Improving Zero-shot Generalization and Robustness of Multi-modal Models

Yunhao Ge[1,2*], **Jie Ren**[1*], **Yuxiao Wang**[1], **Andrew Gallagher**[1], **Ming-Hsuan Yang**[1],
Laurent Itti[2], **Hartwig Adam**[1], **Balaji Lakshminarayanan**[1†], **Jiaping Zhao**[1†]
[1]Google Research [2]University of Southern California, work done while at Google Research
[*]co-first author, †correspondence to {balajiln, jiapingz}@google.com

Abstract

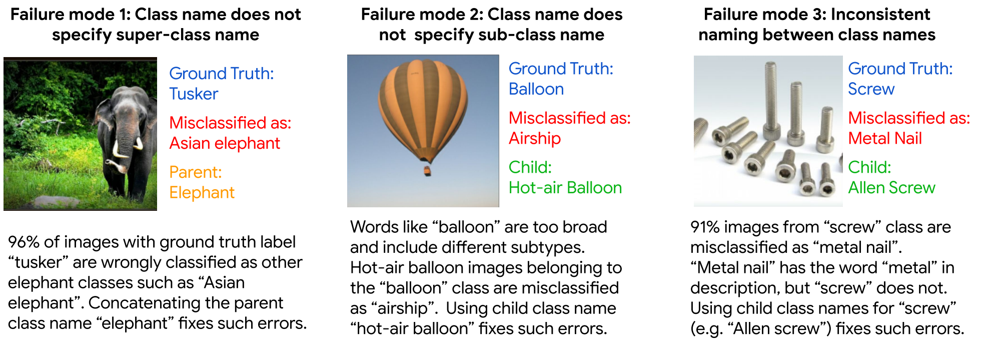
Multi-modal image-text models such as CLIP and LiT have demonstrated impressive performance on image classification benchmarks and their zero-shot generalization ability is particularly exciting. While the top-5 zero-shot accuracies of these models are very high, the top-1 accuracies are much lower (over 25% gap in some cases). We investigate the reason for this performance gap and find that many of the failure cases are caused by ambiguity in the text prompts. First, *we develop a simple and efficient zero-shot post-hoc method to identify images where the top-1 prediction is likely to be incorrect*, by measuring consistency of the predictions w.r.t. multiple prompts and image transformations. We show that our procedure better predicts mistakes, outperforming the popular max logit baseline on selective prediction tasks. Next, *we propose a simple and efficient way to improve accuracy on such uncertain images by making use of the WordNet hierarchy*; specifically we use information from parents in the hierarchy to add superclass to prompts, and use information from children in the hierarchy to devise fine-grained prompts. We conduct experiments on both CLIP and LiT models with five different ImageNet-based datasets. For CLIP, **our method improves the top-1 accuracy by 17.13% on the uncertain subset and 3.6% on the entire ImageNet validation set.** We also show that our method consistently improvement on other ImageNet shifted datasets and other model architectures such as LiT. **Our proposed method is hyperparameter-free, requires no additional model training and can be easily scaled to other large multi-modal architectures.**

1 Introduction

Recently, vision-language multi-modal models trained on large-scale data have achieved significant success in numerous domains and demonstrated excellent zero-shot generalization ability (Radford et al., 2021; Zhai et al., 2022; Pham et al., 2021; Jia et al., 2021). For example, the zero-shot top-1 accuracy for ImageNet using CLIP variants (CLIP ViT-L) matches the performance of the original ResNet model trained from scratch. Recently, CLIP is found to be more robust to distribution shift than the ResNet model, achieving decent performance on ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-Sketch (Wang et al., 2019). Interestingly, we noticed a large gap between top-1 accuracy 64.2% and top-5 accuracy 89.4%, revealing potential headroom for improvement. We investigated the cases where top-1 prediction was incorrect but the top-5 prediction was correct, and identified several failure modes. Most of these failure cases are caused by noise and ambiguity in text prompts which suggest that the text encoder is very sensitive to inputs and as a result, the overall prediction score lacks robustness.

Inspired by the above observation, we first identify the subset of images where the top-1 prediction is likely to be wrong; we use consistency of predictions w.r.t. different text prompts and image augmentations as a signal for uncertainty estimation. For the identified subset, we then propose a principled framework to modify their prompts to improve the accuracy and consequently the robustness. Maximum softmax probability (Hendrycks & Gimpel, 2016) and maximum logit score

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).


Failure mode 1: Class name does not specify super-class name
Ground Truth: Tusker
Misclassified as: Asian elephant
Parent: Elephant
Failure mode 2: Class name does not specify sub-class name
Ground Truth: Balloon
Misclassified as: Airship
Child: Hot-air Balloon
Failure mode 3: Inconsistent naming between class names
Ground Truth: Screw
Misclassified as: Metal Nail
Child: Allen Screw

96% of images with ground truth label "tusker" are wrongly classified as other elephant classes such as "Asian elephant". Concatenating the parent class name "elephant" fixes such errors.

Words like "balloon" are too broad and include different subtypes. Hot-air balloon images belonging to the "balloon" class are misclassified as "airship". Using child class name "hot-air balloon" fixes such errors.

91% images from "screw" class are misclassified as "metal nail". "Metal nail" has the word "metal" in description, but "screw" does not. Using child class names for "screw" (e.g. "Allen screw") fixes such errors.

Figure 1: Typical failure modes in the cases where top-5 prediction was correct but top-1 was wrong.

(Hendrycks et al., 2019) are commonly used confidence scores for classification models. However, we found those scores are not always reliable for the CLIP model due to its poor calibration.

We propose a simple yet efficient zero-shot confidence estimation method better suited for CLIP, based on predictions' self-consistency over different text prompts and image perturbations. Wang et al. (2022) proposed to use *self-consistency between multiple model outputs* to improve the reasoning accuracy of large language models. Here we extend the idea for confidence estimation in multi-modal models by measuring *consistency of predictions to multiple input text prompts and image transformation*. Our method is very effective at predicting mistakes; the low confidence subset identified by our method has significantly lower accuracy (21.58%) than the average accuracy (64.18%). To improve the accuracy for the subset, we develop a prompt augmentation technique using WordNet label hierarchy. Our method leverages information from ancestors (top-down) as well as children (bottom-up) and improves the top-1 accuracy of the subset to 38.71% (17.13% improvement). Our method not only improves model accuracy, but also model robustness, improving on ImageNet variants with distribution shift such as ImageNet-v2, ImageNet-R, ImageNet-Adversarial and Imagenet-Sketch.

2 Zero-shot inference failure case analysis

Given that the top-1 accuracy (64.2%) is much lower than top-5 accuracy (89.4%) for zero-shot ImageNet classification using CLIP, we investigated the failure cases that are "top-5 correct but top-1 wrong" (12605 images, 25.2% of all test images). The most frequent ground-truth classes those images belong to are shown in Table 2. The failure modes can be summarizaed as follows:

(1) Class name does not specify super-class name: Some classes, whose class names do not have their WordNet ancestor (e.g., "tusker", one of 1k ImageNet classes, does not have its parent "elephant" in the class name), may have a relatively lower score than other classes, which explicitly have the ancestor present in the class name (e.g., "Asian elephant"). See examples in Fig. 1 (Left).
(2) Class name does not specify sub-class name: If the class name is too abstract, then its CLIP embedding is not necessarily close to the image embedding: e.g, CLIP wrongly classifies most images from "balloon" class as airship, see Fig. 1 (Middle). That is because there are distinct kinds of balloons, each belonging to a different semantic subgroup. Relying on text embedding of the fine-grained children's class names (e.g., using "hot-air balloon") often fixes such errors. Beyer et al. (2020) reported the similar issue of label ambiguity in ImageNet.
(3) Inconsistent naming between class names: Some ImageNet class names are nouns, but others are adjective-prefixed nouns. This may make CLIP text embedding biased, see one example in Fig. 1 (Right) where images from "screw" class are misclassified as "metal nail".

3 Proposed Method

As shown in Section 2, CLIP models can be very sensitive to different word prompts for images in certain classes. In this section, we first propose a confidence estimation method to identify low confidence predictions. We show that the identified subset has much lower accuracy than the average. We next develop a principled method that utilizes knowledge hierarchy to improve the accuracy of the low confidence subset, and consequently improve the overall accuracy on the whole datasets.

3.1 Zero-shot confidence estimation via self-consistency Given an image $\boldsymbol{x}$ and a candidate class name c, where $c \in \mathcal{C}, |\mathcal{C}| = 1000$, the CLIP model encodes $\boldsymbol{x}$ and c respectively by its image encoder f_{image} and text encoder f_{text}, then we get $\boldsymbol{z}_m = f_{image}(\boldsymbol{x})$ and $\boldsymbol{z}_c = f_{text}(c)$. The prediction logit score is defined as $\text{logit}(\boldsymbol{x}, c) = \cos(\boldsymbol{z}_m, \boldsymbol{z}_c)$, where $\cos(\cdot, \cdot)$ is the cosine similarity between two vectors, and the predicted class is $\arg\max_{c \in \mathcal{C}} \text{logit}(\boldsymbol{x}, c)$. We estimate the confidence by ensembling over text prompts and image augmentations.

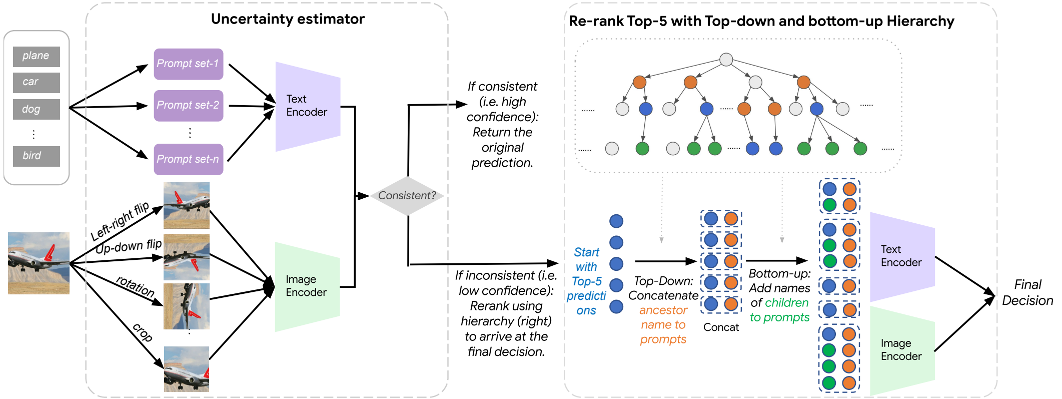

Uncertainty estimator
plane
car
dog
bird
Prompt set-1
Prompt set-2
Prompt set-n
Text Encoder
Left-right flip
Up-down flip
rotation
crop
Image Encoder
Consistent?
If consistent (i.e. high confidence): Return the original prediction.
If inconsistent (i.e. low confidence): Rerank using hierarchy (right) to arrive at the final decision.
Re-rank Top-5 with Top-down and bottom-up Hierarchy
Start with Top-5 predicti ons
Top-Down: Concatenate ancestor name to prompts
Concat
Bottom-up: Add names of children to prompts
Text Encoder
Image Encoder
Final Decision

Figure 2: Overview of our confidence estimation via self-consistency and Top-down and bottom-up prompt augmentation using the WordNet hierarchy. See Algorithms 1 and 2 for pseudocode.

Confidence estimation via text prompts We detect the low confidence subset based on the prediction's self-consistency to prompts. In particular, we make use of the 80 different context prompts (e.g. "`A photo of a big {label}`" and "`A photo of a small {label}`") that were used in CLIP paper (Radford et al., 2021) for prompt ensembling purposes. For a fixed image $\boldsymbol{x}$, given a set of context prompts $\mathcal{T}$, the ensembled logit score is $\text{logit}(\boldsymbol{x}, \mathcal{T}(c)) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{logit}(\boldsymbol{x}, t(c))$, where $t(c)$ denotes the new prompt after applying context prompt $t()$ to c. To estimate the confidence score, we partition the 80 prompts into $L \geq 2$ sets, get the predictions for the L sets $\mathcal{T}_1, \mathcal{T}_2 \ldots \mathcal{T}_L$, and see if the top-1 prediction classes are the same or not. Intuitively, a reliable prediction is one where the top-1 predicted class would be the same among all L predictions. Thus we define a high confidence prediction only if the top-1 predicted class is consistent when L different sets of context prompts $\mathcal{T}_1$, $\mathcal{T}_2$ ... $\mathcal{T}_L$ are applied, i.e. $\arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, \mathcal{T}_1(c)) = \arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, \mathcal{T}_2(c))$ = ... = $\arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, \mathcal{T}_L(c))$. All other cases are considered as low confidence predictions.

Confidence estimation via image perturbation We also estimate the confidence of the prediction based on the score's robustness to image perturbations. Intuitively, if the predicted classes are inconsistent when applying different image perturbations, the prediction is not reliable. Specifically, we consider the common image transformations, left-right flip, up-down flip, rotation, crop, etc., and apply the perturbation method b to the input image, $\text{logit}(b(\boldsymbol{x}), c)$. We define a prediction having low confidence if the top-1 prediction of the perturbed image is inconsistent with the raw image. We find left-right flip works the best among above mentioned perturbations. Finally, we use the union of the two low confidence sets identified by text prompts and image perturbations as the final low confidence subset in the following experiments. Algorithm 1 shows the detailed steps.

3.2 Top-down and bottom-up prompt augmentation using WordNet hierarchy Through extensive analysis of the incorrect predictions among the identified unreliable predictions, we found that many of them are caused by CLIP models' lack of robustness to prompts. A proper prompt that specifies both the general type and the fine-grained attributes of this class are very important for correctly classifying the image. However, the ImageNet class names are not all defined in the same way such that some classes are more abstract than other classes, e.g. 350 classes have children, while the rest of the classes have no children. See Figure 4 in Appendix for more details. To make the ImageNet classification problem better suited for CLIP model, we leverage the underlying WordNet hierarchy and develop a top-down and bottom-up class name augmentation method to improve zero-shot prediction accuracy for unreliable predictions.

Top down: Augmenting class names with super-class attention As shown in failure case analysis, adding the super-class name to reduce ambiguity and to encourage the model's attention on the general concept is helpful for improving the accuracy. Therefore we propose to use WordNet to find the parent node to the raw class name, and concatenate it to the class name, i.e. $\text{logit}(\boldsymbol{x}, c) = \text{logit}(\boldsymbol{x}, [c; p(c)])$ where $p(c)$ is the parent node's name of the class name c, and $[c; p(c)]$ means the string concatenation of the class name and the parent name. We apply the method to top-5 predicted classes. Using the newly defined class names, we are able to re-rank the top-5 predictions for the identified unreliable subset of images. Note that the WordNet contains a few very abstract class names for nodes, such as "physical entity", "artifact", "matter", etc. We found that such parent nodes are not informative, hence we remove them. There are also many academic words in WordNet, for example the parent node of sea anemone is "anthozoan", which can be rare in CLIP training data. Adding those academic words to class name could make the prediction even more non-robust. So we develop a method to simplify the WordNet based on an estimation of the word frequency in CLIP training data by using embedding norm. See Appendix E for more details.

Bottom-up: Augmenting class names with fine-grained subtypes and attribute Some ImageNet class names are generally abstract, but the ImageNet images may belong to a specific subtype of the class. For example, "balloon" is a class name in ImageNet, but most balloon images in ImageNet are actually "hot-air balloon". We have observed that CLIP model's score is very sensitive to prompts and a score for a super class is not necessarily higher than the score for its sub-classes, mismatching with hierarchy prior. To accurately classify the images using CLIP, we need to augment the class name with fine-grained subtypes. Specifically, for each class c that has children in the WordNet hierarchy, we propose to use the score as the maximum of the scores over all its children, $\text{logit}(\boldsymbol{x}, c) = \max\{\text{logit}(\boldsymbol{x}, c), \text{logit}(\boldsymbol{x}, c_1), \ldots, \text{logit}(\boldsymbol{x}, c_r)\}$, where $c_1 \ldots c_r$ are the r children of the node c in the WordNet hierarchy. We apply this bottom-up method to top-5 predicted class names, and re-rank the top predictions. See Algorithm 2 for more details.

4 Experiments and Results

Our proposed method is composed of 2 steps and we conduct experiments to verify the effectiveness of each step: (1) Use zero-shot confidence estimation to identify the subset of samples that have unreliable predictions (see Fig. 3 for the results) (2) Augment the class label prompt using top-down and bottom-up strategies based on the sparsified WordNet on those unreliable subset to improve the accuracy (See Table. 1 for the results).

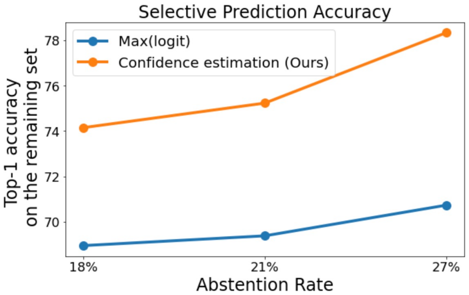


Figure 3: Selective prediction accuracy at different abstention rates for CLIP. See Figure 5 in Appendix D for similar results on LiT.

Our proposed confidence score is better suited for selective prediction than max logit baseline. To compare our proposed confidence score with the baseline regarding their ability for predicting errors, we plot the selective prediction curves (Lakshminarayanan et al., 2017). We use each score to select the least confident 18%, 21%, and 27% samples, and evaluate the accuracy on the remaining. Better confidence estimation allows model to better detect inputs where it is likely to be incorrect and abstain on such examples, as evidenced by the higher accuracy on the remaining set. Figure 3 shows that our proposed confidence score is better than the baseline $\max(\text{logit})$. Note that for our method, we find the 18%, 21%, 27% least confident samples whose top-1 prediction has any inconsistent when applying $B = 2, 4, 8$ different sets of $\mathcal{T}$, respectively. On the least confident subset that our method proposes to abstain, the accuracy is only 21.63%, significantly lower than the overal average (See Table 1).

Using top-down and bottom-up prompt augmentation significantly improves the accuracy on the low confident subset. For the low confident subset, we apply the top-down and bottom-up prompt augmentation method. Table. 1 middle two column shows that we are able to improve 16.97% on the top-1 accuracy (from 21.58% to 38.71%) for the identified low confident subset of samples, and overall 3.6% on the top-1 accuracy (64.18% to 67.78%) for all samples in ImageNet. We show similar improvement on the zero-shot accuracy for ImageNet shifted datasets. To investigate if our method works for other multi-modal models, we apply our method to LiT (Zhai et al., 2022) model and observe that our method improves accuracy and robustness for LiT models as well.

Table 1: CLIP (ViT-B-16) and LiT (ViT-B-32) zero-shot Top-1 accuracy on various datasets

		CLIP	(Ours) Hierarchy-CLIP	LiT	(Ours) Hierarchy-LiT
ImageNet	Low confidence subset	21.58%	**38.71%**	31.18%	**37.25%**
	Full dataset	64.18%	**67.78%**	68.26%	**69.41%**
ImageNet-v2	Low confidence subset	17.77%	**32.50%**	27.08%	**31.45%**
	Full dataset	58.06%	**61.07%**	60.11%	**61.11%**
ImageNet-R	Low confidence subset	16.79%	**27.91%**	21.82%	**22.93%**
	Full dataset	56.88%	**59.46%**	66.54%	**66.75%**
ImageNet-Adversarial	Low confidence subset	10.13%	**18.44%**	7.19%	**8.95%**
	Full dataset	26.12%	**29.23%**	13.93%	**14.56%**
ImageNet-Sketch	Low confidence subset	13.74%	**23.18%**	21.51%	**24.42%**
	Full dataset	44.71%	**47.28%**	52.47%	**53.17%**

References

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Computer Vision and Pattern Recognition*, 2021b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Appendix

A Analyzing the cases where top-5 prediction is correct but top-1 prediction is incorrect

Class name	Error rate
tusker	94%
missile	94%
terrapin	92%
collie	90%
screw	90%
mushroom	88%
Appenzeller Sennenhund	84%
snoek fish	84%
husky	82%
parallel bars	82%
gazelle	82%
sailboat	82%
corn cob	80%
analog clock	78%
cornet	78%
gossamer-winged butterfly	76%
green mamba	76%
tiger cat	74%
hare	74%
canoe	72%

Table 2: List of the most frequent classes where the top-5 prediction is correct but the top-1 prediction is incorrect.

B Locating the 1000 ImageNet classes at WordNet hierarchy

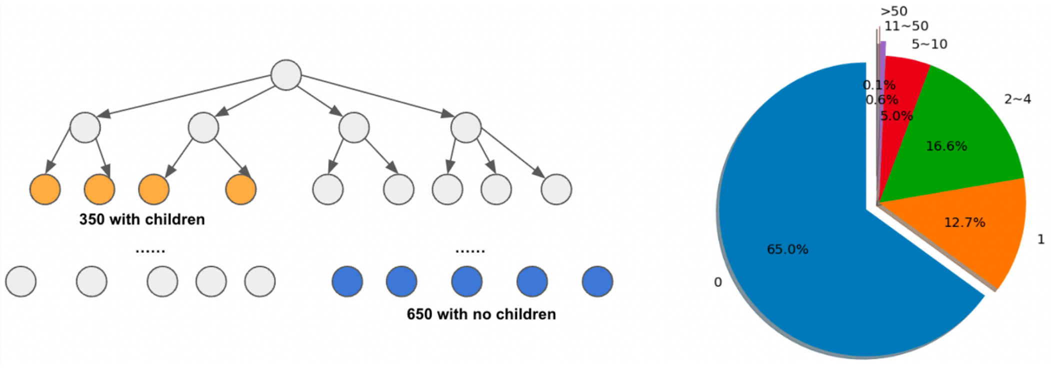


Figure 4: (a) The 1000 ImageNet class names are at different levels of WordNet hierarchy with different degree of abstraction. 350 of them are super-class with sub-classes as the children, while the rest 650 of them have no children. (b) The distribution of the number of children: 12.7% of the classes have one child node, 16.6% of the classes have 2-4 child nodes.

C Algorithm

Algorithm 1 describes the details of Zero-shot confidence estimation; Algorithm 2 describes top-down and bottom-up prompt augmentation using WordNet hierarchy.

Algorithm 1: Zero-shot confidence estimation

```
Input: Input images $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$, Candidate class set $\mathcal{C}$, image encoder $f_{image}$ and text
       encoder $f_{text}$
Output: Low confidence set $\mathcal{O}$
Low confidence set $\mathcal{O}_{text} \leftarrow \emptyset$            ▷Confidence estimation via text prompts
Sample $L$ different context prompt sets $\mathcal{T}_1, \mathcal{T}_2 \dots \mathcal{T}_L$
for $\boldsymbol{x}_i \in \mathcal{X}$ do
    for $\mathcal{T} \in \{\mathcal{T}_1, \mathcal{T}_2 \dots \mathcal{T}_L\}$ do
        $\text{logit}(\boldsymbol{x}_i, \mathcal{T}(c)) \leftarrow \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{logit}(\boldsymbol{x}_i, t(c)) \leftarrow \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \cos(f_{image}(\boldsymbol{x}_i), f_{text}(t(c)))$
    if $\arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, \mathcal{T}_1(c))$ = ... = $\arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}_i, \mathcal{T}_L(c))$ then
        $\boldsymbol{x}_i$ prediction has high confidence
    else
        $\mathcal{O}_{text} \leftarrow \mathcal{O}_{text} \cup \boldsymbol{x}_i$
Low confidence set $\mathcal{O}_{image} \leftarrow \emptyset$      ▷Confidence estimation via image perturbation
Sample $M$ perturbation methods $b_1, \dots, b_M$
for $\boldsymbol{x}_i \in \mathcal{X}$ do
    for $b \in \{b_1, b_2 \dots b_M\}$ do
        $\text{logit}(b(\boldsymbol{x}_i), c) \leftarrow \cos(f_{image}(b(\boldsymbol{x}_i)), f_{text}(c))$
    if $\arg\max_{c\in\mathcal{C}} \text{logit}(b_1(\boldsymbol{x}_i), c)$ = ... = $\arg\max_{c\in\mathcal{C}} \text{logit}(b_M(\boldsymbol{x}_i), c)$ then
        $\boldsymbol{x}_i$ prediction has high confidence
    else
        $\mathcal{O}_{image} \leftarrow \mathcal{O}_{image} \cup \boldsymbol{x}_i$
$\mathcal{O} \leftarrow \mathcal{O}_{text} \cup \mathcal{O}_{image}$
```

Algorithm 2: Top-down and bottom-up prompt augmentation using WordNet hierarchy

```
Input: Input image $\boldsymbol{x} \in \mathcal{O}$, top-5 candidate class set $\mathcal{C}_{top5}$, sparse WordNet hierarchy $H$,
       image encoder $f_{image}$ and text encoder $f_{text}$
Output: Predicted class of $\boldsymbol{x}$
Candidate class set $\mathcal{C} \leftarrow \emptyset$
for $c \in \mathcal{C}_{top5}$ do
    $\mathcal{C} \leftarrow \mathcal{C} \cup [c; \text{parent}(c)]$, where parent($c$) is the parent of $c$ in $H$          ▷Top-down
    if $c$ has $r \geq 1$ children $c_1 \dots c_r$ in $H$ then
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{[c_j; \text{parent}(c)]\}_{j=1}^{r}$                ▷Bottom-up
$\hat{c} \leftarrow \arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, c)$
if $\hat{c} \in \mathcal{C}_{top5}$ then
    final prediction $\leftarrow \hat{c}$
else
    final prediction $\leftarrow$ parent($\hat{c}$)
```

D Additional results

Figure 5 shows selective prediction results for LiT. Similar to CLIP results in Figure 3, we see that our method significantly improves selective prediction.

Figure 6 shows qualitative visualization on more typical failure modes in the cases where our top-down and bottom-up prompt augmentation using WordNet hierarchy method fixes the error.

E Sparsifying WordNet with norm

WordNet contains many academic words that are rarely used in common usage of English, and hence unlikely to occur frequently in the captions used for CLIP training. For example, "anthozoan, actinozoan", "coelenterate", "gastropod", and etc.. Directly using the raw WordNet with academic words as parent is not that helpful for improving zero-shot accuracy, and could even harm the performance. Though we do not have access to the CLIP private data, we study the norm of the word embedding vector and found it is correlated with word frequency. In particular, we compute the L_2 norm of the word embedding when the word is coupled with different context prompts, i.e $\|f_{text}(t(c))\|$, $t \in \mathcal{T}$. For a commonly used word, we found that the variance of the norm, $\text{Var}_{t\in\mathcal{T}}(\|f_{text}(t(c))\|)$,

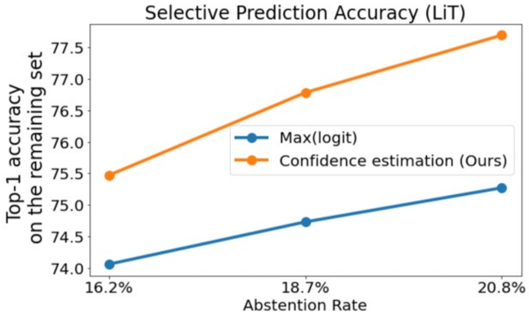
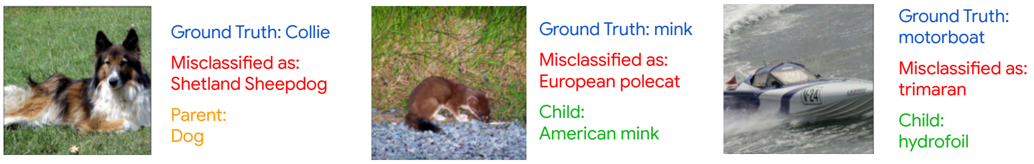


Figure 5: Selective prediction accuracy at different abstention rates for LiT model.



90% of images with ground truth label "collie" are wrongly classified as other dog classes such as "Shetland Sheepdog". Concatenating the parent class name "dog" fixes such errors.

78% of images with ground truth label "mink" are wrongly classified as other animal classes such as "European polecat". Using child class names for "mink" (e.g. "American mink") fixes such errors.

62% of images with ground truth label "motorboat" are wrongly classified as other boat classes such as "trimaran". Using child class names for "motorboat" (e.g. "hydrofoil") fixes such errors.

Figure 6: Qualitative visualization on more typical failure modes in the cases where our top-down and bottom-up prompt augmentation using WordNet hierarchy method fixes the error.

is correlated with word frequency. The rare words tend to have small variances, while the common words tend to have large variances. For example, the variance of the word "anthozoan" is 0.118, while the variance of the word "workplace, work" is 0.724. Thus we use this statistic as a metric to filter out rare words in WordNet. We removed 60% of the nodes in WordNet by setting a 60% quantile as the threshold for the variance. We believe the intuition behind the correlation between the norm variance and the word frequency is that, for a frequent word that is included in have many examples in the CLIP training data, the CLIP model learns a very precise text embedding such that it has the capability to tell the semantic difference under different contexts, e.g. "`a photo of a nice {label}`" and "`a photo of a weird {label}`".

F Related work

Prompt engineering and learning has attracted much attention in vision and learning since the introduction of image-text models (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022). The image-text models align images and their text descriptions into a common space, which facilitates the model to generalize to unseen categories in the inference time. However, it has been observed that the downstream image classification accuracy highly depends on the input prompts. This motivates researchers to either fine-tune or auto-learn prompts when adapting multi-modal models to downstream vision tasks.

Zhou et al. (2022b,a) propose CoOp and CoCoOp to automatically learn the prompt word embeddings in the few-shot settings, and show significant improvements over the vanilla zero-shot image classification based-on prompting. These are learning based approaches, requiring label supervised data from downstream tasks, while our proposed method is zero-shot and post-hoc without using any supervised data. In concurrent work, Shu et al. (2022) propose to learn the prompt embedding in an unsupervised manner by minimizing the entropy of the averaged prediction probability distribution, where each prediction is based on a random augmentation applied to the input image. Our work is different from (Shu et al., 2022) in the sense that we do not learn an input-dependent prompt embedding. Instead we only selectively modify the prompts using knowledge hierarchy for images that have unreliable predictions, and our modified new prompt is natural language not numerical values.