Benchmarking Zero-Shot Recognition with Vision-Language Models: Challenges on Granularity and Specificity

Zhenlin Xu[1*] Yi Zhu[2†] Siqi Deng[1] Abhay Mittal[3†] Yanbei Chen[1] Manchen Wang[3]
Paolo Favaro[1] Joseph Tighe[3†] Davide Modolo[1]
[1]AWS AI Labs [2]Boson AI [3]Meta

Abstract

This paper presents novel benchmarks for evaluating vision-language models (VLMs) in zero-shot recognition, focusing on granularity and specificity. Although VLMs excel in tasks like image captioning, they face challenges in open-world settings. Our benchmarks test VLMs' consistency in understanding concepts across semantic granularity levels and their response to varying text specificity. Findings show that VLMs favor moderately fine-grained concepts and struggle with specificity, often misjudging texts that differ from their training data. Extensive evaluations reveal limitations in current VLMs, particularly in distinguishing between correct and subtly incorrect descriptions. While fine-tuning offers some improvements, it doesn't fully address these issues, highlighting the need for VLMs with enhanced generalization capabilities for real-world applications. This study provides insights into VLM limitations and suggests directions for developing more robust models.

1. Introduction

Vision-language models (VLMs) have shown impressive capabilities in a wide range of tasks, including image captioning [28, 31], visual question answering [1], and notably, zero-shot visual recognition [8, 18, 29, 35]. Models pretrained on large-scale image-caption datasets [10, 21, 22, 32], like CLIP [18], have been at the forefront. These models achieve this by mapping visual and linguistic inputs to a shared latent space, enabling the recognition of novel objects or concepts in a zero-shot manner—a capability critical for developing versatile and intelligent visual systems.

While current VLMs perform well in various tasks, their application in open-world scenarios poses unique challenges. An ideal open-world zero-shot model would recognize any language-defined input, from simple concepts like "an image of flowers" to more complex descriptions like "a person playing with a dog on the beach", and output scores indicating whether the visual input *correctly* implies the semantics of the language input. Existing works often evaluate the zero-shot capability on various classification dataset like ImageNet [19] and domain specific datasets [11] without the notion of granularity of concepts, as well as image and text retrieval on Flickr30K [15] and COCO [13] that are not able to reveal the general failure pattern. These benchmark fall short of replicating the complexities of a realistic open-world setting, leaving a substantial gap in our understanding of the effectiveness of VLMs in such scenarios.

This paper present new benchmarks on the pivotal properties when deploying VLMs for real-world zero-shot recognition: *granularity* and *specificity*. The first benchmark examines VLMs' ability to consistently understand concepts at different levels of *granularity*. For instance, a model should recognize an image of a leopard both when presented with a fine-grained query like "an image of a leopard" and a more coarse-grained query such as "an image of a feline." This consistency is crucial, not just as an indicator of the model's comprehension of concept relationships but also for practical applications. A pertinent example is in autonomous driving, where recognizing "road cone" but failing to identify "barrier" could be problematic. To assess this, we employ an evaluation protocol where we measure the performance discrepancy in recognizing a coarse-grained class, both by directly using the coarse-grained class prompt and by aggregating predictions from its fine-grained children classes. We leverage a dataset with hierarchical labels, adapting ImageNet and its semantic hierarchy from WordNet.

The second benchmark evaluates how the specificity of language inputs affects VLM outputs, even when visual and linguistic inputs align. For example, a simple prompt like "a picture with a dog" may receive a lower score compared to a more detailed but incorrect caption "a dog and cow lying together on an orange couch." This distinction is key in determining whether VLMs can accurately reflect the correctness of the alignment between visual and language inputs, rather than just overall similarity. To test this, we use an image-to-text retrieval task on the MS-COCO dataset,

*Correspondence to: xzhenlin@amazon.com
†Work done while at Amazon

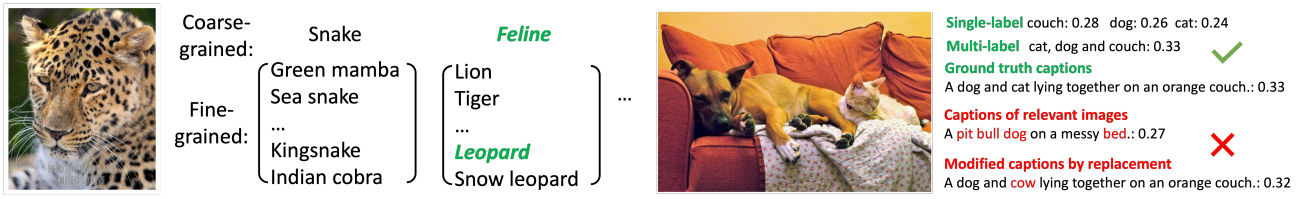


Figure 1. **Left**: Zero-shot models should recognize images with fine-grained (FG) concepts such as "Leopard", as well as coarse-grained (CG) concepts like "Feline" However, they often exhibit performance discrepancies on concepts at different levels of granularity. **Right**: Zero-shot models should recognize whether the text correctly describe the given image. However, vision-language models could be sensitive to the specificity of text and struggle to distinguish between the challenging positive like single-label prompts and hard negatives like poisoned captions with small changes.

designing challenging positive texts with varying levels of specificity, such as single-label prompts with limited information, and hard negative texts like slightly modified but incorrect captions.

Our carefully designed benchmarks led to an extensive evaluation of state-of-the-art vision-language models (VLMs). We focus on contrastive models like CLIP, covering various aspects, including pretraining datasets, architectural designs, cross-modality interactions, and learning objectives. We discovered that VLMs face significant challenges in both benchmarks.

In the granularity evaluation, we find that *VLMs prefer moderately fine-grained concepts over more abstract, coarse-grained ones.* This tendency seems closely tied to the nature of the training data. A detailed analysis of the LAION dataset revealed a higher presence of moderately fine-grained concepts in image alt-text, suggesting that data distribution plays a critical role. In the specificity evaluation, VLMs showed *sensitivity to text specificity*: texts that differ in specificity from the training data, such as straightforward single-label prompts or overly detailed captions, often received lower scores than more precisely detailed but slightly erroneous captions. This challenges the VLMs' ability to accurately distinguish between correct and subtly incorrect descriptions, complicating the retrieval of hard positive texts from hard negatives. The implication is that VLMs' scoring does not always reliably reflect the true alignment between visual and textual inputs. Our exploration into fine-tuning VLMs with these hard text samples revealed that while it offers some improvements, it does not fully resolve the challenges as a complete solution.

To our best knowledge, this is the first comprehensive study that evaluating VLMs from the perspective of semantic granularity and specificity. We believe that the carefully designed benchmark provide a valuable tool to the community to better quantitatively evaluate VLMs. With the proposed benchmark, we observed that all models surprisingly performs significantly worse than what we may hope. The findings and insights from our analysis may shed lights on better understanding the limitations of current VLMs and the challenges of using it for zero-shot recognition, and inspire new models with better generalization.

2. Related Works

Zero-shot visual recognition CLIP-like vision-language foundation models have enabled open vocabulary visual recognition by mapping images with their corresponding language descriptions. Early methods [10, 18] demonstrate the effectiveness of this paradigm on the image classification tasks. For example, CLIP is able to achieve decent zero-shot accuracy on 27 image classification datasets. Given its potential, the language-driven visual recognition paradigm has been extended to tasks including object detection [35], semantic segmentation [29], video action recognition [27], depth estimation [34], etc. Such language guided visual recognition has become the new paradigm in the field of computer vision since it can recognition new objects without any training data. In this paper, we would like to stress test these VLMs in terms of zero-shot visual recognition to better understand their capability and limitation in realistic open-world setting.

Benchmarking vision-language models Thanks to the larger datasets and larger transformer models, many powerful vision-language models have been developed and shown great capability [1, 28, 31]. At the same time, these models are being studied from various perspectives, such as robustness, bias, and other limitations [3, 5–7, 14]. [17] investigates the robustness of nine open-sourced image-text models under common perturbations on five tasks, while [20] studies the robustness of video-text models. [4] further analyzes the robustness of VLMs under challenging natural distribution shifts and show that the more diverse training distribution is the main cause for the robustness gains. [26, 33] systematically evaluates the ability to encode compositional information of the VLMs. [2] investigates the visual reasoning capabilities and social biases of different text-to-image models. To improve transferability, [24] designs an efficient and scalable approach that leverages

external knowledge to learn image representations. In this paper, we study VLMs from two new perspectives: granularity and specificity through the lens of zero-shot visual recognition.

Table 1. An overview of the differences between the vision-language models evaluated in our study by the architecture, pretraining datasets, learning objectives, and if using cross-modality fusion modules. ITC, ITM, MIM, MTM, MMM stands for image-text contrastive, image-text matchng, masked image modeling, masked text modeling and masked multimodal modeling losses.

Model	Architecture	Datasets	Objectives	Fusion
CLIP	ViT-B-32 ViT-L-14	Private400M	ITC	-
OpenCLIP	ViT-B-32	LAION400M	ITC	-
	ViT-B-32 ViT-L-14 VIT-H-14	LAION2B		
UniCL	Swin-B	YFCC14M IN21K IN21K+YFCC14M IN21K+YFCC14M+GCC15M	ITC	-
KLITE	Swin-B	IN21K+YFCC14M+GCC15M	ITC	-
BLIP	ViT-B-16	COCO+VG+CC+SBU +LAION+CapFilt-L	ITC + ITM + Captioning	- ✓
FLAVA	ViT-B/16	PMD70M	ITC+ITM +MMM+MIM +MTM	-

3. Zero-shot Visual Recognition With Vision-Language Models

In this study, we focus on two-stream contrastive vision-language models, such as CLIP, which leverage contrastive pre-training on a large dataset of paired image-text samples to learn cross-modal alignment. These models typically consist of a visual encoder E_v and a text encoder E_t, for encoding visual inputs x_v and textual inputs x_t into aligned representation spaces.

The zero-shot visual recognition task with a vision-language model can be formulated as computing the cross-modality score:

$$f(x_v, x_t) = E_v(x_v) \odot E_t(x_t) \tag{1}$$

Here, the $\odot$ operator computes the score between visual and language embeddings, with cosine similarity being the common choice while some models like FLAVA use an additional module to fuse the multi-modal embeddings. For classification tasks, x_t can be a class prompt, such as "a photo of a car", or it can incorporate additional class-specific knowledge to improve performance. In our subsequent studies, we adopt the prompt templates used in [18] for classification tasks. We simplify $E_t(x_t)$ and $f(x_v, x_t)$ for a class y to $E_t(y)$ and $f_{cls}(x_v, y)$, respectively.

In our study, we evaluate various contrastive vision-language models, each with distinct backbone architectures, training data, and learning objectives, shown in Tab. 1. These variants include CLIP [18], OpenCLIP [9] (trained on the public LAION dataset [23]), UniCL [30] (which incorporates classification annotations into the contrastive learning objective), KLITE [24] (which augments alt-text with extra knowledge during training), FLAVA [25] (trained with both cross-modal and uni-modal data and losses), and BLIP [12] (which includes uni-modal and cross-modal training, along with a captioning head for data bootstrapping). By examining these models, we aim to gain insights into the zero-shot visual recognition capabilities of vision-language models.

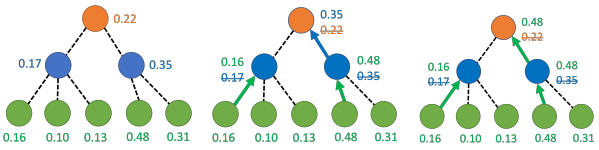
4. Granularity Consistency of Vision-Language Models

In this section, we investigate whether vision-language models (VLMs) perform consistently on visual concepts across different levels of granularity, which indicates their understanding of the relationships between concepts. We propose a benchmark to quantitatively evaluate the performance discrepancy of VLMs on concepts at different granularity levels. Our results show that models trained on image-text pairs exhibit significant performance discrepancies, with better recognition of moderately fine-grained concepts compared to coarse-grained ones. Further analysis suggests that the distribution of training data may account for this discrepancy, with models trained on datasets having more balanced representation across granularity levels showing smaller discrepancies.

4.1. Measure performance discrepancy on a semantic hierarchy

To assess the understanding of vision-language models across different levels of semantic granularity, we use zero-shot classification as our evaluation tool. Directly comparing classification metrics across granularities is not appropriate, as finer-grained classification inherently presents more challenges. Our benchmark focuses on measuring the discrepancy in zero-shot classification performance when using coarse-grained (CG) class prompts directly versus deriving predictions from using finer-grained (FG) children class prompts.

Dataset We expand the popular ImageNet-1K dataset with multi-level label hierarchy. Each of the 1000 fine-grained labels is assigned its ancestor labels based on the WordNet hierarchy, adding 820 ancestor labels. For example, "leopard" images are also labeled as "big cat," "feline," "mammal," "animal," and so on. his expansion allows us to thoroughly investigate how well VLMs perform across varying granularities.



(a) Raw scores (b) Max of direct children (c) Max of leaves

Figure 2. Illustrations on the two ways to propagate scores on the semantic hierarchy. (a) Raw scores without propagation. (b) Propagate the max score from direct children classes. For example, 0.35 = max(0.17, 0.35) (c) Propagate the max score from leaf classes. For example, 0.48 = max(0.16, 0.10, 0.13, 0.48, 0.31)

Evaluation Protocol Given the hierarchical nature of our dataset, each image is associated with multiple labels, transforming our task into a multi-label classification setting. Here, each label is considered for binary classification independently. For ancestor (CG) labels, we employ two distinct methods for score prediction:

1. Direct prediction: Utilize the text prompt of the CG label for cross-modality score with the image.
2. Propagated prediction: Compute the aggregate scores from the FG children class prompts. For instance, the "feline" label score can be derived by aggregating scores from "lion", "tiger", "leopard", etc.

For a class y, the raw cross-modality score between an image x and the textual prompt of y is computed by $S^{\text{raw}}(y) = f(x, y)$. For an ancestor class y_i, we design two specific approaches for score propagation, as illustrated in Figure 2 and formulated below:

1. Propagate the maximum score from direct children classes.

$$S^{\text{child}}(y_i) = \max_{y_j \in Y_C^i} S^{\text{raw}}(y_j) \tag{2}$$

2. Propagate the maximum score from leaf (most fine-grained) chidren classes.

$$S^{\text{leaf}}(y_i) = \max_{y_j \in Y_C^i} S^{\text{leaf}}(y_j) \tag{3}$$

The key idea is that if a VLM has a consistent understanding of concepts across granularities, its performance on directly classifying coarse-grained labels should be similar to the performance obtained by propagating predictions from the fine-grained children labels. We use mean Average Precision (mAP) as the evaluation metric, which does not require score threshold selection. In Tab. 2, we report mAP of leaf classes and ancestor classes. The performance of ancestor classes is evaluated with the raw scores, and scores using propagated from children classes using the two approaches presented above. The critical metric in our analysis is the difference (Δ) between raw scores and propagated scores, serving as an indicator of how well the VLMs bridges the understanding between CG and FG concepts.

Table 2. Zero-shot multi-label classification performance of labels at different levels of granularity on ImageNet. We reported the mean average precision (mAP) of ImageNet-1K fine-grained classes (leaves), and their coarse-grained ancestor classes with raw predictions (Anc_{raw}) and two propagated predictions. The differences (Δ) between the raw and propagated performance of ancestor classes presents the performance discrepancy of vision-language models on concepts at different granularity. Propagating from leaf classes gives the best performance.

Config	Leaves	Anc_{raw}	$\text{Anc}_{\text{child}}$ (Δ)	Anc_{leaf} (Δ)
CLIP				
B-400M	50.10	24.91	45.35 (+20.44)	58.73 (+33.83)
L-400M	65.06	33.64	57.72 (+24.08)	72.25 (+38.61)
OpenCLIP				
B-400M	47.10	20.12	40.66 (+20.54)	54.50 (+34.38)
B-2B	54.97	24.95	47.64 (+22.69)	62.66 (+37.70)
L-2B	65.79	31.59	56.65 (+25.07)	72.53 (+40.94)
H-2B	68.28	32.70	58.70 (+26.00)	74.93 (+42.23)
UniCL(Swin-B)				
YFCC	35.75	20.13	35.90 (+15.77)	47.55 (+27.42)
IN21K	26.28	38.15	39.30 (+1.15)	41.23 (+3.08)
YFCC+IN21K	37.84	35.18	44.84 (+9.65)	51.55 (+16.37)
All	54.49	37.54	54.58 (+17.04)	65.85 (+28.32)
K-LITE	48.40	31.50	49.63 (+18.14)	61.58 (+30.08)
BLIP	41.87	20.31	39.44 (+19.13)	52.08 (+31.77)
$\text{BLIP}_{\text{ft-coco}}$	42.83	22.07	41.45 (+19.38)	54.00 (+31.93)
FLAVA	40.91	21.36	39.32 (+17.96)	51.89 (+30.53)

4.2. Results and Analysis

Tab. 2 displays our granularity benchmark. The Δ values, highlighted in the last two columns, quantify the performance discrepancy between direct predictions using coarse-grained (CG) class prompts and predictions derived from prompts of their finer-grained (FG) children classes.

General Trend Across Models: Our analysis spans a diverse range of models, varying in scale and design. A significant trend emerges across all these models: direct predictions with CG labels consistently yield inferior results compared to those obtained from FG labels. Most notably, score propagation from the leaf (most fine-grained) classes leads to the most substantial performance improvements, decisively outperforming the propagation from direct children classes. This finding robustly indicates that VLMs demonstrate greater reliability and produce more accurate outputs when interacting with more specific, finer-grained concepts.

Granularity-Level Performance Analysis: Delving deeper into how VLMs perform at different granularity levels, we observe distinct patterns in their raw performance. These levels are defined based on the WordNet hierarchy, where level 0 represents the most abstract level "entity"

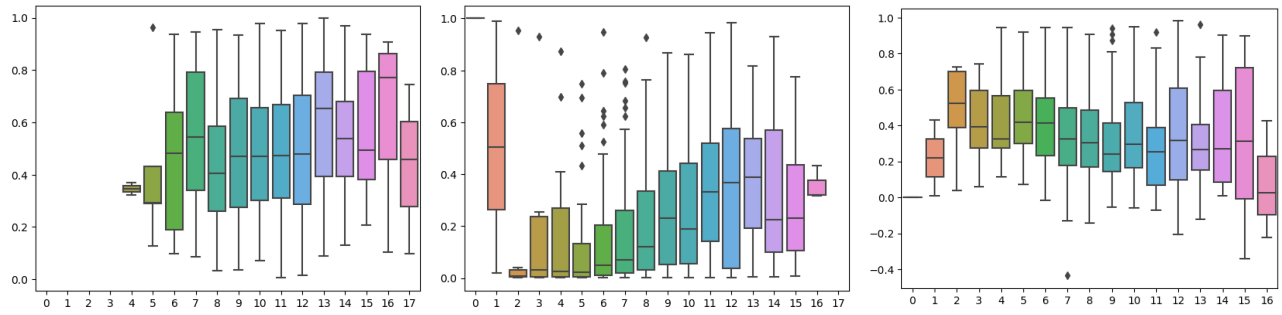


Figure 3. **Left**: The box-plot of zero-shot classification performance (mAP) for leaf class over the level in the semantic hierarchy. **Middle**: The box-plot of classification performance (mAP) for ancestor classes over the level in the semantic tree. Note that level 0 and level 1 have 1 and 2 classes respectively and easy to get high mAP. **Right**:The box-plot of *improved* zero-shot classification performance (mAP) for ancestor class by propagating from leaf classes, over the level in the semantic tree.

where all images are labeled with, and higher numbers indicate increasingly finer-grained concepts. For instance, at Level 3, a class might be as broad as "signal" or "location". At a middle level, such as Level 7, you might find classes like "instrument" or "vehicle." At even deeper levels, say Level 15, the classes are highly specific, like "tiger shark" or "cougar". According to Figure 3-Left and Middle, we find that VLMs are more adept at recognizing higher-level concepts for both leaf and ancestor classes. However, an intriguing dip in performance is noticed at the deepest level e.g. level 17 for leaf classes which consists of extremely fine-grained and likely rare concepts in the training data. These might be specific breeds of animals or types of vehicles not commonly encountered. Interestingly, despite the challenges at the deepest level, propagating scores from leaf classes generally improves performance for the majority of CG ancestor classes. This improvement holds for 775 out of the 820 ancestor classes. The exception is noted at level 16, which does not benefit from propagation due to the underperformance of the level 17 child classes. The results in Figure 3-Right highlight these findings, illustrating that VLMs are best at recognizing *moderatly* finegrained concepts

Influence of Pre-training Data on Granularity Discrepancy: Contrary to expectations, scaling up the alt-text training data or increasing model sizes, such as in the case of OpenCLIP-H-2B, does not seem to effectively address the granularity discrepancy. In fact, such scaling might worsen it. *The distribution of training data content is a more critical factor than its volume in influencing the granularity discrepancy*. This becomes evident when comparing the performance of UniCL models trained on different datasets, such as ImageNet21K, YFCC-14M, and GCC-15M.

1. UniCL models trained on image alt-text data (UniCLYFCC) significantly outperform those trained on ImageNet21K (UniCLIN21K) in FG leaf label classification (35.75 vs. 26.28 mAP).
2. $\text{UniCL}_{\text{IN21K}}$ excels in CG ancestor raw performance due to the comprehensive inclusion of CG classes in IN21K, even surpassing CLIP and OpenCLIP models trained on larger-scale alt-text datasets.
3. However, integrating alt-text data (YFCC14M and GCC15M) with IN21K for UniCL training enhances FG classification at the cost of CG performance, leading to a larger CG-FG discrepancy.

Granularity Bias in Pre-training Data: The distribution of alt-text data, skewed towards fine-grained concepts, appears to contribute significantly to the observed performance discrepancy. The natural inclination to use precise concepts in language descriptions appears to be a driving factor behind this bias. In our further analysis of the OpenCLIP models trained on LAION-2B, we investigate how the distribution of visual concepts in alt-text data correlates with the granularity discrepancy. We use ImageNet samples from each leaf class to find similar images in LAION-2B, determining the frequency of each class name in the training captions. This frequency distribution in Figure 4-Left shows that higher-level (more fine-grained) classes are mentioned more frequently except for the overly fine-grained classes (level$\geq$16), aligning with the performance trends observed in Figure 3. Moreover, we examine the correlation between each ancestor class's performance discrepancy (Δ_{leaf}) and its frequency gap with its leaf children (Δ_{freq}). Our findings, shown in Figure 4-Right, demonstrate a positive correlation (coefficient 0.43 with a significant p-value of $3.4e-39$), reinforcing the notion that the distribution of training data significantly influences the VLMs' granularity bias.

In conclusion, the insights from our granularity benchmark shed light on the significant challenges and limitations inherent in current VLMs, particularly their inconsistent performance across different levels of semantic granularity. These results highlight the imperative for more balanced and diverse training datasets in developing VLMs capable of robust

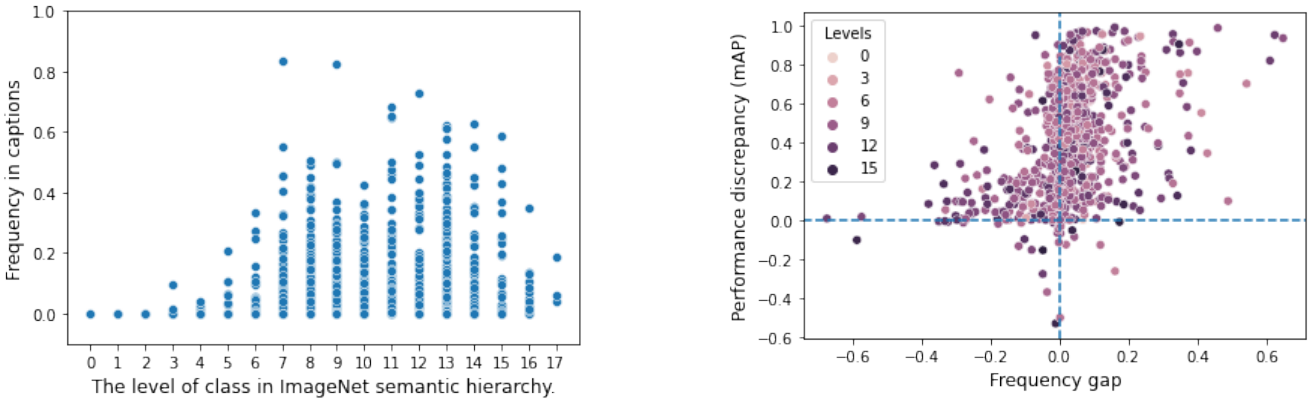


Figure 4. **Left**: The scatter-plot of the frequency of class names in pre-training captions over the level in the semantic tree. Course-grained and overly fine-grained concepts are less presented in captions. **Right**: The scatter-plot of performance discrepancy over the frequency gap between ancestor class names and their leaf children. A positive correlation exists between the performance discrepancy and frequency gap (coefficient 0.43 with p-value 3.4e-39).

5. Evaluate Specificity Robustness

When using vision-language models for open-world zero-shot recognition, the textual describes the visual concepts to recognize and the output score should indicate the chance that the described concepts exist in the visual input. In other words, it is critical to measure the correctness of textual inputs given visual inputs. However, as the example in Figure 1-Right illustrates, the scores of visual language models and do not strictly reflect the correctness of the textual input and thus make it challenging to be useful for open-world visual recognition. Since contrastive vision-language models have been trained on image alt-text pairs, the scores are biased toward the specificity of text as in the pretraining data. In our study, we demonstrated that the specificity of text can distract vision-language scores that VLMs struggle to reflect the correctness faithfully.

Evaluation protocol and dataset We use image-to-text retrieval as the proxy task to demonstrate that the scores of contrastive vision language models can easily be distracted. We build our experiments on images of the MSCOCO2017 dataset and their annotation of captions and bounding boxes. The setup of the image-to-text retrieval task is following. Given a query image and a set of positive and negative text, the score between the query image and each text is used for retrieving the positive text. Average Precision (AP) is the metric for evaluating the performance of each image and we report the mean Average Precision (mAP) of the whole dataset. Typically, the positive text are the captions annotated for the query images (Cap^+), and the negative text is the captions of other images in the data (Cap^-_{rd}). To test our hypothesis, we design the following hard positives and hard negatives.

- *Prompts of a single label* (Prompt^+_s): apply the classification prompts on one label of the query image. For example, "a photo of a dog".
- *Prompts of multiple labels* (Prompt^+_m): apply the classification prompts on all labels in the query image. For example, "a photo of dog, person, ball".
- *Captions from Localized narratives[16]* ($\text{Cap}^+_l n$): the text descriptions that are much longer and more informative than typical captions in MSCOCO and pretraining data.
- *Captions of relevant images* (Cap^-_{rl}): COCO captions of relevant images that have overlapping labels with the query image.
- *Captions with errors* (Cap^-_{er}): modifying true COCO captions of query images with errors by replacing a noun entity in the text with the name of a label that does not exit in the image. We use spaCy [1] for entity recognition.

The hard positives Prompt^+_s, Prompt^+_m and $\text{Cap}^+_l n$ contain less or more information, although still correct, than the true captions Cap^+. They can examine if different specificity of the positive text can reduce the score. The hard negatives Cap^-_{rl} and Cap^-_{er} are similar to true captions but are wrong descriptions. They can examine whether the model can be robust to specificity and indicate the correctness of text input. Note that we use randomly chosen 100 negative texts for each query image for all image-text retrieval experiments and report results of CLIP ViT-B/32.

Results and implications We first plot a normalized histogram of visual-language scores between query images and various textual inputs in Figure 5. Figure 5-Left compares the scores from different types of positive texts. True COCO captions generate higher score than classification prompts

[1] https://spacy.io/

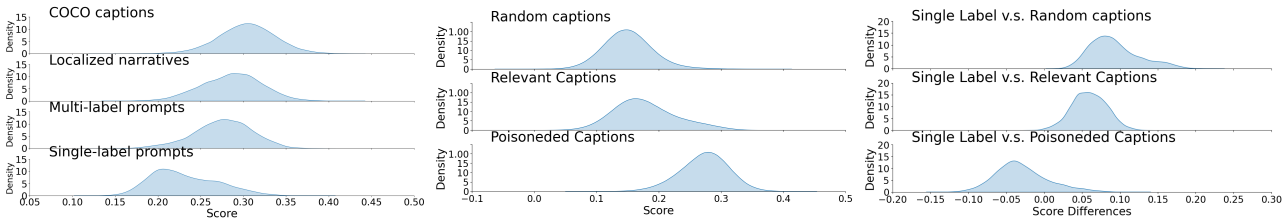


Figure 5. **Left**: Distribution of cross-modality scores with positive text: COCO captions, Localized-narratives captions, single-label, and multi-label prompts. Mismatched specificity in text (either too low or high) results in reduced scores. **Middle**: Distribution of scores with negative text: captions from random images, relevant images, and subtly altered captions. The altered captions attain high scores, similar to positive texts. **Right**: Score differences between single-label prompts and various negative texts, highlighting that correct single-label prompts often score lower than incorrect altered captions.

Table 3. Performance of image-to-text retrieval measured by mean Average Precision (mAP). Columns differentiate between positive and negative text types: Cap^+ (true COCO captions), Cap^+_{ln} (localized-narratives captions), Prompt^+_s and Prompt^+_m (single/multi-label prompts) for positives; Cap^-_{rd}, Cap^-_{rl}, Cap^-_{er} (captions from random images, relevant images, and error-modified true captions) for negatives. This shows how hard positives and negatives can distort vision-language models' similarity scores.

Model	Cap^+			Prompt^+_m			Prompt^+_s			Cap^+_{ln}		
	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}
CLIP-B	94.78	82.77	28.10	74.39	50.12	7.19	55.69	30.17	4.47	81.29	60.00	13.57
CLIP-L	95.64	84.66	30.59	79.84	56.76	8.51	58.18	33.52	4.94	85.51	66.74	16.63
$\text{OpenCLIP}_{\text{B-400M}}$	95.28	84.62	29.61	64.66	39.1	4.49	50.9	25.93	3.63	83.48	63.07	13.85
$\text{OpenCLIP}_{\text{B-2B}}$	96.28	86.73	28.96	75.84	51.83	6.32	61.39	35.89	4.42	88.83	71.76	18.91
$\text{OpenCLIP}_{\text{L-2B}}$	97.09	88.81	33.03	79.22	56.00	6.90	65.44	39.97	4.96	89.50	72.78	18.63
$\text{OpenCLIP}_{\text{H-2B}}$	97.45	89.85	35.82	79.2	57.64	7.49	65.67	42.19	5.75	89.74	73.28	18.09
$\text{UniCL}_{\text{All}}$	94.37	81.76	20.74	82.58	62.33	9.94	**82.45**	**60.02**	8.71	81.96	62.33	12.99
KLITE	92.47	77.67	16.45	75.71	53.60	9.03	69.98	47.06	8.47	79.81	59.24	11.16
BLIP	97.68	90.89	48.53	57.64	32.21	3.13	43.24	20.07	2.81	82.26	63.62	17.94
$\text{BLIP}_{\text{ft-coco}}$	99.07	95.15	**56.44**	74.65	51.02	4.86	65.96	41.77	4.02	89.99	75.92	23.13
$\text{BLIP}_{\text{ft-coco-fusion}}$	**99.26**	**96.08**	38.57	76.59	54.97	3.35	81.62	58.41	2.97	92.51	82.59	22.72
FLAVA	97.73	89.31	29.49	**86.52**	**69.29**	**13.22**	78.35	58.33	**11.22**	**94.83**	**82.87**	**35.09**
NegCLIP	96.6	87.37	51.88	65.32	39.91	6.7	61.32	34.52	6.09	76.70	53.93	13.33

(multiple-label prompts get higher scores than single-label prompts), and Localized-narratives who are overly-detailed captions surprisingly lead to lower scores than normal captions. The observations confirms our hypothesis that *the amount of information (specificity) in text can distort the scores* and the specificity that is closer to the training text leads to higher scores. Shown by Figure 5-Middle, relevant image captions score slightly higher than random ones among negative texts. However, captions with modified errors score as high as true captions, undermining the effectiveness of VL scores. The result verifies our hypothesis that *the similarity scores cannot distinguish the correctness*. When comparing the positive single-label prompts with different types of negative text in Figure 5-Right, single label positives Prompt^+_s are even lower than hard negatives Cap^-_{er}, which is not desired.

Then, we report the image-to-text retrieval results in Tab. 3 when combining different positive and negative text. We can see that using harder positives or harder negatives can degrade image-to-text retrieval performance, and retrieving label prompts from captions with small errors is extremely hard. Comparing the performance of different models, we can see that the BLIP model with the fusion design fine-tuned on COCO is the best when the positive text are true captions which is nature since it is trained on the same data. however, results in worse performance when distinguishing poisoned captions. When the positive text is label prompts, FLAVA is the best or the second best model, probably due to its additional uni-modal training data/loss. UniCL is the best when single-label prompts are the positives, which we think can be explained by the ImageNet21K classification dataset in its training data.

Table 4. Performance of fine-tuned CLIP-B models in MSCOCO-2017's image-to-text retrieval, measured by mean Average Precision.

Model	Cap^+			Prompt^+_m			Prompt^+_s			Cap^+_{ln}		
	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}	Cap^-_{rd}	Cap^-_{rl}	Cap^-_{er}
CLIP-B	94.78	82.77	28.1	74.39	50.12	7.19	55.69	30.17	4.47	**81.29**	**60.00**	13.57
$\text{CLIP-B}_{\text{ft-coco}}$	**96.98**	**88.15**	35.04	71.65	47.16	5.41	60.4	33.53	4.09	80.52	57.96	10.66
NegCLIP	96.6	87.37	51.88	65.32	39.91	6.7	61.32	34.52	6.09	76.70	53.93	13.33
Ours	96.21	85.9	**75.74**	**93.37**	**75.46**	**55.95**	**83.22**	**57.98**	**29.05**	78.54	55.11	**31.74**

6. Limitations of Fine-Tuning VLMs with Hard Samples

In our study, we generate both hard positive and negative text samples to examine whether vision-and-language models (VLMs) accurately interpret text that aptly describes images. In this section, we explore the efficacy of using these hard text samples for training or fine-tuning VLMs. Drawing inspiration from NegCLIP in a recent study [33], which delves into understanding compositional relationships in VLMs, we fine-tune the CLIP-B model using these challenging samples on MSCOCO training data. Our approach differs from NegCLIP in that we focus exclusively on hard text samples for fine-tuning. These include complex positives such as single/multi-label prompts and difficult negatives like captions from relevant images or true captions subtly modified with errors, forming part of our benchmarking strategy. Notably, NegCLIP also uses modified true captions with errors to test compositional understanding.

Following the default fine-tuning hyperparameters cited in [33], we additionally fine-tune CLIP-B with original MSCOCO captions (CLIP-Bft-coco) to establish a fairer baseline. The performance of our models, alongside NegCLIP, is detailed in Table 4. Our findings indicate that fine-tuning with original MSCOCO data without hard samples shows superior performance in distinguishing true captions from simpler negative ones, such as Cap^+ vs Cap^-_{rd} or Cap^-_{rl}. NegCLIP surpasses the COCO-fine-tuned CLIP only in tasks aligned with its data augmentation strategy, e.g., Cap^+ vs Cap^-_{er}, but shows similar or even reduced effectiveness in other tasks. Importantly, our fine-tuned model demonstrates superior performance compared to the baselines and NegCLIP in the most challenging scenarios, Prompt^+_s vs Cap^-_{er}, owing to a broader coverage of harder text sample types.

Despite these enhancements, challenges remain. NegCLIP and our fine-tuned model struggle in complex scenarios like single/multi-label prompts (Prompt^+_s and Prompt^+_m) against subtly altered captions Cap^-_{er}. For tasks with unseen hard text types, like long captions Cap^+_{ln} vs Cap^-_{rd} and Cap^-_{rl}, all fine-tuned models underperform compared to the original CLIP. This indicates that while fine-tuning with hard samples enhances performance on familiar challenging cases, it falls short in addressing the full spectrum of difficulties, especially when encountering cases outside our augmentation strategy. These findings underscore the limitations of solely relying on fine-tuning with hard samples and highlight the urgency for a more comprehensive solution capable of encompassing a wider variety of potential scenarios, as our current methods do not completely resolve the issue.

7. Conclusion and Discussion

As interest in vision-language models grows, we present a novel benchmark and comprehensive study on the behaviors that create challenges to be useful in the open-world settings. First, we demonstrate that VLMs perform inconsistently on concepts of different semantic granularity. Based on our experiments, the performance discrepancy is due to the biased distribution of training data. Second, we show that vision language scores mostly measure similarity rather than correctness and can be distracted by the specificity of text. The scores are higher when the specificity of text are closer to the captions in the training data. This issue cannot be systematically solved by fine-tuning wit hard text mining.

While our study doesn't offer complete solutions for the identified issues, we suggest several promising avenues for improvement. Firstly, addressing the granularity discrepancy and specificity sensitivity could involve enhancing the training data. This can be achieved by augmenting text with a more balanced concept distribution and incorporating hard negatives and positives, possibly using large language models for assistance. Secondly, the dual encoder and embedding similarity approach inherently complicates correct recognition. For instance, true captions and slightly erroneous ones may have similar embeddings from a uni-modal view, resulting in close scores with identical image embeddings. A more advanced cross-modality fusion module could be key in discerning between visual and textual features, enabling distinct outputs for similar textual inputs. Lastly, large language models (LLMs) trained on more diverse text data might help mitigate the challenges we've noted. Our language-only experiment in the Appendix illustrates the potential of using generative LLMs in this context. Exploring and evaluating VLMs integrated with generative LLMs, e.g. vision-LLM, for recognition tasks represents an exciting future direction.

References

[1] Xi Chen et al. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1, 2

[2] Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *arXiv preprint arXiv:2202.04053*, 2022. 2

[3] Giannis Daras and Alexandros G Dimakis. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169*, 2022. 2

[4] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In *ICML*, 2022. 2

[5] Stanislav Fort. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations. 2021. 2

[6] Yuri Galindo and Fabio A. Faria. Understanding clip robustness. 2021.

[7] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. 2

[8] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-Vocabulary Detection via Vision and Language Knowledge Distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

[9] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*, 2021. 1, 2

[11] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022. 1

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[14] David Noever and Samantha E. Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *ArXiv*, abs/2103.10480, 2021. 2

[15] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[16] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 6

[17] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Are Multimodal Models Robust to Image and Text Perturbations? *arXiv preprint arXiv:2212.08044*, 2022. 2

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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 *ICML*, 2021. 1, 2, 3

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[20] Madeline Chantry Schiappa, Shruti Vyas, Hamid Palangi, Yogesh S Rawat, and Vibhav Vineet. Robustness Analysis of Video-Language Models Against Visual and Language Perturbations. In *NeurIPS Datasets and Benchmarks Track*, 2022. 2

[21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu andCade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[24] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. In *NeurIPS*, 2022. 2, 3

[25] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 3

[26] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[27] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A New Paradigm for Video Action Recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[28] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 2

[29] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic Segmentation Emerges from Text Supervision. In *CVPR*, 2022. 1, 2

[30] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space, 2022. 3

[31] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive Captioners are Image-Text Foundation Models. In *CVPR*, 2022. 1, 2

[32] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A New Foundation Model for Computer Vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[33] 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? *arXiv preprint arXiv:2210.01936*, 2022. 2, 8

[34] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can Language Understand Depth? In *ACM MM*, 2022. 2

[35] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*, 2022. 1, 2