Delving into the Openness of CLIP

Shuhuai Ren Lei Li Xuancheng Ren Guangxiang Zhao Xu Sun
National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{shuhuai_ren,lilei}@stu.pku.edu.cn
{renxc,zhaoguangxiang,xusun}@pku.edu.cn

Abstract

Contrastive Language-Image Pre-training (CLIP) formulates image classification as an image-to-text matching task, i.e., matching images to the corresponding natural language descriptions instead of discrete category IDs. This allows for open-vocabulary visual recognition, where the model can recognize images from an open class set (also known as an open vocabulary) in a zero-shot manner. However, evaluating the openness of CLIP-like models is challenging, as the models are open to arbitrary vocabulary in theory, but their accuracy varies in practice. To address this, we resort to an incremental perspective to assess the openness through vocabulary expansions, and define *extensibility* to measure a model's ability to handle novel classes. Our evaluation shows that CLIP-like models are not truly open, and their performance deteriorates as the vocabulary expands. We further dissect the feature space of CLIP from the perspectives of representation alignment and uniformity. Our investigation reveals that the overestimation of openness is due to confusion among competing text features, rather than a failure to capture the similarity between image features and text features of novel classes. We hope that our investigation and analysis will facilitate future research on the CLIP openness issue.[1]

1 Introduction

An intrinsically open mechanism for visual recognition (Deng et al., 2009; He et al., 2016) has always been a shared goal in the computer vision community (Scheirer et al., 2013; Geng et al., 2021; Bendale and Boult, 2015). This mechanism requires models to maintain flexibility to cope with the scaling of the recognition target, where both input images and the corresponding classes will dynamically expand according to actual needs. For example, in medical diagnosis (Razzak et al., 2017), new diseases emerge constantly, and in e-commerce, new categories of products appear daily (Xu et al., 2019), which cannot be predefined in a finite, fixed class set.

Faced with the challenging task of open-world recognition, Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and its open-vocabulary learning paradigm demonstrate superiority over traditional supervised classifiers (He et al., 2016; Dosovitskiy et al., 2021). CLIP pre-trains a vision-language model on web-scale collections of image-text pairs, learning semantic alignment between images and corresponding textual descriptions. During inference, it formulates image classification as an image-to-text matching task, where the set of class names serves as a vocabulary, and textual prompts like "`a photo of a [CLASSNAME]`" are curated as class descriptions for images. By varying the `[CLASSNAME]` placeholder and computing the similarity between class descriptions and images, CLIP can identity the most suitable class name and predict it as the target class. This approach allows CLIP to operate with arbitrary vocabularies and adapt to novel classes by expanding the vocabulary, enabling zero-shot inference for new input images and classes.

Nevertheless, previous evaluation protocols for CLIP models only assess their accuracy on static, closed vocabularies from downstream datasets, leaving their actual performance on open tasks in the shadows (Radford et al., 2021). In this work, we delve into openness, the intriguing yet under-explored property in CLIP-like models (Li et al., 2021b; Mu et al., 2021; Yao et al., 2021; Zhou et al., 2021), and present a novel protocol for evaluating the openness from an incremental view. Specifically, we define a metric of **extensibility** to measure a model's ability to handle new visual concepts through vocabulary expansion. Different from previous metrics, our metric explicitly models the dynamics of the real open world, and formulates the

[1]Our code is available at https://github.com/lancopku/clip-openness

Findings of the Association for Computational Linguistics: ACL 2023, pages 9587–9606

empirical risk of CLIP when new vocabularies incrementally emerge. Additionally, we define a metric of **stability** to explore how stable the model's predictions are for old classes when new classes are introduced, which provides a tool to analyze the compatibility between different classes.

Using our protocol, we conduct a systematic and comprehensive evaluation of CLIP-like models. Our experimental results based on extensibility show that CLIP and its variants have a significant drop in accuracy as the vocabulary size increases. For example, CLIP (RN101) on CIFAR100 experiences a 12.9% drop in accuracy when the vocabulary size expands from 5 to 100. This indicates that the limited zero-shot capability of CLIP-like models is inadequate for supporting their deployment in the open world. What's worse, through an analysis of the prediction shift during vocabulary expansion, we find that the performance of CLIP can be dramatically reduced by adding only three adversarial class names into the vocabulary, exposing the model's poor stability and security risks. Furthermore, we investigate the representation space of CLIP-like models via three metrics: margin, inter-modal alignment, and intra-modal uniformity. Our results show that the small margin between positive and negative class descriptions leads to prediction shifting when competing class features appear. Therefore, enforcing the distinguishability of class features increases the margin and improves the stability of these models.

In summary, our contribution is threefold: **First**, to the best of our knowledge, we are the first to systematically quantify the openness of CLIP, for which we design the evaluation protocol and two indicators of extensibility and stability. **Second**, we conduct extensive experiments on CLIP-like models based on our protocol and find that their openness is overestimated and their performance declines as the vocabulary expands. **Finally**, we analyze the feature space of CLIP from the perspectives of representation alignment and uniformity, observing that the uniformity of the textual space is critical for better extensibility.

2 Related work

Contrastive language-image pre-training and open-vocabulary learning. CLIP (Radford et al., 2021) introduces the paradigm of open-vocabulary learning and learns transferable visual models from natural language supervision. The CLIP model consists of an image encoder and a text encoder, which are utilized to encode image-text pairs into a joint feature space for learning the semantic alignment of vision and language. The paired images and texts are pulled together in the feature space, while the others with dissimilar semantics are pushed apart via a contrastive loss. After pre-training on large-scale image-text pairs, CLIP is able to map images to their corresponding language descriptions, which makes visual recognition generalize in the wild. Recent studies further improve CLIP by using more pre-training data (Jia et al., 2021), incorporating self-supervision (Mu et al., 2021), fine-grained supervision (Yao et al., 2021), and widespread supervision (Li et al., 2021b) to pre-training. Another line of recent studies (Li et al., 2021a; Wang et al., 2022; Yu et al., 2022; Alayrac et al., 2022) adopts seq2seq generation instead of contrastive discrimination framework to achieve open-vocabulary recognition. We leave the investigation of their extensibility for future work.

Open Set and Open-World Visual Recognition. Open Set Recognition (OSR) (Scheirer et al., 2013; Geng et al., 2021) and Open World Recognition (OWR) (Bendale and Boult, 2015) are paradigms aiming to cope with input images from novel classes during inference. OSR requires classifiers to identify images that have not been introduced during training as "unknown". While OWR raises higher demands, models are supposed to incrementally extend and retrain the multi-class classifier as the unknowns are labeled as additional training data. Contrary to the above research, CLIP-based Open-vocabulary Recognition (OVR) aims to identify novel classes in a zero-shot manner by using natural language representations of categories instead of discrete label IDs. This allows CLIP to directly synthesize textual descriptions of novel classes for matching, eliminating the need for relabeling additional training data and re-training the entire model. A more detailed comparison of OSR, OWR, and OVR can be found in Appendix A.1.

3 Openness, Extensibility, and Stability

In this section, we first review CLIP's visual recognition paradigm and demonstrate how it realizes open-vocabulary image classification through vocabulary expansion (§ 3.1). To quantify the actual performance of CLIP-like models as the vocabulary expands, we define the metric of extensibility and propose a systematical evaluation protocol (§ 3.2).

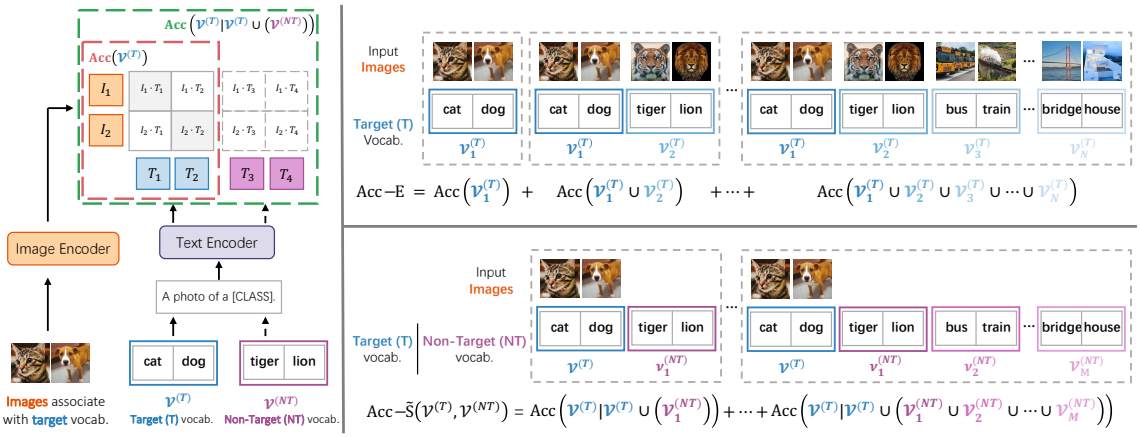


Figure 1: **Left**: the original accuracy of CLIP with target vocabulary (Eq.(1)) and the conditional accuracy of CLIP with non-target vocabulary (Eq.(4)). In the latter, the classes from the non-target vocabulary are involved as distractors for input images restricted in the target vocabulary. **Upper right**: calculation of Acc-E (Eq.(2)). It measures the extensibility of models when recognition targets, including both classes and the associated input images, are scaling simultaneously. **Bottom right**: calculation of Acc-S (Eq.(5)), a sub-problem introduced by Acc-E. It measures the prediction stability on the images from the target vocabulary as the distractors from the non-target vocabularies are incorporated incrementally.

The experimental results and further analysis reveal that, as the vocabulary expands, CLIP's predictions become unstable and prone to drift towards newly introduced competing class descriptions, which limits its extensibility and poses a huge security risk when deployed in real-world applications (§ 3.3).

3.1 Openness of CLIP

CLIP (Radford et al., 2021) models image classification as an image-to-text matching task. Formally, let f be the CLIP model, f_T and f_I be the text and image encoders in CLIP, respectively. The CLIP model takes an image x and a *target vocabulary* $\mathcal{V}^{(T)} = \{w_i\}$ of the class names w_i as inputs, and predicts the image label as:

$$f\left(x, \mathcal{V}^{(T)}\right) = \arg\max_i P\left(y = i \mid x\right) = \arg\max_i \frac{e^{\mathrm{sim}(f_T(t_i), f_I(x))}}{\sum_{j=1}^{|\mathcal{V}^{(T)}|} e^{\mathrm{sim}(f_T(t_j), f_I(x))}},$$

where t_i is the textual description of the class name w_i in a prompt format, e.g., "`a photo of a` w_i", and $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. Such a modeling paradigm can realize open-world image classification in theory by extending the target vocabulary $\mathcal{V}^{(T)}$ to arbitrary degrees. However, in most previous work (Radford et al., 2021; Li et al., 2021b; Mu et al., 2021; Yao et al., 2021; Zhou et al., 2021), CLIP is evaluated with a fixed vocabulary $\mathcal{V}^{(T)}$ of a downstream dataset $\mathcal{D}^{(T)}$:

$$\mathrm{Acc}\left(\mathcal{V}^{(T)}\right) = \frac{1}{|\mathcal{D}^{(T)}|} \sum_{(x,y)\in\mathcal{D}^{(T)}} \mathbb{I}\Big(f\left(x, \mathcal{V}^{(T)}\right) = y\Big), \tag{1}$$

where $|\mathcal{D}^{(T)}|$ is the size of the dataset and $\mathbb{I}(\cdot)$ is the indicator function. This vanilla evaluation setting, utilizing restricted input images and classes, falls short for open recognition tasks. It fails to consider the dynamic expansion of vocabulary during inference and, as a result, cannot accurately reflect CLIP's openness in real-world scenarios where the number of classes may increase.

3.2 Quantifying extensibility for open world

To quantify the model's capability in dealing with newly emerged recognition targets, we propose an evaluation protocol and define a metric of extensibility based on vocabulary expansion. Concretely, we incrementally expand the vocabulary $\mathcal{V}^{(T)}$ in Eq.(1) by introducing new classes and their associated input images, then evaluate the accuracy after each expansion. These accuracy values reflect the model's dynamic performance as openness increases, and the expected average of these values is defined as the model's extensibility. In practice, we achieve this expansion by incrementally unioning N disjoint target vocabularies[2] as shown in the

[2]Since $\mathcal{V}^{(T)}$ is bound with $\mathcal{D}^{(T)}$ in Eq.(1), expanding the target vocabulary also implies expanding $\mathcal{D}^{(T)}$ (including input images and their labels) at the same time, which we omit for brevity.

Model	CIFAR100					ImageNet (Entity13)					ImageNet (Living17)				
	Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability	
		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ
CLIP (RN101)	68.3	55.4	-12.9	54.9	-13.4	80.4	77.4	-3.0	77.3	-3.1	77.6	74.5	-3.1	74.4	-3.2
CLIP (ViT-B/32)	78.0	69.6	-8.4	68.9	-9.1	80.8	78.0	-2.8	77.8	-3.0	78.0	74.4	-3.6	75.0	-3.0
CLIP (ViT-B/16)	**79.7**	**72.6**	**-7.1**	**72.0**	**-7.7**	**83.5**	**81.1**	**-2.4**	**81.0**	**-2.5**	**79.5**	**77.9**	**-1.6**	**77.6**	**-1.9**
SLIP (ViT-B/16)	63.9	51.1	-12.8	50.4	-13.5	65.7	62.3	-3.4	62.0	-3.7	65.7	62.6	-3.1	62.5	-3.2
DeCLIP (ViT-B/32)	**78.7**	**70.8**	**-7.9**	**70.4**	**-8.3**	**81.9**	**79.2**	**-2.7**	**79.1**	**-2.8**	**82.1**	**80.2**	**-1.9**	**80.0**	**-2.1**
PE (ViT-B/32)	78.3	70.3	-8.0	69.9	-8.4	81.9	79.4	-2.5	79.2	-2.7	78.7	76.0	-2.7	75.8	-2.9
PE (ViT-B/16)	**79.6**	**72.6**	**-7.0**	**72.0**	**-7.6**	**85.3**	**83.2**	**-2.1**	**83.1**	**-2.2**	**79.6**	**78.2**	**-1.4**	**78.0**	**-1.6**
CoOp (ViT-B/16)	83.6	76.9	-6.7	76.7	-6.9	87.5	85.3	-2.2	85.5	-2.0	82.7	82.6	-0.1	81.3	-1.4

Table 1: Extensibility and stability of CLIP-like models on CIFAR100 and ImageNet datasets. Δ refers to the decline of Acc-E/Acc-S (%) compared to Acc-C (%). All models exhibit a clear drop in performance as the openness of tasks increases. PE denotes Prompt Ensemble. CoOp requires fine-tuning with the additional training data in downstream datasets (16-shot for all classes), which can be viewed as the upper bound of other zero-shot models.

upper right panel of Figure 1.

Definition 3.1 (Extensibility)**.** Given N disjoint target vocabularies $\{\mathcal{V}_1^{(T)}, \cdots, \mathcal{V}_N^{(T)}\}$, we denote the set of all possible permutations of these vocabularies as $\mathcal{S}_N$, and $\mathcal{V}_{s_i}^{(T)}$ as the $i^{(th)}$ vocabulary in a permutation $s \in \mathcal{S}_N$. When we union the $i^{(th)}$ vocabulary with the previous $i-1$ vocabularies, we achieve a vocabulary expansion and obtain $\mathcal{V}_{s_1}^{(T)} \cup \cdots \cup \mathcal{V}_{s_i}^{(T)}$. The extensibility refers to the averaged classification accuracy across N incremental expansions as i increases from 1 to N:

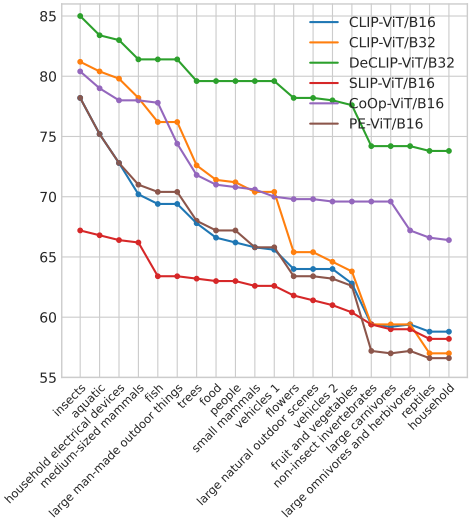
$$\text{Acc-E} = \mathop{\mathbb{E}}_{s \in \mathcal{S}_N} \frac{1}{N} \sum_{i=1}^{N} \text{Acc}\left(\mathcal{V}_{s_1}^{(T)} \cup \cdots \cup \mathcal{V}_{s_i}^{(T)}\right). \tag{2}$$

Experimental settings We evaluate the extensibility of CLIP and its variants, including DeCLIP (Li et al., 2021b), SLIP (Mu et al., 2021), Prompt Ensemble (Radford et al., 2021), CoOp (Zhou et al., 2021), on the CIFAR100 (Krizhevsky and Hinton, 2009) and ImageNet (Deng et al., 2009) datasets. Non-matching methods (Gao et al., 2021; Zhang et al., 2021; Wortsman et al., 2021), such as linear probing, are NOT included since they train a classifier with *finite* class vectors, and thus are not suitable for class scaling in operation. To construct the vocabulary, we leverage the underlying superclass-class hierarchical structure of the two datasets (Krizhevsky and Hinton, 2009; Santurkar et al., 2021) by grouping classes that belong to the same superclass into a vocabulary. Accordingly, CIFAR100 has 20 vocabularies, each with 5 classes. For ImageNet, we utilize two superclass-class structures (Santurkar et al., 2021): Entity13 and Living17. The former has 13 vocabularies, each with 20 classes, while the latter has 17 vocabularies, each with 4 classes. Tables in the Appendix A.2 list all the vocabularies in the two datasets. For each dataset, we calculate Acc-C, the averaged classification accuracy across all single vocabularies, based on Eq.(1):

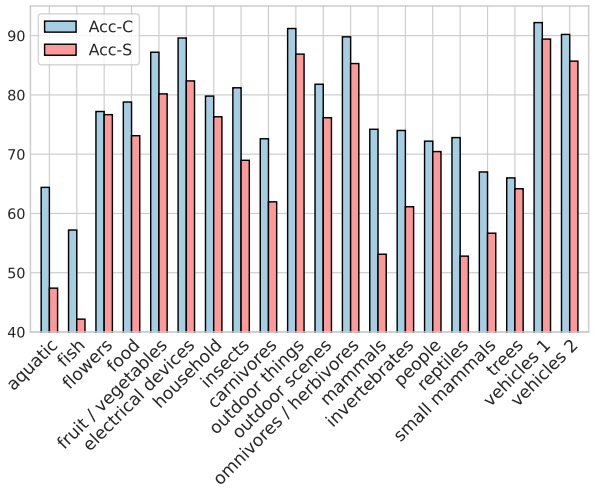
$$\text{Acc-C} = \frac{1}{N} \sum_{i=1}^{N} \text{Acc}\left(\mathcal{V}_i^{(T)}\right). \tag{3}$$

It represents the original model performance on *closed* vocabularies. To calculate the expectation in Acc-E, we sample $100 \times N$ permutations for N vocabularies and take the average.

Results As shown in Table 1, all models exhibit a clear drop in performance as the vocabulary expands. The accuracy of CLIP (RN101) after vocabulary expansion (Acc-E) sharply decreases by 12.9% compared to the accuracy on closed vocabulary (Acc-C). The performance on the data splits in ImageNet is relatively better, with an average decline of 2.7%. Appendix A.3 provides results of expansion at the dataset level, where the expanded vocabularies are from five other datasets. These results show a more dramatic decline of an average of 15.3% on generic dataset expansion. It demonstrate that **the openness of CLIP-like models is overestimated under the vanilla evaluation mechanism**. Besides, there are some interesting findings: **(1)** From the perspective of pre-training, introducing a stronger vision backbone (ViT (Dosovitskiy et al., 2021) vs. ResNet (He et al., 2016)),



(a) Acc-S drops as non-target vocabulary extends (*Insects* as target vocabulary).



(b) Difference between Acc-C and Acc-S of CLIP (ViT-B/32) on different groups.

Figure 2: Acc-C and Acc-S (%) of CLIP and its variants on CIFAR100. The horizontal axis represents the extended non-target vocabularies in order. PE refers to Prompt Ensemble.

widespread supervision (DeCLIP (Li et al., 2021b) vs. CLIP), and more pre-training data (CLIP vs. SLIP (Mu et al., 2021)) can improve the extensibility of models on open tasks. **(2)** During inference, the performance of CLIP can be boosted by ensembling different prompts. **(3)** The most extensible results are obtained by CoOp (Zhou et al., 2021), which performs prompt tuning on all classes of CIFAR100 and ImageNet. However, the prompt tuning method utilizes the additional category information and training data, which cannot be applied to real-world open tasks.

3.3 Stability during vocabulary expansion

As the vocabulary expansion introduces new classes incrementally, some images belonging to previous vocabularies may be incorrectly predicted as new classes, resulting in a drop in accuracy and poor extensibility. To analyze the prediction stability of CLIP during vocabulary expansion, we introduce the *non-target classes*. They do NOT correspond to any input images, and only serving as distractors for the target classes. Based on it, we define conditional classification accuracy as:

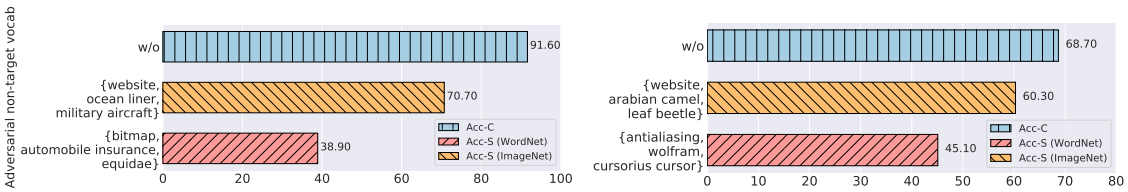
$$\begin{aligned}&\mathrm{Acc}\left(\mathcal{V}^{(T)}\middle|\mathcal{V}^{(T)}\cup\mathcal{V}^{(NT)}\right)\\&=\frac{1}{|\mathcal{D}^{(T)}|}\sum_{(x,y)\in\mathcal{D}^{(T)}}\mathbb{I}\left(f\left(x,\mathcal{V}^{(T)}\cup\mathcal{V}^{(NT)}\right)=y\right),\end{aligned}\tag{4}$$

where $\mathcal{V}^{(NT)}$ is the *non-target vocabulary*, i.e., the vocabulary of non-target classes. The conditional accuracy is depicted in the left panel of Figure 1. In Eq.(4), the categories of the input images are limited to the target vocabulary ($(x,y)\in\mathcal{D}^{(T)}$), but CLIP is asked to distinguish all categories from a larger vocabulary $\mathcal{V}^{(T)}\cup\mathcal{V}^{(NT)}$. In other words, compared to traditional closed-set classification, CLIP is expected to reject all the negative categories from $\mathcal{V}^{(NT)}$. The model is required to distinguish visual concepts stably and robustly, rather than making wrong predictions in the presence of other distractors. Based on Eq.(4), we define the stability of CLIP in the open task as:

Definition 3.2 (Stability)**.** Given a target vocabulary $\mathcal{V}^{(T)}$ and M non-target vocabularies $\{\mathcal{V}_1^{(NT)},\cdots,\mathcal{V}_M^{(NT)}\}$, we denote $\mathcal{S}_M$ as their full permutation, and $\mathcal{V}_{s_i}^{(NT)}$ as the $i^{(th)}$ vocabulary in a permutation $s\in\mathcal{S}_M$. We design the **local stability** to measure the averaged classification accuracy of CLIP on the given target vocabulary when non-target vocabularies are extended incrementally:

$$\begin{aligned}&\text{Acc-}\tilde{\text{S}}\left(\mathcal{V}^{(T)},\mathcal{V}^{(NT)}\right)=\\&\mathop{\mathbb{E}}_{s\in\mathcal{S}_M}\frac{1}{M}\sum_{i=1}^{M}\mathrm{Acc}\left(\mathcal{V}^{(T)}\middle|\mathcal{V}^{(T)}\cup\left(\mathcal{V}_{s_1}^{(NT)}\cup\cdots\cup\mathcal{V}_{s_i}^{(NT)}\right)\right).\end{aligned}\tag{5}$$

As Eq.(5) only reflects the local stability with respect to a single target vocabulary, we further design the **general stability** as an average of local stability over a set of target vocabularies to reduce the bias from data distribution and vocab-



(a) CIFAR10 (b) CIFAR100

Figure 3: Adversarial non-target vocabulary for CIFAR datasets. Adding 3 adversarial non-target classes leads to severe performance (Acc-S) deterioration, revealing the vulnerability of CLIP when faced with malicious vocabulary.

ulary sampling. Specifically, given N vocabularies $\{\mathcal{V}_1, \cdots, \mathcal{V}_N\}$, we regard each vocabulary $\mathcal{V}_i$ as the target vocabulary $\mathcal{V}^{(T)}$ and the rest $\mathcal{V}_{\neq i}$ as the non-target vocabularies $\mathcal{V}^{(NT)}$, and then formulate the general stability as:

$$\text{Acc-S} = \frac{1}{N}\sum_{i=1}^{N} \text{Acc-}\tilde{\text{S}}\left(\mathcal{V}_i, \mathcal{V}_{\neq i}\right). \tag{6}$$

Experimental settings and results The models and datasets adopted for evaluation are consistent with that in § 3.2. For the calculation of stability, take CIFAR100 with $N = 20$ vocabularies as an example, we treat each vocabulary as the target vocabulary and the rest are treated as the non-target vocabularies. To calculate the expectation in Eq.(5), we sample 100 permutations for $M = 19$ non-target vocabularies and report the averaged scores.

Table 1 demonstrates the stability of CLIP-like models. On CIFAR100, the Acc-S of CLIP (RN101) decreased by 13.4%. Figure 2a shows Acc-S on CIFAR100 during non-target vocabulary expansion. Given a closed $\mathcal{V}^{(T)} =$ *Insects*, CLIP (ViT-B/32) achieves an accuracy of 81.2%. However, when the remaining 19 non-target vocabularies are incorporated, the accuracy sharply drops to 57.0%. The decrease of Acc-S brought by the introduction of each non-target vocabulary indicates that more images from *Insects* are incorrectly classified into the new vocabulary. Figure 2b demonstrates the difference between Acc-C and Acc-S for each target vocabulary. When $\mathcal{V}^{(T)} =$ *Medium-sized Mammals*, CLIP is most easily interfered with by the non-target vocabularies, with a 21.08% performance drop. It suggests that **the unstable predictions lead to the poor extensibility of CLIP when new categories are introduced**. Besides, we notice that CLIP performs stably on groups like *Flowers*, where its Acc-S only declines by 0.53% compared to Acc-C. The different behaviors of different groups indicates that **the stability is also influenced by the inherent property of the image categories and naming variation** (Silberer et al., 2020; Takmaz et al., 2022).

3.3.1 Adversarial non-target vocabulary

In order to explore the lower bound of the stability of CLIP, we define the *adversarial non-target vocabulary* $\mathcal{V}^{(ANT)}$ as the non-target vocabulary that reduces Acc-S the most:

$$\mathcal{V}^{(ANT)} = \min_{\mathcal{V}^{(NT)}} \text{Acc}\left(\mathcal{V}^{(T)} \middle| \mathcal{V}^{(T)} \cup \mathcal{V}^{(NT)}\right). \tag{7}$$

To build $\mathcal{V}^{(ANT)}$, we refer to the method of adversarial examples generation (Ren et al., 2019) to traverse the words in a large vocabulary, e.g., the vocabulary of nouns in WordNet (Fellbaum, 2000), which are regarded as non-target classes in order to calculate Acc-S, and then take the most confusing words to form the adversarial non-target vocabulary.

We constrain the size of $\mathcal{V}^{(ANT)}$ to 3. Results in Figure 3 illustrate the performance with nouns in WordNet and class names in ImageNet as the candidate vocabulary, respectively. First, we observe a clear performance degradation on both datasets under adversarial attack, e.g., adding *bitmap*, *automobile insurance* and *equidae* leads to an absolute 52.7% accuracy drop on CIFAR10. Besides, we find that the selected adversarial words are much less concrete than common visual concepts like *Flower*, indicating the potential reason behind is the poor semantic modeling of CLIP on those objects with higher abstraction levels. This investigation reveals that **CLIP is vulnerable when facing malicious non-target vocabulary**, and we hope future work may pay more attention to the robustness of CLIP under open recognition tasks.

4 Dissecting the extensibility of CLIP

Our experimental results in § 3 reveal the poor performance of CLIP on open tasks. In this section,

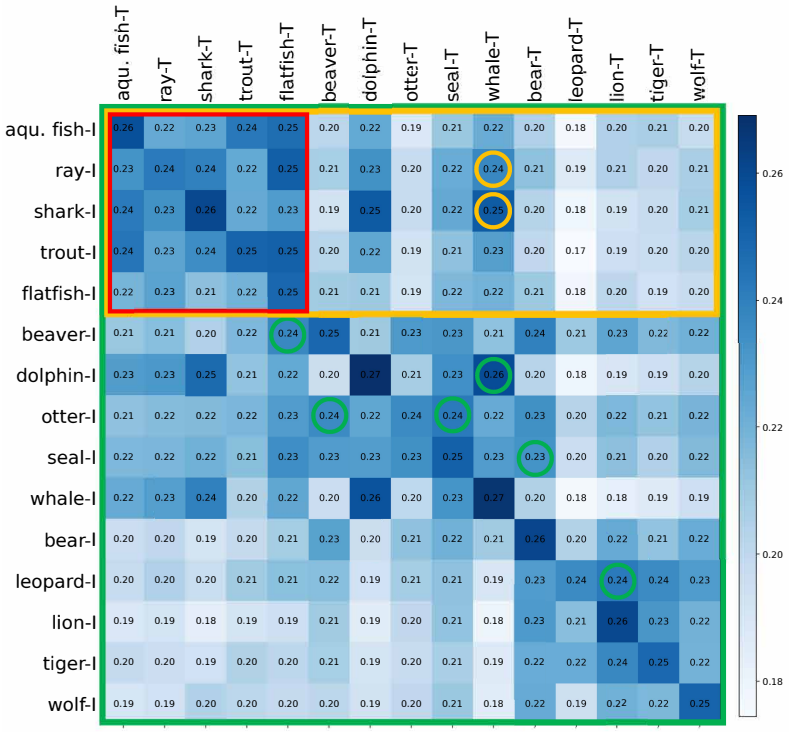


Figure 4: Cosine similarity between image (-I) and text (-T) features of CLIP on CIFAR100. Each value in the matrix are averaged over 100 samples. The expansions from the red box to the green box (diagonal) and the yellow box (horizontal) refer to the calculation of extensibility and stability, respectively. The circle represents that more than 15 wrong predictions have arisen after adding this class.

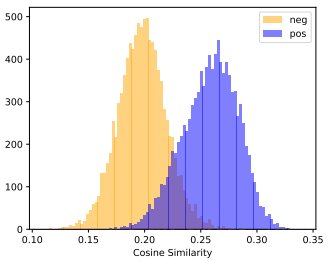


Figure 5: Cosine similarity histogram of positive (pos) and negative (neg) image-text pairs with large overlap.

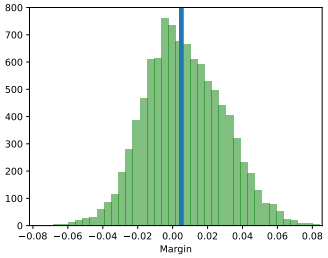


Figure 6: Margin distribution of similarity scores, which are centered around zero with a median value of .005 (the blue vertical line). It indicates that the predictions can be easily inverted with competing classes appearing.

we delve into the representation space of CLIP to understand its extensibility. We first point out that the small margin between positive and negative class descriptions leads to the prediction shifting when competing class features appear, which thus limits the stability of CLIP (§ 4.1). Further, we investigate the representation space of CLIP-like models via two metrics: inter-modal alignment and intra-modal uniformity. The results show that enforcing the distinguishability of class features increases the margin and makes the models scale more stably (§ 4.2).

4.1 Small margin limits the stability of CLIP

Since CLIP formalizes the visual recognition as an image-to-text matching task, each text feature of the class description corresponds to the class vector in traditional classifiers, and the image-text similarity scores are analogous to the logits in classification. Ideally, regardless of vocabulary expansion, for an image, the similarity of the positive pair (the image with the text specifying the ground-truth class) should be higher than that of the negative pairs (the image with the texts specifying other classes) to ensure the correct prediction on open tasks. In other words, the *margin* (Jiang et al., 2019) between positive and the largest negative similarity is a direct contributor to stability.

Unfortunately, the similarity and margin distribution of CLIP do not meet our expectations. Figure 4 illustrates the averaged cosine similarity of CLIP (ViT-B/32) on 15 classes of CIFAR100. The diagonal elements represent the similarity of the positive image-text pairs, while the others represent that of the negative ones. In general, the cosine similarity of image-text pairs is very low, with an average of 0.20. This number is only 0.26 even for the positive pairs. Besides, the similarities of positive and negative pairs are very close, indicating the low distinguishability between different classes. As shown in Figure 5 and Figure 6, the similarity histogram of positive and negative pairs has a large overlap, and the margin is clustered around zero, leaving the predictions of models at risk of being reversed to new non-target classes. For example, as the vocabulary extends from the red box to the green box (diagonal) or the yellow box (horizontal) in Figure 4, more deceptive classes (circles)

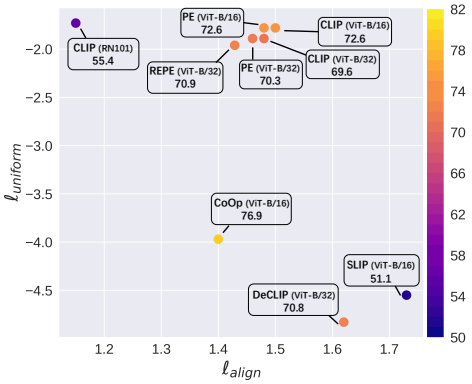


Figure 7: ℓ_{align} and ℓ_{uniform} of CLIP-like models. For both two metrics, lower numbers are better. The color of points and numbers denote the extensibility performance (Acc-E) on CIFAR100 (higher is better).

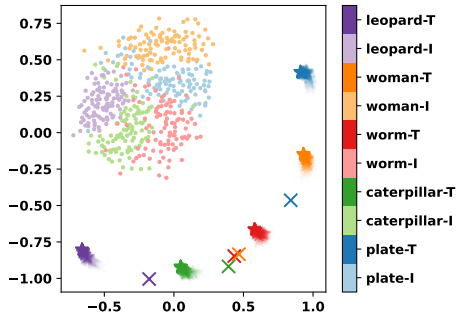


Figure 8: Representation visualization of CLIP and CoOp (ViT-B/16). The five classes with different colors are from CIFAR100. • refers to image features (-I), while $\times$ and $\star$ refers to text features (-T) of CLIP and CoOp, respectively. The color of $\star$ from transparent to opaque indicates the optimization trajectory during the CoOp prompt-tuning process.

with negative margins are added, leading to prediction shift. Particularly, the classes belonging to the same vocabulary[3] have higher similarity and smaller margin, making them more likely to be confused with each other.

4.2 Inter-modal alignment and intra-modal uniformity ground the margin

According to the results in § 4.1, the ideal feature space for CLIP-like models should have a large margin between different classes to ensure stability in open-vocabulary recognition tasks. To achieve this, the text feature of a class name should be close to the features of the images it describes (Ren et al., 2021), and the intra-modal features, especially textual features, should be uniformly distributed to make the descriptions of competing categories more distinguishable (Wang and Isola, 2020). In order to measure the quality of representations in the vision-and-language domain, we propose two metrics, **inter-modal alignment** and **intra-modal uniformity**. Inter-modal alignment calculates the expected distance between features of positive image-text pairs p_{pos} :

$$\ell_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(x,t)\sim p_{\text{pos}}} \|f_I(x) - f_T(t)\|^2, \quad (8)$$

while intra-modal uniformity measures how well the image or text features are uniformly distributed:

$$\begin{aligned}\ell_{\text{uniform}} &\triangleq \ell_{\text{uniform-I}} + \ell_{\text{uniform-T}} \\ &\triangleq \log \mathop{\mathbb{E}}_{x_i,x_j\sim p_{\text{data-I}}} e^{-2\|f_I(x_i)-f_I(x_j)\|^2} + \\ &\quad \log \mathop{\mathbb{E}}_{t_i,t_j\sim p_{\text{data-T}}} e^{-2\|f_T(t_i)-f_T(t_j)\|^2},\end{aligned} \quad (9)$$

where $p_{\text{data-I}}$ and $p_{\text{data-T}}$ denotes the image and text data distribution, respectively. Figure 7 and Table 2 provide quantified loss of alignment and uniformity. CLIP with only cross-modal contrastive learning results in poor intra-modal uniformity ($\ell_{\text{uniform}} > -2.0$), especially on the text side. However, models like SLIP and DeCLIP that incorporate intra-modal contrastive learning in pre-training can better separate image and text features by classes, resulting in a much lower intra-modal uniformity loss ($\ell_{\text{uniform}} < -4.5$). Additionally, the prompt tuning method (CoOp (Zhou et al., 2021)) achieves better inter-modal alignment and the lowest intra-modal uniformity loss on the text side. According to the visualization via Multidimensional Scaling (MDS) (Borg and Groenen, 1997) in Figure 8, the optimization trajectory of prompts in CoOp leads to the cluster center of corresponding image features while also dispersing the position of prompt features, thereby improving both text uniformity and inter-modal alignment and achieving the best extensibility.

4.3 Discussions

After the preliminary explorations on openness of CLIP-like models, we present potential ways to enhance the models's extensibility and stability.

(1) For pre-training: In order to improve the quality of CLIP's feature space and enhance alignment and uniformity, more high-quality pre-training data and effective supervision signals such as ℓ_{align} and ℓ_{uniform} can be introduced during pre-training.

(2) For zero-shot inference: Recall that in vanilla CLIP-like models, the context (hard prompt) for

[3]Every 5 adjacent classes in Figure 4 constitute a vocabulary (superclass), see Table 4 in Appendix A.2

Model	Alignment & Uniformity				Accuracy	
	ℓ_{align} (↓)	$\ell_{\text{uniform-T}}$ (↓)	$\ell_{\text{uniform-I}}$ (↓)	ℓ_{uniform} (↓)	Acc-C (↑)	Acc-E (↑)
CLIP (RN101)	**1.15**	**-1.16**	-0.57	-1.73	68.3	55.4
CLIP (ViT-B/32)	1.48	-0.96	**-0.93**	**-1.89**	78.0	69.6
CLIP (ViT-B/16)	1.50	-0.97	-0.81	-1.78	**79.7**	**72.6**
SLIP (ViT-B/16)	1.73	-2.86	-1.69	-4.55	63.9	51.1
DeCLIP (ViT-B/32)	**1.62**	**-2.96**	**-1.87**	**-4.83**	**78.7**	**70.8**
PE (ViT-B/32)	**1.46**	-0.96	**-0.93**	**-1.89**	78.3	70.3
PE (ViT-B/16)	1.48	**-0.97**	-0.81	-1.78	**79.6**	**72.6**
CoOp (ViT-B/16)	1.40	-3.16	-0.81	-3.97	83.6	76.9

Table 2: Inter-modal alignment (ℓ_{align}), text uniformity ($\ell_{\text{uniform-T}}$), image uniformity ($\ell_{\text{uniform-I}}$), intra-modal uniformity (ℓ_{uniform}), Acc-C (Eq. (3)), and Acc-E (Eq. (2)) of CLIP-like models on CIFAR100. For the first four metrics, lower numbers are better. For the last two metrics, higher numbers are better.

each class name is the same during inference, making it difficult to discriminate between distinct visual categories because the semantics of each cannot be holistically represented. To remedy this, we suggest customizing class descriptions with diverse captions retrieved from the pre-training corpus as a prompt ensemble. The effectiveness of this idea is verified through experiments, details can be found in Appendix A.5.

5 Conclusion

In this paper, we evaluate the extensibility of CLIP-like models for open-vocabulary visual recognition. Our comprehensive study reveals that as the vocabulary expands, the performance of these models deteriorates significantly due to indistinguishable text features among competing classes. We hope that our investigation and analysis will facilitate future research on the CLIP openness issue.

Limitations

To facilitate future research, we analyze the difficulties and possible solutions in this new area. **(1)** As we present extensive empirical results and address the weakness of CLIP on vocabulary expansion, its theoretical risk on open tasks is urged to be investigated. **(2)** The current evaluation protocol is an approximation of the real open world. An evolving benchmark could facilitate future research. **(3)** For various visual categories, their degree of abstraction, the ease of describing them in natural language, and their density in the data distribution can also influence the extensibility and stability of models, which are worth studying.

Acknowledgement

The authors would like to thank the reviewers for their helpful comments. This work is supported by Natural Science Foundation of China (NSFC) No. 62176002. Xu Sun is the corresponding author.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Abhijit Bendale and Terrance E. Boult. 2015. Towards open world recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1893–1902. IEEE Computer Society.

Ingwer Borg and Patrick J. F. Groenen. 1997. Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 40:277–280.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV*.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*.

Christiane D. Fellbaum. 2000. Wordnet : an electronic lexical database. *Language*, 76:706.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *ArXiv*, abs/2110.04544.

Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. 2021. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3614–3631.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. 2019. Predicting the generalization gap in deep networks with margin distributions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561.

A. Krizhevsky and G. Hinton. 2009. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021b. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *ArXiv*, abs/2110.05208.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. *ArXiv*, abs/2112.12750.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society.

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

Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. 2017. Deep learning for medical image processing: Overview, challenges and future. *ArXiv*, abs/1704.06825.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. 2021. Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 514–524, Online. Association for Computational Linguistics.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. 2021. BREEDS: benchmarks for subpopulation shift. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1757–1772.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020. Object naming in language and vision: A survey and a new dataset. In *International Conference on Language Resources and Evaluation*.

Ece Takmaz, Sandro Pezzelle, and R. Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via clip. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *ArXiv*, abs/2109.01903.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Open-world learning and application to product classification. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3413–3419. ACM.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Jiao Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ArXiv*, abs/2111.03930.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *ArXiv*, abs/2109.01134.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. *ArXiv*, abs/2203.05557.

A Appendix

A.1 Comparison of related work

Table 3 provides a more detailed comparison of Closed Set Recognition (OSR) (Scheirer et al., 2013; Geng et al., 2021), Open World Recognition (OWR) (Bendale and Boult, 2015), and Open-vocabulary Recognition (OVR) (Radford et al., 2021) from 5 perspectives of paradigm, goal, signal, classes type in training, and classes type in testing, respectively. Contrary to the above research, CLIP-based OVR aims to identify novel classes in a zero-shot way. Since categories of images in CLIP are represented by natural language rather than discrete label IDs, CLIP can directly synthesize textual descriptions of novel classes for matching, sparing relabeling additional training data and re-training the entire model.

A.2 Superclass-class hierarchy for vocabulary construction

To construct the vocabularies in § 3, we leverage the underlying superclass-class hierarchical structure of CIFAR100 (Krizhevsky and Hinton, 2009) and ImageNet (Deng et al., 2009), and group the classes belonging to the same superclass into a vocabulary. Table 4 lists the vocabularies in CIFAR100, which are specified by (Krizhevsky and Hinton, 2009). There are 20 vocabularies, each with 5 classes. For ImageNet, we utilize two superclass-class structures, Entity13 and Living17 (Santurkar et al., 2021), as shown in Table 5 and Table 6, respectively. Entity13 has 13 vocabularies, each with 20 classes, while Living17 has 17 vocabularies, each with 4 classes.

A.3 Dataset-level extensibility

The evaluation protocol in § 3 estimates the extensibility and stability within a single task dataset,

Task	Paradigm	Goal	Signal	Training	Testing
Closed Set Recognition	Classification	Identifying known classes	Supervised	Known classes	Known classes
Open Set Recognition	Classification	Identifying known classes & rejecting unknown classes	Supervised	Known classes	Known classes & unknown classes
Open World Recognition	Classification	Identifying known classes & detecting unknown classes & labeling unknown data & incrementally learn and extend classifier	Supervised	Incremental known classes	Known classes & unknown classes
Open-vocabulary Recognition	Matching	Identifying classes via natural language	Unsupervised	-	Classes in a vocabulary

Table 3: A comparison of Closed Set Recognition, Open Set Recognition (OSR), Open World Recognition, and Open-vocabulary Recognition (OVR).

where the input images and classes during the vocabulary expansion come from the same data distribution. While the protocol is only an approximation of the real open world, current CLIP-like models have exhibited serious performance degradation. In this section, we take a step further toward real open recognition by conducting a vocabulary expansion setting at the dataset level, where the expanded vocabularies are from different datasets. In this way, the relationship between vocabularies is more uncertain and thus can be viewed as a rigorous stress test for the CLIP-like models. Specifically, we group all categories in a dataset into one vocabulary. Afterward, the inputs and classes of the entire new dataset are introduced at each expansion. Classes in the new vocabulary will be removed if they already exist in the previous vocabularies.

The experiments are conducted with datasets for generic objects, including CIFAR10 (Krizhevsky and Hinton, 2009), CIFAR100 (Krizhevsky and Hinton, 2009), Caltech101 (Fei-Fei et al., 2004), SUN397 (Xiao et al., 2010) and ImageNet (Deng et al., 2009), and specialized datasets focusing on fine-grained categories, including Flowers102 (Nilsback and Zisserman, 2008), OxfordPets (Parkhi et al., 2012) and StanfordCars (Krause et al., 2013). Without loss of the generality, we merge 3 datasets and evaluate the following dataset compositions:

(1) CIFAR100-Caltech101-SUN397

(2) CIFAR10-CIFAR100-ImageNet

(3) Flowers102-OxfordPets-StanfordCars

Composition (1) and (2) probe the performance when all the expanded datasets are generic thus the classes in different datasets are potentially semantic-correlated, while the composition (3) targets at scenarios where the coming datasets have little correlation with previous ones. To eliminate the effect of vocabulary expansion order, we report the average performance of all $\mathrm{A}_3^3 = 6$ possible trials for each composition.

Table 7 demonstrates the result of the dataset-level expansion. **First**, the performance of CLIP-like models on generic dataset expansion drops dramatically. For example, the accuracy (Acc-E) of CLIP (RN101) decreases by an averaged absolute point of 14.2 on the *CIFAR100-Caltech101-SUN397* composition during expansion, and 14.5 on the *CIFAR10-CIFAR100-ImageNet* composition. Due to the existence of subclass-superclass relationship for some classes in different generic datasets, e.g., *cat* in CIFAR10 and *tiger cat* in ImageNet, CLIP is extremely unstable on such expansion across generic datasets. For example, the Acc-S of CLIP (RN101) on the *CIFAR10-CIFAR100-ImageNet* composition is 28.2% lower than Acc-C, indicating the models are prone to be confused about the subclass-superclass relationship. **Meanwhile**, the CLIP-like models exhibit much better extensibility and stability on the dataset-level expansion across specialized datasets, e.g., the *Flowers102-OxfordPets-StanfordCar* composition. The vocabularies of this composition are intrinsically disjoint in semantics, so the model can be stably extended. **In summary**, our investigations on the dataset level expansions along with the task level in the paper show the current CLIP-like models fail to meet the expectation of conducting real open vocabulary recognition.

A.4 Incremental Acc-E and Acc-S on CIFAR100

We record the Acc-E (Eq.(2)) and Acc-S (Eq.(5)) after each vocabulary expansion on CIFAR100 to investigate the openness of CLIP-like models.

Figure 10 shows the Acc-E for 20 trials as new vocabularies are merged incrementally. The falling

Vocabulary (Superclass)	Classes
aquatic	mammals beaver, dolphin, otter, seal, whale
fish	aquarium fish, flatfish, ray, shark, trout
flowers	orchids, poppies, roses, sunflowers, tulips
food	containers bottles, bowls, cans, cups, plates
fruit and vegetables	apples, mushrooms, oranges, pears, sweet peppers
household electrical devices	clock, computer keyboard, lamp, telephone, television
household	furniture bed, chair, couch, table, wardrobe
insects	bee, beetle, butterfly, caterpillar, cockroach
large carnivores	bear, leopard, lion, tiger, wolf
large man-made outdoor things	bridge, castle, house, road, skyscraper
large natural outdoor scenes	cloud, forest, mountain, plain, sea
large omnivores and herbivores	camel, cattle, chimpanzee, elephant, kangaroo
medium-sized mammals	fox, porcupine, possum, raccoon, skunk
non-insect invertebrates	crab, lobster, snail, spider, worm
people	baby, boy, girl, man, woman
reptiles	crocodile, dinosaur, lizard, snake, turtle
small mammals	hamster, mouse, rabbit, shrew, squirrel
trees	maple, oak, palm, pine, willow
vehicles 1	bicycle, bus, motorcycle, pickup truck, train
vehicles 2	lawn-mower, rocket, streetcar, tank, tractor

Table 4: Superclass-class hierarchy in CIFAR100. Each superclass corresponds to a vocabulary, and each vocabulary has 5 classes. There are 20 kinds of vocabulary in total, specified by (Krizhevsky and Hinton, 2009).

lines indicate that the model is either performing poorly on the new input images, or that some images that were correctly identified before are misclassified after introducing the new classes.

Figure 11 shows Acc-S of CLIP-like models during non-target vocabulary expansion. Each subfigure represents the situation when one vocabulary is selected as the target vocabulary. As the remaining 19 non-target vocabularies are incorporated and the model is required to recognize the 5 target classes from 100 potential classes, the accuracy drops sharply. The decrease of Acc-S brought by each introduction of non-target vocabulary indicates that more images from the target vocabulary are incorrectly classified into the new non-target vocabulary by models.

A.5 Retrieval-enhanced prompt engineering

In light of the previous investigations, we propose a simple yet effective method named Retrieval-enhanced Prompt Engineering (REPE) to enforce the distinguishability of class features and the image-class semantic alignment (Cao et al., 2020; Ren et al., 2021). Recall that the context for each class name is the same in vanilla CLIP-like models (e.g., "`a photo of a [CLASSNAME]`"), making it difficult to discriminate between distinct visual categories because the semantics of each cannot be holistically represented (Zhou et al., 2022).

To remedy this, we propose to customize each class description with diverse captions retrieved from the pre-training corpus as a prompt ensemble. Specifically, for each class description based on the original prompt, we utilize CLIP to recall the most similar images from the pre-training dataset via image-text similarity, then obtain their corresponding captions. The retrieved captions with no appearance of the class name are filtered out, yielding K captions. Such a workflow leverages both visual semantics and class names, achieving better performance. Table 8 shows some cases of the captions retrieved by our proposed REPE on CIFAR100. They share the same target of interest with the original prompt, i.e., "`a photo of a [CLASS]`", but provide the context in which the class name is located and thus have richer semantics. For example, given a class like *bridge*, the retrieved captions describe its possible properties (e.g., "golden", "wooded"), connections to other objects (e.g., "over a mountain river"), etc., yielding more expressive and distinguishable text features of the class.

After retrieval, we encode the retrieved captions and conduct a mean pooling operation among them. The final text representation is:

$$f_T^{\text{REPE}}(t_i) = (1-\lambda) f_T(t_i) + \lambda \frac{1}{K} \sum_j^K f_T(rt_{ij}),$$

Vocabulary (Superclass)	Classes
garment	trench coat, abaya, gown, poncho, military uniform, jersey, cloak, bikini, miniskirt, swimming trunks, lab coat, brassiere, hoopskirt, cardigan, pajama, academic gown, apron, diaper, sweatshirt, sarong
bird	African grey, bee eater, coucal, American coot, indigo bunting, king penguin, spoonbill, limpkin, quail, kite, prairie chicken, red-breasted merganser, albatross, water ouzel, goose, oystercatcher, American egret, hen, lorikeet, ruffed grouse
reptile	Gila monster, agama, triceratops, African chameleon, thunder snake, Indian cobra, green snake, mud turtle, water snake, loggerhead, sidewinder, leatherback turtle, boa constrictor, garter snake, terrapin, box turtle, ringneck snake, rock python, American chameleon, green lizard
arthropod	rock crab, black and gold garden spider, tiger beetle, black widow, barn spider, leafhopper, ground beetle, fiddler crab, bee, walking stick, cabbage butterfly, admiral, lacewing, trilobite, sulphur butterfly, cicada, garden spider, leaf beetle, long-horned beetle, fly
mammal	Siamese cat, ibex, tiger, hippopotamus, Norwegian elkhound, dugong, colobus, Samoyed, Persian cat, Irish wolfhound, English setter, llama, lesser panda, armadillo, indri, giant schnauzer, pug, Doberman, American Staffordshire terrier, beagle
accessory	bib, feather boa, stole, plastic bag, bathing cap, cowboy boot, necklace, crash helmet, gasmask, maillot, hair slide, umbrella, pickelhaube, mit- ten, sombrero, shower cap, sock, running shoe, mortarboard, handkerchief
craft	catamaran, speedboat, fireboat, yawl, airliner, container ship, liner, trimaran, space shuttle, aircraft carrier, schooner, gondola, canoe, wreck, warplane, balloon, submarine, pirate, lifeboat, airship
equipment	volleyball, notebook, basketball, hand-held computer, tripod, projector, barbell, monitor, croquet ball, balance beam, cassette player, snorkel, horizontal bar, soccer ball, racket, baseball, joystick, microphone, tape player, reflex camera
furniture	wardrobe, toilet seat, file, mosquito net, four-poster, bassinet, chiffonier, folding chair, fire screen, shoji, studio couch, throne, crib, rocking chair, dining table, park bench, chest, window screen, medicine chest, barber chair
instrument	upright, padlock, lighter, steel drum, parking meter, cleaver, syringe, abacus, scale, corkscrew, maraca, saltshaker, magnetic compass, accordion, digital clock, screw, can opener, odometer, organ, screwdriver
man-made structure	castle, bell cote, fountain, planetarium, traffic light, breakwater, cliff dwelling, monastery, prison, water tower, suspension bridge, worm fence, turnstile, tile roof, beacon, street sign, maze, chain-link fence, bakery, drilling platform
wheeled vehicle	snowplow, trailer truck, racer, shopping cart, unicycle, motor scooter, passenger car, minibus, jeep, recreational vehicle, jinrikisha, golfcart, tow truck, ambulance, bullet train, fire engine, horse cart, streetcar, tank, Model T
produce	broccoli, corn, orange, cucumber, spaghetti squash, butternut squash, acorn squash, cauliflower, bell pepper, fig, pomegranate, mushroom, strawberry, lemon, head cabbage, Granny Smith, hip, ear, banana, artichoke

Table 5: Superclass-class hierarchy in ImageNet (Entity13). Each superclass corresponds to a vocabulary, and each vocabulary has 20 classes. There are 13 kinds of vocabulary in total, specified by BREEDS (Santurkar et al., 2021).

Vocabulary (Superclass)	Classes
salamander	eft, axolotl, common newt, spotted salamander
turtle	box turtle, leatherback turtle, loggerhead, mud turtle
lizard	whiptail, alligator lizard, African chameleon, banded gecko
snake	night snake, garter snake, sea snake, boa constrictor
spider	tarantula, black and gold garden spider, garden spider, wolf spider
grouse	ptarmigan, prairie chicken, ruffed grouse, black grouse
parrot	macaw, lorikeet, African grey, sulphur-crested cockatoo
crab	Dungeness crab, fiddler crab, rock crab, king crab
dog	bloodhound, Pekinese, Great Pyrenees, papillon
wolf	coyote, red wolf, white wolf, timber wolf
fox	grey fox, Arctic fox, red fox, kit fox
domestic cat	tiger cat, Egyptian cat, Persian cat, Siamese cat
bear	sloth bear, American black bear, ice bear, brown bear
beetle	dung beetle, rhinoceros beetle, ground beetle, long-horned beetle
butterfly	sulphur butterfly, admiral, cabbage butterfly, ringlet
ape	gibbon, orangutan, gorilla, chimpanzee
monkey	marmoset, titi, spider monkey, howler monkey

Table 6: Superclass-class hierarchy in ImageNet (Living17). Each superclass corresponds to a vocabulary, and each vocabulary has 4 classes. There are 17 kinds of vocabulary in total, specified by BREEDS (Santurkar et al., 2021).

Model	CIFAR100-Caltech101-SUN397					CIFAR10-CIFAR100-ImageNet					Flowers102-OxfordPets-StanfordCars				
	Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability		Acc-C	Extensibility		Stability	
		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ		Acc-E	Δ	Acc-S	Δ
CLIP (RN101)	65.9	51.7	-14.2	52.7	-13.2	62.4	47.9	**-14.5**	34.2	**-28.2**	65.8	63.1	-2.7	65.7	-0.1
CLIP (ViT-B/32)	72.0	59.4	**-12.6**	61.2	**-10.8**	70.9	52.7	-18.2	41.3	-29.6	65.8	62.0	-3.8	65.8	**-0.0**
CLIP (ViT-B/16)	**74.6**	**60.6**	-14.0	**61.7**	-12.9	**74.7**	**56.6**	-18.0	**43.3**	-31.4	**72.3**	**69.6**	**-2.7**	**72.3**	**-0.0**
SLIP (ViT-B/16)	58.6	44.4	-14.2	46.3	-12.3	55.6	36.7	-18.9	30.5	**-25.1**	35.0	26.0	-9.0	35.0	**-0.0**
DeCLIP (ViT-B/32)	**74.3**	**60.8**	**-13.5**	**63.3**	**-11.0**	**73.0**	**55.4**	**-17.6**	**45.1**	-27.9	**70.2**	**63.3**	**-6.9**	**70.2**	**-0.0**
PE (ViT-B/32)	71.8	59.9	**-11.9**	59.6	**-12.2**	72.2	53.5	**-18.7**	**41.6**	**-30.6**	65.7	62.0	-3.7	65.7	**-0.0**
PE (ViT-B/16)	**75.0**	**61.5**	-13.5	**62.5**	-12.5	**75.4**	**56.7**	**-18.7**	41.3	-34.1	**72.5**	**70.0**	**-2.5**	**72.5**	**-0.0**

Table 7: Extensibility and stability of CLIP and its variants during dataset-level vocabulary expansion. Δ refers to the decline of Acc-E/Acc-S (%) compared to Acc-C (%). PE denotes Prompt Ensemble.

Class	Retrieved captions
apple	"Apple slices stacked on top of each other" "Apples growing on a tree" "Still life with apples in a basket"
woman	"Portrait of a young woman" "Woman standing at the window" "Confident woman in a red dress and gold crown"
bridge	"The golden bridge in Bangkok" "Bridge on the River Kwai ∼Video Clip" "Wooden bridge over a mountain river"
ray	"Stingray in the Grand Cayman, Cayman Islands stock photography" "Common Stingray swimming close to the sea floor." "Sun Rays Tours: Go Pro captured the rays under water"

Table 8: Instances of the captions retrieved by our REPE on CIFAR100.

Model	CIFAR100			ImageNet (Entity13)			ImageNet (Living17)		
	Acc-C	Acc-E	Acc-S	Acc-C	Acc-E	Acc-S	Acc-C	Acc-E	Acc-S
CLIP (RN101)	68.3	55.4	54.9	80.4	77.4	77.3	77.6	74.5	74.4
REPE (RN101)	**68.4** (+0.1)	**55.5** (+0.1)	**55.2** (+0.3)	**81.7** (+1.3)	**79.2** (+1.8)	**79.0** (+1.7)	**77.8** (+0.2)	**75.3** (+0.8)	**75.2** (+0.8)
CLIP (ViT-B/32)	78.0	69.6	68.9	80.8	78.0	77.8	78.0	74.4	75.0
REPE (ViT-B/32)	**78.5** (+0.5)	**70.9** (+1.3)	**70.6** (+1.7)	**82.3** (+1.5)	**79.8** (+1.8)	**79.6** (+1.8)	**79.0** (+1.0)	**76.4** (+2.0)	**76.2** (+1.2)
CLIP (ViT-B/16)	79.7	72.6	72.0	83.5	81.1	81.0	79.5	77.9	77.6
REPE (ViT-B/16)	**79.8** (+0.1)	**72.9** (+0.3)	**72.6** (+0.6)	**85.4** (+1.9)	**83.3** (+2.2)	**83.2** (+2.2)	**79.9** (+0.4)	**78.4** (+0.5)	**78.2** (+0.6)

Table 9: Extensibility and stability of our REPE method on CIFAR100 and ImageNet datasets.

Method	K-shot	CIFAR100	ImageNet
CLIP-Adapter	4	66.6	63.0
CLIP-Adapter + REPE	4	**67.5** (+0.9)	**63.3** (+0.3)
CLIP-Adapter	16	69.0	64.6
CLIP-Adapter + REPE	16	**69.8** (+0.8)	**64.9** (+0.3)

Table 10: Accuracy of CLIP-Adapter and our REPE method with few-shot learning.

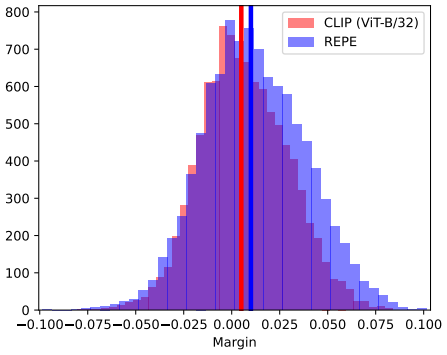


Figure 9: Margin distribution of similarity scores of our REPE (blue) and CLIP (ViT-B/32) (red). The median value of REPE's distribution (the blue vertical line) is larger than that of CLIP (the red line), indicating that the predictions of REPE are harder to be inverted with competing classes than the original CLIP.

where rt_{ij} is the $j^{(th)}$ retrieved caption for class i and λ is a weighting factor. After that, the ensemble text representation $f_T^{\text{REPE}}(t_i)$ is adopted as the class anchor for conducting the image classification. With REPE, the representation of the class description shifts towards that of the representative captions in the pre-training dataset, which alleviates the semantic inconsistency between pre-training and inference.

Experiments We retrieve the images and captions from CC12M (Changpinyo et al., 2021), a subset of the pre-training dataset of CLIP. The images and captions are pre-encoded within **an hour** using a single RTX TITAN GPU, then we build their indices for KNN search with the FAISS framework (Johnson et al., 2019), which also takes about **an hour**. Once the indices are built, we can efficiently search over the dataset according to the query image in less than **5 ms**, which is applicable for query-intensive scenarios.

Table 9 shows the results of REPE. The hyper-parameter K is 100 and λ is 0.25. REPE consistently improves the extensibility and stability of CLIP by an average of **1.2%** across all three datasets. We further evaluate the quality of the enhanced representations by analyzing the loss of text uniformity and inter-modal alignment. As shown in Figure 7, our proposal effectively reduces $\ell_{\text{uniform-T}}$ from -0.8 to -1.0 and ℓ_{align} from 1.5 to 1.4, verifying its effectiveness in improving the class anchor for better extensibility and stability. Additionally, as shown in Figure 9, REPE increases the median value of the margin distribution from 0.005 to 0.01 and pushes the overall distribution towards the positive side compared to vanilla CLIP. It indicates that REPE widens the gap between positive and negative class features, making it more difficult to invert predictions with competing classes. These findings support REPE's effectiveness in alleviating the openness issue.

It is worth noting that compared to the method that requires computation-intensive pre-training procedures (DeCLIP and SLIP), and the prompt-tuning approach (CoOp) demands access to the downstream target dataset, our REPE is a lightweight framework for the zero-shot inference stage without fine-tuning. Besides, since REPE is model-agnostic and orthogonal to parameter-tuning methods, it can also be combined with fine-tuning methods like adapter-tuning (Gao et al., 2021), to achieve a further performance boost of 0.6 on CIFAR100 and ImageNet, which demonstrates the adaptability and superiority of our method. Please refer to Table 10 for details.

Figure 10: Incremental Acc-E of CLIP and its variants on CIFAR100.

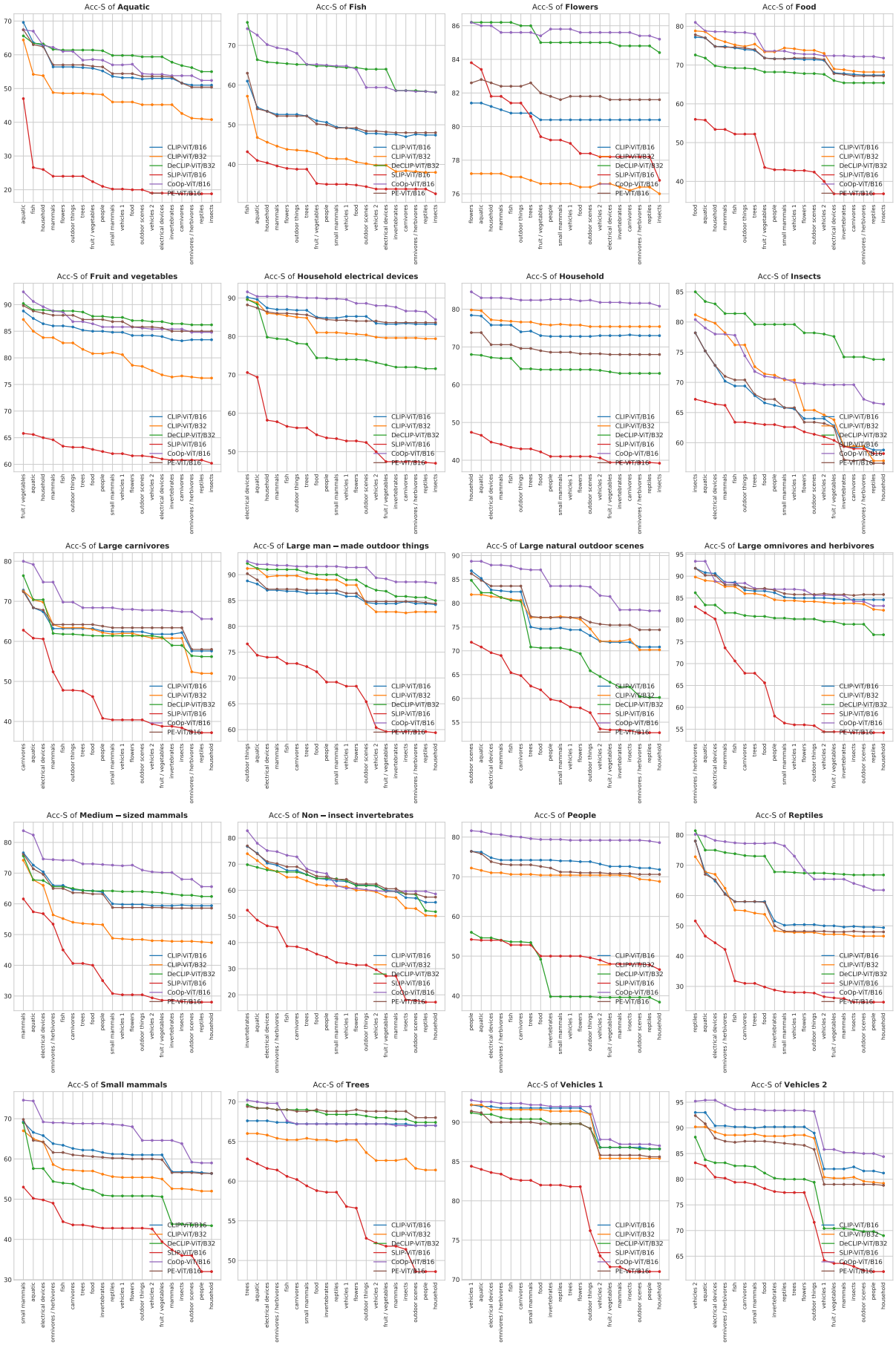


Figure 11: Incremental Acc-S of CLIP and its variants on CIFAR100.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

4.3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.2

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.