BINDSPLIT: BENCHMARKING SEEN AND UNSEEN COMPOSITIONAL GENERALIZATION IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional generalization requires vision-language models (VLMs) to understand novel combinations of familiar concepts, and it is measured with compositional benchmarks. However, if most evaluation samples contain combinations that models already saw during training, high benchmark scores may not reflect generalization to new compositions. In this work, we ask whether compositional benchmark scores reflect generalization to unseen compositions. We focus on attribute–object bindings and introduce the BindSplit protocol, which splits the samples of existing benchmarks into seen and unseen by whether their bindings appear in a reference corpus that approximates the training data. We find that existing benchmarks contain very few samples with unseen bindings: at most 88 per benchmark against CC12M. To address this gap, we introduce BindSplit Bench, a benchmark built to compare performance on seen and unseen bindings directly. Its seen and unseen splits have similar sizes, share their attributes, and are generated and human-reviewed through the same pipeline. Across 75 VLMs, accuracy drops from seen to unseen bindings for 72 models, and seen accuracy, although correlated with unseen accuracy, overstates how much models improve on unseen bindings. Our results suggest that progress on existing compositional benchmarks may not translate directly to unseen compositions. By separating seen and unseen bindings, BindSplit Bench makes this generalization gap measurable.

1 INTRODUCTION

Compositional generalization often refers to the ability to understand novel compositions of familiar parts (Fodor & Pylyshyn, 1988). The real world is made up of combinations of familiar concepts,

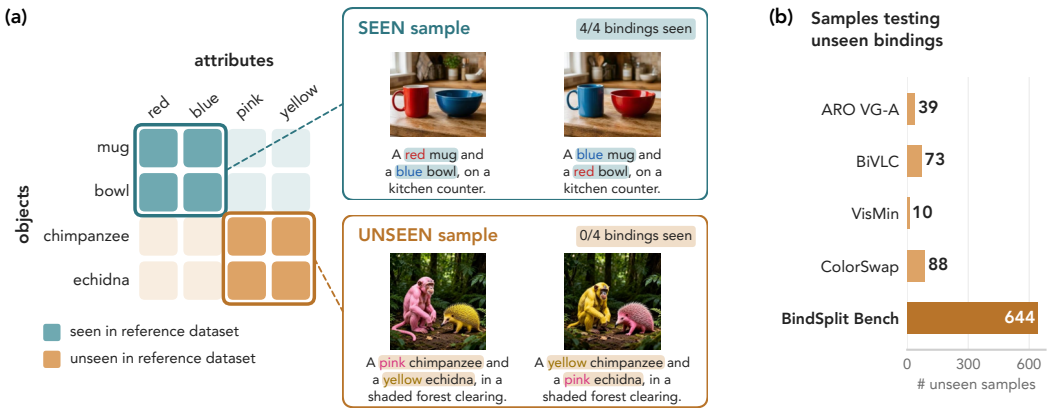

Figure 1: **Existing compositional benchmarks contain few samples that test unseen bindings; `BindSplit` Bench (ours) has 644.** (a) A sample is *seen* when all four of its attribute–object bindings (two per caption) appear in the reference training data, and *unseen* when none of them do. (b) Number of unseen samples in existing compositional benchmarks (no binding in CC12M), and in **`BindSplit` Bench** (no binding in COCO, CC12M, or LAION-400M).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

many of which never appear in the training data, so models need this ability to generalize beyond it. Vision-language models (VLMs) such as CLIP struggle with it: they cannot tell "a red mug and a blue bowl" from "a blue mug and a red bowl" (Figure 1a). These two captions share the same words and differ only in which attribute describes which object (Yuksekgonul et al., 2023). This failure matters, because CLIP is widely used for zero-shot classification, retrieval, and as the vision encoder of multimodal large language models (Radford et al., 2021; Tong et al., 2024). Many methods try to improve CLIP's compositional reasoning (Yuksekgonul et al., 2023; Fan et al., 2023b; Doveh et al., 2023; Patel et al., 2024; Vani et al., 2024; Peleg et al., 2025; Kargi et al., 2026), and their progress is measured by compositional benchmarks, where a higher score indicates better compositionality. We cannot improve what we cannot reliably measure, so it matters whether these benchmarks actually measure compositional generalization.

If a model achieves a higher compositional benchmark score, we would expect it to be better at understanding novel compositions that are unseen in its training data. But are these models really generalizing compositionally? This depends on the compositions in the benchmarks. How often concepts and their combinations appear in training predicts how well models recognize them at evaluation (Udandarao et al., 2024; Wiedemer et al., 2025; Qu & Xie, 2025). So if a benchmark mostly contains compositions that models have already seen, a high score may reflect recognizing familiar compositions rather than generalizing to new ones. We study this for a type of composition that is conceptually simple yet still hard for many VLMs: attribute–object bindings (Yuksekgonul et al., 2023; Lewis et al., 2024). We propose the BindSplit protocol, which splits benchmark samples into **seen** and **unseen** by whether their bindings appear in a reference corpus that serves as a proxy for training data, such as CC12M. The seen split tests how well models recognize familiar compositions; the unseen split tests how well they generalize to novel ones.

Applying the BindSplit protocol to four existing benchmarks, we find that they barely test unseen bindings. Against CC12M, each benchmark has at most 88 samples whose bindings are all unseen (Figure 1b). With so few samples, we cannot reliably measure or compare how well models generalize to unseen bindings.

We therefore build **`BindSplit` Bench** (Figure 1a), a benchmark designed to compare how models perform on seen and unseen bindings. Unlike existing benchmarks, whose unseen samples are the few that happen to be unseen, we construct the seen and unseen splits to be comparable: (1) they have similar sizes (675 seen, 644 unseen), so accuracy on each split is estimated from a similar number of samples; (2) they use the same attributes, and their objects are equally recognizable (mean 0.985 for seen and 0.986 for unseen objects), so a low unseen score does not simply reflect an unknown concept; (3) both come from the same generation and review pipeline, so they do not differ in image or caption style; and (4) in both, the correct and the wrong caption contain equally familiar bindings (all seen or all unseen), so familiarity alone cannot help models pick the correct caption. Each sample follows established practice for testing binding (Thrush et al., 2022; Burapacheep et al., 2024; Udandarao et al., 2025): two images and two captions that swap the attributes of two objects, so a model must bind each attribute to the right object. Human reviewers keep only generated samples whose two images differ minimally beyond the swap.

On **`BindSplit` Bench**, we evaluate 75 models built on CLIP, SigLIP, and EVA-CLIP backbones from ResNet-50 to ViT-E/14, in three regimes: zero-shot, LABCLIP (a linear layer on frozen text embeddings), and compositional fine-tuning. We observe two key results. **(1) VLM accuracy drops from seen to unseen bindings:** 72 of 75 models (96%) drop in text accuracy, by 6.02 percentage points (pp) on average and by up to 13.2 pp, and fine-tuning on hard-negative captions does not remove the drop. The drop also appears in a controlled synthetic setting in which every color and animal is seen in training and only the bindings are new. **(2) Seen accuracy is correlated with unseen accuracy, but overstates progress on unseen bindings.** Across all 75 models, models that score higher on seen bindings also score higher on unseen ones ($R^2 = 0.75$). Earlier work found the same pattern with the one binding setting (Ma et al., 2022; Abbasi et al., 2024), in line with accuracy-on-the-line (Miller et al., 2021); we find that it also holds when each image-caption pair has two bindings. Unseen accuracy, however, rises more slowly than seen accuracy: a model that gains 10 pp on seen bindings gains only about 6 pp on unseen ones. Seen scores, which are mostly what existing benchmarks measure, therefore overstate how much models improve on unseen bindings.

To summarize, our contributions are:

- **BindSplit protocol and benchmark audit.** A protocol that splits compositional benchmark samples into seen and unseen by checking every binding, in both the correct and the wrong caption, against a reference corpus. Applied to four benchmarks, it finds at most 88 unseen samples per benchmark against CC12M.
- **`BindSplit` Bench.** A benchmark with similar-sized seen and unseen splits, generated by one pipeline and human-reviewed, following established benchmark design practices (Udandarao et al., 2025).
- **An evaluation of 75 VLMs.** Nearly all models drop from seen to unseen bindings, including in a controlled synthetic setting, and seen accuracy overstates how much models improve on unseen bindings.

2 Related Work

Improving and measuring VLM compositionality. Most methods that aim to improve compositional understanding in CLIP-style models train the model with hard negatives: captions or images that differ from the correct ones in one compositional detail. These are made by swapping words within a caption (Yuksekgonul et al., 2023), between two captions (Peleg et al., 2025), by editing one part at a time (Kargi et al., 2026), or by generating new captions and images (Patel et al., 2024). Other methods improve the training captions (Doveh et al., 2023; Fan et al., 2023b), change the architecture (Vani et al., 2024), refine text representations at inference (Zhang et al., 2025), or train only a linear layer on top embeddings (Koishigarina et al., 2026). Progress is tracked on benchmarks where a model must match an image to its caption when the wrong option differs in attribute binding, relations, or word order (Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023; Burapacheep et al., 2024; Miranda et al., 2024; Awal et al., 2024; Thrush et al., 2022; Burapacheep et al., 2024; Miranda et al., 2024; Awal et al., 2024). Audits of these benchmarks found visually hard or ambiguous samples (Diwan et al., 2022), implausible negatives (Hsieh et al., 2023), and caption or image differences that let blind baselines match CLIP (Udandarao et al., 2025). However, whether benchmark compositions were seen during training remains largely unexamined. A higher score may therefore reflect recognizing familiar compositions rather than generalizing to new ones.

Training-data exposure shapes VLM behavior. What a VLM sees in training shapes its behavior: pre-training data are long-tailed, how often a concept appears in training predicts how well models recognize it (Fang et al., 2022; Nguyen et al., 2022; Mayilvahanan et al., 2024; Udandarao et al., 2024; Parashar et al., 2024), and covering many distinct combinations matters more than the total amount of data (Uselis et al., 2025). Combinations of concepts matter too: CLIP's zero-shot classification is correlated with pretraining concept co-occurence (Qu & Xie, 2025). A few studies look up test examples in the training data. SVO-Probes counts how often subject-verb-object triplets appear in pre-training captions, but only for its own dataset (Hendricks & Nematzadeh, 2021). Wiedemer et al. (2025) filter existing retrieval benchmarks for unseen object combinations and show that performance depends on individual-object frequency in pretraining data. To our knowledge, no prior work checks existing benchmarks for attribute-object binding in both positive and negative captions.

Evaluating generalization to unseen compositions. Some existing work tests models on bindings held out from training. It differs in what "unseen" is measured against and how many bindings a sample contains. Compositional zero-shot learning (CZSL) holds out bindings from a controlled training split, with one labeled binding per image (Purushwalkam et al., 2019; Naeem et al., 2021). Others hold out color–shape bindings from their fine-tuning data in two-object scenes (Lewis et al., 2024; Pearson et al., 2025); our controlled PUG:SPARE experiment follows this setup. These splits, however, do not account for bindings seen in web-scale pretraining. CREPE (Ma et al., 2022) and ImageNet-AO (Abbasi et al., 2024) instead define unseen combinations by the actual pretraining data of CLIP models, but test one combination at a time: each correct CREPE caption describes a single attribute–object or object–relation–object combination, and each ImageNet-AO image shows a single attribute-object combination. Other benchmarks use two images and two captions per sample (Thrush et al., 2022; Miranda et al., 2024; Awal et al., 2024; Burapacheep et al., 2024; Miranda et al., 2026), but none of them separates bindings seen in web-scale training data from unseen ones. Our benchmark combines the two: it uses this two-image, two-caption format, defines seen and unseen splits against web-scale caption corpora, and builds and scores both splits the same way.

3 DO COMPOSITIONAL BENCHMARKS MEASURE COMPOSITIONAL GENERALIZATION TO UNSEEN BINDINGS?

In this section, we check whether existing benchmarks measure compositional generalization. Compositional generalization is usually understood as the ability to understand novel combinations of known concepts, such as attributes and objects (Fodor & Pylyshyn, 1988). If a model has only ever seen "blue sky" and "white wall", can it still recognize a "blue wall" or a "white sky"? Whether a combination appeared in training matters: how often concepts and their combinations appear in pretraining predicts how well models recognize them (Udandarao et al., 2024; Wiedemer et al., 2025; Qu & Xie, 2025), and more diverse combinations in training improve generalization to new ones (Uselis et al., 2025). So if a benchmark mostly contains combinations that models have already seen, high accuracy may tell us how well models recognize familiar combinations, not how well they generalize to new ones. Before asking whether VLMs generalize compositionally, we therefore ask: *do existing benchmarks contain combinations that are new with respect to the training data?* We focus on attribute–object bindings, which are conceptually simple yet still hard for many VLMs (Yuksekgonul et al., 2023; Lewis et al., 2024; Koishigarina et al., 2026).

3.1 BINDSPLIT PROTOCOL: PARTITIONING BENCHMARK SAMPLES INTO SEEN AND UNSEEN SPLITS

To answer this question, we need to know which bindings each benchmark sample tests, and whether each binding appeared in training. The BindSplit protocol finds this out in four steps.

A *binding* is an (attribute, object) pair, such as (red, car) in "the red car". In an attribute-binding sample, the correct caption and the wrong caption usually swap two attributes ("the white wall and the blue sky" vs. "the blue wall and the white sky"), so together they contain four bindings. Some benchmarks, such as BiVLC and VisMin, also give the wrong caption its own image, as our benchmark does (Figure 1).

Step 1: Find the bindings in each caption. ARO VG-A captions have a fixed form ("the [attribute] [object] and the [attribute] [object]"), so we read the bindings off directly, keeping the 26,508 of 28,748 samples whose attributes and objects are single words. For free-form captions, we extract noun phrases with Qwen3-14B-AWQ, following Kargi et al. (2026), and take an attribute directly in front of a noun as a binding.

Step 2: Choose a reference corpus. We cannot inspect every model's training data, so we check bindings against a public image–caption corpus that serves as a proxy: the COCO training set (2014) or CC12M, each used on its own. *Seen* and *unseen* therefore mean seen or unseen in the reference corpus.

Step 3: Label each binding. A binding is *seen* if its attribute directly precedes its object in some noun phrase of the corpus, treating singular and plural objects as the same, and *unseen* otherwise. For example, "a green pear beside a red apple" counts for (green, pear) but not for (green, apple). A looser rule, which counts an attribute and an object anywhere in the same noun phrase, would label more bindings as seen and leave even fewer unseen samples (Appendix A.9).

Step 4: Partition the samples. A sample is *seen* if all of its bindings, in both captions, are seen, and *unseen* if all of them are unseen; the rest are *mixed*. We leave mixed samples out of both splits, because their two captions can differ in how familiar their bindings are, and a model could then pick the correct caption by familiarity alone.

3.2 EXISTING BENCHMARKS CONTAIN FEW SAMPLES WITH UNSEEN BINDINGS

We apply BindSplit to ARO VG-A, BiVLC, VisMin, and ColorSwap (Table 1). Very few samples test unseen bindings. Even against COCO, a relatively small corpus, only 422 of 26,508 ARO VG-A samples (1.6%) are unseen, and against CC12M only 39 (0.15%) are. BiVLC and VisMin mostly test things other than attribute–object bindings: only 791 of 2,933 BiVLC samples and 294 of 2,084 VisMin samples change an attribute, and of these only 73 and 10 are unseen against CC12M.

ColorSwap has more unseen samples (255 against COCO, 88 against CC12M), but they mostly come from a different image generator than its seen samples: against CC12M, 95% of the unseen

Table 1: **Few samples in existing benchmarks test unseen bindings.** We apply the BindSplit protocol to four benchmarks, with COCO (train 2014) or CC12M as the reference corpus, each used on its own. *Labeled*: samples from which the BindSplit protocol extracts attribute–object bindings. Seen and unseen are given as percentages of labeled samples; the remaining labeled samples are mixed. Full counts are in Appendix A.8.

		Seen (%)		Unseen, n (%)	
Benchmark	Labeled	COCO	CC12M	COCO	CC12M
ARO VG-A	26,508	23.56%	58.10%	422 (1.59%)	39 (0.15%)
BiVLC	536	54.85%	70.15%	97 (18.10%)	73 (13.62%)
VisMin	159	79.25%	82.39%	17 (10.69%)	10 (6.29%)
ColorSwap	923	35.43%	65.01%	255 (27.63%)	88 (9.53%)

samples have DALL-E 3 images, compared with 28% of the seen samples. A drop from seen to unseen accuracy on ColorSwap may therefore mix new bindings with a change of image generator (Udandarao et al., 2025).

These benchmarks were not designed to test unseen bindings, so they contain only the few samples whose bindings happen to be unseen. To measure generalization to unseen bindings, we need a benchmark designed for it: (1) an unseen split large enough to measure accuracy on, (2) a seen split of similar size, (3) both built with the same generation and review pipeline, and (4) unseen bindings checked against large reference corpora.

4 BindSplit Bench: Seen and Unseen Bindings Benchmark

Section 3 ended with what a benchmark for unseen bindings needs: a large unseen split, a seen split of similar size, both built by the same pipeline, and unseen bindings checked against large reference corpora. **BindSplit Bench** meets these needs at the level of its splits: 644 unseen and 675 seen samples, built by one pipeline (Section 4.2), with unseen bindings absent from MS-COCO, CC12M, and LAION-400M. Within each sample, four design choices (D1–D4) make sure that a model has to bind attributes to objects to be correct.

4.1 Design choices

D1: Sample structure. Each sample has two image–caption pairs whose captions swap the attributes of two objects (Figure 1a). A sample is seen if all four of its bindings appear in the reference corpora, and unseen if none of them do.

D2: Two bindings per caption. With one binding, as in "a pink cheeseburger" versus "a turquoise cheeseburger", a model can be correct by recognizing the color alone. With two, both captions mention the same things (pink, turquoise, a cheeseburger, and mashed potato) and differ only in which color binds to which object, so the model must bind them correctly. CLIP matches a single object to its attribute well, but does much worse when two objects each need their own attribute (Lewis et al., 2024).

D3: Minimal difference beyond the swap. Both captions share one LLM-written scene, both images come from the same generator and random seed, and human reviewers keep a sample only when its two images differ minimally beyond the swap. This removes shortcuts that do not require binding (Udandarao et al., 2025).

D4: Winoground-style bidirectional evaluation. Following Udandarao et al. (2025), we evaluate in both directions (image-to-text, text-to-image), in the two-image, two-caption format of Winoground (Thrush et al., 2022). **Text accuracy** counts a sample as correct when each image picks its own caption over the swapped one; **image accuracy** counts when each caption picks its own image over the other image; **group accuracy** needs both to be correct.

The closest benchmarks meet only some of these design choices: ColorSwap (Burapacheep et al., 2024) swaps colors between two objects but generates its two images independently, and VisMin (Awal et al., 2024) keeps its image pairs minimal but replaces one attribute instead of swapping attributes between two objects.

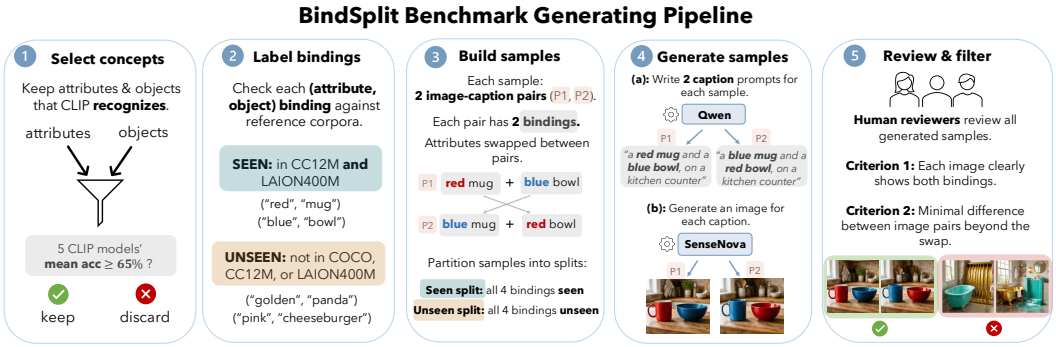

Figure 2: **The `BindSplit` Bench generation pipeline.** (1) Keep attributes and objects that CLIP models recognize. (2) Label each binding seen (in both CC12M and LAION-400M) or unseen (in none of COCO, CC12M, and LAION-400M). (3) Build samples whose four bindings are all seen or all unseen. (4) Generate two captions (Qwen3) and two images from the same seed (SenseNova). (5) Human reviewers keep samples whose images show their bindings and differ minimally beyond the swap (left: kept; right: removed).

4.2 BindSplit BenchGENERATING PIPELINE

Our benchmark's generating pipeline consists of five steps (Figure 2): select concepts, label bindings, build samples, generate samples, and review and filter them.

Step 1: Select concepts. We take attributes and objects from ImageNet-AO (Abbasi et al., 2024), built on VAW attributes (Pham et al., 2021) and ImageNet-1K objects (Russakovsky et al., 2014). So that a low unseen score reflects binding rather than an unknown concept, we keep only those that five CLIP models recognize on their own (mean accuracy of at least 65% against each incorrect concept): 60 of 84 attributes and all 643 objects.

Step 2: Label bindings. Following the BindSplit protocol, we label every pair of a kept attribute and a kept object. A binding is *seen* if it appears in both CC12M and LAION-400M. Unseen candidates are ImageNet-AO bindings, which Abbasi et al. (2024) found absent from LAION, CommonPool, YFCC, and CC12M; we keep one as *unseen* only if its attribute and object never appear together in any noun phrase of COCO, CC12M, or LAION-400M (a 173M-sample subset), which is stricter than the rule in Section 3.1.

Step 3: Build samples. Each sample combines two attributes of the same type (e.g., two colors) with two objects; it is seen if all four of its bindings are seen, and unseen if all four are unseen. We draw the same number of candidates per attribute pair for both splits and leave out colors that are hard to tell apart (Appendix A.2).

Step 4: Generate samples. We turn each sample into two captions and two images. Qwen3-14B-AWQ (Yang et al., 2025) writes a short scene that fits both objects, and both captions share it: "A pink balloon and a turquoise bow tie, on a wooden table in a living room." SenseNova-U1.5-8B-MoT (Diao et al., 2026) then generates a 2048×2048 image for each caption from the same random seed, which keeps the scene and layout similar so that the swap is the main visual difference.

Step 5: Human review and filter. The same seed does not guarantee a good sample: an image can show the wrong binding, or the two images can differ beyond the swap. Each sample was checked by one of twelve human reviewers. Our main set (T1) keeps samples whose images both

show their bindings and differ minimally beyond the swap; a companion set (T2) also allows small extra differences. Unseen images show their bindings less often than seen ones (80.8% vs. 86.0% of reviewed samples), which is why every sample is reviewed (Table 2). After keeping only samples whose attributes appear in both splits, T1 has 675 seen and 644 unseen samples (T2: 969 and 914). We report results on T1, and on T2 in the appendix.

What the two splits share. The two filtered splits use the same 37 attributes but mostly different objects (9 shared, 113 seen-only, and 121 unseen-only object names). This follows from the all-four rule (Step 3): two objects that form a seen sample with two colors usually cannot also form an unseen sample with the same colors. We match attributes rather than objects because attributes are the harder concepts to recognize: across the 60 attributes and 643 objects that passed the recognizability screen (Step 1), mean accuracy is 0.74 for attributes and 0.98 for objects. The objects are equally recognizable in both splits (mean 0.985 over the seen object names and 0.986 over the unseen ones). Section 5.4 removes the object difference altogether: on the synthetic PUG:SPARE dataset, the seen and unseen splits share every object and attribute.

We release **`BindSplit` Bench** with its seen and unseen labels. The labels are relative to our reference corpora, not to any particular model's training data; researchers who have a model's training data can re-run our labeling (Step 2) on it to label the bindings for that model.

5 Evaluating VLMs on Seen and Unseen Bindings

In this section, we evaluate 75 models on **`BindSplit` Bench**, check that the benchmark measures what it is meant to (Section 5.3), and test the seen-to-unseen drop in a controlled synthetic setting (Section 5.4).

5.1 Experimental setup

Models evaluated We evaluate 75 models built on CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and EVA-CLIP (Sun et al., 2023) backbones, plus two CLIP variants, LaCLIP (Fan et al., 2023a) and SPARO (Vani et al., 2024). The backbones range from ResNet-50 to ViT-E/14 and were pretrained on corpora from YFCC-15M and CC12M to LAION-2B and WebLI. We use each backbone in up to three regimes, which change the pretrained model by increasing amounts: zero-shot (no change; 20 models), LABCLIP (one linear layer on frozen text embeddings; 16 models), and compositional fine-tuning (all weights; 39 models across three methods). We train each LABCLIP and fine-tuned model with 3 random seeds and report the mean. Not every backbone appears in every regime. Full list of models for each regime is listed in Table 3 (Appendix).

Which models the seen and unseen splits apply to. We label a binding seen or unseen by checking it against reference corpora (Section 4.2), not against each model's own training data. For 7 of our 20 backbones, those pretrained on CC12M or LAION-400M, the reference corpora include the model's own pretraining data, so for them we have checked exposure: a seen binding appears in the model's pretraining data, and an unseen binding does not. For the other backbones, the splits are an approximation: our seen binding never appear in a model's training data, and unseen binding may appear in it. For the two YFCC-15M backbones, only the unseen bindings are checked (by ImageNet-AO), and CC15M adds CC3M, which neither check covers. For the remaining 10 backbones, pretrained on OpenAI's WIT, LAION-2B, WebLI, or Merged-2B, we cannot check exposure. Compositional fine-tuning uses COCO, which we checked, so it adds no unseen binding as a positive caption.

LABCLIP and compositional fine-tuning. LABCLIP (Koishigarina et al., 2026) keeps both vision and text encoders frozen and trains a single linear layer on the text embeddings. Because the encoders are frozen, LABCLIP models differ only in the pretrained representations they start from. For compositional fine-tuning, we fine-tune all weights of the pretrained backbones with three compositional methods: NegCLIP (Yuksekgonul et al., 2023), CLIC (Peleg et al., 2025), and CS-CLIP (Kargi et al., 2026), using each method's default hyperparameters. We train both regimes on COCO (train2014).

Controlled synthetic setting (PUG:SPARE). PUG:SPARE (Koishigarina et al., 2026) has rendered images of two animals, each in one color, with captions such as "Red Lion and Blue Zebra" (12 animals, 8 colors). We hold out a block of 3 animals $\times$ 3 colors (9 animal–color bindings) from training; every animal and every color still appears in training in other combinations. Images whose two bindings are both seen are split into training, validation, and seen test sets (80/10/10). Test images with one or two held-out bindings form the partially unseen and fully unseen test sets. We train LABCLIP (20 backbones) and full fine-tuning (17 backbones) on the training set, repeat this with 3 different held-out blocks, and report image-to-text R@1 over all 7,392 two-animal captions, which include the attribute-swapped caption. A caption that lists the same two bindings in the other order also counts as correct.

5.2 BINDSPLIT BENCHRESULTS

Result 1: VLM accuracy drops from seen to unseen bindings. If a model generalizes compositionally, it should do about as well on unseen bindings as on seen ones. Most models do not: they have lower accuracy on the unseen split than on the seen split, in text, image, and group accuracy (Table 5). We call this decrease the *seen-to-unseen drop* and report it in percentage points (pp); our tables show it as a negative Δ, where Δ = unseen $-$ seen accuracy. The drop is largest for text accuracy (Figure 3): 72 of 75 models (96%) drop, by 6.02 pp on average and by up to 13.2 pp. The three models that do not drop change by only 0.5 to 2.8 pp, and two of them are YFCC-15M ResNets with some of the lowest seen text accuracies of all models, leaving little room to drop.

Compositional fine-tuning does not remove the seen-to-unseen drop. Text accuracy drops for 18 of 20 zero-shot models, 15 of 16 LABCLIP models, and all 39 compositionally fine-tuned models. The drop is also largest for the models whose exposure we actually checked. For the 22 models built on backbones pretrained on CC12M or LAION-400M, seen bindings appear in the model's own pretraining data and unseen bindings do not; for the other models, seen and unseen are an approximation (Section 5.1). All 22 of these models drop in text accuracy, by 8.13 pp on average. Image and group accuracy drop less (by 2.04 and 2.05 pp on average, for 83% and 91% of models), but both start low, which leaves less room to drop (Section 5.3).

Result 2: Seen accuracy is correlated with unseen accuracy. If seen accuracy rose one-to-one with unseen accuracy, existing benchmarks, which contain mostly seen bindings (Section 3), would be enough to track progress on unseen ones. We test this on text accuracy, the only metric on which any model scores above chance on seen bindings. Across all 75 models, seen and unseen text accuracy are strongly correlated ($R^2 = 0.75$; Figure 3), just as earlier work found for images with a single attribute–object pair (Abbasi et al., 2024; Ma et al., 2022); we extend this to two bindings per image, with seen and unseen samples built and scored the same way. The fitted slope, however, is well below 1: 0.60, with a 95% confidence interval of [0.45, 0.70] from resampling backbones (Appendix A.5). A model that gains 10 pp on seen bindings gains only about 6 pp on unseen ones. Seen accuracy is therefore correlated with unseen accuracy, but it overstates how much models improve on unseen bindings; only an unseen split measures the size of the accuracy gap. On the 22 models whose exposure we checked, the correlation is much weaker ($R^2 = 0.33$), over a narrower range of seen accuracy (18–33%), and the slope again stays below 1 (0.40 [0.23, 0.64]).

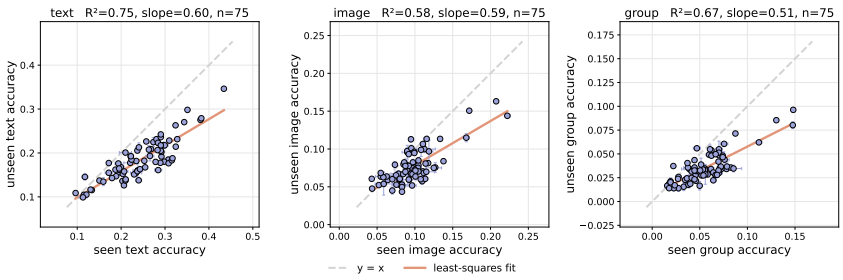

Figure 3: **Seen and unseen accuracies are correlated on `BindSplit` Bench** (T1; text, image, group). Each point is one of 75 models (trained models: mean $\pm$ SD over 3 seeds); solid line: least-squares fit, dashed: $y = x$. Probit version: Figure 11.

5.3 CHECKING BENCHMARK VALIDITY

Models score low even on seen bindings: no model exceeds 23% image accuracy or 15% group accuracy on the seen split (Table 3). Is this a limitation of the models, or a flaw in our benchmark? We check three things.

Low image and group accuracy is common. Existing benchmarks with real images show the same pattern. On Winoground, CLIP ViT-B/32 scores 30.75% in text but only 10.50% in image and 8.00% in group accuracy (Thrush et al., 2022), and on BiVLC every model's image accuracy is more than 20 pp below its text accuracy (Miranda et al., 2024).

Models can see attributes in our images. If our images did not clearly show their attributes, models would struggle even when each caption names only one binding. Instead, when we shorten each caption to its first binding ("A pink sewing machine, on a wooden workbench ...") and keep the images unchanged, image and group accuracy increase for all 7 zero-shot models we test, by 17.7 pp and 12.0 pp on average combining seen and unseen splits (Table 13; per split results in Appendix A.6). This is consistent with our images being readable, and with binding two objects at once being what models find hard. The seen-to-unseen drop also remains with one-binding captions: all 7 models still drop in text and group accuracy (Appendix A.6).

Our seen split agrees with the closest existing benchmark. ColorSwap (Burapacheep et al., 2024) also swaps the colors of two objects and gives each caption its own image. Across 125 checkpoints of 55 models[1], seen accuracy on **`BindSplit` Bench** is correlated with ColorSwap accuracy ($R^2 = 0.76$ for text and 0.70 for group accuracy). The correlation is weaker with VisMin (0.52 and 0.43) and BiVLC (0.53 and 0.22), whose samples also change objects, counts, spatial relations, or add new content (Appendix A.10).

Together, these checks suggest that the low scores reflect the models rather than our benchmark.

5.4 CONTROLLED SYNTHETIC SETTING (PUG:SPARE)

Result 4: When every attribute and object is seen in training, accuracy still drops as more bindings are unseen. In **`BindSplit` Bench**, the seen and unseen splits mostly use different objects (Section 4), so part of the drop could come from the objects rather than the bindings. PUG:SPARE removes this difference. The drop remains. R@1 drops as the number of unseen bindings in a sample goes from 0 (seen) to 1 (partially unseen) to 2 (fully unseen), for all 17 backbones with full fine-tuning and for 19 of 20 backbones with LABCLIP (Figure 15); the exception is SigLIP L/16 with LABCLIP.

Result 5: After full fine-tuning on PUG:SPARE, all models are at ceiling on seen bindings but vary widely on unseen ones. After full fine-tuning, all 17 backbones reach 99.7–100.0% seen R@1 (mean over the 3 data splits), while their fully unseen R@1 ranges from 42.8% to 96.5% (Figure 15). A test with only seen bindings would score these models as equal; only the unseen split shows that they differ.

6 DISCUSSION AND CONCLUSION

Recommendations. First, before reporting compositional progress on a benchmark, check how many of its samples test bindings that are unseen in the model's training data. The BindSplit protocol does this. Second, report accuracy on seen and unseen bindings separately, for example on **`BindSplit` Bench**, so that the gap between them is visible.

Limitations. Our seen and unseen labels are relative to reference corpora (COCO, CC12M, and LAION-400M), which match the pretraining data of only 7 of our 20 backbones. The two splits share attributes but mostly not objects (9 shared, 113 seen-only, and 121 unseen-only object names). The objects are equally recognizable, and PUG:SPARE shows the drop when all objects are shared, but part of the drop on **`BindSplit` Bench** may still come from the objects. Finally, samples can pair objects that rarely appear together in the real world (e.g., a shark and a mask), and unseen bindings are unusual by construction, so **`BindSplit` Bench** may be harder than benchmarks built from real photos.

[1] NegCLIP and 6 CS-CLIP backbones lack ColorSwap, VisMin, and BiVLC evaluations.

FINAL INSTRUCTIONS

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools, including AI coding agents, to implement parts of the method's code, to write evaluation and visualization scripts, and to assist in experiment analysis. We also used generative AI to create and process data: Qwen3-14B-AWQ (Yang et al., 2025) writes the benchmark captions from two attribute–object bindings, SenseNova-U1.5-8B-MoT (Diao et al., 2026) generates the corresponding images, and Qwen3-14B-AWQ extracts the noun phrases from existing benchmark and reference-corpus captions for BindSplit (Section 3). We have not used generative AI tools to propose or refine hypotheses; formulating mathematical claims, developing theoretical models, writing proofs, qualitative data analysis, and translation are not applicable to this work. Additionally, we used generative AI tools to search related literature and to draft and edit parts of the paper. The research direction, the method, and every decision on what to run and report were made by the authors, who reviewed all AI-assisted work: AI-written code was reviewed and tested, reported numbers come from reviewed scripts, and AI-drafted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

REFERENCES

Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Deciphering the role of representation disentanglement: Investigating compositional generalization in clip models. In *European Conference on Computer Vision*, 2024. URL https://api.semanticscholar.org/CorpusID:271051442.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=bJddXCyosA.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorswap: A color and word order dataset for multimodal evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 1716–1726. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.99. URL https://doi.org/10.18653/v1/2024.findings-acl.99.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. Sensenova-u1.5: Towards native unified visual intelligence, 2026. URL https://arxiv.org/abs/2609.11929.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:253255481.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ARrwf7Ev2T.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=SVjDiiVySh.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=SVjDiiVySh.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yu Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248505704.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28 1-2:3–71, 1988. URL https://api.semanticscholar.org/CorpusID:29043627.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics*, 2021. URL https://api.semanticscholar.org/CorpusID:235458570.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Jsc7WSCZd4.

Bora Kargi, Arnas Uselis, and Seong Joon Oh. Half-truths break similarity-based retrieval. *ArXiv*, abs/2602.23906, 2026. URL https://api.semanticscholar.org/CorpusID:286171185.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DldwXCCP25.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1487–1500, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.101. URL https://aclanthology.org/2024.findings-eacl.101/.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921, 2022. URL https://api.semanticscholar.org/CorpusID:254685851.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel. Does CLIP's generalization performance mainly stem from high train-test similarity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tnBaiidobu.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/miller21b.html.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. BiVLC: Extending vision-language compositionality evaluation with text-to-image retrieval. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=WMzQIP70O0.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Revisiting compositionality in dual-encoder vision-language models: The role of inference. *ArXiv*, abs/2604.11496, 2026. URL https://api.semanticscholar.org/CorpusID:287432648.

Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 953–962, 2021. URL https://api.semanticscholar.org/CorpusID:231786709.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of CLIP. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=LTCBavFWp5C.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. The neglected tails in vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12988–12997, 2024. URL https://api.semanticscholar.org/CorpusID:267094920.

Maitreya Patel, Naga Sai Abhiram kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZfRGRK5Kxl.

Beth Pearson, Bilal Boulbarss, Michael Wray, and Martha Lewis. Evaluating compositional generalisation in vlms and diffusion models. *ArXiv*, abs/2508.20783, 2025. URL https://api.semanticscholar.org/CorpusID:280949932.

Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=PTLIM88XYR.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13013–13023, 2021. doi: 10.1109/CVPR46437.2021.01282.

Senthil Purushwalkam, Maximillian Nickel, Abhinav Kumar Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3592–3601, 2019. URL https://api.semanticscholar.org/CorpusID:155089387.

Helen Qu and Sang Michael Xie. Impact of pretraining word co-occurrence on compositional generalization in multimodal models. *ArXiv*, abs/2507.08000, 2025. URL https://api.semanticscholar.org/CorpusID:280232343.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina

Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014. URL https://api.semanticscholar.org/CorpusID:2930547.

Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *ArXiv*, abs/2303.15389, 2023. URL https://api.semanticscholar.org/CorpusID:257766387.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *ArXiv*, abs/2007.00644, 2020. URL https://api.semanticscholar.org/CorpusID:220280805.

Damien Teney, LIN Yong, Seong Joon Oh, and Ehsan Abbasnejad. ID and OOD performance are sometimes inversely correlated on real-world datasets. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HZQZli6amV.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238, 2022. URL https://api.semanticscholar.org/CorpusID:248006414.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. URL https://api.semanticscholar.org/CorpusID:266976992.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=9VbGjXLzig.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good crepe needs more than just sugar: Investigating biases in compositional vision-language benchmarks, 2025. URL https://arxiv.org/abs/2506.08227.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does data scaling lead to visual compositional generalization? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=M2WMUuwoh5.

Ankit Vani, Bac Nguyen, Samuel Lavoie, Ranjay Krishna, and Aaron Courville. SPARO: Selective attention for robust and compositional transformer encodings for vision. In *European Conference on Computer Vision*, pp. 233–251, 2024. URL https://arxiv.org/abs/2404.15721. Added: 2026-05-07.

Thaddaus Wiedemer, Yash Sharma, Ameya Prabhu, Matthias Bethge, and Wieland Brendel. Pretraining frequency predicts compositional generalization of clip on real-world tasks. *ArXiv*, abs/2502.18326, 2025. URL https://api.semanticscholar.org/CorpusID:276580329.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,

Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. URL https://api.semanticscholar.org/CorpusID:257767223.

Qi Zhang, Yuxu Chen, Lei Deng, and Lili Shen. Abe-clip: Training-free attribute binding enhancement for compositional image-text matching. *ArXiv*, abs/2512.17178, 2025. URL https://api.semanticscholar.org/CorpusID:284058026.

A Appendix

A.1 Extended Related Works

Out-of-distribution generalization and accuracy-on-the-line. In image classification, models that are more accurate in distribution tend to also be more accurate out of distribution, a pattern called accuracy-on-the-line (Miller et al., 2021; Taori et al., 2020), although it is not universal and can break (Teney et al., 2023). These studies mostly concern image classification, with one label per image. In our setting, recognizing each individual attribute and object is not enough – the model must also bind them correctly. Earlier work has made an analogous correlation plot for one binding per image: Abbasi et al. (2024) plots ImageNet accuracy against accuracy on unseen bindings (Imagenet-AO) for CLIP models, and the two rise together; Ma et al. (2022) plots unseen against seen composition scores for CLIP models. For object combinations, in contrast, Wiedemer et al. (2025) find that unseen object combinations are about as easy as familiar combinations. We ask whether models with higher seen accuracy also have higher unseen accuracy in a 2×2 setting, where each sample has two image-caption pairs and the second pair swaps the attributes of the first (e.g., "a red mug and a blue cup" and "a blue mug and a red bowl"). We test this across zero-shot and compositionally fine-tuned models (Section 5).

A.2 BindSplit BenchGeneration Pipeline

Supplementary details of our benchmark generating pipeline.

Step 1: Select concepts. We take our attributes and objects from ImageNet-AO (Abbasi et al., 2024), whose attribute–object pairs are built on the attributes from Visual Attributes in the Wild (VAW) (Pham et al., 2021) and objects from ImageNet-1K classes (Russakovsky et al., 2014). Before building any samples, we check that models can recognize each attribute and object on its own. Otherwise, a low score on an unseen sample could mean that a model does not know an attribute or an object, rather than that it cannot bind them. We use five CLIP models to test each object and attribute's recognizability. Each object is evaluated on its ImageNet validation images and each attribute on labelled object crops from VAW. For each image, we compare the correct concept with each wrong concept (all-pairs binary classification): an object against all 642 other objects, and an attribute against the attributes that VAW marks as absent for that image. We keep attributes and objects with mean binary classification accuracy of at least 65% across all five models. This gives us 60 of 84 attributes and all 643 objects.

Step 3: Build samples. A sample combines two attributes of the same type, such as two colors or two materials, with two different objects. The two attributes and two objects create four distinct bindings. Each of the two image-caption pairs in a sample has two of the four bindings, and the pairs differ by the swapped attributes. As in Step 4 of the BindSplit protocol, a sample is seen if all four of its bindings are seen, and unseen if all four are unseen. We only build seen and unseen samples, so the benchmark has no mixed samples. For each pair of attributes, we select the same number of candidate samples for the seen and the unseen split. Human review later removes some samples, so the final counts per attribute pair are no longer equal. We leave out four color pairs that are hard to tell apart in an image: purple/violet, teal/turquoise, aqua/teal, and aqua/turquoise.

Step 5: Human review and filter. Using the same image-generating seed with careful Qwen-written captions does not guarantee a good sample that meets our criteria: an image can show an incorrect binding, and the two images can still differ in style and scenary. To make ensure our sample quality, we asked human reviewers to review every sample. For each image, reviewers must confirm it clearly shows both bindings in its caption. For the two images, they must check whether they can clearly tell the images apart by the swap and, if so, how much they differ beyond the swap: minimally, with some differences that are still acceptable, or a lot. Our main set (T1) keeps a sample when both images clearly show their bindings and visual difference is minimal beyond the attribute swap. A larger set (T2) also keeps samples with a bit more visual differences. The 2,362 samples were checked by thirteen reviewers; each sample was checked by one reviewer. The visually similar (T1) set has 677 seen and 646 unseen samples; the with-some-more-visual-differences (T2) set has 973 seen and 915 unseen samples.

Finally, we keep a sample only if both of its attributes appear in both splits, so that the seen and unseen splits use the same attributes. We call the result the *attribute-matched* T1 and T2. This removes 4 samples from T1, leaving 675 seen and 644 unseen, and 5 from T2, leaving 969 and 914. We report results on attribute-matched T1, and on attribute-matched T2 as a companion set. Most attribute-matched T1 samples swap two colors (976, 74% of both splits) or two materials (335, 25% of both splits); 8 samples swap patterns. The two samples with state (e.g., wet, snowy) as the attribute type that passed human review were removed by this filter. Not every word kept in Step 1 ends up in the benchmark: shape and texture attributes form almost no usable samples, some words have no sample in one of the two splits, and the near-synonym exclusion, the review, and the attribute filter remove more. Attribute-matched T1 uses 37 of the 60 kept attributes and 243 of the 643 kept objects.

Review per split. Table 2 follows the samples of each split through the review. The steps are applied in order, and each percentage is out of the reviewed samples of that split. Unseen samples pass every step slightly less often than seen samples: both images show their bindings for 80.8% of the unseen samples and 86.0% of the seen samples, and 55.0% of the unseen samples reach T1, against 57.0% of the seen samples. Every sample was reviewed by one reviewer. A reviewer could save a sample again; we drop the 12 samples whose answer changed between saves. The samples that were not reviewed were not rejected: the review tool shows samples in a fixed random order, and the review stopped before every sample was shown.

More examples. Figures 5–7 show more samples. Each row pairs a seen and an unseen sample that swap the same two attributes.

Table 2: Human review per split. From "Consistent verdict" to "T1, attribute-matched", each row counts the samples that also pass every row above it; percentages are out of the reviewed samples of each split. "T1, attribute-matched" is the set used for the main results. The last row counts samples rejected because the two images differ a lot beyond the swap.

Stage	Seen	Unseen
Selected pairs	1,586	1,586
After near-synonym filter	1,362	1,362
Reviewed	1,187	1,175
Consistent verdict	1,184 (99.8%)	1,166 (99.2%)
Image 1 shows its bindings	1,094 (92.2%)	1,035 (88.1%)
Both images show their bindings	1,021 (86.0%)	949 (80.8%)
Images can be told apart	1,013 (85.3%)	948 (80.7%)
Differ minimally beyond the swap, or borderline (T2)	973 (82.0%)	915 (77.9%)
Differ minimally beyond the swap (T1)	677 (57.0%)	646 (55.0%)
T1, attribute-matched (reported)	675 (56.9%)	644 (54.8%)
Rejected: images differ a lot	40 (3.4%)	33 (2.8%)

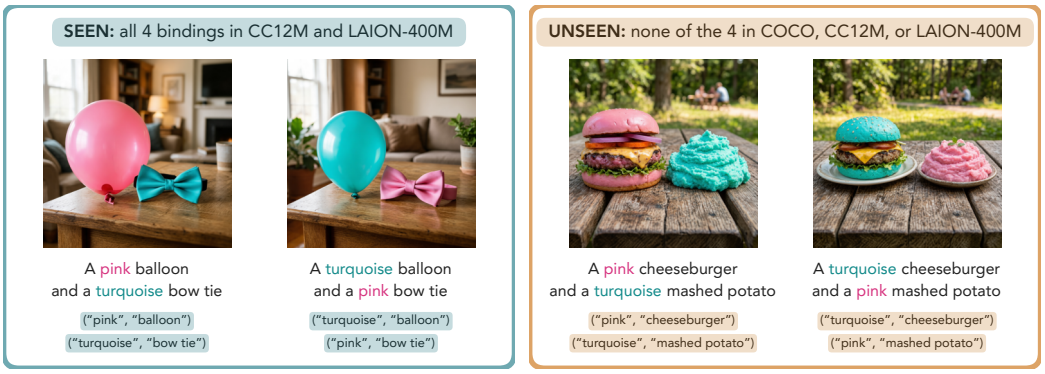

Figure 4: **A seen and an unseen sample of `BindSplit` Bench that swap the same two colors.** Each sample has two captions that swap the attributes of two objects, and one generated image per caption; below each caption are its two attribute–object bindings. In the seen sample (left), all four bindings appear in CC12M and LAION-400M. In the unseen sample (right), none of the four appear in COCO, CC12M, or LAION-400M. The captions shown omit the short scene at the end of each evaluated caption (e.g., *"…, on a wooden table in a living room"*). More shown in Figures 5–7.

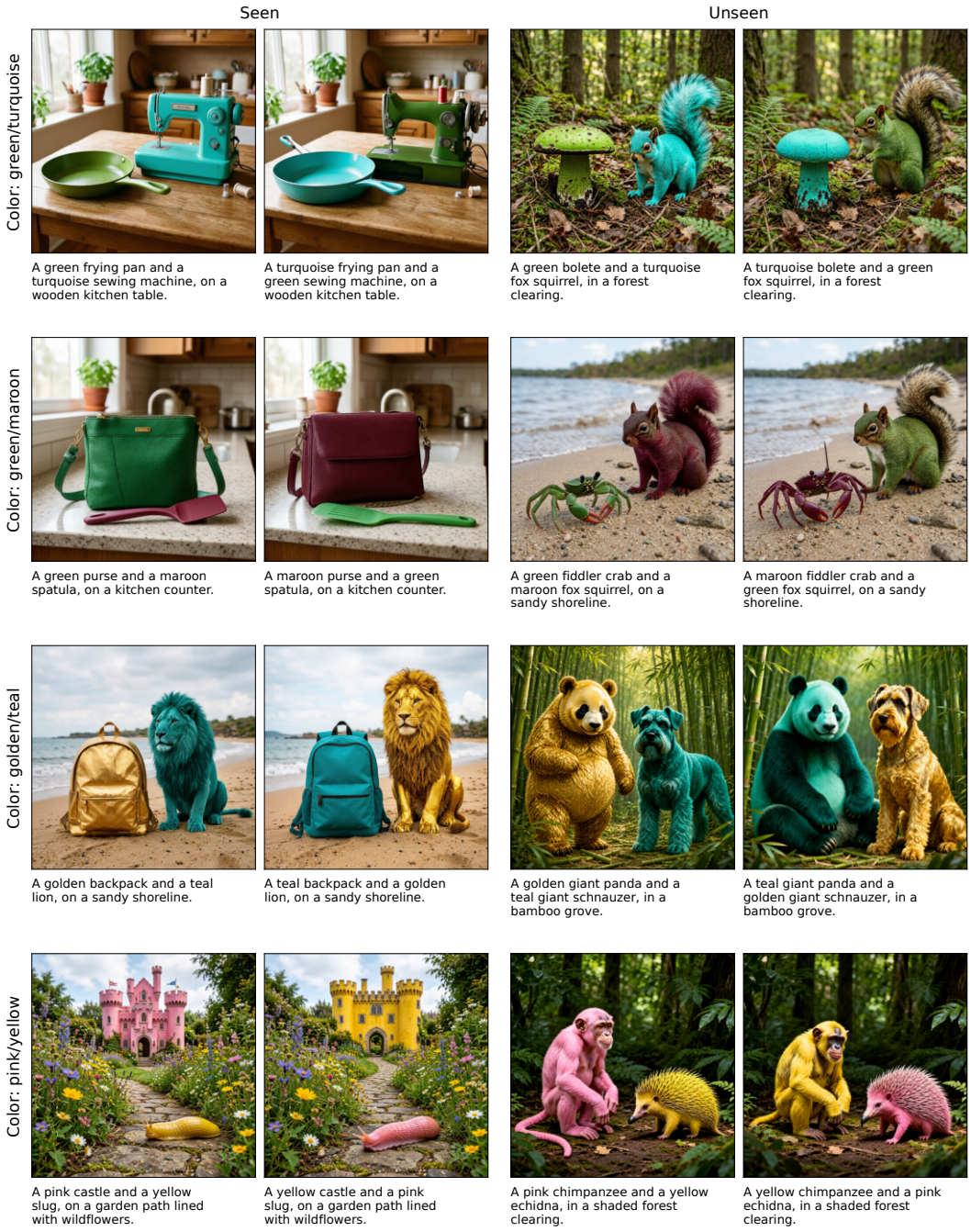

Figure 5: More **`BindSplit` Bench** samples (part 1 of 3). Each row shows one attribute swap on a seen sample (left) and an unseen sample (right); each sample has two images, each with the caption it is evaluated with.

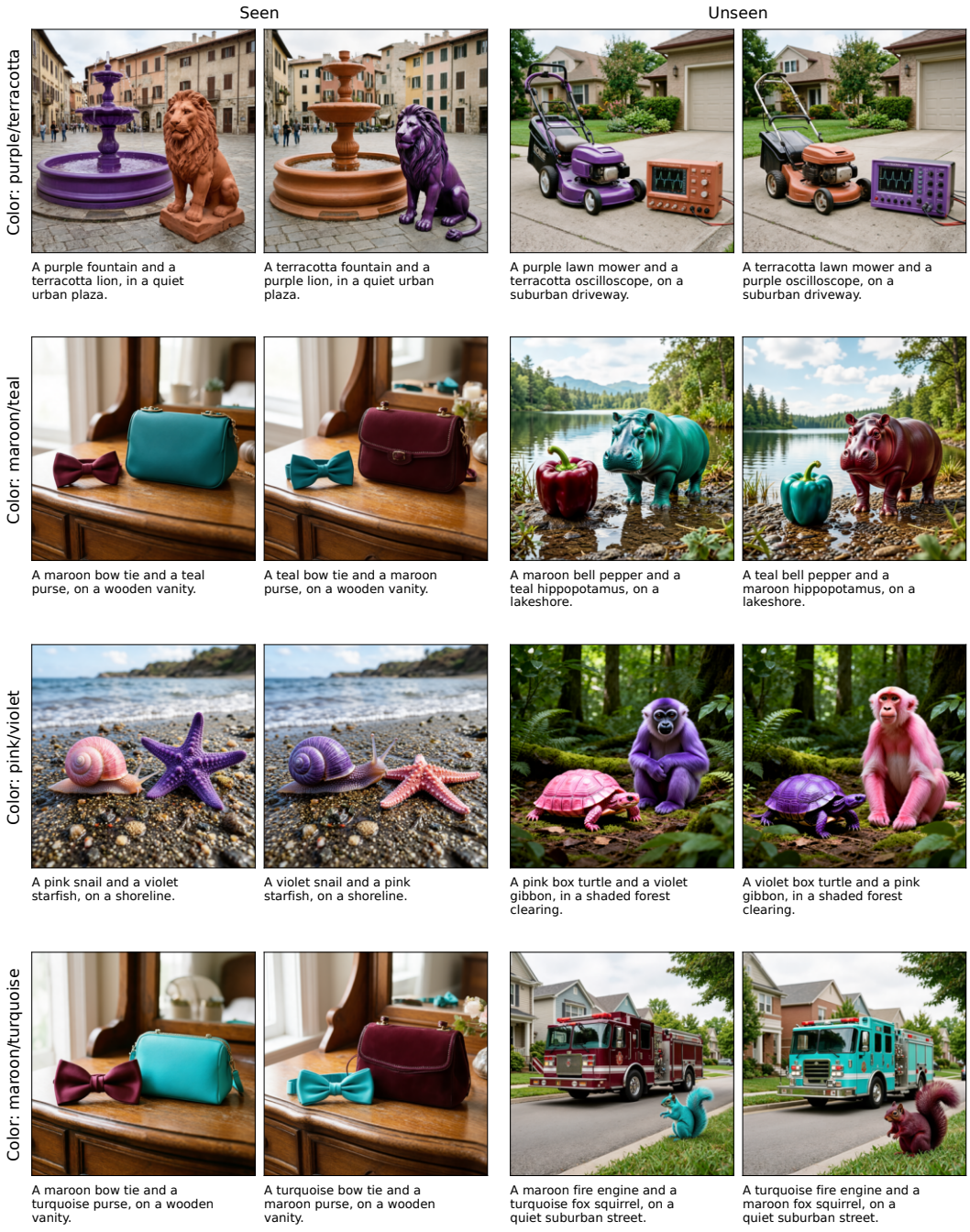

Figure 6: More **`BindSplit` Bench** samples (part 2 of 3). Each row shows one attribute swap on a seen sample (left) and an unseen sample (right); each sample has two images, each with the caption it is evaluated with.

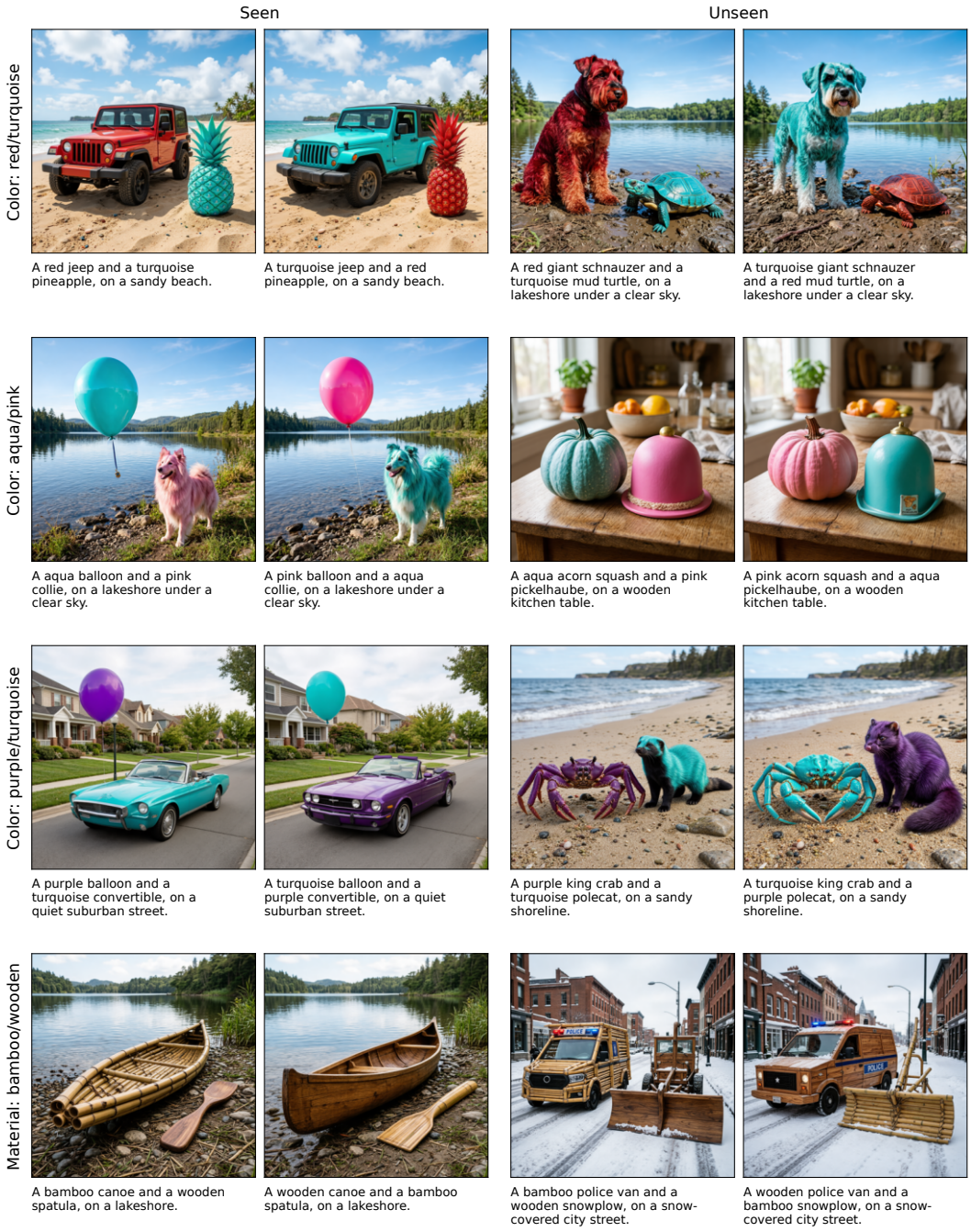

Figure 7: More **BindSplit Bench** samples (part 3 of 3). Each row shows one attribute swap on a seen sample (left) and an unseen sample (right); each sample has two images, each with the caption it is evaluated with.

Table 3: **`BindSplit` Bench results on the attribute-matched visually-similar split (T1).** Text, image, and group accuracy (%; defined in Section 4) on seen (675 samples) and unseen (644 samples) bindings for all 75 models, grouped by regime. NegCLIP, CLIC, and CS-CLIP fully fine-tune the backbone on COCO; LABCLIP trains one linear layer on frozen text embeddings. Trained models show the mean ± standard deviation over 3 training seeds; zero-shot models have a single value. Random chance is 25% for text and image accuracy and 16.7% for group accuracy. The COCO-trained SigLIP LABCLIP models are excluded because their heads did not converge (Section 5.1).

		Text		Image		Group	
Model	Pretraining	Seen	Unseen	Seen	Unseen	Seen	Unseen
Zero-shot							
CLIP B/32	CC12M	19.6	16.5	9.2	7.8	4.4	3.3
CLIP B/16	CC15M	21.0	14.0	8.3	5.0	4.1	2.5
CLIP B/32	OpenAI WIT	22.4	18.5	8.3	4.3	4.0	1.9
CLIP L/14	OpenAI WIT	17.2	17.7	9.9	9.6	3.7	4.0
EVA01 g/14	LAION-400M	29.0	23.6	11.4	8.1	7.4	4.7
EVA02 B/16	Merged-2B	24.3	16.3	8.0	6.1	4.4	2.3
EVA02 E/14	LAION-2B	34.4	27.0	13.3	11.3	8.7	7.1
EVA02 L/14	Merged-2B	22.2	19.3	9.2	9.3	4.4	4.3
LaCLIP B/32	CC12M	17.9	14.3	5.5	6.8	2.2	3.7
OpenCLIP B/16	LAION-2B	31.0	22.8	8.7	7.0	6.7	2.8
OpenCLIP B/32	LAION-2B	32.4	21.4	10.1	7.1	7.0	3.3
OpenCLIP L/14	LAION-2B	29.0	24.2	8.6	8.2	6.2	4.8
OpenCLIP B/16	LAION-400M	31.6	18.3	10.1	6.4	6.8	2.8
OpenCLIP B/32	LAION-400M	31.7	18.8	8.7	7.8	5.0	3.3
OpenCLIP L/14	LAION-400M	29.0	21.7	9.9	9.5	6.7	5.0
OpenCLIP RN101	YFCC-15M	9.6	10.9	7.6	5.3	2.4	1.7
OpenCLIP RN50	YFCC-15M	13.0	11.6	4.3	6.1	1.8	1.4
SigLIP B/16	WebLI	35.1	29.8	17.2	15.1	13.0	8.5
SigLIP L/16	WebLI	43.4	34.6	20.7	16.3	14.8	9.6
SPARO B/32	CC12M	28.3	20.7	13.8	8.4	7.0	4.2
NegCLIP							
CLIP B/16	CC15M	23.0 ± 0.1	15.1 ± 0.5	8.5 ± 0.2	8.0 ± 0.1	4.4 ± 0.3	3.6 ± 0.3
CLIP L/14	OpenAI WIT	20.3 ± 0.3	13.6 ± 0.3	9.0 ± 0.3	9.7 ± 0.4	4.6 ± 0.4	3.2 ± 0.2
EVA02 B/16	Merged-2B	25.3 ± 0.4	18.2 ± 0.3	10.4 ± 0.5	7.5 ± 0.6	6.2 ± 0.1	3.4 ± 0.3
EVA02 L/14	Merged-2B	21.0 ± 1.4	20.1 ± 0.5	11.6 ± 0.4	11.3 ± 0.2	4.7 ± 0.4	5.6 ± 0.0
OpenCLIP B/16	LAION-2B	26.8 ± 0.5	20.9 ± 0.4	11.8 ± 0.2	8.1 ± 0.2	7.4 ± 0.3	4.8 ± 0.4
OpenCLIP B/32	LAION-2B	29.0 ± 0.4	20.5 ± 0.6	11.5 ± 1.1	7.0 ± 0.8	7.6 ± 0.5	4.4 ± 0.4
OpenCLIP L/14	LAION-2B	28.4 ± 0.3	22.4 ± 0.3	9.5 ± 0.3	10.1 ± 0.4	6.2 ± 0.3	5.4 ± 0.2
OpenCLIP B/16	LAION-400M	28.9 ± 0.2	17.8 ± 0.2	10.8 ± 0.1	6.0 ± 0.4	6.6 ± 0.3	2.8 ± 0.4
OpenCLIP B/32	LAION-400M	30.8 ± 0.1	17.7 ± 0.1	10.0 ± 0.5	7.2 ± 0.8	7.2 ± 0.2	3.7 ± 0.2
OpenCLIP L/14	LAION-400M	28.0 ± 0.2	23.1 ± 0.3	11.1 ± 0.3	10.7 ± 0.3	7.6 ± 0.4	6.1 ± 0.2
OpenCLIP RN101	YFCC-15M	12.1 ± 0.1	10.5 ± 0.4	5.5 ± 0.7	6.4 ± 0.4	1.6 ± 0.2	1.7 ± 0.1
OpenCLIP RN50	YFCC-15M	15.1 ± 0.5	13.7 ± 0.2	4.3 ± 0.3	4.8 ± 0.2	1.9 ± 0.1	2.1 ± 0.1
SigLIP B/16	WebLI	32.4 ± 0.5	26.3 ± 0.3	16.8 ± 0.4	11.5 ± 0.5	11.2 ± 0.3	6.2 ± 0.2
SigLIP L/16	WebLI	38.3 ± 0.2	27.9 ± 0.3	22.2 ± 0.4	14.4 ± 0.2	14.8 ± 0.1	8.0 ± 0.4
LABCLIP							
CLIP B/16	CC15M	20.6 ± 0.1	12.6 ± 0.3	7.5 ± 0.8	6.5 ± 0.5	3.7 ± 0.2	2.4 ± 0.3
CLIP B/32	OpenAI WIT	17.2 ± 0.8	15.5 ± 0.6	7.9 ± 0.8	5.4 ± 0.2	3.3 ± 0.7	1.6 ± 0.4
CLIP L/14	OpenAI WIT	16.0 ± 0.5	13.5 ± 0.9	6.3 ± 0.6	6.4 ± 0.2	2.0 ± 0.2	1.8 ± 0.1
EVA01 g/14	LAION-400M	26.3 ± 0.5	20.8 ± 0.9	10.8 ± 0.6	9.9 ± 0.9	6.0 ± 0.3	4.9 ± 0.1
EVA02 B/16	Merged-2B	21.0 ± 0.3	13.7 ± 0.7	5.9 ± 1.2	5.4 ± 1.5	2.8 ± 0.2	1.4 ± 0.3
EVA02 E/14	LAION-2B	28.4 ± 1.3	22.3 ± 0.3	9.5 ± 0.3	11.3 ± 0.4	6.1 ± 0.1	5.5 ± 0.4
EVA02 L/14	Merged-2B	18.7 ± 0.2	16.2 ± 0.2	7.0 ± 0.3	6.0 ± 0.8	2.7 ± 0.2	2.4 ± 0.1
OpenCLIP B/16	LAION-2B	20.4 ± 0.6	15.2 ± 0.2	8.1 ± 0.6	5.2 ± 0.8	3.8 ± 0.5	1.8 ± 0.3
OpenCLIP B/32	LAION-2B	23.5 ± 0.4	18.2 ± 0.4	9.7 ± 0.9	6.7 ± 0.4	5.3 ± 0.8	3.4 ± 0.6
OpenCLIP L/14	LAION-2B	24.0 ± 0.6	21.3 ± 1.0	8.9 ± 0.5	7.9 ± 0.5	5.1 ± 0.2	3.9 ± 0.7
OpenCLIP B/16	LAION-400M	24.0 ± 0.7	13.8 ± 0.5	8.3 ± 0.8	5.3 ± 0.7	5.0 ± 0.2	1.7 ± 0.4
OpenCLIP B/32	LAION-400M	23.7 ± 0.6	15.8 ± 0.3	9.2 ± 1.0	6.0 ± 0.5	4.6 ± 0.1	2.2 ± 0.2
OpenCLIP L/14	LAION-400M	27.9 ± 0.3	18.0 ± 0.5	10.5 ± 0.1	8.5 ± 0.4	6.4 ± 0.5	3.1 ± 0.5
OpenCLIP RN101	YFCC-15M	11.3 ± 0.9	9.9 ± 0.3	7.0 ± 0.2	4.5 ± 0.5	2.3 ± 0.3	1.4 ± 0.3
OpenCLIP RN50	YFCC-15M	11.8 ± 0.7	14.5 ± 0.7	5.8 ± 0.5	6.3 ± 0.7	1.9 ± 0.3	2.0 ± 0.5
SPARO B/32	CC12M	20.7 ± 0.9	15.8 ± 0.4	10.5 ± 0.5	7.0 ± 0.1	5.4 ± 0.5	2.7 ± 0.1
CS-CLIP							
CLIP B/32	CC12M	20.0 ± 0.7	16.8 ± 0.7	9.9 ± 0.6	7.6 ± 0.3	3.9 ± 0.8	3.4 ± 0.3
CLIP B/16	CC15M	23.4 ± 1.3	15.5 ± 0.8	8.4 ± 0.4	6.7 ± 1.0	5.5 ± 0.1	3.3 ± 0.4
CLIP B/32	OpenAI WIT	22.1 ± 1.5	17.8 ± 0.3	9.0 ± 0.1	7.2 ± 0.9	4.1 ± 0.3	3.2 ± 0.4
CLIP L/14	OpenAI WIT	18.9 ± 0.6	14.7 ± 0.3	9.7 ± 0.7	7.5 ± 0.2	4.5 ± 0.7	2.7 ± 0.1
EVA02 B/16	Merged-2B	23.8 ± 1.1	16.0 ± 1.5	9.5 ± 0.4	6.7 ± 0.9	5.6 ± 0.4	2.8 ± 1.0
EVA02 L/14	Merged-2B	19.4 ± 0.8	17.6 ± 0.5	9.0 ± 0.7	9.6 ± 0.9	3.8 ± 0.5	3.6 ± 0.2
OpenCLIP B/16	LAION-2B	26.6 ± 0.9	19.3 ± 1.1	12.7 ± 0.2	7.6 ± 0.6	7.9 ± 0.5	3.7 ± 0.5
OpenCLIP B/32	LAION-2B	29.9 ± 0.8	21.7 ± 0.7	12.5 ± 1.0	7.5 ± 0.2	8.5 ± 0.9	3.5 ± 0.4
OpenCLIP L/14	LAION-2B	27.4 ± 0.8	22.5 ± 1.0	11.3 ± 1.4	10.0 ± 0.5	7.0 ± 0.4	5.5 ± 0.9
OpenCLIP B/16	LAION-400M	25.5 ± 0.5	16.7 ± 0.1	11.4 ± 0.4	6.1 ± 0.3	7.0 ± 0.4	3.0 ± 0.2
OpenCLIP B/32	LAION-400M	28.4 ± 0.8	16.1 ± 0.4	11.2 ± 0.2	6.8 ± 0.5	7.1 ± 0.3	3.5 ± 0.4
OpenCLIP L/14	LAION-400M	25.7 ± 0.4	22.6 ± 0.4	11.9 ± 0.5	9.7 ± 1.4	7.0 ± 0.6	5.1 ± 0.2
CLIC							
CLIP B/32	OpenAI WIT	19.3 ± 0.1	15.5 ± 0.7	8.0 ± 0.5	6.0 ± 0.6	3.5 ± 0.2	2.3 ± 0.4
CLIP L/14	OpenAI WIT	19.6 ± 0.9	16.1 ± 0.6	7.4 ± 0.2	7.9 ± 0.4	2.8 ± 0.3	2.6 ± 0.1
EVA02 B/16	Merged-2B	23.7 ± 0.1	20.5 ± 0.5	8.3 ± 0.5	5.2 ± 0.9	5.2 ± 0.1	2.7 ± 0.2
EVA02 L/14	Merged-2B	22.2 ± 0.5	19.2 ± 0.2	9.2 ± 0.3	9.5 ± 0.2	4.7 ± 0.3	3.1 ± 0.2
OpenCLIP B/16	LAION-2B	29.2 ± 0.1	19.2 ± 0.2	9.2 ± 0.2	6.0 ± 0.2	6.4 ± 0.3	2.2 ± 0.1
OpenCLIP B/32	LAION-2B	30.8 ± 0.3	18.5 ± 0.6	11.0 ± 0.4	5.7 ± 0.4	7.8 ± 0.2	3.4 ± 0.0
OpenCLIP L/14	LAION-2B	29.6 ± 0.4	21.7 ± 0.2	10.5 ± 0.1	7.9 ± 0.2	7.3 ± 0.0	3.4 ± 0.3
OpenCLIP B/16	LAION-400M	29.2 ± 0.3	20.2 ± 0.2	7.9 ± 0.4	6.7 ± 0.4	5.5 ± 0.2	3.5 ± 0.3
OpenCLIP B/32	LAION-400M	29.8 ± 0.2	18.0 ± 0.3	9.6 ± 0.5	5.2 ± 0.3	5.8 ± 0.1	2.8 ± 0.2
OpenCLIP L/14	LAION-400M	32.8 ± 0.2	23.1 ± 0.3	11.1 ± 0.1	7.0 ± 0.1	8.2 ± 0.3	3.6 ± 0.2
OpenCLIP RN101	YFCC-15M	11.5 ± 0.1	10.5 ± 0.1	6.4 ± 0.4	5.9 ± 0.2	1.9 ± 0.1	1.7 ± 0.1
OpenCLIP RN50	YFCC-15M	13.3 ± 0.1	11.5 ± 0.1	5.2 ± 0.3	5.2 ± 0.1	1.8 ± 0.1	1.9 ± 0.0

Table 4: **`BindSplit` Bench results on the attribute-matched T2 split**, which also keeps samples whose two images differ somewhat beyond the attribute swap. Text, image, and group accuracy (%) on seen (969 samples) and unseen (914 samples) bindings. Models, regimes, and notation as in Table 3.

		Text		Image		Group	
Model	Pretraining	Seen	Unseen	Seen	Unseen	Seen	Unseen
Zero-shot							
CLIP B/32	CC12M	20.8	16.6	9.5	6.6	4.4	3.0
CLIP B/16	CC15M	20.9	15.8	7.7	7.3	4.0	3.7
CLIP B/32	OpenAI WIT	22.6	18.6	7.6	4.5	4.0	1.9
CLIP L/14	OpenAI WIT	16.9	18.3	10.0	9.4	3.9	4.2
EVA01 g/14	LAION-400M	29.1	24.6	11.0	8.2	7.1	4.4
EVA02 B/16	Merged-2B	24.0	17.5	7.9	5.5	4.0	1.9
EVA02 E/14	LAION-2B	34.2	26.7	13.8	10.6	9.1	6.1
EVA02 L/14	Merged-2B	22.9	19.7	9.6	8.6	5.2	4.2
LaCLIP B/32	CC12M	18.4	14.1	5.7	5.9	3.0	3.2
OpenCLIP B/16	LAION-2B	29.4	22.2	9.2	7.1	6.1	3.0
OpenCLIP B/32	LAION-2B	31.8	22.9	9.7	6.8	6.6	3.4
OpenCLIP L/14	LAION-2B	27.8	23.9	9.1	7.4	6.7	4.2
OpenCLIP B/16	LAION-400M	29.7	18.7	10.1	6.3	6.5	2.6
OpenCLIP B/32	LAION-400M	29.4	19.3	8.7	7.9	5.3	3.5
OpenCLIP L/14	LAION-400M	28.4	21.7	9.9	9.3	6.8	4.6
OpenCLIP RN101	YFCC-15M	10.8	11.6	8.5	5.0	2.7	1.5
OpenCLIP RN50	YFCC-15M	12.7	12.6	5.0	7.2	1.9	2.2
SigLIP B/16	WebLI	36.2	30.4	17.9	14.6	13.1	8.5
SigLIP L/16	WebLI	42.7	37.9	20.6	16.2	14.6	10.4
SPARO B/32	CC12M	28.6	21.4	13.9	9.4	7.6	4.4
NegCLIP							
CLIP B/16	CC15M	22.0 ± 0.2	17.3 ± 0.5	8.1 ± 0.3	7.9 ± 0.1	4.3 ± 0.3	3.7 ± 0.1
CLIP L/14	OpenAI WIT	20.3 ± 0.3	15.0 ± 0.2	9.0 ± 0.4	10.0 ± 0.3	4.7 ± 0.4	3.3 ± 0.2
EVA02 B/16	Merged-2B	24.1 ± 0.2	18.3 ± 0.5	10.5 ± 0.3	7.3 ± 0.6	6.1 ± 0.2	3.2 ± 0.2
EVA02 L/14	Merged-2B	20.1 ± 1.0	18.7 ± 0.4	10.7 ± 0.4	10.4 ± 0.5	4.4 ± 0.2	4.8 ± 0.1
OpenCLIP B/16	LAION-2B	26.1 ± 0.5	20.8 ± 0.4	11.7 ± 0.3	7.8 ± 0.2	6.9 ± 0.2	4.4 ± 0.2
OpenCLIP B/32	LAION-2B	29.2 ± 0.2	22.4 ± 0.5	11.1 ± 0.6	7.0 ± 0.5	7.6 ± 0.3	4.5 ± 0.3
OpenCLIP L/14	LAION-2B	27.1 ± 0.2	22.7 ± 0.3	9.9 ± 0.1	9.7 ± 0.3	6.3 ± 0.2	5.1 ± 0.2
OpenCLIP B/16	LAION-400M	27.4 ± 0.2	18.7 ± 0.1	10.3 ± 0.2	6.2 ± 0.1	6.0 ± 0.2	3.1 ± 0.2
OpenCLIP B/32	LAION-400M	29.4 ± 0.1	17.0 ± 0.1	10.3 ± 0.6	7.5 ± 0.5	6.8 ± 0.3	3.6 ± 0.2
OpenCLIP L/14	LAION-400M	28.2 ± 0.1	24.5 ± 0.3	11.5 ± 0.4	10.3 ± 0.4	8.2 ± 0.2	6.1 ± 0.3
OpenCLIP RN101	YFCC-15M	12.8 ± 0.2	10.6 ± 0.3	6.3 ± 0.6	6.0 ± 0.2	2.0 ± 0.2	1.7 ± 0.1
OpenCLIP RN50	YFCC-15M	13.5 ± 0.3	13.7 ± 0.2	4.6 ± 0.2	5.2 ± 0.1	1.9 ± 0.2	2.3 ± 0.1
SigLIP B/16	WebLI	32.9 ± 0.3	26.1 ± 0.2	17.2 ± 0.4	11.4 ± 0.5	11.0 ± 0.4	6.1 ± 0.1
SigLIP L/16	WebLI	38.5 ± 0.2	29.8 ± 0.3	23.0 ± 0.4	15.8 ± 0.3	15.1 ± 0.1	9.4 ± 0.3
LABCLIP							
CLIP B/16	CC15M	20.5 ± 0.1	14.7 ± 0.5	7.4 ± 0.4	6.4 ± 0.4	3.6 ± 0.2	2.7 ± 0.2
CLIP B/32	OpenAI WIT	17.9 ± 0.7	15.1 ± 0.5	8.3 ± 0.3	5.3 ± 0.3	3.6 ± 0.5	1.5 ± 0.3
CLIP L/14	OpenAI WIT	15.9 ± 0.2	14.7 ± 0.5	6.7 ± 0.8	6.4 ± 0.2	2.4 ± 0.3	1.9 ± 0.2
EVA01 g/14	LAION-400M	26.8 ± 0.5	21.1 ± 1.0	11.1 ± 0.3	9.0 ± 0.6	6.5 ± 0.1	4.7 ± 0.2
EVA02 B/16	Merged-2B	20.9 ± 0.7	14.6 ± 0.6	5.9 ± 0.5	4.8 ± 0.8	2.9 ± 0.2	1.4 ± 0.2
EVA02 E/14	LAION-2B	27.9 ± 0.7	22.5 ± 0.2	9.9 ± 0.6	11.5 ± 0.4	6.6 ± 0.5	5.8 ± 0.4
EVA02 L/14	Merged-2B	18.8 ± 0.4	16.3 ± 0.4	8.0 ± 0.2	5.9 ± 0.8	3.1 ± 0.2	2.4 ± 0.3
OpenCLIP B/16	LAION-2B	21.8 ± 0.4	16.0 ± 0.3	8.2 ± 0.5	5.1 ± 0.5	4.0 ± 0.4	1.8 ± 0.4
OpenCLIP B/32	LAION-2B	24.5 ± 0.2	19.3 ± 0.4	10.4 ± 0.9	6.6 ± 0.2	6.2 ± 0.7	3.5 ± 0.6
OpenCLIP L/14	LAION-2B	24.3 ± 0.3	21.0 ± 0.9	9.7 ± 0.2	7.5 ± 0.4	5.7 ± 0.3	3.8 ± 0.6
OpenCLIP B/16	LAION-400M	23.1 ± 0.9	13.9 ± 0.4	7.9 ± 0.5	5.4 ± 0.4	4.6 ± 0.3	1.8 ± 0.3
OpenCLIP B/32	LAION-400M	23.2 ± 0.4	16.6 ± 0.2	9.0 ± 0.9	6.0 ± 0.5	4.4 ± 0.4	2.6 ± 0.3
OpenCLIP L/14	LAION-400M	28.5 ± 0.2	18.8 ± 0.3	9.9 ± 0.3	8.1 ± 0.2	6.1 ± 0.6	3.0 ± 0.5
OpenCLIP RN101	YFCC-15M	12.3 ± 0.7	10.6 ± 0.2	7.2 ± 0.5	5.1 ± 0.4	2.3 ± 0.4	1.8 ± 0.3
OpenCLIP RN50	YFCC-15M	11.9 ± 0.9	14.4 ± 0.4	6.0 ± 0.3	6.2 ± 0.5	2.2 ± 0.2	2.1 ± 0.2
SPARO B/32	CC12M	21.4 ± 0.7	16.0 ± 0.1	11.2 ± 0.6	6.6 ± 0.2	5.6 ± 0.5	2.6 ± 0.1
CS-CLIP							
CLIP B/32	CC12M	21.1 ± 0.4	17.0 ± 0.8	9.4 ± 0.3	7.8 ± 0.5	4.1 ± 0.4	3.4 ± 0.2
CLIP B/16	CC15M	22.8 ± 1.1	15.8 ± 0.7	8.7 ± 0.6	6.7 ± 0.9	5.4 ± 0.4	3.1 ± 0.4
CLIP B/32	OpenAI WIT	22.8 ± 1.1	18.5 ± 0.6	9.0 ± 0.2	6.5 ± 0.6	4.1 ± 0.4	2.9 ± 0.3
CLIP L/14	OpenAI WIT	18.4 ± 0.2	16.5 ± 0.2	9.6 ± 0.4	8.0 ± 0.4	4.6 ± 0.4	3.4 ± 0.1
EVA02 B/16	Merged-2B	23.7 ± 1.5	16.6 ± 0.9	9.8 ± 0.4	7.3 ± 0.5	5.4 ± 0.5	2.8 ± 0.3
EVA02 L/14	Merged-2B	19.4 ± 0.3	16.4 ± 0.5	9.8 ± 1.0	8.7 ± 0.5	4.5 ± 0.3	3.0 ± 0.2
OpenCLIP B/16	LAION-2B	24.3 ± 0.7	19.4 ± 1.0	11.7 ± 0.2	7.3 ± 0.7	6.8 ± 0.3	3.6 ± 0.3
OpenCLIP B/32	LAION-2B	28.7 ± 0.6	22.6 ± 0.7	11.7 ± 0.7	7.4 ± 0.2	8.1 ± 0.7	3.8 ± 0.3
OpenCLIP L/14	LAION-2B	26.3 ± 0.7	22.9 ± 0.2	10.5 ± 1.5	10.2 ± 0.6	6.4 ± 0.5	5.6 ± 0.6
OpenCLIP B/16	LAION-400M	24.5 ± 0.4	17.5 ± 0.4	10.4 ± 0.3	6.0 ± 0.5	6.5 ± 0.2	3.0 ± 0.1
OpenCLIP B/32	LAION-400M	27.6 ± 0.5	16.7 ± 0.2	11.2 ± 0.4	6.7 ± 0.8	6.9 ± 0.4	3.9 ± 0.1
OpenCLIP L/14	LAION-400M	27.1 ± 0.6	22.9 ± 0.3	11.3 ± 0.5	10.2 ± 0.7	7.3 ± 0.4	5.4 ± 0.2
CLIC							
CLIP B/32	OpenAI WIT	19.3 ± 0.2	16.4 ± 0.6	7.9 ± 0.3	5.2 ± 0.5	3.5 ± 0.1	1.9 ± 0.3
CLIP L/14	OpenAI WIT	19.8 ± 0.6	16.8 ± 0.6	7.7 ± 0.3	8.1 ± 0.2	3.2 ± 0.4	2.9 ± 0.2
EVA02 B/16	Merged-2B	23.6 ± 0.2	21.0 ± 0.5	8.8 ± 0.5	4.6 ± 0.6	5.0 ± 0.2	2.4 ± 0.2
EVA02 L/14	Merged-2B	21.8 ± 0.4	19.7 ± 0.2	9.5 ± 0.3	8.8 ± 0.1	4.4 ± 0.3	3.3 ± 0.1
OpenCLIP B/16	LAION-2B	28.0 ± 0.1	19.1 ± 0.3	9.3 ± 0.2	5.7 ± 0.3	6.4 ± 0.3	2.3 ± 0.2
OpenCLIP B/32	LAION-2B	29.7 ± 0.3	20.0 ± 0.3	10.1 ± 0.1	6.0 ± 0.3	7.1 ± 0.2	3.6 ± 0.1
OpenCLIP L/14	LAION-2B	27.7 ± 0.3	22.6 ± 0.1	11.0 ± 0.1	7.3 ± 0.1	7.0 ± 0.1	3.3 ± 0.2
OpenCLIP B/16	LAION-400M	27.0 ± 0.3	20.5 ± 0.1	7.5 ± 0.3	5.9 ± 0.2	5.1 ± 0.1	3.3 ± 0.2
OpenCLIP B/32	LAION-400M	29.2 ± 0.2	18.9 ± 0.2	9.1 ± 0.4	5.6 ± 0.3	5.5 ± 0.1	3.3 ± 0.1
OpenCLIP L/14	LAION-400M	31.9 ± 0.4	23.6 ± 0.2	11.1 ± 0.1	6.9 ± 0.2	7.8 ± 0.0	3.7 ± 0.1
OpenCLIP RN101	YFCC-15M	12.8 ± 0.1	10.6 ± 0.1	7.0 ± 0.2	5.7 ± 0.1	2.1 ± 0.2	1.7 ± 0.1
OpenCLIP RN50	YFCC-15M	11.7 ± 0.2	12.3 ± 0.1	4.9 ± 0.3	5.3 ± 0.1	1.4 ± 0.1	2.3 ± 0.0
SigLIP B/16	WebLI	38.2 ± 0.4	27.9 ± 0.2	10.9 ± 0.2	9.0 ± 0.2	7.7 ± 0.2	4.9 ± 0.1

Table 5: **Seen-to-unseen drops** on the attribute-matched visually-similar split (T1) of **BindSplit Bench**, by method and by whether the seen/unseen labels could be checked against the backbone's own training data (Checked: CC12M, LAION-400M, 7 backbones; Partly checked: YFCC-15M, CC15M, 3 backbones; Not checkable: OpenAI WIT, LAION-2B, WebLI, Merged-2B, 10 backbones). N counts models (backbone $\times$ method; trained models averaged over 3 seeds). "Dropping" counts models that score lower on unseen than on seen text accuracy; mean Δ is unseen minus seen accuracy in percentage points (pp), averaged over models. A negative Δ is a drop.

		Text		Image	Group
	N	Dropping	Mean Δ (pp)	Mean Δ (pp)	Mean Δ (pp)
By method					
Zero-shot	20	18/20	-6.01	-1.87	-1.98
LABCLIP	16	15/16	-4.90	-1.54	-1.52
NegCLIP	14	14/14	-6.55	-2.16	-2.22
CLIC	13	13/13	-6.71	-2.08	-2.29
CS-CLIP	12	12/12	-6.14	-2.78	-2.39
By training-data check					
Checked	22	22/22	-8.13	-2.64	-2.45
Partly checked	12	10/12	-2.95	-0.70	-0.64
Not checkable	41	40/41	-5.78	-2.10	-2.24
All	75	72/75	-6.02	-2.04	-2.05

Table 6: **Seen-to-unseen drops in text, image, and group accuracy.** For each of the 75 models in Figure 8, Δ is unseen minus seen accuracy on the attribute-matched visually similar (T1) split of **BindSplit Bench**, in percentage points (pp); trained models are averaged over 3 seeds. "Models dropping" counts models with $\Delta < 0$, and Range gives the smallest and largest Δ. Most models drop on all three metrics, but the mean drop in text accuracy (-6.02 pp) is about three times the mean drop in image and group accuracy (-2.04 and -2.05 pp).

Metric	Models dropping	Mean Δ (pp)	Median Δ (pp)	Range (pp)
Text	72/75 (96%)	-6.02	-6.03	$[-13.2, +2.8]$
Image	62/75 (83%)	-2.04	-2.00	$[-7.8, +1.8]$
Group	68/75 (91%)	-2.05	-1.80	$[-6.7, +1.5]$

Table 7: **Seen-to-unseen drops by regime** on the attribute-matched visually-similar split (T1) of **BindSplit Bench**. For each regime, "Dropping" counts the models that score lower on unseen than on seen bindings, and mean Δ is unseen minus seen accuracy in percentage points (pp), averaged over models; trained models are first averaged over 3 seeds. The All row matches Table 6.

		Text		Image		Group	
Regime	N	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)
Zero-shot	20	18/20	-6.01	17/20	-1.87	18/20	-1.98
LABCLIP	16	15/16	-4.90	13/16	-1.54	15/16	-1.52
Compositional fine-tuning							
NegCLIP	14	14/14	-6.55	10/14	-2.16	11/14	-2.22
CLIC	13	13/13	-6.71	11/13	-2.08	12/13	-2.29
CS-CLIP	12	12/12	-6.14	11/12	-2.78	12/12	-2.39
All	75	72/75	-6.02	62/75	-2.04	68/75	-2.05

Table 8: **Seen-to-unseen drops by regime** on the attribute-matched T2 split of **`BindSplit` Bench**. For each regime, "Dropping" counts the models that score lower on unseen than on seen bindings, and mean Δ is unseen minus seen accuracy in percentage points (pp), averaged over models; trained models are first averaged over 3 seeds.

		Text		Image		Group	
Regime	N	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)
Zero-shot	20	18/20	-5.16	18/20	-2.08	17/20	-2.10
LABCLIP	16	15/16	-4.61	14/16	-1.94	16/16	-1.65
Compositional fine-tuning							
NegCLIP	14	13/14	-5.43	12/14	-2.27	12/14	-2.13
CLIC	13	12/13	-5.46	11/13	-2.36	12/13	-2.10
CS-CLIP	12	12/12	-5.31	12/12	-2.51	12/12	-2.18
All	75	70/75	-5.17	67/75	-2.20	69/75	-2.02

Table 9: **Seen-to-unseen drops by whether the labels were checked against the model's pretraining corpus**, on the attribute-matched visually-similar split (T1) of **`BindSplit` Bench**. Checked: backbones pretrained on CC12M or LAION-400M; partly checked: YFCC-15M and CC15M; not checkable: OpenAI WIT, LAION-2B, WebLI, and Merged-2B (Section 5.1). N counts models across all regimes. "Dropping" and mean Δ are as in Table 7.

		Text		Image		Group	
Labels checked	N	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)
Checked (7 backbones)	22	22/22	-8.13	21/22	-2.64	21/22	-2.45
Partly checked (3 backbones)	12	10/12	-2.95	8/12	-0.70	8/12	-0.64
Not checkable (10 backbones)	41	40/41	-5.78	33/41	-2.10	39/41	-2.24
All (20 backbones)	75	72/75	-6.02	62/75	-2.04	68/75	-2.05

Table 10: **Seen-to-unseen drops by whether the labels were checked against the model's pretraining corpus**, on the attribute-matched T2 split of **`BindSplit` Bench**. Groups and columns as in Table 9.

		Text		Image		Group	
Labels checked	N	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)	Dropping	Mean Δ (pp)
Checked (7 backbones)	22	22/22	-7.29	21/22	-2.65	21/22	-2.42
Partly checked (3 backbones)	12	8/12	-2.07	8/12	-0.60	9/12	-0.42
Not checkable (10 backbones)	41	40/41	-4.94	38/41	-2.43	39/41	-2.28
All (20 backbones)	75	70/75	-5.17	67/75	-2.20	69/75	-2.02

A.3 FULL RESULTS ON BINDSPLIT BENCH

A.4 SEEN-TO-UNSEEN DROPS: DETAILS

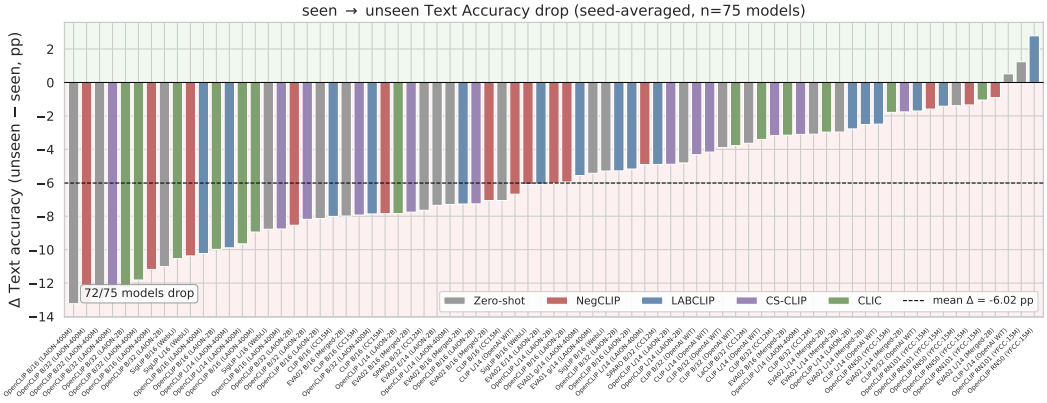

Figure 8: **VLM text accuracy drops from seen to unseen bindings.** Each bar is one of 75 models: zero-shot, LABCLIP (a linear map trained on frozen CLIP text embeddings), and compositionally fine-tuned (NegCLIP, CS-CLIP, CLIC); trained models are averaged over 3 seeds. Bars show unseen minus seen text accuracy on the attribute-matched visually-similar split (T1) of **BindSplit Bench**. 72 of the 75 models score lower on the unseen split (bars below zero), with a mean drop of 6.02 percentage points (dashed line). More details in Table 6.

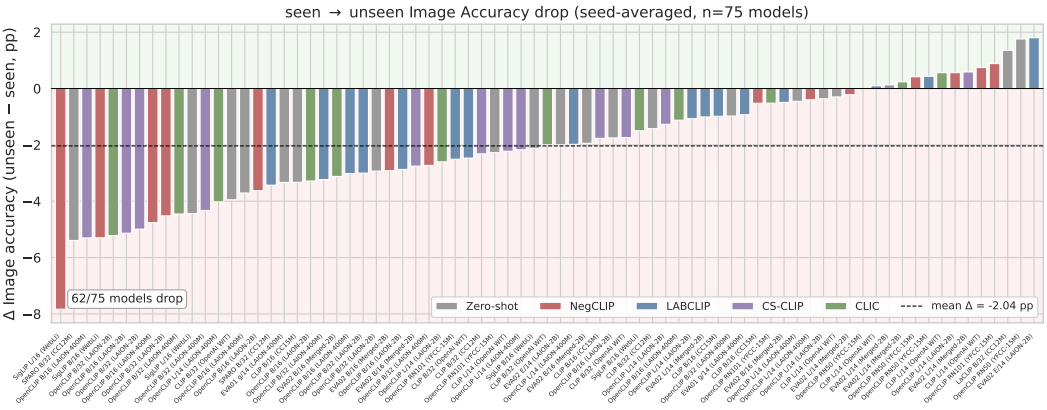

Figure 9: **VLM image accuracy drops from seen to unseen bindings.** Each bar is one of 75 models: zero-shot, LABCLIP (a linear map trained on frozen CLIP text embeddings), and compositionally fine-tuned (NegCLIP, CS-CLIP, CLIC); trained models are averaged over 3 seeds. Bars show unseen minus seen image accuracy on the attribute-matched visually-similar split (T1) of **BindSplit Bench**. 62 of the 75 models score lower on the unseen split (bars below zero), with a mean drop of 2.04 percentage points (dashed line). More details in Table 6.

Table 11: **Least-squares fit of unseen on seen accuracy** on the attribute-matched visually-similar split (T1) of **`BindSplit` Bench**. Slope with a 95% confidence interval, and R^2, for all 75 models and for the 22 models on the 7 backbones whose pretraining corpora we checked (Section 5.1); with 7 backbones, the latter intervals are approximate. Intervals come from 10,000 bootstrap resamples of pretrained backbones (each keeping all models built on the same pretrained backbone; accuracies averaged over 3 seeds).

		Text		Image		Group	
Models	N	Slope [95% CI]	R^2	Slope [95% CI]	R^2	Slope [95% CI]	R^2
All (20 backbones)	75	0.60 [0.45, 0.70]	0.75	0.59 [0.38, 0.68]	0.58	0.51 [0.36, 0.58]	0.67
Checked (7 backbones)	22	0.40 [0.23, 0.64]	0.33	0.33 [0.05, 0.65]	0.14	0.28 [−0.02, 0.64]	0.14

Table 12: **Least-squares fit of unseen on seen accuracy** on the attribute-matched T2 split of **`BindSplit` Bench**. Rows, columns, and intervals as in Table 11.

		Text		Image		Group	
Models	N	Slope [95% CI]	R^2	Slope [95% CI]	R^2	Slope [95% CI]	R^2
All (20 backbones)	75	0.65 [0.50, 0.78]	0.80	0.62 [0.42, 0.72]	0.66	0.56 [0.40, 0.65]	0.76
Checked (7 backbones)	22	0.60 [0.33, 0.84]	0.48	0.54 [0.28, 0.84]	0.35	0.49 [0.16, 0.81]	0.43

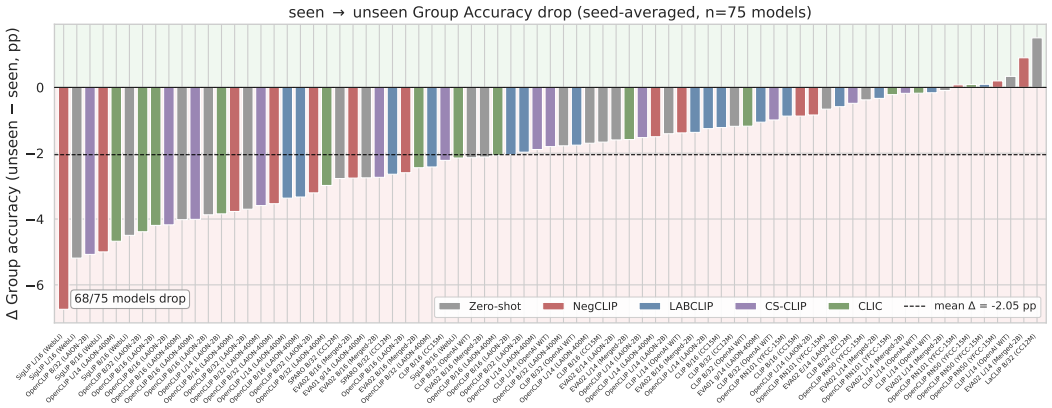

Figure 10: **VLM group accuracy drops from seen to unseen bindings.** Each bar is one of 75 models: zero-shot, LABCLIP (a linear map trained on frozen CLIP text embeddings), and compositionally fine-tuned (NegCLIP, CS-CLIP, CLIC); trained models are averaged over 3 seeds. Bars show unseen minus seen group accuracy on the attribute-matched visually-similar split (T1) of **`BindSplit` Bench**. 68 of the 75 models score lower on the unseen split (bars below zero), with a mean drop of 2.05 percentage points (dashed line). More details in Table 6.

A.5 Seen versus unseen accuracy: fitted lines

Tables 11 and 12 give the least-squares fit of unseen on seen accuracy for T1 and T2 splits of **`BindSplit` Bench**. To estimate uncertainty, we resample backbones rather than models, because models (zero-shot, LABCLIP, compositional fine-tuning) built on the same (pretrained) backbone are not independent: each of 10,000 resamples draws backbones with replacement and keeps every model built on them, with each model's accuracy averaged over its 3 seeds (the intervals therefore do not include seed-to-seed variation). Because the 22 checked models come from only 7 backbones, their intervals are approximate.

Table 13: **Image and group accuracy rise when each caption names one binding.** Accuracy (%) with two-binding captions ("two", used throughout the paper) and one-binding captions ("one": each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = one-binding − two-binding in percentage points (pp), on the attribute-matched T1 split of **`BindSplit` Bench** (seen and unseen pooled, 1,319 samples), for 7 zero-shot models.

		Text			Image			Group		
Model	Pretraining	two	one	Δ	two	one	Δ	two	one	Δ
CLIP B/32	CC12M	18.0	19.9	+1.8	8.5	21.8	+13.3	3.9	10.9	+7.1
CLIP B/32	OpenAI WIT	20.5	21.8	+1.3	6.4	21.2	+14.8	3.0	12.7	+9.8
CLIP L/14	OpenAI WIT	17.4	25.9	+8.5	9.8	28.6	+18.8	3.9	18.4	+14.6
OpenCLIP B/32	LAION-2B	27.1	29.7	+2.7	8.6	31.0	+22.4	5.2	22.2	+17.1
OpenCLIP L/14	LAION-2B	26.7	30.2	+3.5	8.4	34.6	+26.2	5.5	22.1	+16.6
SigLIP B/16	WebLI	32.5	34.0	+1.4	16.1	37.1	+21.0	10.8	25.4	+14.6
SPARO B/32	CC12M	24.6	20.0	−4.5	11.1	18.4	+7.3	5.6	9.9	+4.3
Mean (7 models)		23.8	25.9	+2.1	9.9	27.5	+17.7	5.4	17.4	+12.0

Table 14: **The seen-to-unseen drop persists with one-binding captions.** Δ is unseen minus seen accuracy (percentage points) on the attribute-matched T1 split of **`BindSplit` Bench** for 7 zero-shot models. *Two-binding*: the captions used throughout the paper (e.g., "A pink sewing machine and a turquoise wig, on a wooden workbench …"). *One-binding*: each caption shortened to its first binding, with the scene kept ("A pink sewing machine, on a wooden workbench …"); images unchanged. "Models dropping" counts models with Δ < 0. A negative Δ is a drop. Two-binding values are those in Table 3; one-binding values come from a separate evaluation run.

		Text		Image		Group	
Model	Pretraining	Two-binding	One-binding	Two-binding	One-binding	Two-binding	One-binding
CLIP B/32	CC12M	−3.1	−3.9	−1.4	−2.5	−1.2	−4.0
CLIP B/32	OpenAI WIT	−3.9	−4.6	−3.9	−3.1	−2.1	−4.6
CLIP L/14	OpenAI WIT	+0.5	−7.6	−0.3	−1.2	+0.3	−6.6
OpenCLIP B/32	LAION-2B	−11.0	−11.3	−2.9	−12.7	−3.7	−13.4
OpenCLIP L/14	LAION-2B	−4.8	−5.0	−0.4	+0.3	−1.4	−4.4
SigLIP B/16	WebLI	−5.3	−6.0	−2.1	−3.7	−4.5	−6.5
SPARO B/32	CC12M	−7.6	−4.5	−5.4	−4.4	−2.8	−4.2
Models dropping		6/7	7/7	7/7	6/7	6/7	7/7
Mean Δ		−5.0	−6.1	−2.4	−3.9	−2.2	−6.2

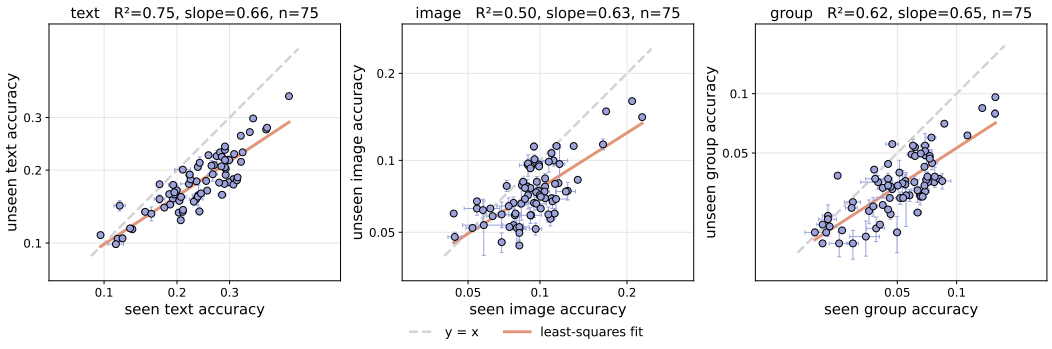

Figure 11: **`BindSplit` Benchseen versus unseen accuracy on probit-scaled axes.** Same models and data as Figure 3, with both axes probit-transformed (tick labels show raw accuracy), as in accuracy-on-the-line analyses (Miller et al., 2021). The line is a least-squares fit in probit space; panel titles give its R^2 and slope.

Table 15: **The seen-to-unseen drop persists with one-binding captions (companion to Table 14).** Δ is unseen minus seen accuracy (percentage points) on the attribute-matched T2 split of **`BindSplit` Bench** (969 seen / 914 unseen samples) for 7 zero-shot models. *Two-binding*: the captions used throughout the paper (e.g., "A pink sewing machine and a turquoise wig, on a wooden workbench ..."). *One-binding*: each caption shortened to its first binding, with the scene kept ("A pink sewing machine, on a wooden workbench ..."); images unchanged. "Models dropping" counts models with $\Delta < 0$. Two-binding values are those in Table 3; one-binding values come from a separate evaluation run.

		Text		Image		Group	
Model	Pretraining	Two-binding	One-binding	Two-binding	One-binding	Two-binding	One-binding
CLIP B/32	CC12M	−4.2	−4.0	−2.9	−4.1	−1.5	−3.9
CLIP B/32	OpenAI WIT	−4.0	−2.7	−3.2	−2.4	−2.2	−4.2
CLIP L/14	OpenAI WIT	+1.3	−5.1	−0.6	−0.4	+0.2	−4.5
OpenCLIP B/32	LAION-2B	−8.9	−9.3	−2.9	−10.8	−3.2	−11.9
OpenCLIP L/14	LAION-2B	−3.9	−5.1	−1.6	−1.6	−2.6	−5.1
SigLIP B/16	WebLI	−5.8	−4.7	−3.3	−2.6	−4.6	−5.1
SPARO B/32	CC12M	−7.1	−2.4	−4.5	−4.2	−3.3	−3.3
Models dropping		6/7	7/7	7/7	7/7	6/7	7/7
Mean Δ		−4.7	−4.8	−2.7	−3.7	−2.4	−5.4

Table 16: **Two- vs. one-binding captions, seen split (T1).** Accuracy (%) with two-binding captions (S2, used throughout the paper) and one-binding captions (B1: each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = B1 − S2 in percentage points (pp), on the seen split of attribute-matched T1 (675 samples), for 7 zero-shot models. Values are rounded independently.

		Text			Image			Group		
Model	Pretraining	S2	B1	Δ	S2	B1	Δ	S2	B1	Δ
CLIP B/32	CC12M	19.6	21.8	+2.2	9.2	23.0	+13.8	4.4	12.9	+8.4
CLIP B/32	OpenAI WIT	22.4	24.0	+1.6	8.3	22.7	+14.4	4.0	15.0	+11.0
CLIP L/14	OpenAI WIT	17.2	29.6	+12.4	9.9	29.2	+19.3	3.7	21.6	+17.9
OpenCLIP B/32	LAION-2B	32.4	35.3	+2.8	10.1	37.2	+27.1	7.0	28.7	+21.8
OpenCLIP L/14	LAION-2B	29.0	32.6	+3.6	8.6	34.5	+25.9	6.2	24.3	+18.1
SigLIP B/16	WebLI	35.1	36.9	+1.8	17.2	39.0	+21.8	13.0	28.6	+15.6
SPARO B/32	CC12M	28.3	22.2	−6.1	13.8	20.6	+6.8	7.0	12.0	+5.0
Mean (7 models)		26.3	28.9	+2.6	11.0	29.4	+18.4	6.5	20.4	+14.0

Table 17: **Two- vs. one-binding captions, unseen split (T1).** Accuracy (%) with two-binding captions (S2, used throughout the paper) and one-binding captions (B1: each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = B1 − S2 in percentage points (pp), on the unseen split of attribute-matched T1 (644 samples), for 7 zero-shot models. Values are rounded independently.

		Text			Image			Group		
Model	Pretraining	S2	B1	Δ	S2	B1	Δ	S2	B1	Δ
CLIP B/32	CC12M	16.5	17.9	+1.4	7.8	20.5	+12.7	3.3	8.9	+5.6
CLIP B/32	OpenAI WIT	18.5	19.4	+0.9	4.3	19.6	+15.2	1.9	10.4	+8.5
CLIP L/14	OpenAI WIT	17.7	22.0	+4.3	9.6	28.0	+18.3	4.0	15.1	+11.0
OpenCLIP B/32	LAION-2B	21.4	23.9	+2.5	7.1	24.5	+17.4	3.3	15.4	+12.1
OpenCLIP L/14	LAION-2B	24.2	27.6	+3.4	8.2	34.8	+26.6	4.8	19.9	+15.1
SigLIP B/16	WebLI	29.8	30.9	+1.1	15.1	35.2	+20.2	8.5	22.0	+13.5
SPARO B/32	CC12M	20.7	17.7	−3.0	8.4	16.1	+7.8	4.2	7.8	+3.6
Mean (7 models)		21.3	22.8	+1.5	8.7	25.5	+16.9	4.3	14.2	+9.9

Table 18: **Two- vs. one-binding captions, T2 (pooled).** Accuracy (%) with two-binding captions (S2, used throughout the paper) and one-binding captions (B1: each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = B1 − S2 in percentage points (pp), on the attribute-matched T2 split (seen and unseen pooled), for 7 zero-shot models. Values are rounded independently.

		Text			Image			Group		
Model	Pretraining	S2	B1	Δ	S2	B1	Δ	S2	B1	Δ
CLIP B/32	CC12M	18.8	20.8	+2.0	8.1	21.3	+13.3	3.7	11.2	+7.5
CLIP B/32	OpenAI WIT	20.7	22.3	+1.6	6.1	21.6	+15.5	3.0	12.8	+9.8
CLIP L/14	OpenAI WIT	17.6	25.6	+8.0	9.7	28.1	+18.4	4.0	18.1	+14.0
OpenCLIP B/32	LAION-2B	27.5	29.3	+1.9	8.3	29.5	+21.2	5.0	20.8	+15.7
OpenCLIP L/14	LAION-2B	25.9	29.3	+3.5	8.3	33.8	+25.5	5.5	21.9	+16.4
SigLIP B/16	WebLI	33.4	33.9	+0.5	16.3	37.4	+21.2	10.9	25.4	+14.5
SPARO B/32	CC12M	25.1	20.7	−4.4	11.7	19.0	+7.3	6.1	10.1	+4.1
Mean (7 models)		24.1	26.0	+1.9	9.8	27.3	+17.5	5.5	17.2	+11.7

Table 19: **Two- vs. one-binding captions, seen split (T2).** Accuracy (%) with two-binding captions (S2, used throughout the paper) and one-binding captions (B1: each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = B1 − S2 in percentage points (pp), on the seen split of attribute-matched T2, for 7 zero-shot models. Values are rounded independently.

		Text			Image			Group		
Model	Pretraining	S2	B1	Δ	S2	B1	Δ	S2	B1	Δ
CLIP B/32	CC12M	20.8	22.7	+1.9	9.5	23.3	+13.8	4.4	13.1	+8.7
CLIP B/32	OpenAI WIT	22.6	23.6	+1.0	7.6	22.7	+15.1	4.0	14.9	+10.8
CLIP L/14	OpenAI WIT	16.9	28.1	+11.1	10.0	28.3	+18.3	3.9	20.2	+16.3
OpenCLIP B/32	LAION-2B	31.8	33.8	+2.1	9.7	34.8	+25.1	6.6	26.5	+19.9
OpenCLIP L/14	LAION-2B	27.8	31.8	+4.0	9.1	34.6	+25.5	6.7	24.4	+17.6
SigLIP B/16	WebLI	36.2	36.2	+0.0	17.9	38.7	+20.8	13.1	27.9	+14.8
SPARO B/32	CC12M	28.6	21.9	−6.7	13.9	21.1	+7.1	7.6	11.8	+4.1
Mean (7 models)		26.4	28.3	+1.9	11.1	29.1	+18.0	6.6	19.8	+13.2

Table 20: **Two- vs. one-binding captions, unseen split (T2).** Accuracy (%) with two-binding captions (S2, used throughout the paper) and one-binding captions (B1: each caption shortened to its first binding, scene kept; images unchanged), and the change Δ = B1 − S2 in percentage points (pp), on the unseen split of attribute-matched T2, for 7 zero-shot models. Values are rounded independently.

		Text			Image			Group		
Model	Pretraining	S2	B1	Δ	S2	B1	Δ	S2	B1	Δ
CLIP B/32	CC12M	16.6	18.7	+2.1	6.6	19.3	+12.7	3.0	9.2	+6.2
CLIP B/32	OpenAI WIT	18.6	20.9	+2.3	4.5	20.4	+15.9	1.9	10.6	+8.8
CLIP L/14	OpenAI WIT	18.3	23.0	+4.7	9.4	27.9	+18.5	4.2	15.8	+11.6
OpenCLIP B/32	LAION-2B	22.9	24.5	+1.6	6.8	24.0	+17.2	3.4	14.7	+11.3
OpenCLIP L/14	LAION-2B	23.9	26.7	+2.8	7.4	32.9	+25.5	4.2	19.3	+15.1
SigLIP B/16	WebLI	30.4	31.5	+1.1	14.6	36.1	+21.6	8.5	22.8	+14.2
SPARO B/32	CC12M	21.4	19.5	−2.0	9.4	16.8	+7.4	4.4	8.4	+4.0
Mean (7 models)		21.7	23.5	+1.8	8.4	25.3	+17.0	4.2	14.4	+10.2

A.6 CAPTION FORMAT: ONE-BINDING CAPTIONS

A.7 COMPARING VISION-LANGUAGE COMPOSITIONAL BENCHMARKS

A.8 BINDSPLIT AUDIT: FULL COUNTS

Table 21: **How many benchmark samples test an unseen attribute–object binding?** Each row applies BindSplit to one benchmark with one reference corpus, COCO (train 2014) or CC12M, each used alone. Read a row from left to right as a funnel. **Samples:** all samples of the benchmark (ARO VG-A: one image with a true and a swapped caption; BiVLC, VisMin and ColorSwap: two images, each with its own caption). **Attribute-change samples:** samples whose two captions differ in an attribute, according to the benchmark's own sample types (BiVLC: replace-, swap- and add-attribute; VisMin: the attribute category); every ARO VG-A and ColorSwap sample is one. **Labeled:** attribute-change samples from which BindSplit extracts at least one attribute–object binding (ARO VG-A: all four bindings have one-word attributes and objects; BiVLC, VisMin and ColorSwap: at least one binding formed by the words that differ between the two captions). **Seen / Mixed / Unseen:** labeled samples whose bindings are all seen, partly seen, or all unseen in the reference corpus, as a count and a percentage of *Labeled*. A binding is seen if its attribute directly precedes its object in some noun phrase of the corpus, with singular and plural objects counted as the same. **Not labeled:** attribute-change samples without an extractable binding (ARO VG-A: a multi-word attribute or object; BiVLC, VisMin and ColorSwap: the change lies inside a phrase with a preposition or conjunction, or replaces a whole word); they are not included in the percentages. Example: of BiVLC's 2,933 samples, 791 are attribute-change samples, 536 of them are labeled, and 97 of those (18.10%) are unseen against COCO. For BiVLC, VisMin and ColorSwap, Unseen is an upper estimate, because an object can span several words that must all appear together in the corpus.

Benchmark	Corpus	Samples	Attribute-change samples	Labeled	Seen	Mixed	Unseen	Not labeled
ARO VG-A	COCO	28,748	28,748	26,508	6,246 (23.56%)	19,840 (74.85%)	422 (1.59%)	2,240
ARO VG-A	CC12M	28,748	28,748	26,508	15,400 (58.10%)	11,069 (41.76%)	39 (0.15%)	2,240
BiVLC	COCO	2,933	791	536	294 (54.85%)	145 (27.05%)	97 (18.10%)	255
BiVLC	CC12M	2,933	791	536	376 (70.15%)	87 (16.23%)	73 (13.62%)	255
VisMin	COCO	2,084	294	159	126 (79.25%)	16 (10.06%)	17 (10.69%)	135
VisMin	CC12M	2,084	294	159	131 (82.39%)	18 (11.32%)	10 (6.29%)	135
ColorSwap	COCO	1,000	1,000	923	327 (35.43%)	341 (36.94%)	255 (27.63%)	77
ColorSwap	CC12M	1,000	1,000	923	600 (65.01%)	235 (25.46%)	88 (9.53%)	77

A.9 BINDSPLIT: SENSITIVITY TO THE MATCHING RULE

Table 22: ARO VG-A samples that are seen, mixed or unseen (count, % of the 26,508 labeled samples) against COCO (train 2014) and CC12M, each used alone, under three rules for when a binding counts as seen in the reference corpus. **Loose:** the attribute and the object both occur in the same noun phrase of the corpus, not necessarily next to each other. **Reported** (the rule used in the main text, Table 1): the attribute directly precedes the object somewhere in a noun phrase; singular and plural objects count as the same. **Strict:** as reported, and the object is also the last word of the noun phrase. E.g., for the binding (red, chair): "a chair with a red cushion" counts only under loose; "red chair cushion" counts under loose and reported, but not strict; "an old red chair" counts under all three. Looser rules call more bindings seen, so fewer samples are unseen.

Corpus	Rule	Seen	Mixed	Unseen
COCO	Loose	8,932 (33.70%)	17,375 (65.55%)	201 (0.76%)
COCO	Reported	6,246 (23.56%)	19,840 (74.85%)	422 (1.59%)
COCO	Strict	5,407 (20.40%)	20,633 (77.84%)	468 (1.77%)
CC12M	Loose	21,068 (79.48%)	5,433 (20.50%)	7 (0.03%)
CC12M	Reported	15,400 (58.10%)	11,069 (41.76%)	39 (0.15%)
CC12M	Strict	12,779 (48.21%)	13,666 (51.55%)	63 (0.24%)

A.10 COMPARISON WITH EXISTING BENCHMARKS

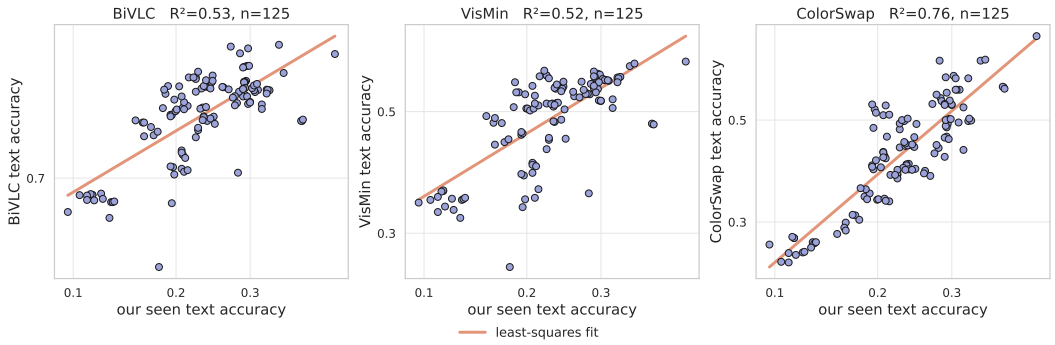

Figure 12: **Seen-split text accuracy on `BindSplit` Bench versus text accuracy on BiVLC (left), VisMin (middle), and ColorSwap (right)**, on probit-scaled axes. Each point is one of 125 model checkpoints (trained models: one point per seed). The x-axis uses the seen split of T1 (677 samples). R^2: BiVLC 0.53, VisMin 0.52, ColorSwap 0.76.

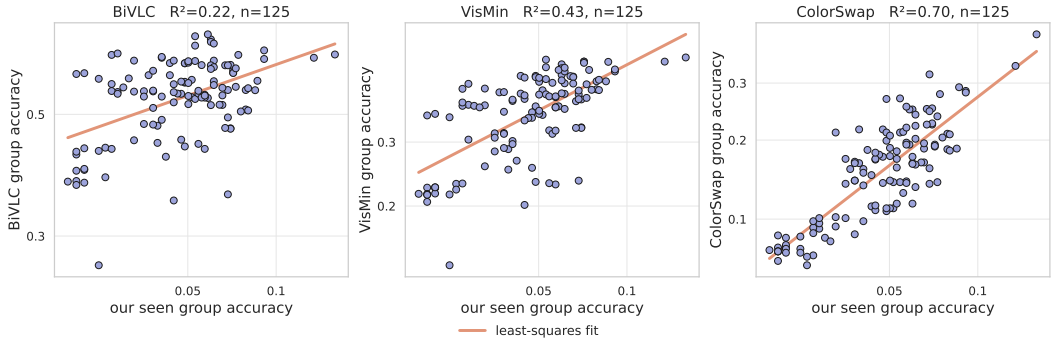

Figure 13: **Seen-split group accuracy on `BindSplit` Bench versus group accuracy on BiVLC (left), VisMin (middle), and ColorSwap (right)**, on probit-scaled axes. Each point is one of 125 model checkpoints (trained models: one point per seed). The x-axis uses the seen split of T1 (677 samples). R^2: BiVLC 0.22, VisMin 0.43, ColorSwap 0.70.



Figure 14: **Seen-split image accuracy on `BindSplit` Bench versus image accuracy on BiVLC (left), VisMin (middle), and ColorSwap (right)**, on probit-scaled axes. Each point is one of 125 model checkpoints (trained models: one point per seed). The x-axis uses the seen split of T1 (677 samples). R^2: BiVLC 0.15, VisMin 0.32, ColorSwap 0.63.

Table 23: VLM performance on three bi-directional vision language compositional benchmarks: BiVLC, VisMin, and ColorSwap. Three accuracies are shown for each benchmark: **text** (image to text retrieval), **image** (text to image retrieval), and **group** (both directions).

		BiVLC			VisMin			ColorSwap		
Model	Pretraining	Text	Image	Group	Text	Image	Group	Text	Image	Group
Zero-shot										
CLIP B/32	CC12M	66.1	39.3	35.5	34.0	30.6	20.2	48.4	21.5	18.1
CLIP B/16	CC15M	73.0	46.7	42.8	40.7	36.8	26.9	46.6	23.5	18.5
CLIP B/32	OpenAI WIT	75.8	52.4	49.1	43.6	38.1	29.7	38.2	17.7	13.7
CLIP L/14	OpenAI WIT	77.2	54.5	51.1	48.0	43.1	35.4	31.3	13.7	8.6
EVA01 g/14	LAION-400M	81.9	56.7	54.7	54.4	46.3	40.6	42.5	17.6	14.1
EVA02 B/16	Merged-2B	81.9	54.0	52.1	53.5	45.2	38.5	39.9	16.3	10.9
EVA02 E/14	LAION-2B	86.4	57.2	55.7	57.8	50.9	45.0	62.0	34.8	29.2
EVA02 L/14	Merged-2B	83.1	56.8	55.2	57.0	47.5	41.9	38.9	16.5	11.4
LaCLIP B/32	CC12M	55.6	30.4	25.8	25.1	20.8	12.8	36.3	11.4	7.1
OpenCLIP B/16	LAION-2B	82.4	55.1	53.3	55.4	49.3	42.0	49.5	23.1	18.7
OpenCLIP B/32	LAION-2B	81.3	51.3	49.5	55.0	46.7	39.4	58.5	28.9	24.8
OpenCLIP L/14	LAION-2B	84.2	56.8	55.0	56.9	50.9	43.8	53.3	27.9	22.3
OpenCLIP B/16	LAION-400M	80.0	56.0	54.2	52.1	45.2	37.8	43.9	18.9	13.9
OpenCLIP B/32	LAION-400M	79.5	54.3	51.8	50.6	43.8	35.6	52.6	24.6	20.7
OpenCLIP L/14	LAION-400M	81.8	56.5	54.6	56.0	47.5	40.7	53.1	24.1	19.5
OpenCLIP RN101	YFCC-15M	64.7	44.5	39.3	34.7	31.9	22.4	26.1	12.1	6.8
OpenCLIP RN50	YFCC-15M	63.8	43.2	38.0	32.3	31.2	20.6	25.6	15.4	8.5
SigLIP B/16	WebLI	83.6	61.7	59.6	58.2	51.4	45.0	62.2	40.2	33.4
SigLIP L/16	WebLI	85.5	62.6	60.2	58.5	52.9	45.8	66.7	46.1	40.0
SPARO B/32	CC12M	70.8	39.9	36.5	36.2	34.6	23.7	62.0	38.6	31.7
LABCLIP										
CLIP B/16	CC15M	73.7 ± 0.3	52.7 ± 0.0	48.2 ± 0.2	41.1 ± 0.2	37.7 ± 0.1	28.7 ± 0.3	43.2 ± 0.4	22.5 ± 0.2	16.9 ± 0.3
CLIP B/32	OpenAI WIT	76.2 ± 0.3	56.8 ± 0.5	53.2 ± 0.3	44.8 ± 0.5	38.7 ± 0.5	30.5 ± 0.3	30.5 ± 0.7	15.9 ± 0.4	9.8 ± 0.3
CLIP L/14	OpenAI WIT	77.9 ± 0.2	59.7 ± 0.5	56.7 ± 0.5	48.7 ± 0.6	42.2 ± 0.2	34.7 ± 0.3	28.5 ± 0.6	13.3 ± 0.6	8.0 ± 0.7
EVA01 g/14	LAION-400M	82.0 ± 0.1	61.2 ± 0.5	58.6 ± 0.4	52.8 ± 0.5	47.7 ± 0.3	41.0 ± 0.3	39.2 ± 0.5	17.7 ± 0.8	12.6 ± 1.0
EVA02 B/16	Merged-2B	79.6 ± 0.2	58.3 ± 0.8	55.3 ± 0.8	51.0 ± 0.4	44.1 ± 0.6	36.6 ± 0.5	34.0 ± 0.3	13.1 ± 0.7	8.3 ± 0.4
EVA02 E/14	LAION-2B	86.0 ± 0.3	64.4 ± 0.7	62.6 ± 0.8	55.0 ± 0.2	50.3 ± 0.2	43.4 ± 0.3	53.7 ± 0.4	30.0 ± 1.1	24.8 ± 0.5
EVA02 L/14	Merged-2B	81.7 ± 0.4	61.9 ± 0.5	59.9 ± 0.6	53.9 ± 1.0	46.4 ± 0.2	40.1 ± 0.5	35.2 ± 1.0	14.7 ± 0.3	10.0 ± 0.2
OpenCLIP B/16	LAION-2B	80.6 ± 0.3	60.0 ± 0.5	56.9 ± 0.4	53.4 ± 0.9	46.5 ± 0.3	39.0 ± 0.2	42.1 ± 0.3	21.9 ± 0.9	15.8 ± 0.8
OpenCLIP B/32	LAION-2B	79.6 ± 0.1	58.6 ± 0.1	55.6 ± 0.1	51.2 ± 0.3	44.9 ± 0.2	36.1 ± 0.3	50.0 ± 0.3	25.7 ± 0.8	20.4 ± 0.7
OpenCLIP L/14	LAION-2B	82.6 ± 0.4	61.5 ± 0.3	59.0 ± 0.1	54.6 ± 0.5	48.3 ± 0.3	41.5 ± 0.4	44.7 ± 0.2	24.0 ± 0.5	18.2 ± 0.6
OpenCLIP B/16	LAION-400M	78.3 ± 0.3	58.4 ± 0.6	55.6 ± 0.3	51.0 ± 0.6	42.5 ± 0.2	35.1 ± 0.2	40.0 ± 0.1	18.9 ± 0.1	14.1 ± 0.2
OpenCLIP B/32	LAION-400M	78.3 ± 0.2	56.6 ± 0.3	53.6 ± 0.1	48.4 ± 0.2	41.9 ± 0.1	34.1 ± 0.3	45.2 ± 0.3	24.4 ± 0.2	17.5 ± 0.3
OpenCLIP L/14	LAION-400M	80.3 ± 0.5	60.6 ± 0.4	58.1 ± 0.3	53.3 ± 0.4	46.2 ± 0.3	38.8 ± 0.1	44.1 ± 0.9	21.2 ± 0.4	16.3 ± 0.3
OpenCLIP RN101	YFCC-15M	67.5 ± 0.2	49.2 ± 0.1	44.2 ± 0.1	35.4 ± 0.2	32.5 ± 0.3	23.1 ± 0.3	23.8 ± 0.9	11.6 ± 0.6	6.7 ± 0.5
OpenCLIP RN50	YFCC-15M	66.7 ± 0.1	48.6 ± 0.3	43.5 ± 0.3	33.7 ± 0.4	30.5 ± 0.2	21.6 ± 0.0	24.7 ± 0.3	14.0 ± 0.2	7.0 ± 0.5
SPARO B/32	CC12M	70.9 ± 0.3	49.3 ± 0.2	44.5 ± 0.4	35.9 ± 0.9	33.2 ± 0.4	23.3 ± 0.2	51.9 ± 1.0	35.2 ± 0.7	26.6 ± 0.8
CS-CLIP										
CLIP B/32	CC12M	71.6 ± 0.2	49.4 ± 0.3	45.4 ± 0.3	39.4 ± 0.2	35.9 ± 0.2	25.7 ± 0.3	52.1 ± 1.1	25.0 ± 0.2	21.5 ± 0.3
CLIP B/16	CC15M	77.2 ± 0.4	57.5 ± 0.6	53.4 ± 0.6	45.3 ± 0.2	40.0 ± 0.2	31.8 ± 0.3	48.9 ± 0.7	26.3 ± 0.6	21.5 ± 0.4
CLIP B/32	OpenAI WIT	81.9 ± 0.3	62.4 ± 0.6	59.5 ± 0.3	51.0 ± 0.7	45.9 ± 0.4	37.6 ± 0.2	40.2 ± 0.3	21.4 ± 0.5	16.1 ± 0.9
EVA02 B/16	Merged-2B	83.2 ± 0.6	64.9 ± 0.9	62.5 ± 0.8	54.7 ± 0.9	52.8 ± 0.4	44.6 ± 0.8	45.2 ± 1.1	22.3 ± 0.6	16.9 ± 0.8
OpenCLIP B/32	LAION-2B	82.2 ± 0.3	63.3 ± 0.7	60.1 ± 0.7	55.7 ± 0.3	51.2 ± 0.4	43.3 ± 0.2	58.7 ± 0.3	33.4 ± 0.3	28.2 ± 0.3
OpenCLIP B/32	LAION-400M	81.5 ± 0.5	61.1 ± 0.3	58.2 ± 0.3	53.2 ± 0.2	45.0 ± 1.0	37.7 ± 0.3	55.1 ± 0.6	30.3 ± 0.4	24.8 ± 0.8
CLIC										
CLIP B/32	OpenAI WIT	78.7 ± 0.1	57.0 ± 0.1	54.0 ± 0.1	46.3 ± 0.2	39.2 ± 0.1	31.5 ± 0.1	40.4 ± 0.4	17.7 ± 0.1	14.2 ± 0.2
CLIP L/14	OpenAI WIT	79.4 ± 0.2	56.1 ± 0.1	53.9 ± 0.1	50.6 ± 0.1	43.5 ± 0.1	36.6 ± 0.1	34.3 ± 0.3	13.7 ± 0.1	9.2 ± 0.1
EVA02 B/16	Merged-2B	82.1 ± 0.0	54.8 ± 0.3	52.9 ± 0.3	54.2 ± 0.2	45.4 ± 0.3	38.9 ± 0.5	41.0 ± 0.2	16.8 ± 0.3	11.3 ± 0.3
EVA02 L/14	Merged-2B	83.1 ± 0.1	60.6 ± 0.0	58.7 ± 0.0	56.1 ± 0.1	47.9 ± 0.1	42.1 ± 0.1	39.0 ± 0.2	15.7 ± 0.3	11.1 ± 0.3
OpenCLIP B/16	LAION-2B	82.7 ± 0.1	53.5 ± 0.1	51.6 ± 0.0	55.6 ± 0.2	49.6 ± 0.1	42.2 ± 0.2	48.8 ± 0.4	21.2 ± 0.1	16.7 ± 0.1
OpenCLIP B/32	LAION-2B	81.6 ± 0.1	52.3 ± 0.2	50.7 ± 0.2	55.4 ± 0.1	46.5 ± 0.2	39.6 ± 0.1	56.2 ± 0.0	25.0 ± 0.2	20.8 ± 0.3
OpenCLIP L/14	LAION-2B	84.3 ± 0.1	59.1 ± 0.3	57.2 ± 0.3	56.3 ± 0.1	52.1 ± 0.2	44.9 ± 0.1	50.1 ± 0.1	24.0 ± 0.4	19.1 ± 0.2
OpenCLIP B/16	LAION-400M	81.0 ± 0.1	55.1 ± 0.3	53.4 ± 0.2	53.9 ± 0.1	45.0 ± 0.1	38.2 ± 0.1	46.4 ± 0.3	18.1 ± 0.3	14.1 ± 0.1
OpenCLIP B/32	LAION-400M	80.2 ± 0.1	55.2 ± 0.0	52.9 ± 0.1	51.9 ± 0.0	42.9 ± 0.2	36.3 ± 0.1	53.2 ± 0.2	23.4 ± 0.4	19.4 ± 0.2
OpenCLIP L/14	LAION-400M	81.7 ± 0.0	56.4 ± 0.2	54.3 ± 0.2	55.8 ± 0.1	48.5 ± 0.1	41.3 ± 0.2	50.0 ± 0.3	22.6 ± 0.2	18.6 ± 0.2
OpenCLIP RN101	YFCC-15M	67.5 ± 0.1	44.8 ± 0.1	40.6 ± 0.2	36.6 ± 0.1	31.7 ± 0.1	22.6 ± 0.0	27.3 ± 0.1	11.2 ± 0.2	7.6 ± 0.1
OpenCLIP RN50	YFCC-15M	66.3 ± 0.1	43.4 ± 0.1	38.6 ± 0.1	35.3 ± 0.2	30.0 ± 0.0	21.7 ± 0.0	26.4 ± 0.1	11.4 ± 0.2	7.3 ± 0.1
SigLIP B/16	WebLI	78.0 ± 0.1	50.5 ± 0.1	47.6 ± 0.1	47.9 ± 0.1	38.9 ± 0.0	32.5 ± 0.1	56.6 ± 0.2	28.1 ± 0.1	22.6 ± 0.1

Table 24: R@1 on the seen, partially unseen, and fully unseen splits of PUG:SPARE in the LAB-CLIP setting. 3x3 bindings are held out.

Model ID	Model	Seen	Partially unseen	Fully unseen
clip_cc12m_b32	CLIP (CC12M)	51.5 ± 0.4	23.5 ± 2.1	13.3 ± 1.9
clip_cc15m_b16	CLIP (CC15M) B/16	67.0 ± 0.9	36.6 ± 3.2	24.7 ± 9.0
clip_openai_b32	CLIP (OpenAI) B/32	69.0 ± 1.4	35.1 ± 9.4	21.6 ± 11.4
clip_openai_l14	CLIP (OpenAI) L/14	85.4 ± 1.5	47.6 ± 7.8	22.9 ± 9.0
eva01_g14	EVA01 (LAION-2B) g/14	75.4 ± 0.5	55.8 ± 1.3	41.3 ± 6.4
eva02_b16	EVA02 (Merged-2B) B/16	66.3 ± 0.9	41.5 ± 0.7	27.2 ± 4.7
eva02_e14	EVA02 (LAION-2B) E/14	80.9 ± 1.3	64.2 ± 6.1	50.6 ± 16.1
eva02_l14	EVA02 L/14	73.8 ± 0.9	55.7 ± 2.8	41.7 ± 8.8
negclip_coco_ft_b32	NegCLIP (COCO ft)	69.2 ± 3.4	39.2 ± 5.7	27.8 ± 8.0
openclip_laion2b_b16	OpenCLIP (LAION-2B) B/16	80.7 ± 1.2	55.7 ± 6.2	41.2 ± 9.7
openclip_laion2b_b32	OpenCLIP (LAION-2B) B/32	72.0 ± 2.1	49.3 ± 6.3	38.8 ± 10.0
openclip_laion2b_l14	OpenCLIP (LAION-2B) L/14	87.3 ± 1.0	60.4 ± 7.3	43.0 ± 13.2
openclip_laion400m_b16	LAION-400M B/16	83.7 ± 0.1	57.2 ± 10.3	41.4 ± 19.0
openclip_laion400m_b32	LAION-400M B/32	75.7 ± 3.4	47.0 ± 11.7	32.1 ± 16.1
openclip_laion400m_l14	LAION-400M L/14	88.9 ± 1.2	59.6 ± 5.5	38.7 ± 5.6
openclip_yfcc15m_rn101	OpenCLIP (YFCC15M) RN101	53.4 ± 1.3	24.8 ± 3.2	11.7 ± 4.4
openclip_yfcc15m_rn50	OpenCLIP (YFCC15M) RN50	53.9 ± 0.7	27.4 ± 2.4	16.6 ± 2.6
siglip_b16	SigLIP (WebLI) B/16	49.4 ± 1.9	47.2 ± 4.3	42.6 ± 7.2
siglip_l16	SigLIP L/16	49.3 ± 2.1	52.8 ± 4.7	46.2 ± 7.3
sparo_cc12m_b32	SPARO (CC12M)	87.0 ± 3.4	58.4 ± 8.6	42.7 ± 14.9

A.11 Controlled synthetic PUG:SPARE setting results

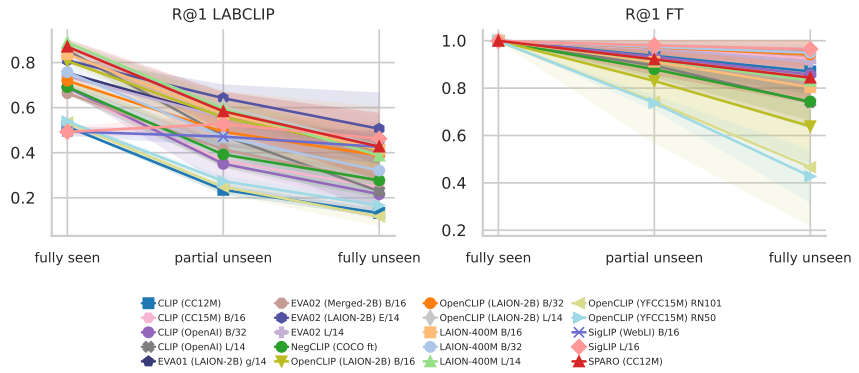

Figure 15: **R@1 drops as more bindings are unseen on PUG:SPARE.** Across 20 backbones with LABCLIP alignment (left) and 17 with full backbone fine-tuning (FT) (right), R@1 image-to-text retrieval accuracy drops with the number of held-out attribute-object bindings per sample, at each step for all backbones except SigLIP L/16 with LABCLIP (0 = fully-seen, 1 = partially-unseen, 2 = fully-unseen). Lines: mean across 3 data-split seeds; ribbons: standard deviation. Random-chance order-invariant R@1 is $0.0271\%(2/7392)$, so all plotted values lie above chance.

Table 25: R@1 on the seen, partially unseen, and fully unseen splits of PUG:SPARE in the full fine-tuning setting. 3x3 bindings are held out.

Model ID	Model	Seen	Partially unseen	Fully unseen
clip_cc12m_b32	CLIP (CC12M)	100.0 ± 0.0	93.6 ± 5.1	87.0 ± 9.6
clip_cc15m_b16	CLIP (CC15M) B/16	100.0 ± 0.0	96.8 ± 3.3	92.5 ± 6.4
clip_openai_b32	CLIP (OpenAI) B/32	100.0 ± 0.0	93.2 ± 5.1	85.7 ± 10.2
clip_openai_l14	CLIP (OpenAI) L/14	99.9 ± 0.0	89.5 ± 6.0	74.0 ± 11.2
eva02_b16	EVA02 (Merged-2B) B/16	100.0 ± 0.0	97.6 ± 2.7	96.1 ± 4.6
eva02_l14	EVA02 L/14	99.9 ± 0.1	96.7 ± 1.1	91.1 ± 6.0
negclip_coco_ft_b32	NegCLIP (COCO ft)	99.7 ± 0.5	87.9 ± 3.5	74.3 ± 8.0
openclip_laion2b_b16	OpenCLIP (LAION-2B) B/16	100.0 ± 0.0	83.0 ± 8.5	63.9 ± 16.1
openclip_laion2b_b32	OpenCLIP (LAION-2B) B/32	100.0 ± 0.0	97.2 ± 3.1	94.2 ± 5.9
openclip_laion2b_l14	OpenCLIP (LAION-2B) L/14	100.0 ± 0.0	98.2 ± 1.8	96.5 ± 3.3
openclip_laion400m_b16	LAION-400M B/16	100.0 ± 0.0	91.1 ± 7.2	80.4 ± 15.1
openclip_laion400m_b32	LAION-400M B/32	100.0 ± 0.0	97.4 ± 2.2	94.9 ± 3.4
openclip_laion400m_l14	LAION-400M L/14	100.0 ± 0.0	91.9 ± 3.0	83.2 ± 3.0
openclip_yfcc15m_rn101	OpenCLIP (YFCC15M) RN101	99.7 ± 0.2	74.3 ± 17.6	46.6 ± 25.0
openclip_yfcc15m_rn50	OpenCLIP (YFCC15M) RN50	99.9 ± 0.0	73.7 ± 3.2	42.7 ± 11.3
siglip_l16	SigLIP L/16	99.9 ± 0.2	98.0 ± 2.2	96.5 ± 3.5
sparo_cc12m_b32	SPARO (CC12M)	100.0 ± 0.0	92.1 ± 6.7	84.4 ± 11.0

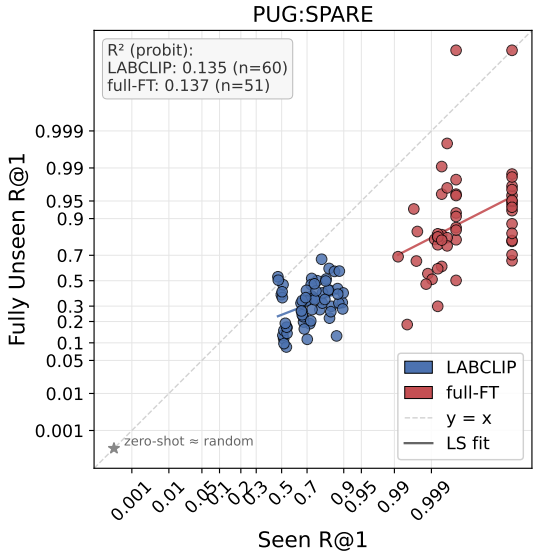

Figure 16: **Seen versus fully unseen R@1 on PUG:SPARE**, on probit-scaled axes (tick labels show raw R@1). Each point is one backbone on one of 3 data splits, under LABCLIP (20 backbones) or full fine-tuning (17 backbones). Solid lines are least-squares fits per regime and the dashed line is $y = x$. Under full fine-tuning, 17 of the 51 seen values are exactly 100%, which a probit axis cannot show; they are clipped for plotting, so the fitted line and R^2 for full fine-tuning depend on the clipping value and are not meaningful.