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

Supplementary Material

A. Extended experimental details

A.1. Datasets

Open Images. The Open Images V7 dataset [32] contains 61.4M images with image-level labels spanning 20.6K object classes. The images are web-crawled from Flickr and the classes include items of clothing, food types, animals, vehicles and more. We motivate the choice of this dataset because of its scale and diversity, and because it is widely used for training and benchmarking models within the computer vision community. We only sample images from the validation and test splits that have been verified by humans to contain the labeled object (i.e. all false positives are removed). This is 390,797 validation and 1,319,751 test images, respectively.

MS-COCO. The Microsoft COCO dataset [33] contains 328K images with instance labels spanning 80 object classes. The images are also web-crawled from Flickr and include "common" objects like people, animals, vehicles, furniture, food and more. We motivate this choice of dataset because, like Open Images, it is widely used for training and benchmarking models. We only sample images from the val2017 split (5K images) as the test split does not have ground-truth labels that are publicly available.

A.2. CLIP variants

We include all CLIP variants we study in Tab. A.1, including their pre-training dataset (and size) and model checkpoint. All checkpoints are taken from `open_clip` [29].

A.3. Disability and exclusive disability objects

Three annotators manually categorized the 486 ORBIT objects into 55 disability objects, 42 exclusive disability objects (a subset of disability objects) and 431 non-disability objects. Of these, 39, 30 and 310 were unique disability, exclusive disability and non-disability objects, respectively. We include the object lists for each category below:

Unique disability objects [39 objects]: folded cane, solo audiobook player, orbit braille reader and notetaker, victor stream book reader, cane, white cane, digital recorder, magnifier, long cane, pen friend, braille note, dog poo, symbol cane, pocket magnifying glass, glasses, folded long guide cane, insulin pen, dictaphone, white mobility came, dog lead, retractable dog lead, braille orbit reader, victor reader stream, dogs lead, my hearing aid, water level sensor, braillepen slim braille keyboard, guide dog play cola, black mobility cane, my braille displat, visibility stick, leash, inhaler, liquid level indicator, hearing aid, guide dog harness, orbit reader 20 braille display, folded white cane, my cane

Unique exclusive disability objects [30 objects]: folded cane, solo audiobook player, orbit braille reader and notetaker, victor stream book reader, cane, white cane, digital recorder, magnifier, long cane, pen friend, braille note, dog poo, symbol cane, pocket magnifying glass, dictaphone, folded long guide cane, white mobility came, braille orbit reader, victor reader stream, my hearing aid, water level sensor, braillepen slim braille keyboard, black mobility cane, my braille displat, visibility stick, liquid level indicator, hearing aid, orbit reader 20 braille display, folded white cane, my cane

Unique non-disability objects [310 objects]: cushion, tred mill, apple airpods, headphones, ipod stand, wallet for bus pass cards and money, handheld police scanner, shelf unit with things, av tambourine, tea, toothbrush, door, door keys, lotion bottle, pint glass, favourite earings, prosecco, apple mobile phone, hat, tumble dryer, wall plug, risk watch, green water bottle, apple earpods, hole punch, phone stand, aspirin, tablets, garden shed, desk, knitting basket, dark glasses, headphone case, bin, chap stick, blue headphones, ottawa bus stop, fire stick remote, perfume, hair clip, pink himalayan salt, my purse, yellow marker, ipod in wallet, deodorant, mobile phone, iphone stand, apple phone charger, pencil case, one cup kettle, phone charger, adaptive dryer, skip prep, sunglasses case, eyewear case, apple headphones, front door, cranberry cream tea, backpack, keychain, 13 measuring cup, microwave, apple wireless keyboard, my tilly hat, dog toy, speaker, water bottle, my airpods, garden table, ruler, journal, stairgate, sleep mask, coffee mug, radar key, lighter, trainer shoe, toaster, vape pen, banana, house keys, winter gloves, cannabis vape battery, my tilly hat upside down, cap, small space screwdriver, dab radio, watering can, wheely bin, litter and dog waste bin, my headphones, my muse s headband, airpods, set of keys, wireless earphones, iphone in case, pink marker, scissors, blue tooth keyboard, remote control, my wraparound sunglasses, finger nail clipper, vagabond ale bottle, face mask, screwdriver, sock, front door to house, my mug, single airpod, back patio gate, earphones, 14 measuring cup, sky q remote, tv unit, lip balm, reptile green marker, coin purse, post box, watch, t-shirts, bus stop sign, buckleys, ladies purse, iphone air pods, recycling bin, black bin, key, black small wallet, table fan, exercise bench, keyboard, hand gel, purse, vase with flowers, white came, house door, wallet, reading glasses, orange skullcap, baked bean tin, migenta marker, my purple mask, condom box, mediterranean sea

Table A.1. All CLIP variants with their pre-training dataset, pre-training dataset size, and checkpoint (taken from open_clip [29]).

CLIP variant	Pre-training dataset	Dataset size	Checkpoint
ViT-B/16	WIT [52]	400M	openai
ViT-B/16	LAION-80M [30]	80M	Data-80M_Samples-34B_lr-1e-3_bs-88k
ViT-B/16	LAION-400M [57]	400M	laion400m_e32
ViT-B/16	LAION-2B [58]	2B	laion2b_s34b_b88k
ViT-B/16	DataComp-L [23]	140M	datacomp_l_s1b_b8k
ViT-B/16	CommonPool-L [23]	1.28B	commonpool_l_s1b_b8k
ViT-B/16	CommonPool-L (CLIP-Score filt.) [23]	384M	commonpool_l_clip_s1b_b8k
ViT-B/32	WIT [52]	400M	openai
ViT-B/32	LAION-80M [30]	80M	Data-80M_Samples-34B_lr-1e-3_bs-88k
ViT-B/32	LAION-400M [57]	400M	laion400m_e32
ViT-B/32	LAION-2B [58]	2B	laion2b_s34b_b79k
ViT-B/32	DataComp-S [23]	1.4M	datacomp_s_s13m_b4k
ViT-B/32	DataComp-M [23]	14M	datacomp_m_s128m_b4k
ViT-B/32	CommonPool-S [23]	12.8M	commonpool_s_s13m_b4k
ViT-B/32	CommonPool-S (CLIP-Score filt.) [23]	3.8M	commonpool_s_clip_s13m_b4k
ViT-B/32	CommonPool-M [23]	128M	commonpool_m_s128m_b4k
ViT-B/32	CommonPool-M (CLIP-Score filt.) [23]	38M	commonpool_m_clip_s128m_b4k
ViT-L/14	WIT [52]	400M	openai
ViT-L/14	LAION-80M [30]	80M	Data-80M_Samples-34B_lr-1e-3_bs-88k
ViT-L/14	LAION-400M [57]	400M	laion400m_e32
ViT-L/14	LAION-2B [58]	2B	laion2b_s32b_b82k
ViT-L/14	DataComp-XL/1B [23]	1.4B	datacomp_xl_s13b_b90k
ViT-L/14	CommonPool-XL (CLIP-Score filt.) [23]	3.8B	commonpool_xl_clip_s13b_b90k
ViT-H/14	LAION-2B [58]	2B	laion2b_s32b_b79k
ViT-g/14	LAION-2B [58]	2B	laion2b_s34b_b88k

salt, my work backpack, personal mug, bottle opener, my slate, clickr, measuring spoon, rice, mug, iphone 6, presentation remote, secateurs, ps4 controller, remote tv, necklace, wardrobe, aspirin vs tylenol, mouse, small screwdriver, socks, eye drops, mustard, hand saw, lipstick, bose wireless headphones, hair brush, hairbrush, pinesol cleaner, memory stick, glasses case, knitting needle, pepper shaker, cup again, bone conducting headset, fridge, usb stick, compact disc, work phone, wine glass, my front door, work bag, headband, airpod pro, walletv, my laptop, money pouch, remote, jd whisky bottle, paperclips, pex plumbers pliers, samsung tv remote control, my airpod pros case, portable keyboard, money clip, flat screen television, clear nail varnish, usb c dongle, amazon remote control, digital dab radio, 1 cup, measuring cup, tissue box, baseball cap, earpods, gloves, p939411 white cane, smarttv, skipping rope, back door, i d wallet, bluetooth keyboard, sunglasses, headset, my pill dosette, fridge freezer indicator, usb, apple pencil, black strappy vest, my apple watch, cell phone, apple wath, airpods pro charging case, slippers, dog streetball, corkscrew, airpod case, veg peeler, local post box, brown leather bracelet, pill bottle, my wallet, medication, mayonnaise jar, sofa, bottle, virgin remote control, money, slipper, fish food, styrofoam cup, blue facemask, i phone 11 pro, my keyboard, ipad, nobile phone stand, glasses cleaning wipe, bottle of alcoholic drink, cooker, tv remote, front door keys, tweezers, shed door, kettle, alcohol wipe, make up, battery drill, spanner, apple tv remote, bag, phone case, mini bluetooth keyboard, stylus, shoulder bag, comb, my keys, mirror, my clock, eye glasses, nike trainers, my water bottle, garden wall, sharp knife, my shoes, back pack, grinder, 12 measuring cup, iphone, phone, covid mask, mountain dew can, wheelie bin, car, headphone, keys, large sewing needle, miter saw, apple watch, chicken instant noodles, tv remote control, adaptive tennis ball, embroidery thread cone, washing basket, wrist watch, lime green marker, glass, boot, bed, bose earpods, television remote control, dining table setup, toddler cup, tape measure, adaptive washing machine, pop bottle, electric sanding disc, washing machine, my sennheiser pxc 350-2, ladies silver bracelet

A.4. Colors and materials

Three annotators manually annotated the ORBIT validation and test objects (208 objects) with their color and material. In most cases, each object was labeled with one color and one material, but in some cases up to two labels were selected (*e.g.* a water bottle with a plastic body and metal lid was assigned "plastic metal" as its material). The labels were iterated until all three annotators agreed. All colors and materials were selected from the following lists:

Colors [20 colors]: red, silver, yellow, grey, dark, pink, multicolour, purple, white, beige, burgundy, maroon, blue, green, black, gold, brown, light, transparent, orange

Materials [23 materials]: rubber, crystal, cardboard, denim, material, styrofoam, stone, glass, foam, cloth, leather, ceramic, plastic, wood, paper, embroidered, wooden, suede, canvas, patterned, metal, cotton, lacquered

A.5. Textual analysis of LAION-400M, LAION-2B and DataComp-1B

Our aim is to quantify the prevalence of disability content in large-scale datasets used to pre-train LMMs – specifically, LAION-400M [57], LAION-2B [58] and DataComp-1B (or XL) [23]. To do this, we first extract all visual concepts from the captions of each dataset (see Algorithm 1). We define a visual concept as a noun phrase that contains a physical object (*e.g*. "park bench"). We consider a noun phrase as a phrase that contain a common noun and optional adjectives (*e.g*. "green park bench"). We consider the common noun to be a physical object if it traverses the "entity", "physical_entity", and "object" hypernyms and then either the "artifact", "whole", "part", or "living_thing" hypernym in the WordNet tree hierarchy [40] (see Algorithm 2). In Tab. A.2 we report the top ten noun phrases extracted from LAION-400M, LAION-1B and DataComp-1B. We see that many of these are shared across all three datasets, including "image", "photo", "man", and "woman".

A.5.1 Prevalence of disability vs non-disability objects

In Sec. 4.1.2 of the main paper, we quantify how often disability and non-disability objects occur in the extracted visual concepts. We use the ORBIT objects as a seed set (see lists in App. A.3). Three annotators first grouped the object labels into clusters based on object similarity (*e.g*. all guide canes, all spectacles). Two synonyms were then assigned per cluster to account for different ways objects can be described. Early results showed that the disability clusters' synonyms occurred extremely rarely in the visual concepts, so we broadened this to 5-16 synonyms per cluster. The 222 and 312 synonyms for the disability and non-disability clusters are provided in Tab. A.3 and Tab. A.4, respectively.

We then count how many times each of these synonyms appears in the extracted visual concepts (see counts in Tabs. A.3 and A.4). We do this using direct string matching allowing for partial matches (*e.g*. "braille note taker" is marked as present in the visual concept "cheap braille note taker"). Before matching, we lower-case and remove punctuation from all synonyms and visual concepts following typical VQA practices (see `processPunctuation` function in the GT-Vision-Lab/VQA repo). For synonyms that contain multiple words, we also allow for multiple spellings (*e.g*. tread mill and treadmill).

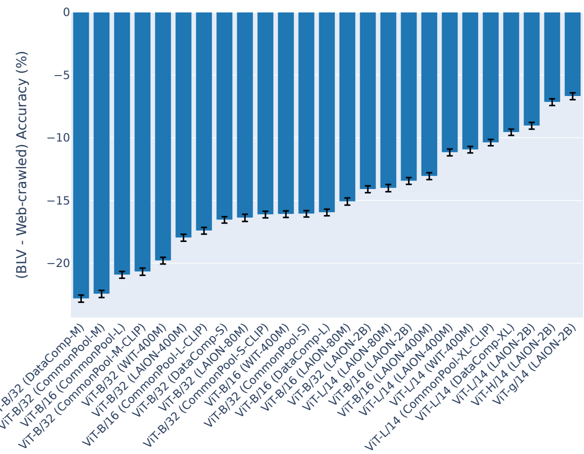


Figure B.1. **No CLIP variant achieves parity between BLV and web-crawled datasets on a standarized zero-shot image classification task (see Sec. 3.1).** Each bar represents the variant's delta in average accuracy (with 95% c.i.) between all images sampled from BLV versus web-crawled datasets.

A.5.2 Prevalence of colors vs materials

In Sec. 4.3.2 of the main paper, we quantify how often colors and materials occur in each dataset. We use the color and materials in App. A.4 and directly count their frequency (via partial string matching, as above) in each dataset's extracted visual concepts (see counts in Tab. A.5).

B. Extended results

B.1. BLV versus web-crawled data

We extend Fig. 1 in the main paper, with the delta in average accuracy between BLV and web-crawled datasets, reported per CLIP variant in Fig. B.1. We see that no model achieves parity, with smaller architectures (*e.g*. ViT-B/32) pre-trained on smaller datasets (*e.g*. DataComp-M, CommonPool-M) generally having a larger delta than larger architectures (*e.g*. ViT-g/14, ViT-H/14) pre-trained on larger datasets (*e.g*. LAION-2B).

B.2. Robustness to image content from BLV users

B.2.1 Disability objects are less well recognized than non-disability objects

Fig. B.2 shows the difference in average accuracy for disability, exclusive disability and non-disability objects for each CLIP variant, ordered by pre-training dataset size, on the ORBIT Clean (Fig. B.2a) and ORBIT Clutter (Fig. B.2b) datasets. Fig. B.3a and Fig. B.3b show the same, but with the CLIP variants ordered by architecture size. From these figures, we see that the accuracy difference between disability/exclusive disability objects and

Algorithm 1 Pseudocode for `extract_noun_phrases(captions: List[str]) -> List[str]`

```
regex_pattern = r"""NP: <DT|PRP$\$>?<JJ>*<NN|NNS>""" # a noun phrase (NP) contains a
  singular or plural noun (NN/NNS) which may be prefixed by an article (DT)/possessive
  pronoun (PRP) and/or adjectives (JJ)
chunker = nltk.RegexpParser(regex_pattern)
noun_phrases = []
for caption in captions do
    tokens = nltk.word_tokenize(caption.lower()) # tokenize
    pos_tags = nltk.pos_tag(tokens) # extract parts of speech
    np_tree = chunker(pos_tags) # extract noun phrase tree
    noun = extract_noun(np_tree) # extract noun (NN or NNS) from tree
    if is_physical_object(noun) then
        cleaned_np = clean_np(np_tree) # remove DT/PRP and singularize noun
        noun_phrases.extend(cleaned_np)
    end if
end for
return noun_phrases
```

Algorithm 2 Pseudocode for `is_physical_object(word: str) -> bool:`

```
synsets = wordnet.synsets(word, "n") # get WordNet noun synsets
for synset in synsets do
    paths = synset.hypernym_paths() # get hypernym paths
    for path in paths do
        # path is a list e.g. [Synset("entity.n.01"), ..., Synset("bench.n.01")]
        if (path contains "entity.n" AND "physical_entity.n" AND "object.n" \
          AND ("artifact.n" OR "whole.n" OR "part.n" OR "living_thing.n")):
          return True
    end for
end for
return False
```

non-disability objects remains largely constant regardless of test dataset, pre-training dataset size, and architecture size.

B.2.2 A few-shot approach can *sometimes* reduce the disability and non-disability accuracy gap

In Sec. 4.1.3 of the main paper, we show how a few-shot approach can be effective at reducing the accuracy difference between disability and non-disability objects in some scenarios. We use ProtoNets [59] as the few-shot approach, which computes an average embedding (or prototype) for each object class by simply averaging the embeddings of K training images for each class. A test image is then classified as the class whose prototype is most similar to the image's embedding, where similarity is measured by Euclidean distance. We extend Tab. 3 in the main paper with Figs. B.4a and B.4b here. Fig. B.4a shows ProtoNets can reduce the accuracy difference between disability/exclusive disability and non-disability objects on the ORBIT Clean dataset. Fig. B.4b shows ProtoNets' results on the ORBIT Clutter dataset, however, the few-shot adaptation is less effective, even with 40 shots per object.

In Figs. B.5a and B.5b, we examine how CLIP's pre-training dataset size influences the few-shot adaptation on ORBIT Clean and Clutter, respectively. We split the CLIP variants into three groups: those pre-trained on 0-100M examples, 100-1000M examples, and 1B+ examples. For each group, we average the delta in accuracy between disability and non-disability objects for all CLIP variants in that group, for each shot setting. For ORBIT Clean, we see that as the pre-training dataset and the number of shots increase, the delta generally decreases – with 1B+ pre-training examples and a 40-shot setting achieving the lowest delta (-0.08) between disability and non-disability object accuracy. For ORBIT Clutter, however, this trend is less pronounced. Increasing the number of pre-training examples does reduce the delta generally, but the best setting (1B+ pre-training examples, 40 shots) still sees a delta of -10.98 percentage points. Furthermore, for under 100M pre-training examples, the delta remains largely constant (around -18 percentage points) suggesting that a few-shot approach is less effective if the model has not seen enough pre-training data.

Table A.2. Top 10 noun phrases extracted from the captions of the LAION-400M [57], LAION-2B [58] and DataComp-1B [23] datasets. See extraction protocol in App. A.5.

LAION-400M		LAION-2B		DataComp-1B	
Noun phrase	**Occurrence count**	**Noun phrase**	**Occurrence count**	**Noun phrase**	**Occurrence count**
image	8,930,057	image	69,739,546	image	35,784,938
photo	7,559,650	photo	55,970,047	photo	28,612,087
vector	4,523,146	vector	29,203,691	vector	14,157,166
man	3,458,074	stock	22,209,987	stock	13,062,936
design	3,052,442	man	21,261,979	background	10,829,331
background	2,760,736	background	20,873,562	design	10,473,440
woman	2,513,375	picture	19,322,717	home	8,731,799
home	2,446,557	design	18,335,833	picture	8,523,946
stock	2,236,958	home	17,747,460	man	8,453,533
picture	2,235,123	woman	17,001,793	view	7,595,762

Table A.3. Disability object clusters, their synonyms, and their prevalence in the LAION-400M (L400M), LAION-2B (L2B) and DataComp-1B (DC1B) datasets. Numbers reported are the total number of times each cluster's synonyms appeared in the dataset's extracted visual concepts (total visual concepts – LAION-400M: 384,468,921; LAION-2B: 2,737,763,447; and DataComp-1B: 1,342,369,058).

Object cluster	Synonyms	L400M	L2B	DC1B
braille readers	braille note taker, braille reader, braille display, braille notetaker, braille tablet, braille computer, braille keyboard, orbit reader, braillepen slim braille keyboard, braillepen slim keyboard	4	8	6
dictaphones	dictaphone, digital recorder, voice recorder, dictation machine, audio recorder, voice recording device, dictation recorder, audio dictation device, voice transcription device, handheld recorder	40	185	168
digital book readers	digital book player, digital book reader, victor stream, victor reader stream, talking book, humanware reader, solo audiobook player, audiobook player	0	2	1
dog leads	dog lead, dogs lead, dog leash, dogs leash, leash, dog tether, dogs tether	434	1,663	1,402
dog poo	dog poo, dog poop, dog waste, dog scat, dog dung, canine faeces, canine feces, canine faeces	3	11	9
glasses	glass, sight glass, spectacle, eyeglass, reading glass, prescription glass, optical glass, corrective lens, bi focal, eyewear, frame, multi focal, optical, vision aid, spec	17,620	68,169	46,259
guide canes	guide cane, symbol cane, mobility cane, long cane, white cane, blind cane, white mobility cane, vision cane, assistive cane, visibility stick	21	79	50
hearing aids	hearing aid, hearing device, hearing amplifier, assistive listening device, hearing implant, cochlear implant, audio prosthesis, auditory prosthesis	23	122	77
inhalers	inhaler, asthma pump, asthma puffer, aerosol inhaler, inhalant delivery system	81	282	315
insulin pens	insulin pen, insulin injector, insulin delivery pen, insulin auto-injector, insulin syringe pen, insulin dispenser, insulin delivery system, insulin applicator, insulin dosing pen, diabetes pen	2	4	4
liquid level sensors	liquid level sensor, liquid level indicator, liquid level detector, liquid level gauge, water level sensor, water level indicator, water level detector, water level gauge	1	5	8
magnifiers	magnifier, magnifying glass, magnification aid, magnifying lens	89	390	362
audio labelers	penfriend, pen friend, audio labeller, audio labelling device, audio labelling pen, audio labelling tool, voice labeller, voice labelling pen, voice labelling device, voice labelling tool, speech-enabled labeller, speech-enabled labelling device, speech-enabled labelling pen, speech-enabled labelling tool, talking label maker, speech-based label printer	8	19	11

B.3. Robustness to image quality from BLV users

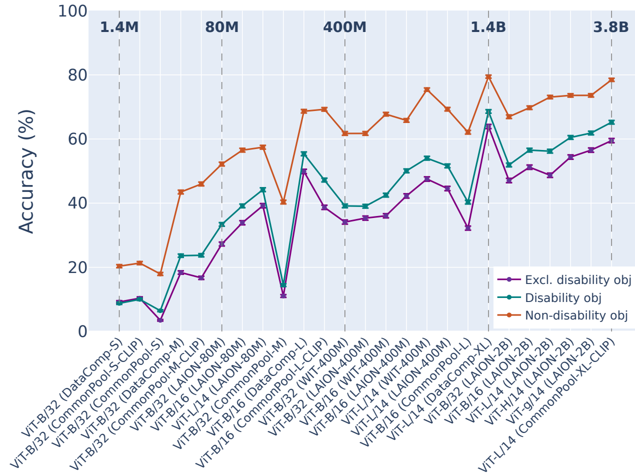
We include the raw marginal effects of each quality issue on model accuracy for all CLIP variants in Tabs. B.3 to B.5. These correspond to Fig. 3 in the main paper, with experimental details provided in Sec. 4.2. We note that the same image may be sampled multiple times as a result of the episodic sampling procedure (see Sec. 3.1). No two tasks share the same set of N objects, however, so for a given image, the model is always presented a different classification problem. The logistic regression is sensitive to input-output similarities, however, so we filter out all duplicate images to avoid biasing our sample. This resulted in 93,698 images for ORBIT Clean and 6,764 for VizWiz-Classification. We report the prevalence of each quality issue in these datasets in Tabs. B.1a and B.1b.

Table A.4. Non-disability object clusters, their synonyms, and their prevalence in the LAION-400M (L400M), LAION-2B (L2B) and DataComp-1B (DC1B) datasets. Numbers reported are the total number of times each cluster's synonyms appeared in the dataset's extracted visual concepts (total visual concepts – LAION-400M: 384,468,921; LAION-2B: 2,737,763,447; and DataComp-1B: 1,342,369,058).

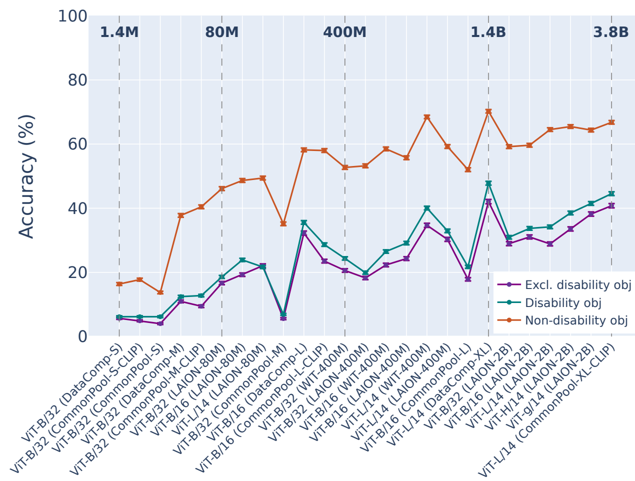
Object cluster	**Synonyms**	**L400M**	**L2B**	**DC1B**	**Object cluster**	**Synonyms**	**L400M**	**L2B**	**DC1B**
airpods	airpod, ear phone	655	2,077	1,763	make-up	make-up, make up	6,202	23,424	15,361
airpods cases	airpods case, airpods pro case	0	1	2	markers	marker, felt-tip pen	1,254	5,525	3,634
alcohol wipes	alcohol wipe, alcohol pad	1	1	0	measuring spoons	measuring spoon, measuring cup	10	50	55
bags	bag, backpack	15,250	52,351	34,650	medications	medication, pill	689	2,477	2,415
balls	ball, dog toy	6,540	23,289	14,888	mirrors	mirror, looking glass	4,755	18,788	13,327
bananas	banana, fruit	7,631	25,879	21,884	mice	bluetooth mouse, wireless mouse	3	13	16
baskets	basket, crate	3,375	12,380	8,821	mugs	mug, cup	10,360	39,145	30,334
beds	bed, mattress	8,087	35,714	23,575	nail clippers	nail clipper, tweezers	10	48	76
beers	beer, alcohol	458	2,145	1,564	nail polishes	nail polish, nail varnish	916	2,918	1,787
bins	bin, trash can	1,265	4,526	3,370	needles	needle, pin	7,531	27,235	23,560
bottles	bottle, thermos	6,113	21,684	20,062	paper clips	paper clip, paper fastener	106	386	366
bottle openers	bottle opener, cork screw	127	515	683	peelers	peeler, scraper	294	1,085	1,259
bracelets	bracelet, necklace	18,438	63,242	61,734	pencil cases	pencil case, pen case	33	161	152
brushes	brush, comb	3,866	13,624	11,820	phones	phone, iphone	5,217	19,911	11,396
bus stops	bus stop, bus station	36	163	103	phone chargers	phone charger, charging cable	2	13	17
cans	can, tin	2,648	10,906	7,629	phone stands	phone stand, ipad stand	9	32	17
cars	car, vehicle	22,867	84,568	57,283	plugs	plug, socket	3,340	12,519	10,972
CDs	compact disc, cd	9,675	31,511	14,609	police scanners	police scanner, radio scanner	1	2	2
cleaners	cleaner, surface spray	956	3,767	2,850	pops	pop, soda	5,795	20,591	13,506
clocks	clock, timekeeper	2,783	9,560	7,310	post boxes	post box, mail box	608	2,010	1,625
condom boxes	condom box, durex box	0	1	0	purses	purse, wallet	6,221	21,201	15,256
cookers	cooker, air fryer	654	2,612	2,718	radios	radio, receiver	4,638	15,645	12,138
cushions	cushion, pillow	9,081	33,694	24,195	rice	rice, noodle	4,725	16,489	13,833
deodorants	deodorant, perfume	1,564	5,875	5,556	rulers	ruler, tape measure	772	2,963	2,118
dog waste bins	dog waste bin, dog waste container	0	0	0	sauces	mustard, mayonnaise	1,050	3,877	2,941
doors	door, entrance	15,627	54,395	44,418	scissors	secateur, scissor	85	284	268
drills	drill, power tool	1,055	4,115	3,082	screwdrivers	screwdriver, spanner	383	1,466	2,109
electric saws	electric saw, chain saw	368	1,340	1,007	sheds	shed, tool shed	786	3,112	2,270
eye drops	eye drop, eye gel	6	35	18	shoes	shoe, sneaker	11,078	43,179	21,112
face masks	face mask, face covering	43	255	181	skipping ropes	skipping rope, jump rope	7	19	17
fans	fan, air cooler	7,638	24,013	15,585	sleep masks	sleep mask, eye mask	44	158	144
fish foods	fish food, fish flake	4	15	12	socks	sock, sockwear	2,959	8,765	6,443
fridges	fridge, freezer	1,784	5,516	4,926	sofas	sofa, couch	11,588	47,305	31,437
game controllers	wireless controller, game controller	5	29	36	spices	salt, pepper	2,814	9,365	9,068
gates	gate, gateway	3,133	11,797	8,287	styluses	apple pen, stylus	151	688	681
glasses	glass, tumbler	5,022	19,732	13,595	sunglasses	sunglass, shade	6,964	24,927	16,626
glasses cases	glasses case, sunglasses case	2	5	4	tables	desk, table	23,387	104,138	69,152
glasses cleaners	glasses cleaner, lens wipe	0	0	0	tambourines	tambourine, tamborine	89	362	375
gloves	glove, mitten	3,866	13,548	8,228	tea	tea, teabag	7,739	25,126	20,249
grinders	grinder, food processor	421	1,645	1,525	thread cones	thread cone, thread spool	15	49	51
hair clips	hair clip, headband	1,546	5,115	4,381	tissue boxes	tissue box, kleenex	12	29	26
hand sanitizers	hand sanitizer, hand santiser	0	10	5	toothbrushes	toothbrush, dental brush	537	2,329	2,443
hand saws	hand saw, hack saw	46	206	245	tread mills	tread mill, running machine	263	953	627
hats	hat, cap	10,100	34,840	24,185	t-shirts	t-shirt, tee	30,770	100,369	71,408
headphones	headphone, headset	2,241	7,843	6,396	TVs	tv, television	15,350	53,606	35,370
headphone cases	headphone case, headphones case	0	1	1	TV remotes	tv remote, remote control	203	866	658
hole punches	hole punch, paper punch	47	174	142	USB sticks	usb stick, flash drive	221	651	690
iPads	ipad, tablet	6,365	21,157	14,534	vapes	vape, e-cigarette	160	540	469
journals	journal, notebook	6,201	24,084	15,205	vases	vase, jug	4,853	16,954	16,971
kettles	kettle, toaster	1,128	4,879	4,348	walls	wall, fence	14,538	60,387	39,221
keys	key, key chain	8,770	31,370	22,409	wardrobes	wardrobe, cupboard	2,101	8,193	5,926
keyboards	keyboard, keypad	2,491	8,639	9,582	washing machines	washing machine, dryer	767	3,040	2,194
knives	knife, blade	4,012	15,521	17,972	watches	watch, smart watch	9,651	36,043	24,128
laptops	laptop, chromebook	7,522	24,913	17,948	watering cans	watering can, water can	14	110	66
lipsticks	lipstick, lip balm	1,453	5,213	4,635	weight benches	weight bench, gym bench	10	32	33

Table A.5. Colors/materials and their prevalence in the LAION-400M (L400M), LAION-2B (L2B) and DataComp-1B (DC1B) datasets. Numbers reported are the total number of times each color/material appeared in the dataset's extracted visual concepts (total visual concepts – LAION-400M: 384,468,921; LAION-2B: 2,737,763,447; and DataComp-1B: 1,342,369,058).

		L400M	**L2B**	**DC1B**
Colors	beige	2,462	8,857	5,265
	black	87,366	323,959	207,730
	blue	53,947	193,504	131,665
	brown	21,904	84,994	55,553
	burgundy	1,261	4,127	2,446
	dark	10,888	36,735	25,818
	gold	5,882	19,564	12,894
	green	38,876	136,448	92,922
	grey	12,816	47,470	28,847
	light	31,413	120,232	92,203
	maroon	613	2,367	1,495
	multicolour	278	652	404
	orange	10,138	30,733	23,262
	pink	14,925	60,226	35,658
	purple	10,985	37,673	25,270
	red	45,267	165,576	104,289
	silver	8,789	33,614	27,766
	transparent	2,513	10,098	6,807
	white	94,014	366,224	236,456
	yellow	20,723	73,049	49,121
Materials	canvas	897	3,254	1,920
	cardboard	451	1,546	1,154
	ceramic	12,808	48,649	43,339
	cloth	3,498	13,528	8,968
	cotton	12,619	47,188	28,071
	crystal	12,663	46,968	36,329
	denim	5,130	15,265	8,720
	embroidered	2,090	8,026	3,904
	foam	675	2,234	1,894
	glass	3,264	11,420	7,889
	lacquered	351	1,679	1,077
	leather	1,214	4,234	2,505
	material	9,287	44,848	25,692
	metal	3,183	11,780	9,198
	paper	14,615	58,584	40,500
	patterned	1,057	3,841	1,947
	plastic	4,136	14,748	10,997
	rubber	4,620	18,765	12,234
	stone	7,262	28,321	20,570
	styrofoam	83	294	257
	suede	2,787	10,929	5,197
	wood	10,947	43,897	30,791
	wooden	18,239	73,016	51,445



(a) ORBIT Clean



(b) ORBIT Clutter

Figure B.2. **CLIP's difference in accuracy between disability and non-disability objects remains largely constant as its pre-training dataset increases**. Zero-shot accuracy is averaged (with 95% c.i.) over images from ORBIT Clean/Clutter of each object type. Experimental details in Sec. 4.1.1.

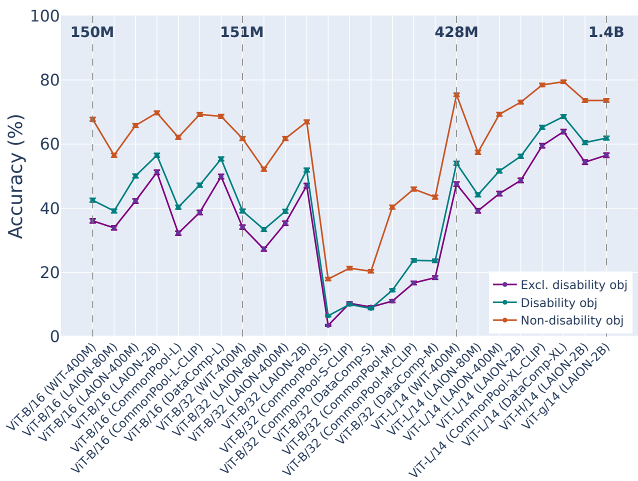
Table B.1. Prevalence of each quality issue in the (a) ORBIT Clean and (b) VizWiz-Classification datasets. Numbers reported as the raw counts of each issue and as a percentage of the total non-disability/disability images (ORBIT Clean) and total images (VizWiz-Classification).

	Total frames	Framing	Blur	Viewpoint	Occlusion	Lighting
Non-disability object	86,185	52,275 (60.7%)	28,235 (32.8%)	14,253 (16.5%)	11,302 (13.1%)	3,788 (4.4%)
Disability object	7,513	3,290 (43.8%)	2,237 (29.8%)	394 (5.2%)	976 (13.0%)	205 (2.7%)

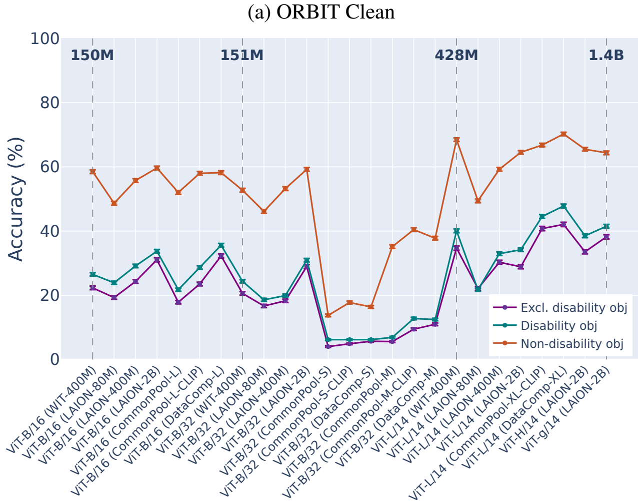
(a) ORBIT Clean

	Total frames	Framing	Blur	Viewpoint	Occlusion	Overexposed	Underexposed	Other
Non-disability object	6,764	3,715 (54.9%)	2,544 (37.6%)	1,118 (16.5%)	142 (2.1%)	327 (4.8%)	288 (4.3%)	16 (0.2%)

(b) VizWiz-Classification

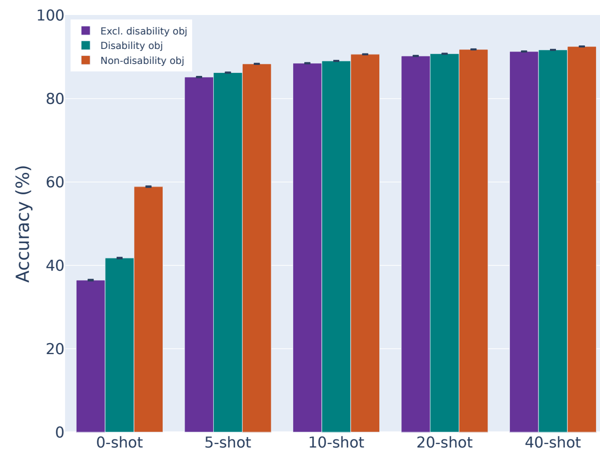


(a) ORBIT Clean

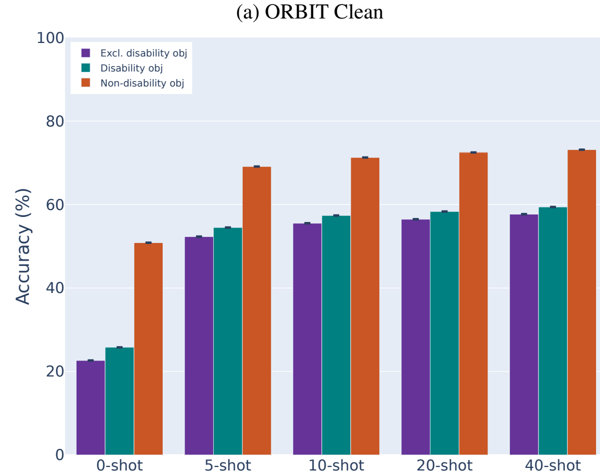


(b) ORBIT Clutter

Figure B.3. **CLIP's difference in accuracy between disability and non-disability objects remains largely constant as its architecture size increases**. Zero-shot accuracy is averaged (with 95% c.i.) over images from ORBIT Clean/Clutter of each object type. Experimental details in Sec. 4.1.1.

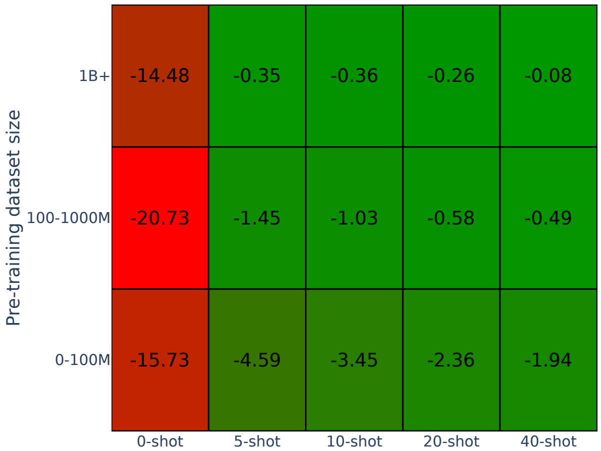


(a) ORBIT Clean

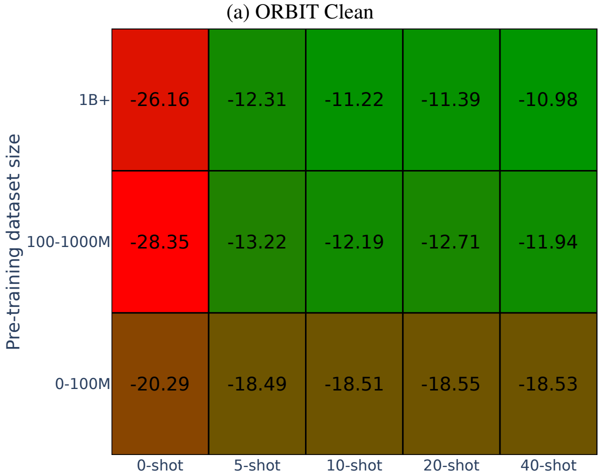


(b) ORBIT Clutter

Figure B.4. **A few-shot approach (ProtoNets [59]) can reduce the accuracy gap between disability and non-disability objects, but not for realistic, cluttered images.** Bars represent the average accuracy (with 95% c.i.) over all test frames for each shot setting ($K = [5, 10, 20, 40]$). $K = 0$ is equivalent to the zero-shot setting described in Sec. 4.1.1. Experimental details in Sec. 4.1.3.



(a) ORBIT Clean



(b) ORBIT Clutter

Figure B.5. **The larger the dataset used to pre-train CLIP, the more effective a few-shot approach is at closing the accuracy gap between disability and non-disability objects on ORBIT Clean, but this is less so for ORBIT Clutter especially for pre-training datasets $<$100M examples**. Each block reports the average delta in accuracy between disability and non-disability objects for the models that fall within that group. Models: 25 CLIP variants.

Table B.2. **Including an object's material in its prompt leads to text embeddings that are the least aligned with the 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 Clutter.

Prompt	**Obj. name**	**Material + obj. name**	**Color + obj. name**	**Color + material + obj. name**
CLIP Score	**22.81** $\pm$ **0.02**	21.92 $\pm$ 0.02	22.73 $\pm$ 0.02	21.86 $\pm$ 0.02

Table B.3. **Marginal effects of explanatory variables on CLIP's zero-shot classification accuracy (with ViT-B/16 vision encoders) on the ORBIT Clean and VizWiz-Classification datasets**. Main values are marginal effects, while values in brackets are p-values. */**/*** indicates within 90/95/99% confidence interval, respectively. Experimental details in Sec. 4.2. L-80M=LAION-80M, L-400M=LAION-400M, L-2B=LAION-2B, DC-L=DataComp-L, CP-L=CommonPool-L, CP-L-CLIP=CommonPool-L-CLIP.

Dataset	Explanatory variable	ViT-B/16 (WIT)	ViT-B/16 (L-80M)	ViT-B/16 (L-400M)	ViT-B/16 (L-2B)	ViT-B/16 (DC-L)	ViT-B/16 (CP-L)	ViT-B/16 (CP-L-CLIP)
ORBIT Clean	framing	0.044***	0.044***	0.043***	0.042***	0.02***	0.054***	0.033***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	blur	-0.095***	-0.139***	-0.117***	-0.139***	-0.153***	-0.135***	-0.146***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	viewpoint	-0.115***	-0.139***	-0.122***	-0.1***	-0.092***	-0.094***	-0.086***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	occlusion	-0.054***	-0.099***	-0.088***	-0.086***	-0.088***	-0.104***	-0.099***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	lighting	-0.213***	-0.212***	-0.262***	-0.245***	-0.226***	-0.297***	-0.246***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj	-0.328***	-0.341***	-0.278***	-0.245***	-0.258***	-0.352***	-0.319***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj:framing	0.143***	0.057***	0.108***	0.097***	0.141***	0.06***	0.097***
		(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj:blur	0.018	0.074***	-0.015	-0.017	-0.021	0.044***	0.011
		(0.764)	(1.0)	(0.673)	(0.747)	(0.843)	(0.991)	(0.551)
	excl. disability obj:viewpoint	0.113***	0.207***	0.173***	0.18***	0.069**	0.07*	0.065**
		(1.0)	(1.0)	(1.0)	(1.0)	(0.969)	(0.947)	(0.959)
	excl. disability obj:occlusion	-0.058***	0.007	-0.051**	-0.004	-0.006	0.032	0.03
		(0.994)	(0.22)	(0.984)	(0.169)	(0.239)	(0.837)	(0.872)
	excl. disability obj:lighting	-0.239***	-0.128**	-0.238***	-0.208***	-0.294***	-0.204**	-0.214***
		(1.0)	(0.969)	(1.0)	(1.0)	(1.0)	(0.986)	(0.999)
VizWiz-Classification	framing	0.001	0.001	-0.009	-0.035***	-0.01	-0.028**	-0.018
		(0.051)	(0.093)	(0.542)	(0.995)	(0.599)	(0.979)	(0.862)
	blur	-0.028**	-0.019	-0.009	0.0	-0.014	-0.015	-0.02
		(0.976)	(0.867)	(0.52)	(0.006)	(0.727)	(0.761)	(0.885)
	rotation	-0.079***	-0.116***	-0.055***	-0.086***	-0.07***	-0.092***	-0.09***
		(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)	(1.0)
	occlusion	-0.096**	-0.144***	-0.138***	-0.187***	-0.142***	-0.209***	-0.186***
		(0.978)	(0.999)	(0.999)	(1.0)	(0.999)	(1.0)	(1.0)
	overexposure	-0.034	0.023	-0.033	-0.011	-0.07**	-0.064**	-0.029
		(0.777)	(0.592)	(0.758)	(0.301)	(0.987)	(0.974)	(0.688)
	underexposure	-0.02	-0.088***	-0.084***	-0.065**	-0.051*	-0.073**	-0.074**
		(0.498)	(0.996)	(0.995)	(0.969)	(0.91)	(0.984)	(0.985)
	other	0.129	0.099	0.009	0.243*	0.017	0.115	0.04
		(0.668)	(0.579)	(0.059)	(0.911)	(0.112)	(0.63)	(0.251)

Table B.4. **Marginal effects of explanatory variables on CLIP's zero-shot classification accuracy (with ViT-B/32 vision encoders) on the ORBIT Clean and VizWiz-Classification datasets**. Main values are marginal effects, while values in brackets are p-values. */**/*** indicates within 90/95/99% confidence interval, respectively. Experimental details in Sec. 4.2. L-80M=LAION-80M, L-400M=LAION-400M, L-2B=LAION-2B, DC-S=DataComp-S, DC-M=DataComp-M, CP-S=CommonPool-S, CP-S-CLIP=CommonPool-S-CLIP, CP-M=CommonPool-M, CP-M-CLIP=CommonPool-M-CLIP.

Dataset	Explanatory variable	ViT-B/32 (WIT)	ViT-B/32 (L-80M)	ViT-B/32 (L-400M)	ViT-B/32 (L-2B)	ViT-B/32 (DC-S)	ViT-B/32 (DC-M)	ViT-B/32 (CP-S)	ViT-B/32 (CP-S-CLIP)	ViT-B/32 (CP-M)	ViT-B/32 (CP-M-CLIP)
ORBIT Clean	framing	0.062***	0.046***	0.054***	0.051***	0.035***	0.101***	0.046***	0.05***	0.11***	0.097***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	blur	-0.106***	-0.128***	-0.142***	-0.132***	-0.025***	-0.113***	-0.029***	-0.05***	-0.103***	-0.119***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	viewpoint	-0.096***	-0.111***	-0.111***	-0.099***	-0.036***	-0.058***	-0.011***	-0.032***	-0.045***	-0.092***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(0.999)	(1.0)	(1.0)	(1.0)
	occlusion	-0.075***	-0.095***	-0.094***	-0.088***	-0.088***	-0.142***	-0.072***	-0.08***	-0.109***	-0.123***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	lighting	-0.174***	-0.297***	-0.292***	-0.261***	-0.117***	-0.228***	-0.209***	-0.217***	-0.296***	-0.294***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj	-0.33***	-0.359***	-0.311***	-0.265***	-0.297***	-0.311***	-0.124***	-0.204***	-0.415***	-0.576***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj:framing	0.137***	0.118***	0.124***	0.113***	-0.016	-0.094***	-0.259***	-0.147***	-0.064***	0.049*
		(1.0)	(1.0)	(1.0)	(1.0)	(0.479)	(1.0)	(1.0)	(1.0)	(0.99)	(0.949)
	excl. disability obj:blur	0.072***	-0.012	0.011	0.013	0.028	0.045**	-0.01	0.043*	0.062**	0.062**
		(1.0)	(0.458)	(0.479)	(0.593)	(0.759)	(0.959)	(0.347)	(0.948)	(0.987)	(0.984)
	excl. disability obj:viewpoint	0.128***	0.166***	0.06*	0.129***	0.219***	0.143***	0.148**	0.225***	0.259***	0.268***
		(1.0)	(1.0)	(0.911)	(1.0)	(1.0)	(0.999)	(0.99)	(1.0)	(1.0)	(1.0)
	excl. disability obj:occlusion	-0.151***	0.109***	-0.022	-0.025	0.176***	0.036	-0.013	0.038	0.166***	0.31***
		(1.0)	(1.0)	(0.659)	(0.779)	(1.0)	(0.758)	(0.313)	(0.791)	(1.0)	(1.0)
	excl. disability obj:lighting	-0.188***	0.025	-0.255***	-0.285***	0.17***	-0.313**	0.188***	0.123*	-0.156	-0.273*
		(1.0)	(0.312)	(0.998)	(1.0)	(1.0)	(0.98)	(0.998)	(0.945)	(0.768)	(0.904)
VizWiz-Classification	framing	0.011	-0.024**	-0.002	-0.034***	-0.015*	-0.034***	0.014*	-0.005	-0.021*	-0.003
		(0.634)	(0.963)	(0.115)	(0.995)	(0.945)	(0.997)	(0.908)	(0.441)	(0.936)	(0.19)
	blur	-0.009	-0.003	-0.029**	0.007	-0.005	-0.026**	-0.001	-0.02**	-0.033***	-0.03**
		(0.517)	(0.202)	(0.98)	(0.433)	(0.455)	(0.97)	(0.106)	(0.97)	(0.995)	(0.984)
	rotation	-0.072***	-0.153***	-0.052***	-0.113***	-0.067***	-0.092***	-0.075***	-0.117***	-0.16***	-0.111***
		(1.0)	(1.0)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	occlusion	-0.144***	-0.121***	-0.17***	-0.132***	-0.057*	-0.175***	-0.098**	-0.128***	-0.182***	-0.216***
		(0.999)	(0.993)	(1.0)	(0.997)	(0.907)	(1.0)	(0.987)	(0.996)	(1.0)	(1.0)
	overexposure	0.035	-0.063**	-0.059**	-0.025	-0.004	-0.082***	-0.013	-0.027	-0.11***	-0.063**
		(0.775)	(0.974)	(0.961)	(0.618)	(0.17)	(0.995)	(0.489)	(0.778)	(1.0)	(0.972)
	underexposure	-0.009	-0.034	-0.095***	-0.046	-0.019	-0.081***	-0.008	-0.049*	-0.011	-0.094***
		(0.237)	(0.754)	(0.998)	(0.868)	(0.648)	(0.992)	(0.307)	(0.948)	(0.294)	(0.997)
	other	-0.13	0.176	0.06	0.063	0.002	0.084	-0.106	0.107	-0.101	0.016
		(0.705)	(0.877)	(0.368)	(0.388)	(0.023)	(0.543)	(0.635)	(0.876)	(0.577)	(0.103)

Table B.5. **Marginal effects of explanatory variables on CLIP's zero-shot classification accuracy (with ViT-L/14, ViT-H/14 and ViT-g/14 vision encoders) on the ORBIT Clean and VizWiz-Classification datasets**. Main values are marginal effects, while values in brackets are p-values. */**/*** indicates within 90/95/99% confidence interval, respectively. Experimental details in Sec. 4.2. L-80M=LAION-80M, L-400M=LAION-400M, L-2B=LAION-2B, DC-Xl=DataComp-XL, CP-XL-CLIP=CommonPool-XL-CLIP.

Dataset	Explanatory variable	ViT-L/14 (WIT)	ViT-L/14 (L-80M)	ViT-L/14 (L-400M)	ViT-L/14 (L-2B)	ViT-L/14 (DC-XL)	ViT-L/14 (CP-XL-CLIP)	ViT-H/14 (L-2B)	ViT-g/14 (L-2B)
ORBIT Clean	framing	-0.013***	0.027***	0.03***	0.019***	-0.016***	-0.018***	0.025***	0.025***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	blur	-0.098***	-0.139***	-0.131***	-0.129***	-0.099***	-0.114***	-0.113***	-0.123***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	viewpoint	-0.117***	-0.15***	-0.114***	-0.124***	-0.087***	-0.09***	-0.115***	-0.106***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	occlusion	-0.065***	-0.091***	-0.085***	-0.078***	-0.068***	-0.068***	-0.078***	-0.082***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	lighting	-0.174***	-0.288***	-0.248***	-0.223***	-0.189***	-0.184***	-0.173***	-0.222***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj	-0.26***	-0.235***	-0.257***	-0.258***	-0.223***	-0.245***	-0.267***	-0.283***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj:framing	0.085***	0.138***	0.085***	0.112***	0.16***	0.127***	0.103***	0.138***
		(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)
	excl. disability obj:blur	-0.012	-0.04**	-0.013	-0.032**	-0.083***	-0.051***	-0.039***	-0.027**
		(0.644)	(0.982)	(0.635)	(0.98)	(1.0)	(1.0)	(0.997)	(0.956)
	excl. disability obj:viewpoint	0.101***	0.113***	0.144***	0.108***	0.041	0.119***	0.125***	0.118***
		(1.0)	(0.999)	(1.0)	(1.0)	(0.884)	(1.0)	(1.0)	(1.0)
	excl. disability obj:occlusion	-0.052***	-0.023	0.014	-0.005	-0.004	-0.018	0.011	0.041**
		(0.998)	(0.696)	(0.56)	(0.237)	(0.218)	(0.765)	(0.492)	(0.982)
	excl. disability obj:lighting	-0.192***	-0.369***	-0.185***	-0.147***	-0.133***	-0.153***	-0.176***	-0.243***
		(1.0)	(1.0)	(0.999)	(0.998)	(1.0)	(1.0)	(1.0)	(1.0)
VizWiz-Classification	framing	0.007	-0.04***	-0.038***	-0.014	-0.002	0.013	-0.006	-0.008
		(0.434)	(0.999)	(0.998)	(0.748)	(0.138)	(0.737)	(0.402)	(0.503)
	blur	-0.005	0.015	-0.014	-0.002	-0.007	-0.009	-0.012	-0.019
		(0.317)	(0.773)	(0.737)	(0.131)	(0.404)	(0.552)	(0.676)	(0.87)
	rotation	-0.052***	-0.093***	-0.076***	-0.031*	-0.065***	-0.048***	-0.066***	-0.101***
		(0.999)	(1.0)	(1.0)	(0.948)	(1.0)	(0.998)	(1.0)	(1.0)
	occlusion	-0.118***	-0.074*	-0.143***	-0.126***	-0.134***	-0.126***	-0.136***	-0.106**
		(0.996)	(0.915)	(0.999)	(0.998)	(0.999)	(0.999)	(0.999)	(0.987)
	overexposure	-0.004	-0.017	-0.052*	0.001	0.015	-0.018	0.019	0.004
		(0.105)	(0.449)	(0.934)	(0.015)	(0.416)	(0.499)	(0.499)	(0.101)
	underexposure	-0.023	-0.034	-0.092***	-0.076***	-0.023	-0.056*	-0.035	-0.028
		(0.57)	(0.743)	(0.998)	(0.99)	(0.562)	(0.949)	(0.761)	(0.642)
	other	0.093	0.281**	0.08	0.194	0.235	0.051	0.066	0.022
		(0.525)	(0.972)	(0.473)	(0.83)	(0.876)	(0.314)	(0.395)	(0.137)

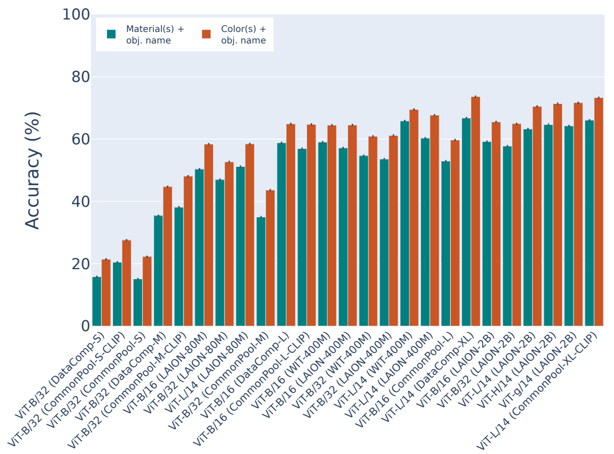


Figure B.6. **All CLIP variants classify objects more accurately when objects are described by their color rather than their material**. Each bar is the average accuracy (with 95% c.i.) over 200K images (100K ORBIT Clean, 100K ORBIT Clutter) for that CLIP variant when given either a material or color prompt. Variants ordered by pre-training dataset size.

B.4. Robustness to language used by BLV users

B.4.1 CLIP classifies objects more accurately when they are described by color rather than material

We extend Tab. 4 in the main paper, with Tab. B.2 for the ORBIT Clutter dataset here. We see that the CLIP scores for the lower bound prompt (*i.e.* just the object name) are the highest, followed by the color prompt. Similar to Tab. 4, we see that both the material prompt and the upper bound prompt which includes the object's material have the lowest CLIP scores, suggesting that including the object's material in the prompt harms embedding alignment.

We explore the impact this has on classifier accuracy by combining the textual prompts with the standard zero-shot set-up described in Sec. 3.1. Specifically, rather than embedding the raw ORBIT object labels for each task's N classes, we instead embed their textual prompts. In the first experiment, we embed all N objects as their color prompts, and in the second as their material prompts. For both experiments, we use $T = 50$, $N = 20$, $M = 100$. In Fig. B.6, we see that across all CLIP variants, objects are classified more accurately when they are described by their color rather than their material – by 7.1 percentage points more, on average. We see that this difference is largely constant regardless of both architecture and pre-training dataset size (see Fig. B.7).

C. Example-based analysis

C.1. Standardized image selection

We run our analysis on 180 images spanning 20 objects which are selected through a standardized process as a way to systematically assess failure cases. Specifically, we select the 5 top- and bottom-performing (disability and non-disability) objects from the ORBIT dataset using the standardized zero-shot classification set-up (see objects in Tab. C.1). We take performance to be the average accuracy per object, computed over all the CLIP variants we considered. For each object, we extract the noun phrase from its raw label and apply simple pre-processing to ensure that it is unambiguous and concise (see cleaned phrases in Tab. C.1). These cleaned noun phrases are used as the text prompts for all three downstream models we study. We then sample 9 images for each of the 20 objects – 6 from its clutter videos and 3 from its clean videos. We only sample images where the object is tagged as present. To increase image diversity, we ensure that images are sampled from all videos available for each object, and are sampled at even intervals. Specifically, for clean videos we sample the 3 frames at 25%, 50% and 75% positions, alternating the video we sample from each time (*e.g.* if an object has 2 clean videos, then we sample 1 frame at 25% of video 1, 1 frame at 50% of video 2, and 1 frame at 75% of video 1). For clutter videos, we sample 6 frames at 25%, 35%, 45%, 55%, 65% and 75%, also alternating the video for each sample. We limit frame sampling to between 25% and 75% of each video as ORBIT data collectors were instructed to start each video with the camera close to the object and then move it further away, so we wanted to exclude frames where the camera might be too close/far from the object.

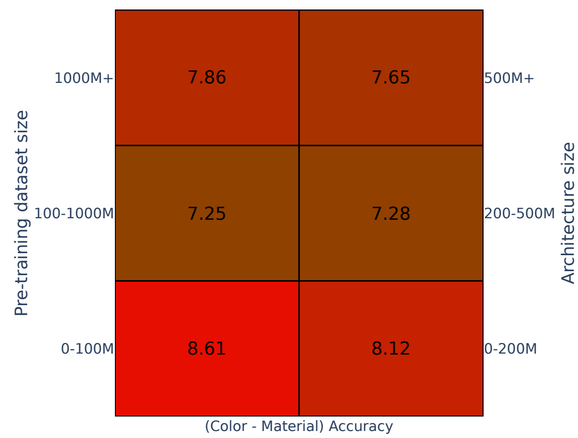


Figure B.7. **Increasing the pre-training dataset and architecture size only marginally reduces the difference in zero-shot accuracy between prompts describing an object by its color versus material.** Numbers reported are the delta in zero-shot accuracy between when a color versus material prompt is used as the text input to CLIP on the ORBIT Clean dataset. Each block averaged the delta for all CLIP variants that fall within that group. Experimental details in Sec. 4.3.1. Models: All 25 CLIP variants

C.2. Object detection with OWL-ViT

We extend Fig. 4 in the main paper with Fig. C.1 here, where we show one example of OWL-ViT's bounding box

Table C.1. **Top- and bottom-performing disability and non-disability objects from the ORBIT dataset sed for the example-based analysis.** The noun phrases are extracted from the raw label and cleaned to ensure they are unambiguous and concise.

<table>
<tr><th></th><th></th><th>Raw object label</th><th>Cleaned noun phrase</th></tr>
<tr><td rowspan="10">Disability objects</td><td rowspan="5">Top 5</td><td>my braille displat</td><td>braille sense display</td></tr>
<tr><td>dog poo</td><td>dog poo</td></tr>
<tr><td>dog lead</td><td>dog lead</td></tr>
<tr><td>white cane</td><td>guide cane</td></tr>
<tr><td>folded long guide cane</td><td>guide cane</td></tr>
<tr><td rowspan="5">Bottom 5</td><td>victor reader stream</td><td>victor reader stream</td></tr>
<tr><td>braille note</td><td>braille notetaker</td></tr>
<tr><td>liquid level indicator</td><td>liquid level indicator</td></tr>
<tr><td>liquid level indicator</td><td>liquid level indicator</td></tr>
<tr><td>dictaphone</td><td>dictaphone</td></tr>
<tr><td rowspan="10">Non-disability objects</td><td rowspan="5">Top 5</td><td>back patio gate</td><td>gate</td></tr>
<tr><td>local post box</td><td>post box</td></tr>
<tr><td>wine glass</td><td>wine glass</td></tr>
<tr><td>tv remote control</td><td>remote control</td></tr>
<tr><td>remote control</td><td>remote control</td></tr>
<tr><td rowspan="5">Bottom 5</td><td>digital dab radio</td><td>digital radio</td></tr>
<tr><td>my clock</td><td>digital clock</td></tr>
<tr><td>grinder</td><td>tobacco grinder</td></tr>
<tr><td>dog streetball</td><td>ball</td></tr>
<tr><td>shoulder bag</td><td>shoulder bag</td></tr>
</table>

Table C.2. **OWL-ViT's mean intersection-over-union (IOU) is ∼2x lower for disability compared to non-disability objects.** Mean IOU (with 95% c.i.) is computed between the predicted and ground-truth bounding box for each object.

	mean IOU
Disability objects	0.1323 (0.0947)
Non-disability objects	**0.2488 (0.1829)**

detections for each of the 10 disability and 10 non-disability objects. Specifically, for each object we show the image that had the bounding box with the highest confidence score across all 9 images analyzed for that object. We see that the confidence scores in these images are ∼3x lower for disability than non-disability objects, on average. We also see that for 4/10 disability objects, the incorrect object is detected (versus 2/10 non-disability objects).

We also report the mean intersection-over-union (IOU) between OWL-ViT's predicted and ground-truth bounding box for each object in Tab. C.2. Since ground-truth bounding boxes are only publicly available for the clutter images, we manually annotated the remaining 6 clean images per object. Our results show that the mean IOU is ∼2x lower for disability compared to non-disability objects. Taken together, these results suggest that overall, OWL-ViT performs less reliably and confidently for disability content.

C.3. Semantic segmentation with CLIPSeg

Semantic segmentation models are also highly likely to be integrated into assistive applications to help BLV users localize objects. We examine CLIPSeg [37] which trains a decoder on top of CLIP's frozen vision and text encoders to enable zero-shot image segmentation from text prompts. Unlike OWL-ViT, CLIPSeg does not fine-tune the CLIP encoders, and its pre-trained embeddings are used directly. As before, we run all 180 images through the model with the cleaned noun phrases as text prompts. We find:

Table C.3. **CLIPSeg segments non-disability objects with higher confidence than non-disability objects**. The average confidence (with 95% c.i.) is reported over all pixels with a confidence above 0.1 within the object's ground-truth bounding box.

	Avg in-box confidence
Disability objects	0.2181 (0.0842)
Non-disability objects	**0.4276 (0.1286)**

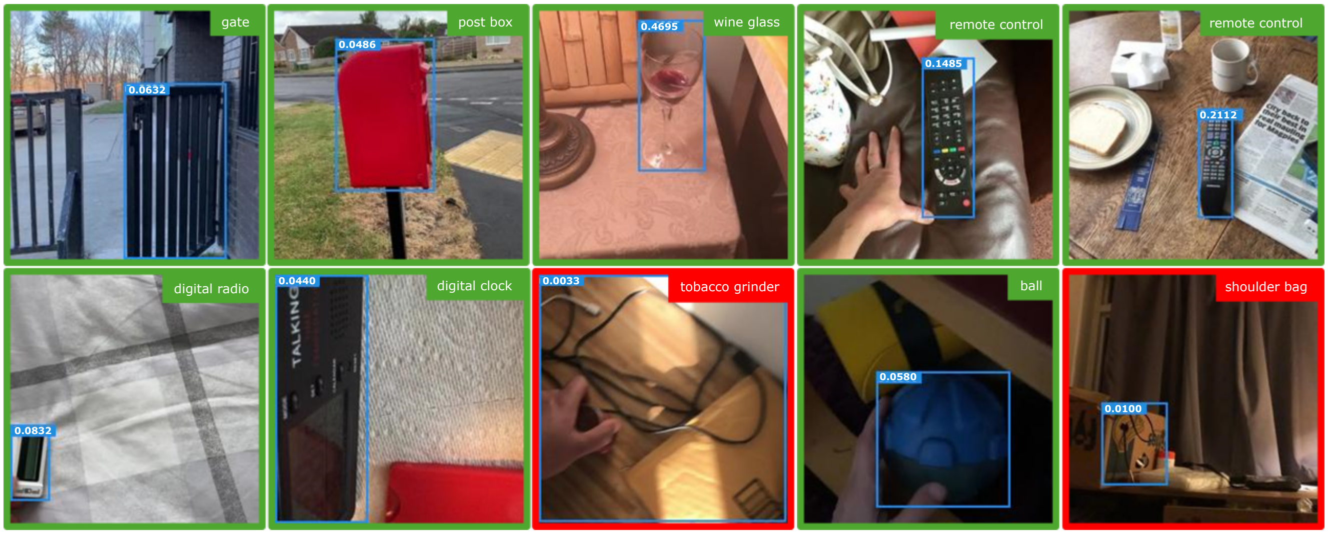
Table C.4. **CLIPSeg incorrectly segments disability objects more often than non-disability objects on the ORBIT Clutter dataset**. Numbers are the average confidence over all pixels *outside* the object's ground-truth bounding box divided by the average confidence over all pixels in the image (with 95% c.i.), considering only pixels above a 0.1 confidence threshold.

	Avg confusion score
Disability objects	**0.2698 (0.1990)**
Non-disability objects	0.1292 (0.0749)

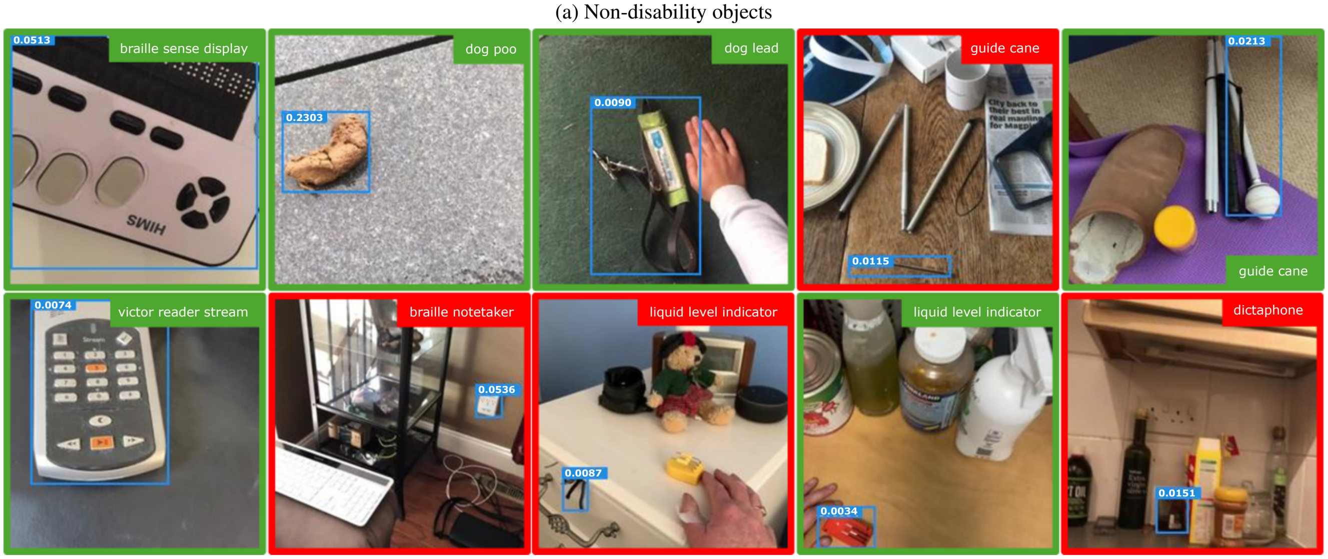
Segmentation maps of disability objects are less confident than those of non-disability objects. In Tab. C.3, we compute the average confidence value over all pixels in the segmentation map that fall within the ground-truth bounding box of the target object. To control for the degree of background present across bounding boxes (especially for irregular-shaped objects), we only consider pixels that have a confidence score greater than 0.1. With this, we find that CLIPSeg's segmentation maps are ∼2x more confident for non-disability objects compared to disability objects. In Fig. C.2, we show CLIPSeg's segmentations for a guide cane versus a TV remote, two objects for which the confidence score difference was most pronounced.

Disability objects are more likely to be segmented as the incorrect object in realistic settings compared to non-disability objects. In Tab. C.4, we compute a confusion score per image: the average confidence score of all the pixels that fall *outside* the object's ground-truth bounding box, divided by the average confidence score over all pixels in the image. This gives us a measure of how confidently the model is segmenting objects besides the ground-truth object, where a high score indicates the segmentation may be a false positive. Here we also only include confidence scores above a 0.1 threshold. We see that CLIPSeg is ∼2x more likely to confuse a disability object with another object compared to a non-disability object in ORBIT Clutter images where multiple objects are present.

We include examples of this in Fig. C.3. Here we see that



(a) Non-disability objects



(b) Disability objects

Figure C.1. **OWL-ViT detects disability objects less confidently than non-disability objects.** For each of the (a) 10 non-disability and (b) 10 disability objects, we show the image with the highest-scoring bounding box out of the 9 images analyzed for that object.

CLIPSeg fails to segment prominent disability objects (see liquid level indicators, guide canes, and Braille notetakers in Fig. C.3b) but succeeds in segmenting non-disability objects in similarly cluttered scenes (see shoulder bags and wine glasses in Fig. C.3a).

C.4. Text-to-image generations with DALL-E2

Prompt templates. Three annotators manually created two prompts for each of the 20 objects. The first prompt was just the cleaned noun phrase (taken from Tab. C.1). The second prompt combined the cleaned noun phrase with a surface and up to two adjacent objects. The surface and adjacent objects were selected to match an image from the ORBIT Clutter dataset of that object. The image was selected such that it had at least one adjacent object present. The prompt was then created with the template: "*<object_name > on <surface> next to <adjacent-object-1> and <adjacent-object-2>*" (*e.g.* "wine glass on a wooden table next to a bottle of wine and a candle").

We extend Fig. 5 in the main paper with Fig. C.4 here. We show DALL-E2's generations for the two prompt types for non-disability (Fig. C.4a) and disability (Fig. C.4b) objects. Overall, we see that DALL-E2 does not generate correct representations for many of the disability objects, either defaulting to a common object or fabricating an object entirely. In contrast, the generations for non-disability objects are highly realistic and mostly correct.

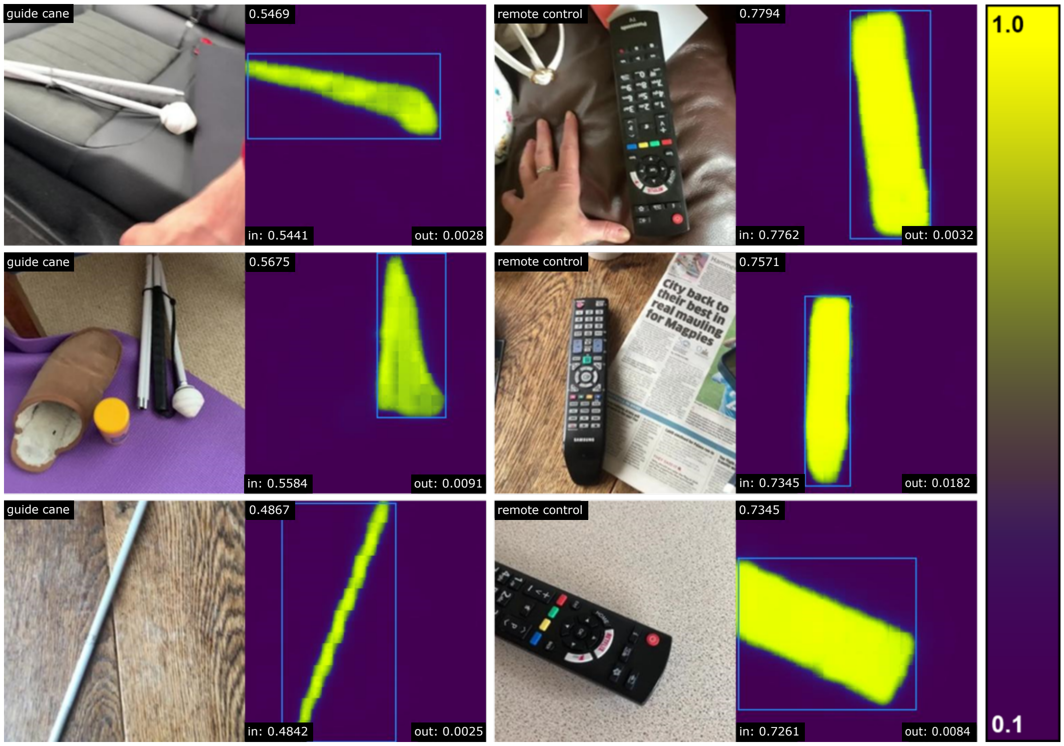
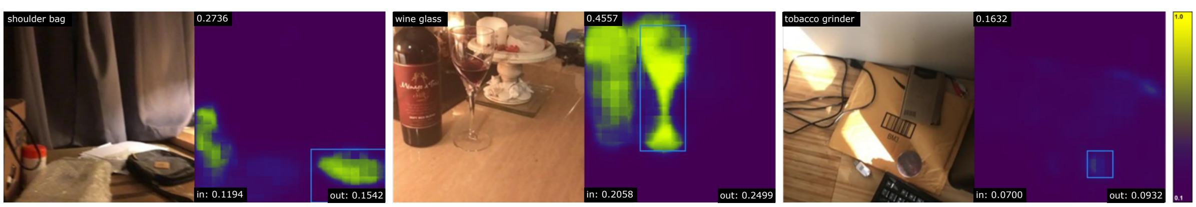
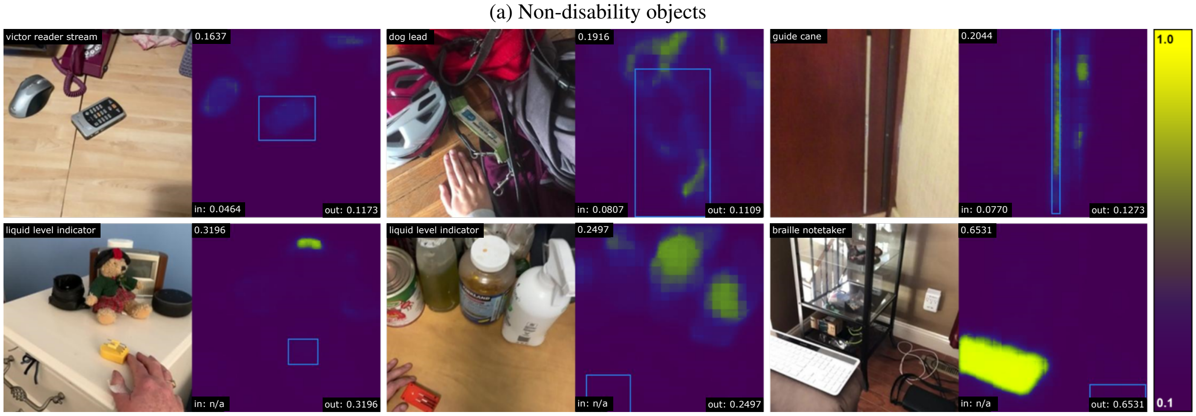


Figure C.2. **CLIPSeg segments non-disability objects (right: TV remote) with higher confidence than disability objects (left: guide cane).** For each image, we report the average confidence score over all pixels inside and outside the object's ground-truth bounding box ("in" and "out", respectively), considering only pixels above a 0.1 confidence threshold. See quantitative results in Tab. C.3.

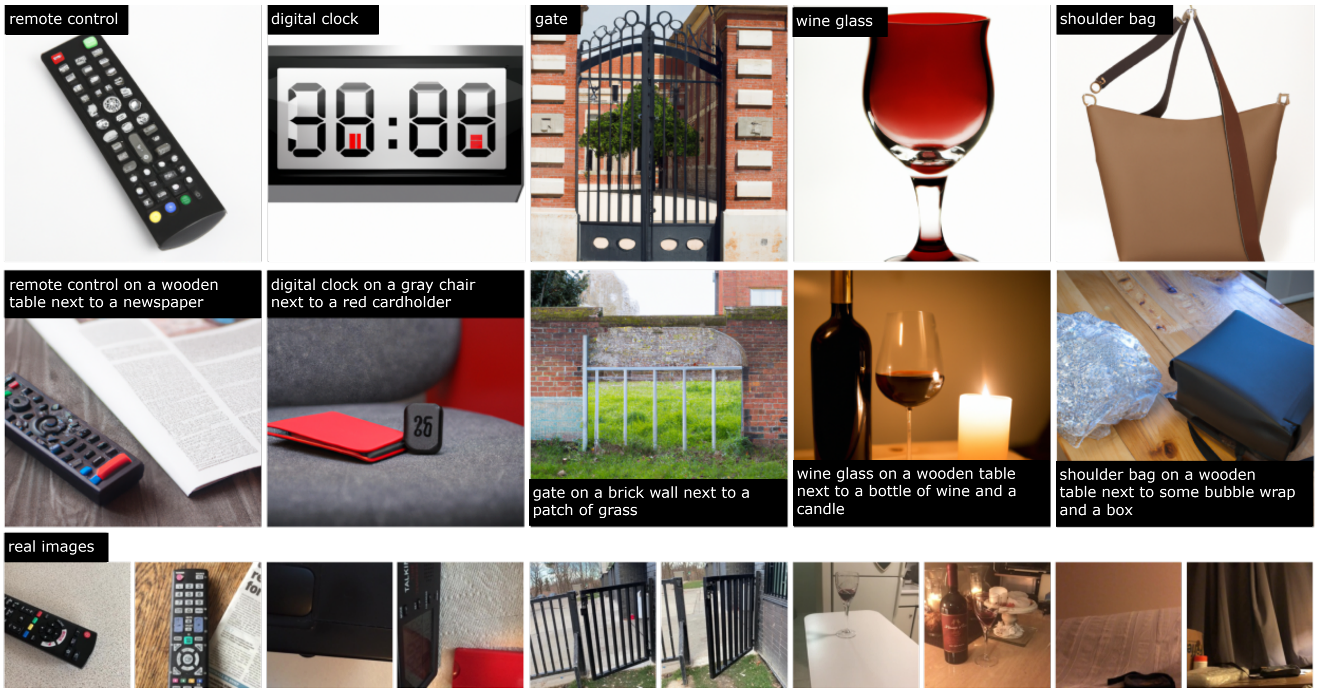


(a) Non-disability objects

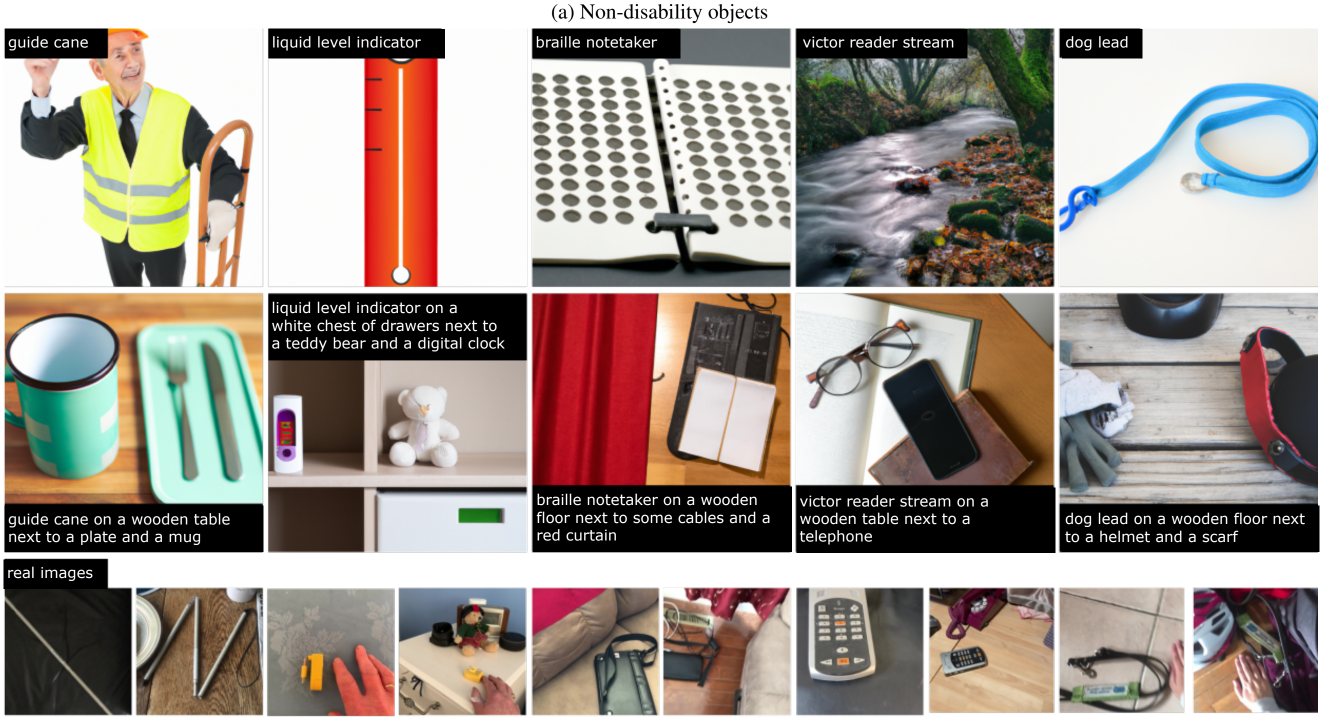


(b) Disability objects

Figure C.3. **CLIPSeg is more likely to segment disability objects (bottom) incorrectly in cluttered scenes compared to disability objects (top).** For each image, we report the average confidence score over all pixels inside and outside the object's ground-truth bounding box ("in" and "out", respectively), considering only pixels above a 0.1 confidence threshold. The correct object is marked by the bounding box. See quantitative results in Tab. C.4.



(a) Non-disability objects



(b) Disability objects

Figure C.4. **DALL-E2 generates high-quality images of non-disability objects (a), but defaults to more common objects or fabrications for disability objects (b).** For each sub-figure, the top row shows generations for a simple prompt containing just the object name, while the second row shows generations for the richer prompt where a surface and adjacent objects are also specified. The bottom row shows real images of each object.