy; 2) being sensitive to image quality, particularly occlusion and lighting issues; and 3) recognizing objects described by material less well than color, with discrepancies of 7 percentage points. In all cases, a larger pre-training dataset or architecture does not lead to parity.

To further understand our results, we examine the upstream source and downstream impact of these disparities. First, we conduct a textual analysis of the captions in LAION-400M/2B and DataComp1B and find that disability objects and materials are mentioned $\sim$17x and $\sim$4x less frequently than non-disability objects and colors, respectively. Second, we find performance disparities on BLV data persist in three downstream models that use CLIP: OWL-ViT [41] for object detection, CLIPSeg [37] for semantic segmentation, and DALL-E2 [53] for text-to-image generation. We close by discussing a set of possible mitigations, including few-shot model adaption and application-level solutions, toward making automated visual assistance for BLV users more equitable.

In summary, our work contributes to the literature on how LMMs perform for users in the margins, specifically highlighting how CLIP may underperform for BLV users if integrated into assistive applications. Our contributions are:

- An empirical study of CLIP's performance on BLV image content, image quality and textual content.
- The first quantification of BLV content representation in LAION-400M, LAION-2B, and DataComp-1B.
- An example-based analysis that illustrates how performance disparities on BLV data persist in three downstream models that use CLIP.

2. Related Works

Large multi-modal models. LMMs now have impressive capabilities in analyzing and synthesizing images [7, 16, 19, 30, 49, 52, 68]. Contrastive models [30, 52, 68], a prominent sub-class, learn joint image and text embeddings by training on massive web-crawled data using a contrastive loss [15, 48]. They are unique in their architecture scale, and in the way they are trained on web-crawled data in an unsupervised manner. Unlike previous models, the rich embeddings they learn are leveraged by a wide range of downstream models – either directly [21, 52], or as part of a larger system [10, 35, 37, 41–43, 46, 53, 61–63, 67].

LMMs and fairness. LMMs are known to have social biases across gender, race, age, and geographic location [6, 36, 45, 66]. CLIP, for example, has been shown to classify people of color as non-human and males as criminal more often than white people and females, respectively [6]. Some works have studied these representational harm for people with disabilities, however only in natural language [28]. Quality-of-service harms arise when an application underperforms or fails for a particular user group [13, 18, 65] – *e.g.* a facial recognition system that does not detect women with darker skin tones [14]. These can be systematically identified and mitigated through disaggregated reporting of a model's performance [9, 47]. This has not been well studied for people with disabilities generally or BLV people specifically, with the evidence either anecdotal (*e.g.* GPT-4Vision model card [49]) or not specific to LMMs [8].

3. Methodology

Our work investigates CLIP's robustness to image and text data from BLV users in the context of a zero-shot image classification task. This provides a worse-case estimate of how CLIP will perform out-of-the-box in downstream assistive applications. Here we describe the experimental set-up, CLIP variants, and datasets used in our analyses.

3.1. Episodic zero-shot image classification

An image classifier selects which object $c \in \mathcal{C}$ is present in an image, where $\mathcal{C}$ is the set of possible object classes and $|\mathcal{C}|$ is the task's "way". A zero-shot classifier does this without seeing any training images of the classes beforehand.

Our first analysis compares CLIP's performance on different datasets (rather than the more typical multiple models on a single dataset), requiring our classification task set-up to be standardized across datasets. We take inspiration from the episodic sampling used in meta-learning [22]: for each dataset j annotated with $\mathcal{C}_j$ object classes, we sample T fixed N-way classification tasks, where for each task we randomly sample N classes from $\mathcal{C}_j$. For each task, we randomly sample M test images per class. The classification accuracy is then computed for all $T*M*N$ images and the average (and 95% confidence interval) is reported. We repeat this for each dataset, with T, N and M held constant. We use variations of this to compare CLIP's performance

between object types (Sec. 4.1) and text prompts (Sec. 4.3) with details provided in each section, respectively.

3.2. Logistic Regression

We also aim to understand which characteristics *within* images and text affect CLIP's performance. We use logistic regression, a common tool for hypothesis testing, to estimate the marginal effect of each characteristic on the model's accuracy. This approach avoids the need for careful experimental set-up which controls for all factors except the variable of interest. Logistic regression extends Ordinary Least Squares (OLS) regression to the case when the output variable is binary, as is our case where the model correctly identifies the ground-truth object or not. Formally, we use:

$$p(z_i) = \frac{1}{1 + e^{-z_i}} \tag{1}$$

where $z_i = \alpha_1 + \beta_1 X_i + \alpha_2 D_i + \beta_2 D_i X_i + \epsilon_i$. The output variable is $p(z_i) \in [0, 1]$, the probability that the model correctly identifies the ground-truth object in image i, with 1 for correct, and 0 otherwise. The explanatory variables are X_i, a vector of binary variables that encode whether a particular characteristic is present in image i, and D_i is a binary variable indicating whether the ground-truth object is a disability object (*e.g.* a guide cane). The interaction term $\beta_2 D_i X_i$ measures whether the marginal effect of each characteristic in X_i is compounded or mitigated for disability objects relative to non-disability objects. ϵ_i are residuals which are assumed homoskedastic and uncorrelated.

The coefficients $\alpha_1, \beta_1, \alpha_2, \beta_2$ are estimated through maximum likelihood. In OLS the coefficients directly represent the marginal effect of each X_i variable on the dependent variable. In contrast, here they represent the marginal effect on the log-odds ratio, which is linear in X_i:

$$\ln\left(\frac{p(z_i)}{1 - p(z_i)}\right) = \alpha_1 + \beta_1 X_i + \alpha_2 D_i + \beta_2 D_i X_i + \epsilon_i \tag{2}$$

This makes the coefficients difficult to interpret so we instead report them as $\partial p / \partial x$, the marginal effect of each characteristic $x \in X$ on the model's probability of being correct, p. We report the average of this marginal effect across all observations in the sample. We interpret each effect through its sign, magnitude, and significance. A negative sign means the model is less likely to be correct when that characteristic is present in an image – on average and holding all other characteristics constant. Its magnitude measures the extent of this impact. Its significance indicates its reliability based on a two-sided t-test that estimates the probability that the marginal effect is different from zero.

3.3. CLIP variants

We study 25 CLIP variants spanning architecture size, pre-training dataset, and pre-training dataset size (see Tab. A.1 for summary). We focus on variants that use a Transformer [64] and Vision Transformer (ViT) [20] as the text and vision encoders respectively as they are most widely used. Specifically, we consider ViT-B/16, ViT-B/32, ViT-L/14, ViT-H/14 and ViT-g/14 vision encoders with associated text encoders. For datasets, we consider OpenAI's closed-source WIT [52] and open-source LAION (80M/400M/2B) [30, 57, 58], DataComp (S/M/L/XL) [23], and CommonPool (S/M/L/XL) [23] with and without CLIP Score filtering [26]. These span 80M-3.8B image-text pairs.

We use CLIP as a zero-shot classifier by embeddding a task's class labels using its text encoder, and each task image with its vision encoder. An image's prediction is taken to be the class whose embedding has the highest cosine similarity (after a softmax) with the image's embedding.

3.4. Datasets

Our analyses are based on two large-scale datasets captured by BLV users: ORBIT [38] and VizWiz-Classification [8]. Both datasets were collected through real-world assistive applications: a personalizable object recognizer app for ORBIT [44]; and a visual question-answering app for VizWiz-Classification [12]. Both are therefore highly representative of typical BLV user data. We contrast these with two common web-crawled datasets – MS-COCO [33] and Open Images [32] – which are typical of the data used to pre-train LMMs, and widely used for benchmarking. We consider only the test and validation sets of these datasets. Below we provide descriptions of the BLV datasets, with the web-crawled datasets described in the appendix.

ORBIT [38] contains 3,822 videos (2.68M frames) of 486 objects collected by 67 BLV users on their mobile phones. For each object, users captured videos which show the object alone, and in a realistic scene alongside other items, which we call the Clean and Clutter datasets, respectively. ORBIT Clean frames are annotated with 6 quality issues (*e.g.* framing, blur) following the categories in [17].

VizWiz-Classification [8] contains 8,900 images from the original VizWiz dataset [25], a dataset of images taken by over 11,000 BLV users via a visual assistance mobile app [12]. All images are annotated with 200 ImageNet object categories and the 6 quality issues of [17] (including an extra "other" quality issue).

4. Experimental Results

Our first finding is that CLIP's accuracy is 15.0 percentage points lower on BLV datasets (ORBIT and VizWiz-Classification) than web-crawled datasets (MS-COCO and Open Images) (see Fig. 1). We use the standardized zero-shot set-up (see Sec. 3.1) and average the $T*N*M$ predictions per dataset from each of the 25 CLIP variants. While the accuracy difference is less for larger CLIP architectures than smaller ones, no model achieves parity (see Fig. B.1).

Table 1. **CLIP underperforms on disability and exclusive disability objects by significant margins compared to non-disability objects.** Zero-shot accuracy is averaged (with 95% c.i.) over 27.5K images of each object type processed by each of the 25 CLIP variants. Experimental details in Sec. 4.1.1.

Object Category	ORBIT Clean	ORBIT Clutter
Excl. disability	$36.5\% \pm 0.1\%$	$22.6\% \pm 0.1\%$
Disability	$41.8\% \pm 0.1\%$	$25.8\% \pm 0.1\%$
Non-disability	$\mathbf{58.9\% \pm 0.1\%}$	$\mathbf{50.9\% \pm 0.1\%}$

In the best case, the gap is 6.7 percentage points (ViT-g/14, LAION-2B) while in the worst, it is 22.8 percentage points (ViT-B/32, DataComp-M). This preliminary result hints at deeper issues. In the following sections, we aim to identify potential sources of this discrepancy and why it occurs.

4.1. Robustness to image content from BLV users

To understand why accuracy is lower, we first examine BLV image content. The BLV community uses a range of assistive objects, like guide canes and Braille displays [31, 38, 44] (see Fig. 2), which are not included in popular benchmarks [33, 54, 56]. We assess CLIP's performance on such "disability" objects versus more common objects.

We define disability objects as those that assist BLV people (*e.g.* dog collar); exclusive disability objects as the subset exclusively used by BLV people (*e.g.* guide cane); and non-disability objects as those used by everyone (*e.g.* keys). Three annotators categorized the ORBIT Clean and Clutter datasets[2] resulting in 55 disability, 42 exclusive disability, and 431 non-disability objects (see App. A.3 for lists).

4.1.1 Disability objects are less well recognized than non-disability objects

We compare zero-shot classification accuracy between disability and non-disability objects using a variant of the episodic set-up described in Sec. 3.1. Specifically, for each disability object we sample two N-way tasks with a "target" object and N-1 non-disability "distractor" objects. The first task contains a disability target object and the second task contains a non-disability target. The distractors are randomly sampled from the non-disability objects, each coming from a unique object cluster. We repeat T times for each disability object, sampling a pair of tasks with a different set of distractor objects and non-disability target object. For each task, we randomly sample M frames of the target object, and ask CLIP to classify them from the task's N possible objects. We report the average accuracy of all frames with a disability and a non-disability object as the target, respectively (T*55*M each). We also report the average accuracy over the subset of frames that are exclusive disability objects. We use $T = 5$, $N = 20$, $M = 100$.

[2]We do not consider VizWiz-Classification, as none of its 200 ImageNet labels are disability objects.

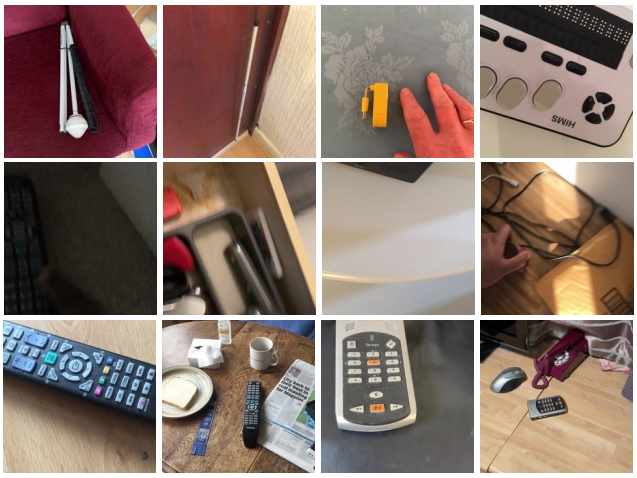

Figure 2. **Examples from the ORBIT Dataset**. (top) Disability objects: guide canes, liquid level sensor, electronic Braille device. (middle) Quality issues typical in BLV images: underexposure, blur, camera viewpoint, and framing. (bottom) A remote control and a Victor Reader Stream in a clean and clutter frame.

Under this setting, we find that disability and exclusive disability objects have accuracies of 21.1 and 25.3 percentage points less than non-disability objects, respectively, on average across the ORBIT Clean dataset (see Tab. 1). The gap widens by a further 3-4 percentage points when more realistic scenarios are presented from ORBIT Clutter. We find that the worst performing objects include Braille note-takers, talking book devices and liquid level indicators.

We also investigate the role of CLIP's pre-training dataset size on this finding. We find that accuracy increases with pre-training dataset size generally, but the delta between non-disability and disability objects stays roughly constant (see Fig. B.2). This suggests that web-crawling more data may not be enough to improve performance on potentially long-tailed objects. We see similar trends for increasing architecture sizes (see Fig. B.3).

4.1.2 Disability objects are under-represented in large-scale datasets compared to non-disability objects

To better understand why more pre-training data does not improve performance on disability objects, we analyze the composition of three of CLIP's large-scale pre-training datasets for the presence of disability content – LAION-400M [57], LAION-2B [58], and DataComp-XL [23] (also called DataComp-1B). These datasets are used for pre-training LMMs more broadly, with DataComp-XL achieving the highest accuracies on ORBIT.

Given the scale of the datasets, we conduct a text-based analysis of their captions as a more computationally tenable approach than analyzing their images. We first extract all noun phrases that contain a physical object[3] from the

[3]A physical object traverses the entity → physical-entity → object → OR(artifact, whole, part, living-thing) hypernym path in WordNet [40].

Table 2. **Disability objects occur 16-17x less frequently in the captions of popular large-scale image-text datasets compared to non-disability objects.** The mentions of 222 disability object synonyms and 312 non-disability synonyms were counted in noun phrases (NPs) extracted from these datasets. Details in Sec. 4.1.2.

	LAION-400M	LAION-2B	DataComp-1B
Captions	401,300,000	2,322,161,808	1,387,173,656
NPs	384,468,921	2,737,763,447	1,342,369,058
Unique NPs	5,984,181	22,657,632	15,071,341
Disability obj. mentions	18,326 (0.0048%)	70,939 (0.0026%)	48,672 (0.0036%)
Non-disability obj. mentions	425,046 (0.1106%)	1,550,043 (0.0566%)	1,126,356 (0.0839%)
Normalized non-dis/dis ratio	**16.8**	**15.6**	**16.5**

captions, referred to as "visual concepts"[4]. We then compute how prevalent ORBIT's disability and non-disability objects are contained in these visual concepts. We use ORBIT to contextualize our previous results as it is a realistic representation of the types of objects important to BLV users, however, other object lists could be used.

To do this, we first group similar objects from the ORBIT dataset into higher-level clusters (*e.g.* all guide canes). As each cluster could be described in several ways (*e.g.* "symbol canes", "guide canes"), we assign each two relevant synonyms. This was expanded to 15 synonyms for disability objects based on initial experimentation, resulting in 222 disability object synonyms, and 312 non-disability synonyms overall. We then count how many times each synonym appears within the visual concepts using string matching, allowing partial matches after simple pre-processing (see App. A.5 for details).

We find that disability objects occur 16-17x less frequently than non-disability objects across all three datasets (Tab. 2). We compute this by normalizing the number of mentions by the number of synonyms for disability and non-disability objects, respectively, and taking their ratio. We also see that LAION-2B has 7x the number of noun phrases as LAION-400M, but <4x the unique noun phrases, suggesting that it contains more of the same rather than new visual concepts (see App. A.5 for further statistics).

4.1.3 A few-shot approach can *sometimes* reduce the disability and non-disability accuracy gap

As CLIP is also known to be a good few-shot learner [60], we investigate whether providing several examples of an object can equalize performance between disability and non-disability content. We integrate a ProtoNets approach [59] with the "distractor" set-up described in Sec. 4.1.1, using embeddings directly from CLIP's vision encoder[5]. Specifically for each disability object, we sample pairs of N-way tasks in the same way, except now we additionally sample K training shots of each class which we use to compute the class prototypes. As before, we evaluate the model on M test images for the disability and non-disability target object in each task pair, with the prediction taken to be the closest prototype. We consider $K = [5, 10, 20, 40]$. We compare this to recent methods [39, 51] which improve CLIP's zero-shot performance by embedding LLM-generated descriptions of objects (rather than just the raw labels). We use GPT-4 as the LLM and the same generation hyperparameters as [39, 51].

Table 3. **A few-shot method using ProtoNets [59] (5-shot) achieves the highest accuracy and lowest accuracy gap between disability and non-disability objects, versus vanilla CLIP (0-shot) and CLIP with LLM-generated object descriptions [39, 51].** Averaged over 25 CLIP variants.

Obj type	ORBIT Clean Acc (%)				ORBIT Clutter Acc (%)			
	0-shot	[39]	[51]	5-shot	0-shot	[39]	[51]	5-shot
Disability	41.8	48.3	50.1	**86.2**	25.8	32.1	34.2	**54.5**
Non-disability	58.9	57.0	57.0	**88.3**	50.9	50.2	49.4	**69.1**
Accuracy gap	17.1	8.7	6.9	**2.1**	25.1	18.1	15.2	**14.6**

We find that augmenting CLIP with LLM-generated object descriptions [39, 51] outperforms vanilla CLIP (0-shot) which just embeds the raw object labels, but not a few-shot approach (5-shot) which embeds a few image examples of each object (see Tab. 3). This holds for both the ORBIT Clean and Clutter datasets. Crucially, the accuracy gap between disability and non-disability objects is lowest with a few-shot approach, though this accuracy gap quickly saturates, with no significant gains coming from more than 5 shots (see Fig. B.4). We also note that while a few-shot approach can reduce the accuracy gap to 2% in the simple images from ORBIT Clean, it is less effective in the more realistic images from ORBIT Clutter, with disability objects performing 14-15% points worse than non-disability objects, even when scaled to 40 shots (see Fig. B.4b).

Furthermore, a few-shot approach is only effective as a mitigation if CLIP is pre-trained on a large enough dataset. We find that for pre-training datasets of less than 100M examples, the accuracy difference is 3-4x larger than that for 100-1000M examples, and 9-10x larger than that for 1B+ examples (see Figs. B.5a and B.5b). These factors are roughly constant across the number of shots. Overall, this speaks to the power of large-scale pre-training, even if a small amount of extra effort is required.

4.2. Robustness to image quality from BLV users

Images captured by BLV users are of more variable quality than those captured by sighted users. These issues include atypical framing, camera blur, camera viewpoint (rotation), occlusion, overexposure, and underexposure [17, 31], which are annotated in the ORBIT Clean and VizWiz Classification datasets. We run the standardized zero-shot set-up (see Sec. 3.1) on these datasets for all CLIP variants. We then use the statistical tools described in Sec. 3.2 to

[4]We release these publicly at [REMOVED FOR REVIEW]

[5]We note that this few-shot set-up does not use CLIP's text encoder.

disentangle the marginal effect of each quality issue on model performance, both in general and for disability objects specifically. For ORBIT, we treat X_i as a binary vector indicating the presence of five quality issues[6] in image i, D_i as a binary indicating the presence of an exclusive disability object, and $D_i X_i$ as the interactions between them. For VizWiz, we encode seven quality issues (including the "other" category) in X_i, but exclude D_i or $D_i X_i$ as VizWiz labels do not include disability objects.

4.2.1 Blur, viewpoint, occlusion and lighting issues significantly reduce model accuracy.

In Fig. 3, we show that the marginal effects of blur, viewpoint (rotation), occlusion, and lighting issues on model accuracy are negative, large, and statistically significant for most models. All else equal, blur reduces model accuracy by 11 percentage points and 1 percentage point in the ORBIT and VizWiz datasets, respectively, on average. Viewpoint issues by 9 and 8 percentage points on each dataset respectively; occlusion by 9 and 14 percentage points; and lighting issues by 23 and 8 percentage points. We note that these effects are cumulative meaning that the impact on model accuracy is summed if multiple issues occur in the same image. We also note that pre-training on larger datasets, in general, does not guarantee robustness (*e.g.* variants pre-trained on LAION-2B, one of the largest datasets, are negatively affected by viewpoint and occlusion issues by 3-12 and 8-19 percentage points, respectively). We include the raw marginal effects in Tabs. B.3 to B.5.

Framing issues in the ORBIT dataset stand as the exception, with the marginal effect being positive and statistically significant. This can be explained by how the ORBIT videos were collected. To orient the camera, BLV users were instructed to hold it close to the object initially, and then move away. So, the initial frames in the video tend to be at close range – an easier recognition task – but also have framing issues. This is supported by the VizWiz results where framing issues, which occur at further distances from the object, have a negative marginal effect on accuracy.

4.2.2 The impact of quality issues is typically not worse for disability compared to non-disability objects.

Fig. 3 further shows that accuracy is 29 percentage points lower for exclusive disability objects than non-disability objects in the ORBIT Clean dataset, on average across all models, supporting the findings in Sec. 4.1.1. The marginal effect of a quality issue, however, typically affects disability objects no worse than non-disability ones. This can be seen by comparing the net effect of a quality issue on each object type. Let the baseline be the accuracy for non-disability objects. The accuracy for a disability object with no quality

[6]We combine over- and underexposure into a joint "lighting issue" due to low incidence rates of each of these issues.

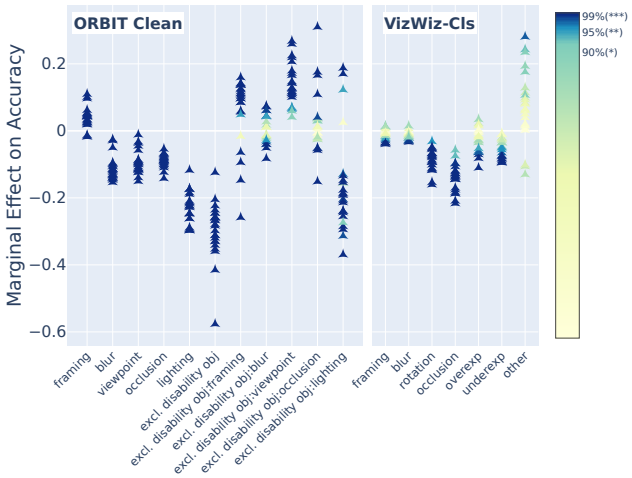

Figure 3. **Blur, viewpoint/rotation, occlusion and lighting issues all have large negative marginal effects on model accuracy, with high statistical significance, but these are not compounded for exclusive disability objects.** Each dot represents a CLIP variant, with its color showing the significance level.

issues will be 29 percentage points lower. Introducing occlusion will reduce the accuracy for non-disability objects by 9 percentage points on average. For disability objects, occlusion will reduce accuracy by this, plus the marginal effect of the interaction term (+2 percentage points), for a net effect of -7. The positive and significant interaction term indicates that having an occlusion issue and being a disability object has an effect that is slightly less than the sum of its parts. The only exception is overexposure issues, which do compound if they co-occur with a disability object.

4.3. Robustness to language used by BLV users

Assistive applications are likely to leverage the multi-modal capabilities of LMMs, so it is important to understand how CLIP performs on the range of language used by BLV people. For example, BLV users commonly use tactile rather than visual words to describe their objects [44]. In this section, we study one instantiation of this – CLIP's robustness to recognizing objects described by their color, "yellow mug", versus their material, "plastic cup".

To do this, three annotators manually labeled the ORBIT validation and test objects (208 objects) with a color and a material[7]. Each adjective was selected from a predefined list of 20 colors and 23 materials (see App. A.4). A text prompt was then created for each object using the template "<*adjective*> <*object_name*>", where <*adjective*> was the object's color or material, and <*object_name*> was the noun extracted from the raw object label. We use these templates – referred to as color and material prompts – to examine CLIP's sensitivity to different object descriptions.

[7]We assigned up to 2 adjectives per object in some cases where objects were multiple colors or materials.

Table 4. **Describing an object by its color (rather than material, or color and material) leads to text embeddings that are most aligned with that object's image embeddings.** CLIP scores [26] between image and prompt embeddings are averaged (with 95% c.i.) for 100 images per object per prompt type on ORBIT Clean.

Prompt	Obj. name	Material + obj. name	Color + obj. name	Color + material + obj. name
CLIP Score	24.07 ± 0.02	23.88 ± 0.02	**25.20 ± 0.02**	24.76 ± 0.02

4.3.1 CLIP classifies objects more accurately when they are described by color rather than material

We compute CLIP scores [26] between an image and four different prompt embeddings, for 100 randomly sampled images of each object in ORBIT Clean. We consider the color and material prompts, a lower bound containing just the object name, and an upper bound adding both color and material adjectives. We expect that the lower bound prompt, which provides the least detail about the object, should align less strongly with the object's image embedding than the upper bound prompt, which provides the most specific detail. In Tab. 4, however, we see this is not the case. Rather, color prompts have the highest CLIP scores and material prompts the lowest. Interestingly, the upper bound has a lower average CLIP score than the color prompt, suggesting that adding the object's material is harming alignment.

To quantify the impact of this on accuracy, we run the standard zero-shot set-up (Sec. 3.1), embedding these textual prompts instead of the raw object labels. We see that across all variants, CLIP classifies objects 7.1 percentage points more accurately when they are described by their color rather than their material (see Fig. B.6).

4.3.2 Materials are under-represented in large-scale datasets compared to colors

We further examine this result by measuring how frequently colors versus materials appear in the captions of LAION-400M, LAION-2B and DataComp-1B. We use the extracted noun phrases from Sec. 4.1.2, and count the number of times the 20 material and 23 color annotations are mentioned. In Tab. 5, we see that colors are mentioned ∼4x more frequently than materials across both datasets, once normalized. This helps to explains some of the results

Table 5. **Materials occur ∼4x less frequently than colors in the captions of popular large-scale image-text datasets.** The mentions of 20 colors and 23 materials were counted in the noun phrases extracted in Tab. 2.

	LAION-400M	LAION-2B	DataComp-1B
Color mentions	475,060 (0.12%)	1,756,102 (0.06%)	1,165,871 (0.09%)
Material mentions	131,876 (0.03%)	513,014 (0.02%)	354,598 (0.03%)
Norm'ed color/ material ratio	**4.1**	**3.9**	**3.8**

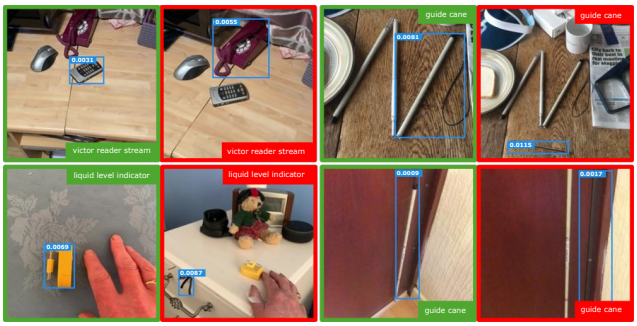

Figure 4. **OWL-ViT [41] detects disability objects less consistently than non-disability objects**. Disability objects are often mistaken for other objects, sometimes with higher confidence.

in Sec. 4.3.1. Taken together, this suggests that models pre-trained on these datasets may perform worse for BLV users who describe their objects by their material, with the potential that this may extend to other tactile-based descriptions.

5. Example-based impact analysis

Sec. 4 broadly shows that CLIP is sensitive to image and textual data provided by BLV users in a zero-shot classification task. We investigate whether these performance disparities persist in three downstream models that use CLIP – OWL-ViT [42], CLIPSeg [37], and DALL-E2 [53]. We run our analysis on 180 BLV images which are systematically selected for 20 objects – the 5 top- and bottom-performing disability and non-disability objects from the ORBIT dataset (see App. C for full protocol). For space reasons, we include CLIPSeg results in App. C.3.

5.1. Object detection with OWL-ViT

Object detection is already widely available in BLV assistive applications [3, 5], and in future, many may rely on models that use CLIP, such as OWL-ViT [41]. OWL-ViT predicts bounding boxes for objects specified in free-form text prompts. It does this by appending a bounding box regression and class-wise layer to CLIP's (pre-trained) encoders and then fine-tuning on an object detection dataset. We run all 180 images through OWL-ViT (with a ViT-B/32 vision encoder) with the (cleaned) noun phrase extracted from the raw object label as the text prompt. A team of three annotators then manually evaluated the detections. We find:

Disability objects are less consistently detected than non-disability objects. Our results show that 6/10 non-disability objects were correctly detected (taken as the box with the highest confidence) in all 9 frames showing that object, compared to 3/10 disability objects. In many of these failed frames, the model mistook the disability object for another object, often with a higher confidence (Fig. 4). This behavior would have a large negative effect on the user experience of an object detection app.

Table 6. **OWL-ViT [41]'s correct bounding box predictions have confidence scores that are ∼5x lower for disability than non-disability objects on average**. The confidence score of the predicted box per image is averaged (with 95% c.i.) over 90 images for disability and non-disability objects, respectively.

Object	Correct boxes	Incorrect boxes
Dis. objs	0.016 ± 0.008	0.008 ± 0.003
Non-dis. objs	0.084 ± 0.030	0.008 ± 0.003

The model is less confident about disability object detections than non-disability object detections. In Tab. 6, we see that OWL-ViT's confidence for the correct bounding box is ∼5x lower for disability objects compared to non-disability objects. We see that incorrect boxes have similar confidence scores between disability and non-disability objects, which is expected. See examples in Fig. C.1.

5.2. Text-to-image generation with DALL-E2

DALL-E2 [53] also uses CLIP: during training its decoder is conditioned on image embeddings from frozen CLIP. We investigate the downstream impacts of this by examining if DALL-E2 can generate disability content. We create two prompts for each of the 20 objects using the templates: i) "*<object_name>*" ii) "*<object_name> on <surface> next to a <adjacent-object>*". The object name was the object label's cleaned noun phrase, and the surface/adjacent object was chosen to match a randomly sampled clutter image of that object (see App. C for details). Three annotators then manually evaluated four generations from DALL-E2 per prompt. A generated image was considered correct if it contained the object specified in the prompt. We find:

Generations of disability objects are more likely to be incorrect compared to non-disability objects. DALL-E2 correctly generated the object in the prompt for 18/80 images of disability objects, versus 74/80 images of non-disability objects. For some disability objects, no generations contained a valid representation of the object – including guide canes, electronic Braille devices, and liquid level indicators (see Figs. 5 and C.4). In these cases, the generations either defaulted to a more common object (*e.g.* a walking stick for "guide cane") or fabricated an object entirely (*e.g.* random dot patterns for "Braille sense display", colorful thermometers for "liquid level sensor"). It also failed to generate specific instances of assistive devices (*e.g.* "Victor Reader Stream", a talking book device, resulted in images of books or river streams). In contrast, DALL-E2 generates highly realistic of non-disability objects (see Fig. C.4a).

6. Discussion

Our evaluation of CLIP reveals that it consistently underperforms on BLV data across visual content, visual quality, and textual content, irrespective of architecture size, pre-training dataset, or pre-training dataset size. We discuss mitigation strategies to make LMMs more equitable for BLV users and marginalized groups more generally.

Our results suggest that the performance disparities come in part from the distribution shift between web-crawled and BLV user data. This highlights the importance of systematic reporting of the contents of large-scale datasets used for pre-training, in the spirit of datasheets for datasets [24]. Our analysis in Secs. 4.1.2 and 4.3.2 provides a starting point, but this should be extended to other datasets and marginalized content. With the data composition known, mitigation strategies can then be developed. For example, assistive device websites and disability dataset platforms like IncluSet [4] could explicitly be crawled.

We also show that a few-shot approach can mitigate performance disparities relating to image content – a more cost-effective alternative than re-training a LMM. The few-shot model adaptation could be done when the LMM is developed, when the application is developed, or by the end-users themselves as part of a teachable paradigm [31, 38]. Each of these options is an open research question with the need to more deeply explore interaction paradigms and light-weight model adaptation techniques [11, 27].

Finally, application-level mitigations should also be considered. For BLV users, auxiliary models could support users to reduce image variance, helping them stabilize the camera or alerting about the lighting conditions, for example. We could also leverage data augmentation techniques that are personalized to individual users or user groups. For BLV users who tend to take blurry images, for example, we could automatically inject blur into the few-shot images so that the model becomes more robust to this quality issue.

The findings in this paper prompt a critical look at the development cycle of current LMMs. Greater transparency and disaggregation in dataset reporting is needed, regardless of the proprietary nature of a dataset. Future work should also explore lightweight model adaption techniques that allow application developers and users to bring equity to their experiences. We must continue to work with marginalized communities – "nothing about us without us" – to equalize the benefit of LMMs and their extraordinary capabilities.

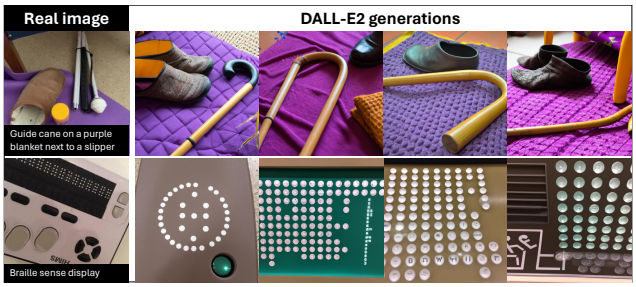

Figure 5. **DALL-E2 [53] either defaults to common objects or fabrications when prompted with disability objects like guide canes and electronic Braille devices**. Instead, it generates high-quality images of non-disability objects (see Fig. C.4a).

References

[1] Be My Eyes. https://www.bemyeyes.com, . Accessed: 2023-10-11. 1

[2] Be My Eyes uses GPT-4 to transform visual accessibility. https://openai.com/customer-stories/be-my-eyes, . Accessed: 2023-10-11. 1

[3] Google Lookout. https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal. Accessed: 2023-11-06. 1, 7

[4] IncluSet. https://incluset.com/. Accessed: 2023-11-09. 8

[5] Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai. Accessed: 2023-10-11. 1, 7

[6] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 1, 2

[7] 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. In *Advances in Neural Information Processing Systems*, 2022. 2

[8] Reza Akbarian Bafghi and Danna Gurari. A New Dataset Based on Images Taken by Blind People for Testing the Robustness of Image Classification Models Trained for ImageNet Categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3

[9] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021. 2

[10] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 2

[11] Samyadeep Basu, Daniela Massiceti, Shell Xu Hu, and Soheil Feizi. Strong Baselines for Parameter Efficient Few-Shot Fine-tuning. *arXiv preprint arXiv:2304.01917*, 2023. 8

[12] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. VizWiz: Nearly real-time answers to visual questions. In *Annual ACM Symposium on User Interface Software and Technology*, 2010. 3

[13] Sarah Bird, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, Microsoft, Tech. Rep. MSR-TR-2020-32, 2020. 2

[14] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 2

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. PMLR, 2020. 2

[16] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2

[17] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 5

[18] Sam Corbett-Davies, J Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal Machine Learning Research*, 2023. 2

[19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXic:2305.06500*, 2023. 1, 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representation*, 2020. 3

[21] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. CLIP2Video: Mastering video-text retrieval via image CLIP. *arXiv preprint arXiv:2106.11097*, 2021. 2

[22] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017. 2

[23] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1, 2, 3, 4, 13, 14, 16

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 8

[25] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz Grand Challenge: Answering visual questions from blind people. In *IEEE/CVR conference on computer vision and pattern recognition*, 2018. 3

[26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1, 3, 7, 20

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRa: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[28] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. In *Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2020. 2

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 12, 13

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 2021. 2, 3, 13

[31] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *CHI Conference on Human Factors in Computing Systems*, 2017. 4, 5, 8

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7), 2020. 3, 12

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014. 3, 4, 12

[34] Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The clear benchmark: Continual learning on real-world imagery. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1

[35] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, 2022. 1, 2

[36] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable Bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 1, 2

[37] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 7, 25

[38] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *IEEE/CVF International Conference on Computer Vision*, 2021. 3, 4, 8

[39] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 5

[40] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. 4, 14

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, 2022. 1, 2, 7, 8

[42] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *arXiv preprint arXiv:2306.09683*, 2023. 1, 7

[43] Ron Mokady, Amir Hertz, and Amit H Bermano. CLIP-Cap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[44] Cecily Morrison, Rita Marques, Martin Grayson, Daniela Massiceti, Camilla Longden, Linda Yilin Wen, and Edward Cutrell. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision. In *International ACM SIGACCESS Conference on Computers and Accessibility*, 2023. 1, 3, 4, 6

[45] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023. 1, 2

[46] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[47] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards Accountable AI: Hybrid human-machine analyses for characterizing system failure. In *AAAI Conference on Human Computation and Crowdsourcing*, 2018. 2

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[49] OpenAI. GPT-4V(ision) System Card, 2023. 1, 2

[50] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1

[51] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *International Conference on Computer Vision*, 2023. 5

[52] 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*. PMLR, 2021. 1, 2, 3, 13

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 7, 8

[54] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 4

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image syn-

thesis with latent diffusion models. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 1

[56] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. 4

[57] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2, 3, 4, 13, 14, 16

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2, 3, 4, 13, 14, 16

[59] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 2017. 5, 15, 19

[60] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022. 5

[61] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022. 1, 2

[62] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. CLIP4caption: CLIP for video caption. In *ACM International Conference on Multimedia*, 2021.

[63] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 3

[65] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey. *arXiv preprint arXiv:2111.03015*, 2021. 2

[66] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022. 1, 2

[67] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 2

[68] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2