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









Yunhao Ge* | Jie Ren * | Andrew Gallagher | Yuxiao Wang | Ming-Hsuan Yang | Hartwig Adam | Laurent Itti | Balaji Lakshminarayanan | Jiaping Zhao

* co-first author   correspondence to {balajiln, jiapingz}@google.com

Google Research

USC

Paper   Code   Project page

session and poster ID

WED-AM-272

JUNE 18-22, 2023

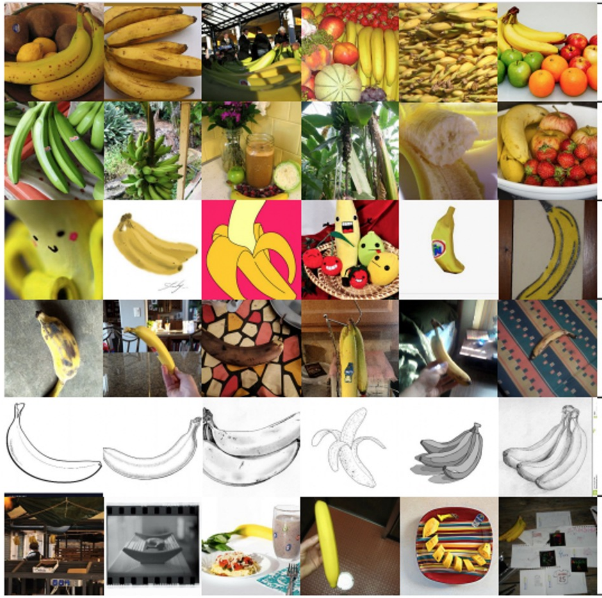
CVPR

VANCOUVER CANADA

Motivation: Multi-modal Models (Vision-Language Models)

1. CLIP[1] has high zero-shot accuracy and more robustness to distribution shifts.

Dataset Examples



Dataset	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

[1] Radford, Alec, et al. "**Learning transferable visual models from natural language supervision**." International conference on machine learning. PMLR, 2021.

Vision-language models, like CLIP, have high zero-shot classification accuracy and robustness to distribution shifts.

Finding 1: Multi-modal Models are sensitive to prompts

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	**91.83**

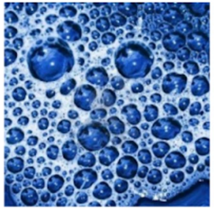
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a $\boxed{\text{flower}}$ photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

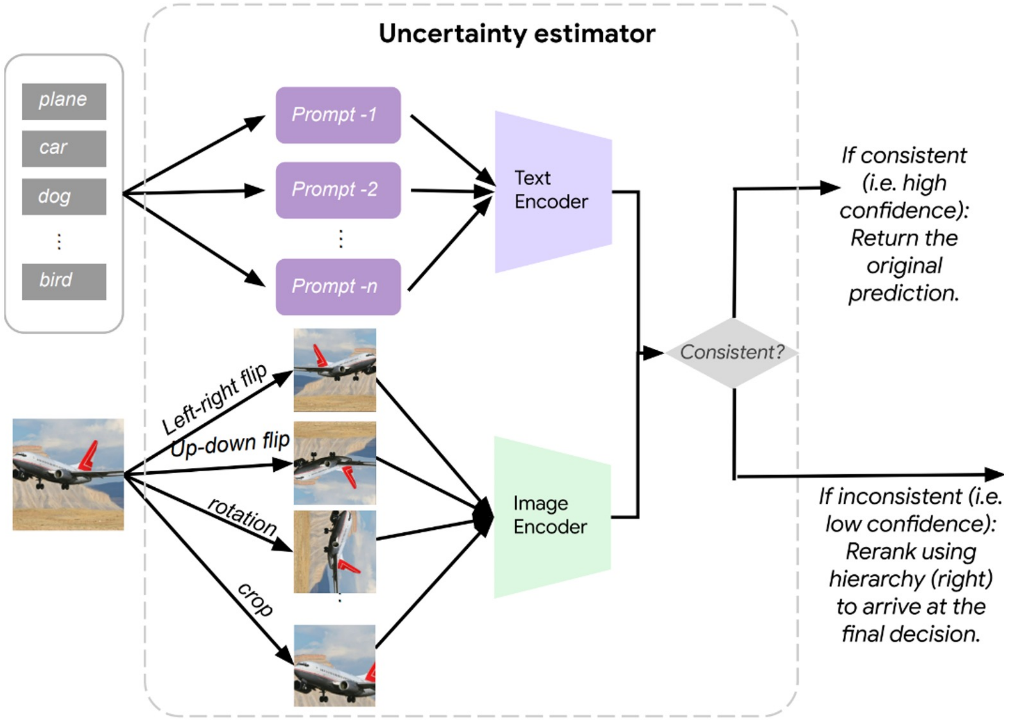
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	**83.53**

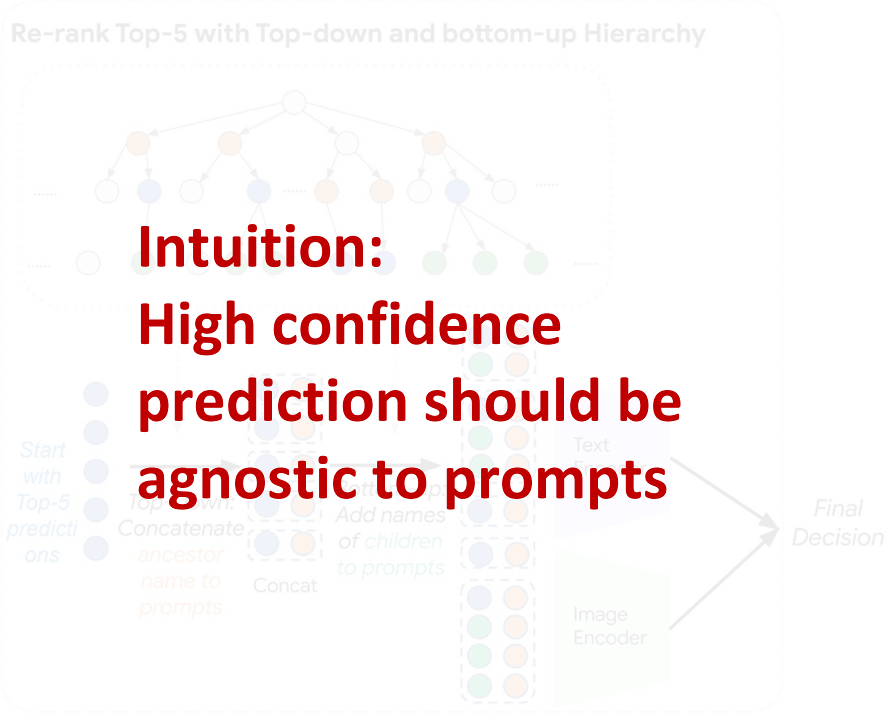
(d)

Zhou, Kaiyang, Jingkang Yang, Chen Change Loy, and Ziwei Liu. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130, no. 9 (2022): 2337-2348.

However, they are sensitive to prompts. Different prompts lead to different performances.

Step 1: Uncertainty Estimation

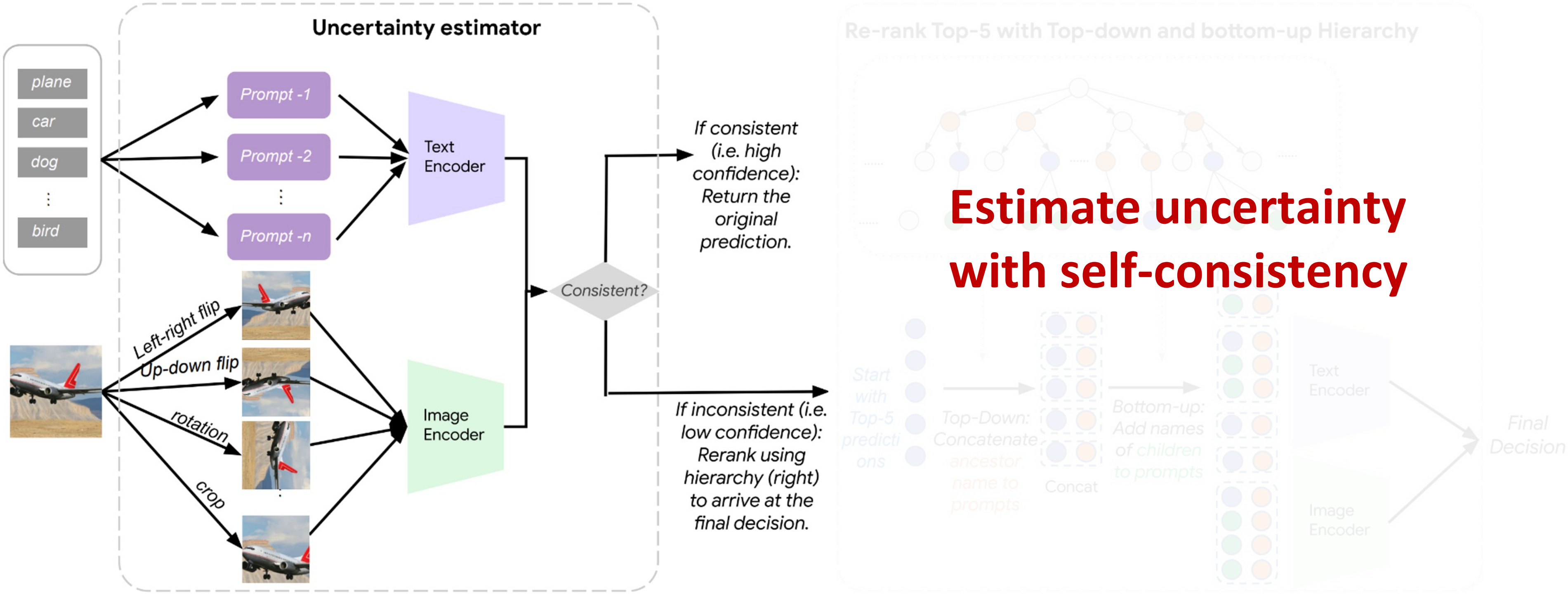




Intuition:
High confidence prediction should be agnostic to prompts

Our first goal is to estimate the model's uncertainty, which allows the model to say "I do not know", when it has low confidence. Our intuition is: a high-confidence prediction should be agnostic to different prompts.

Step 1: Uncertainty Estimation



Estimate uncertainty with self-consistency

In other words, we estimate model uncertainty by measuring the self-consistency when applying different class-agnostic text prompts.

Finding 2: Explaining the accuracy gap between top-1 (64.2%) and top-5(89.4%)

Failure mode 1: Class name does not specify super-class name

Ground Truth:
Tusker

Misclassified as:
Asian elephant

Parent:
Elephant

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.

Failure mode 2: Class name does not specify sub-class name

Ground Truth:
Balloon

Misclassified as:
Airship

Child:
Hot-air Balloon

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.

Failure mode 3: Inconsistent naming between class names

Ground Truth:
Screw

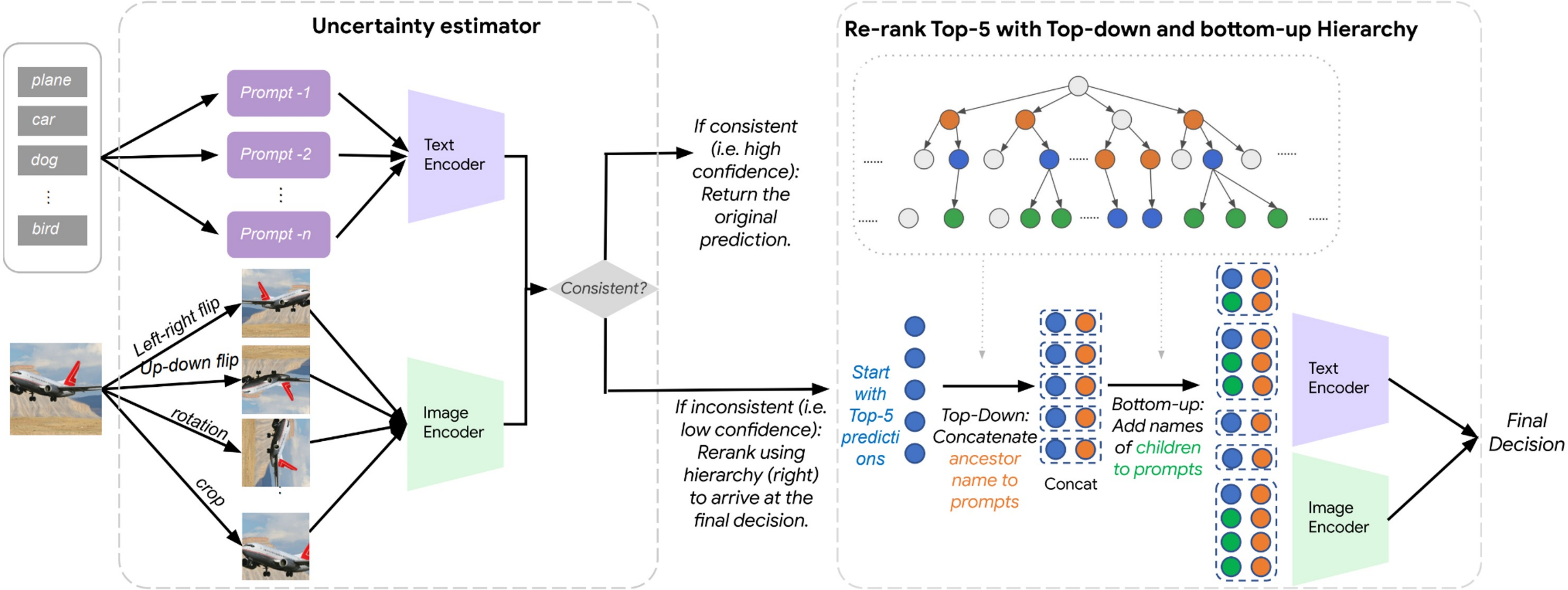
Misclassified as:
Metal Nail

Child:
Allen Screw

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.

We also conduct failure case analysis. Most of the errors are due to the class name lacks information from WordNet hierarchy.

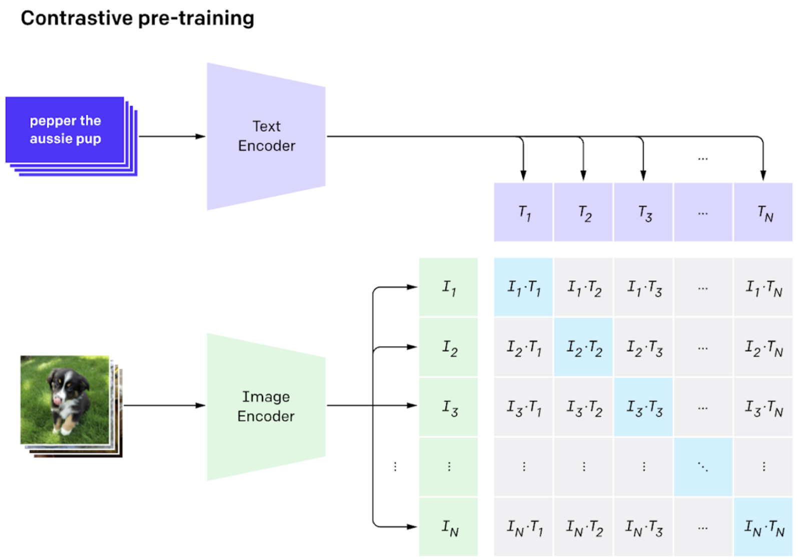
Step 2: Top-down and bottom-up label augmentation using WordNet hierarchy



So we augment the original class name to borrow the WordNet hierarchy knowledge during decision. Our method hyperparameter-free, requires no additional model training and can be easily scaled to other models.

Background: Multi-modal Models (Vision-Language Models)

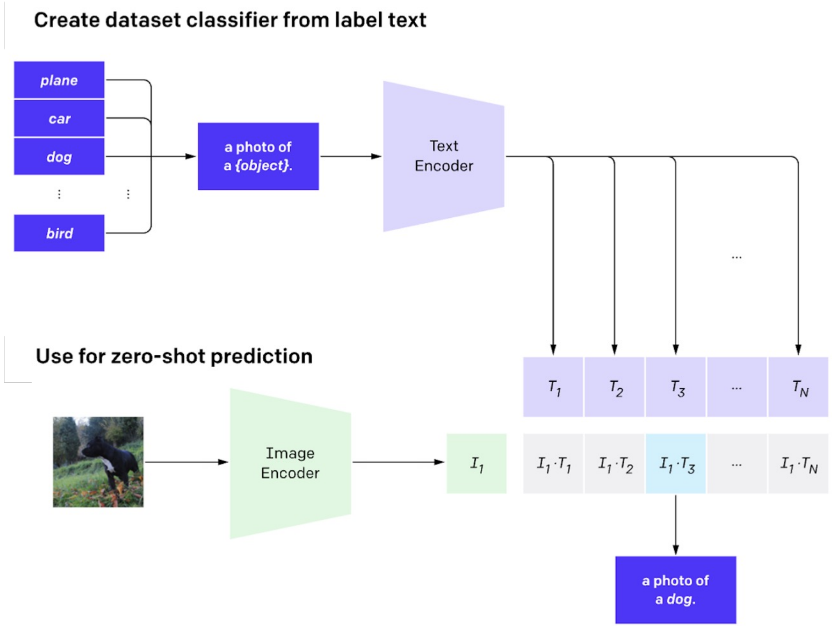
Training with large scale (easy to access) image-text pairs



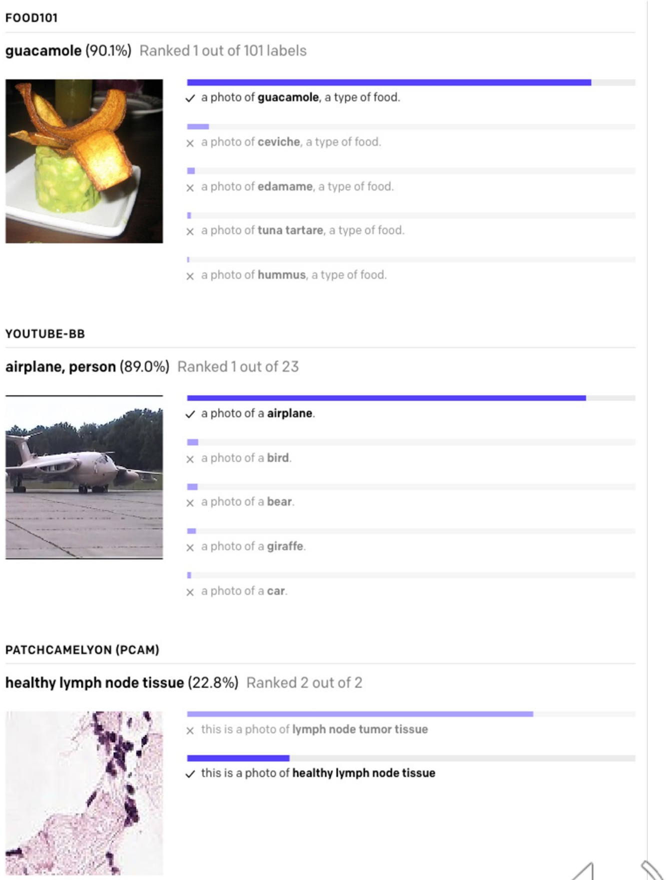
Shared Vision and language embedding

Some background of CLIP: it is trained using large-scale image-text pairs with contrastive loss. The images go through the image encoder, and the text goes through the text encoder. If they are from the same pair, their distance should be small; otherwise, they should have a large distance. CLIP created a shared vision and language embedding.

Background: Multi-modal Models

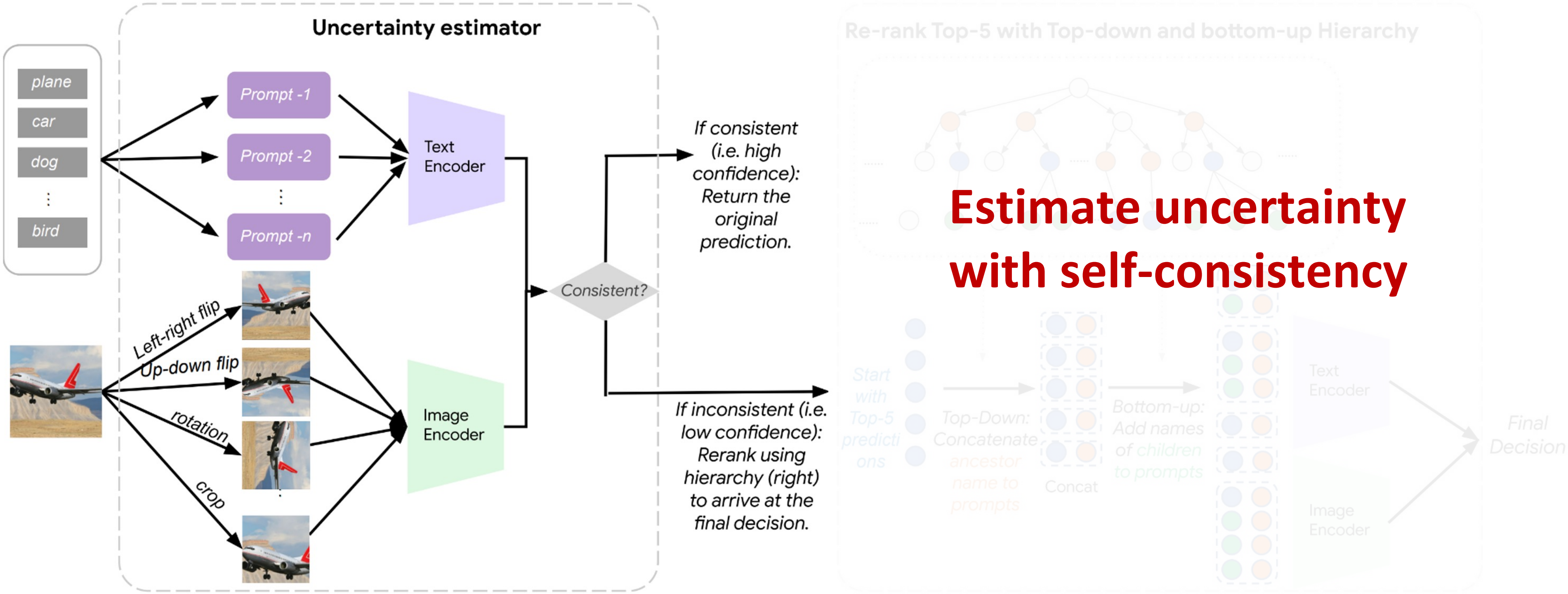


logit = cos(zimg, ztext)



During zero-shot inference, given a test image and candidate class names, they will compare the cosine similarity of the image embedding and all candidate class embedding in the shared latent space, and select the class name with the largest cosine similarity as prediction. On some dataset, they perform well, while in some dataset that requires domain expert knowledge, like medical image, they may make mistakes.

Step 1: Uncertainty Estimation



Estimate uncertainty with self-consistency

To estimate uncertainty, given a test image, we made multiple times decisions by applying different class-agnostic prompts to the candidate classes. For instance, "a good image of, a bad image of…". We calculate the decision consistency as the confidence score. The intuition is, if the decision is not influenced by different prompts, it has high confidence.

Result 1: Our proposed confidence score is better suited for selective prediction than baselines

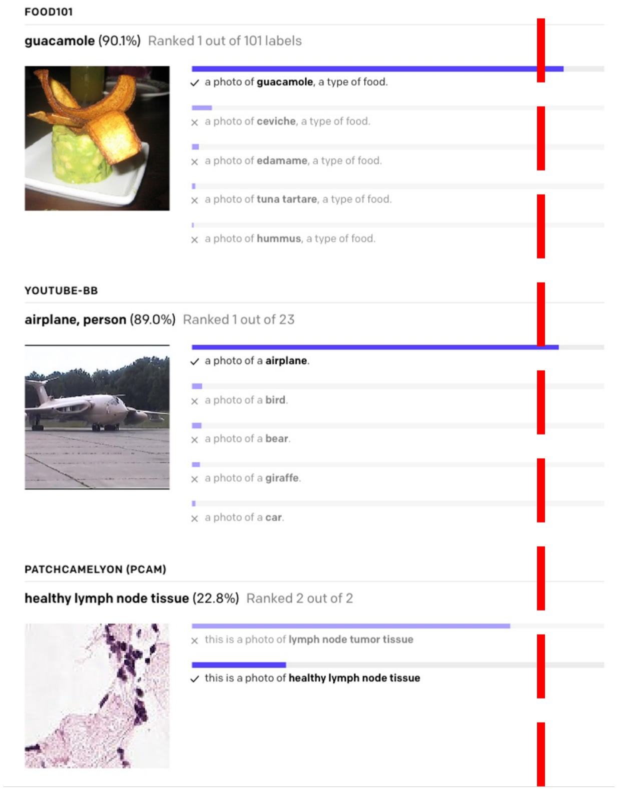
Goal:

high confidence → correct
low confidence → wrong

Baseline:

Max Logits: set threshold of the max logit

max_{K classes} logit_k

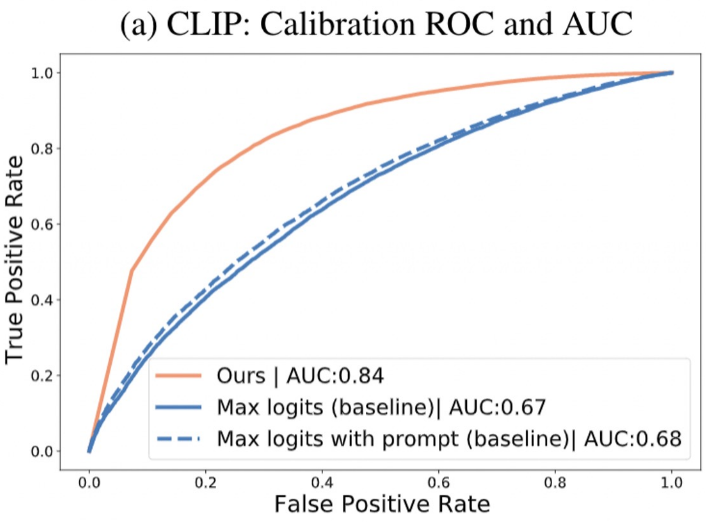


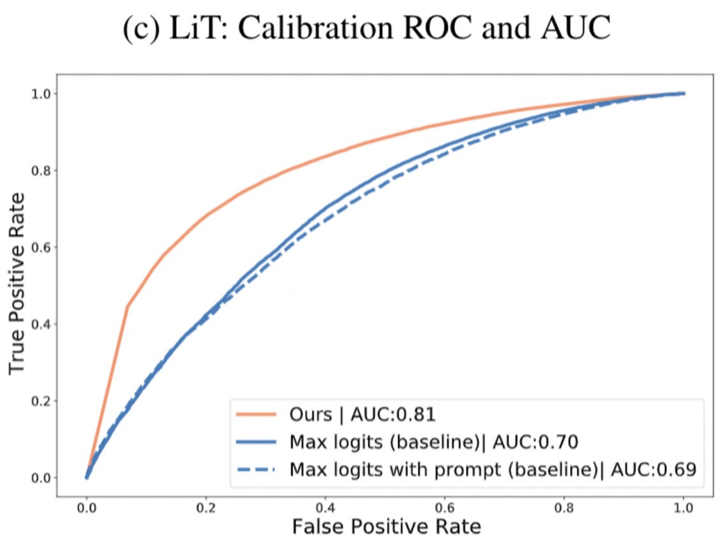
For evaluation, a good confidence score is a signal of the correctness of model prediction: the prediction with high confidence score is correct, and the prediction with a low confidence score is wrong. One baseline is Max Logits, which uses a fixed threshold to estimate confidence.

Result 1: Our proposed confidence score is better suited for selective prediction than baselines

Goal:
high confidence → correct
low confidence → wrong

Baseline:
Max Logits: set threshold of the max logit
max_{K classes} logit_k



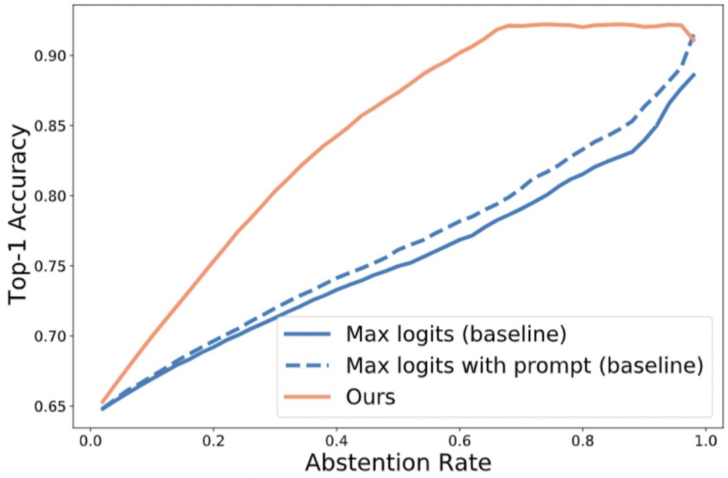


We can compute the AUC score, where we use the confidence score to predict the correctness of the model prediction.
Our self-consistency confidence score (orange curve) is better suited for vision-language models.

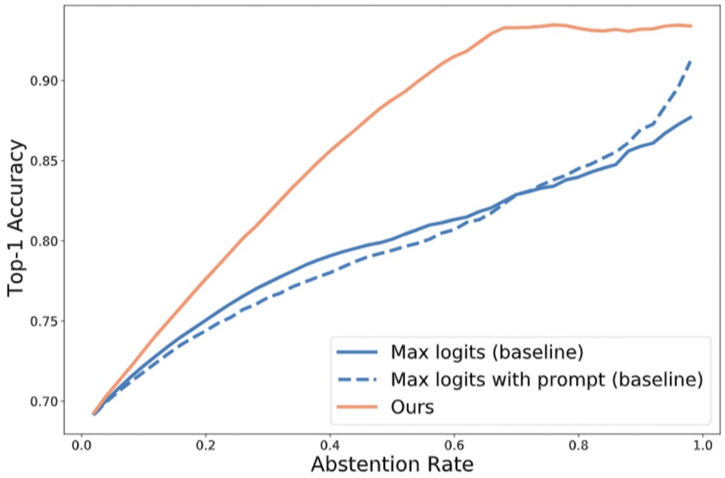
Result 1: Our proposed confidence score is better suited for selective prediction than baselines

Goal: High accuracy on the high confidence set.

(b) CLIP: Selective prediction

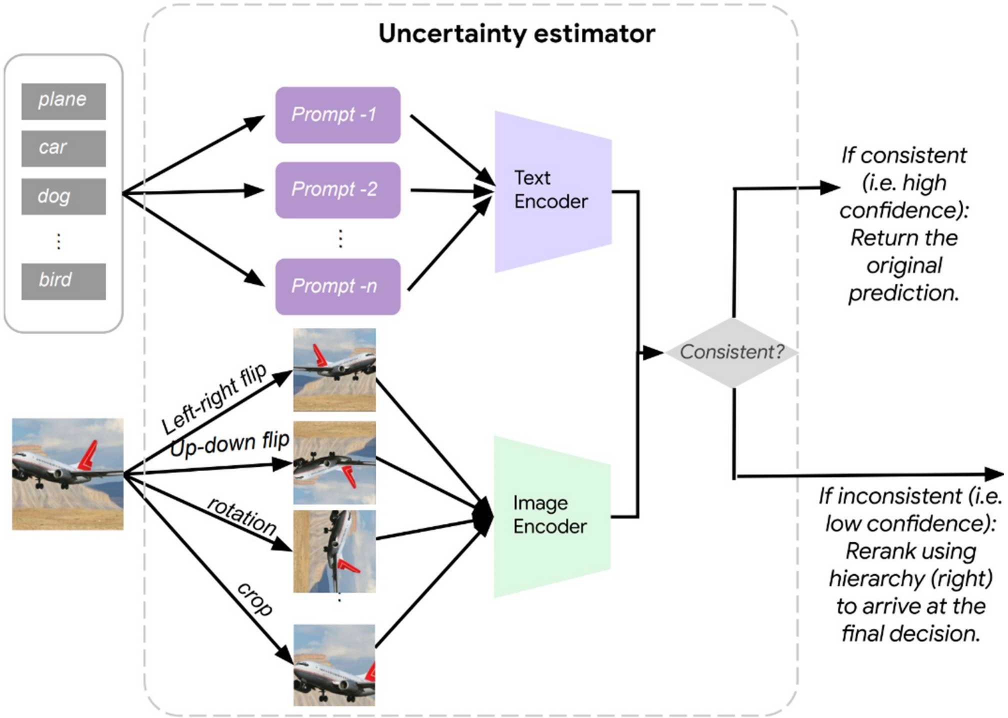


(d) LiT: Selective Prediction



We also evaluate Selective prediction, where we give the model a rejection budget to say “I do not know” on the low confident decision, and we only calculate the accuracy of the high confidence set. Ours performs better on both CLIP and LiT models.

Step 1: Uncertainty Estimation



How to improve the performance on low confidence set?

With uncertainty estimation, we allow the model to say "I do not know" on low confidence set. What if we still want the model to make a decision even though the confidence is low? How to improve the performance on the low confidence set?

Finding 2: Explaining the accuracy gap between top-1 (64.2%) and top-5(89.4%)

Failure mode 1: Class name does not specify super-class name

Ground Truth: Tusker

Misclassified as: Asian elephant

Parent: Elephant

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.

Failure mode 2: Class name does not specify sub-class name

Ground Truth: Balloon

Misclassified as: Airship

Child: Hot-air Balloon

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.

Failure mode 3: Inconsistent naming between class names

Ground Truth: Screw

Misclassified as: Metal Nail

Child: Allen Screw

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.

While the top-5 zero-shot accuracies of these models are very high, the top-1 accuracies are much lower (over a 25% gap in some cases). We conduct failure case analysis. For instance, we find most of the tuskers are wrongly classified as Asian elephants by CLIP. But if we explicitly concatenate the parent class name "elephant" to "tusker" as a prompt, the error is fixed.

Finding 2: Explaining the accuracy gap between top-1 (64.2%) and top-5(89.4%)

Failure mode 1: Class name does not specify super-class name

Ground Truth: Tusker

Misclassified as: Asian elephant

Parent: Elephant

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.

Failure mode 2: Class name does not specify sub-class name

Ground Truth: Balloon

Misclassified as: Airship

Child: Hot-air Balloon

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.

Failure mode 3: Inconsistent naming between class names

Ground Truth: Screw

Misclassified as: Metal Nail

Child: Allen Screw

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.

Most of the balloon are wrongly classified as "airship". But if we check the image, we find actually they are hot-air balloon. Class names like "balloon" are too broad and include different subtypes. Using the child class name "hot-air balloon" fixes such errors.

Finding 2: Explaining the accuracy gap between top-1 (64.2%) and top-5(89.4%)

Failure mode 1: Class name does not specify super-class name

Ground Truth:
Tusker

Misclassified as:
Asian elephant

Parent:
Elephant

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.

Failure mode 2: Class name does not specify sub-class name

Ground Truth:
Balloon

Misclassified as:
Airship

Child:
Hot-air Balloon

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.

Failure mode 3: Inconsistent naming between class names

Ground Truth:
Screw

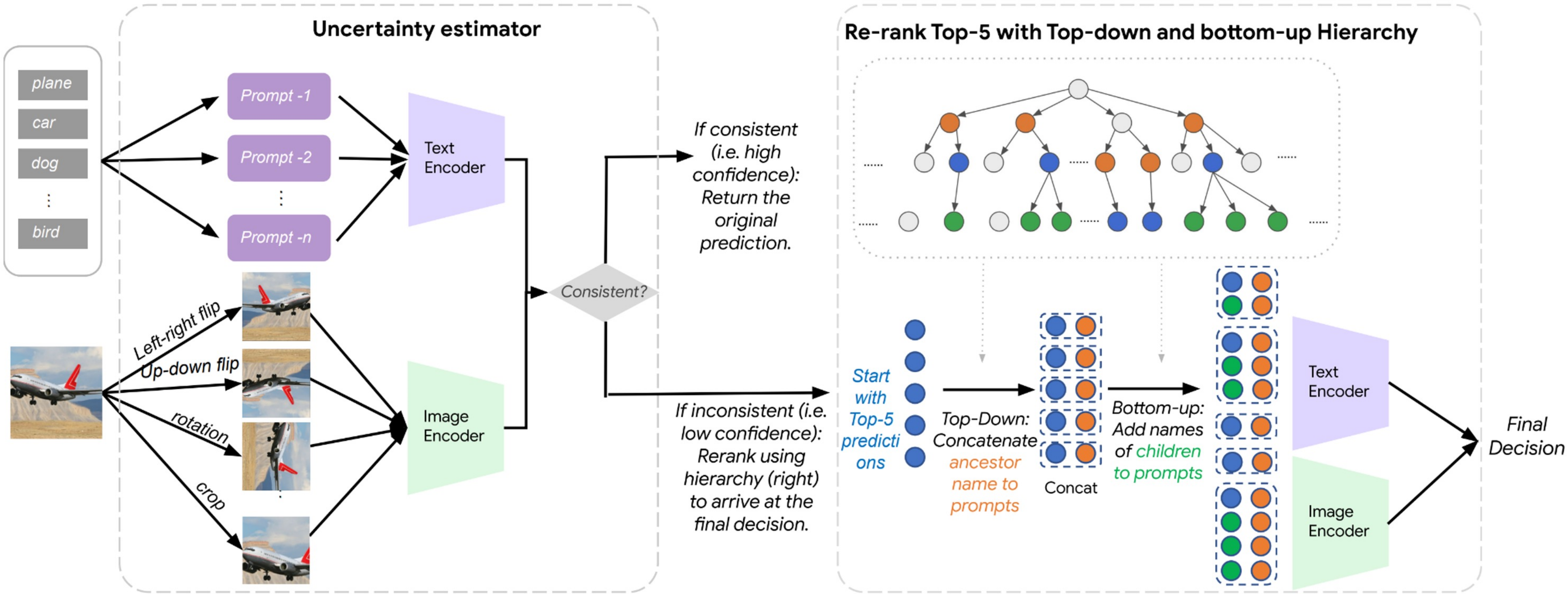
Misclassified as:
Metal Nail

Child:
Allen Screw

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.

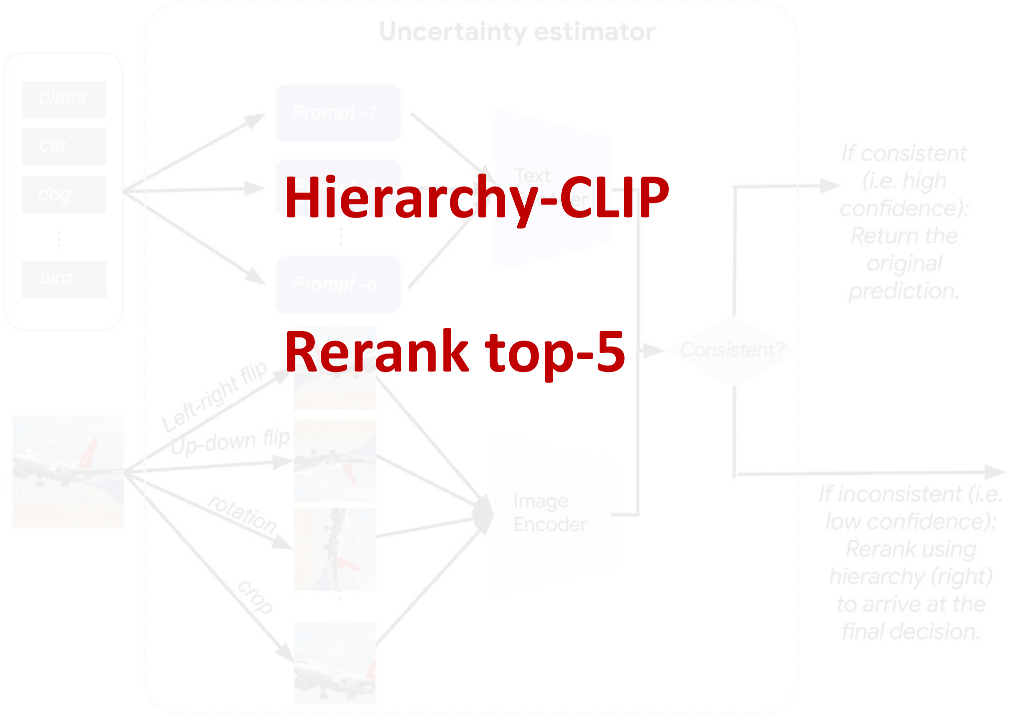
Most of the errors are due to the class name itself may not align well with the image meaning. In other words, class name lacks context information from the WordNet hierarchy.

Step 2: Top-down and bottom-up label augmentation using WordNet hierarchy



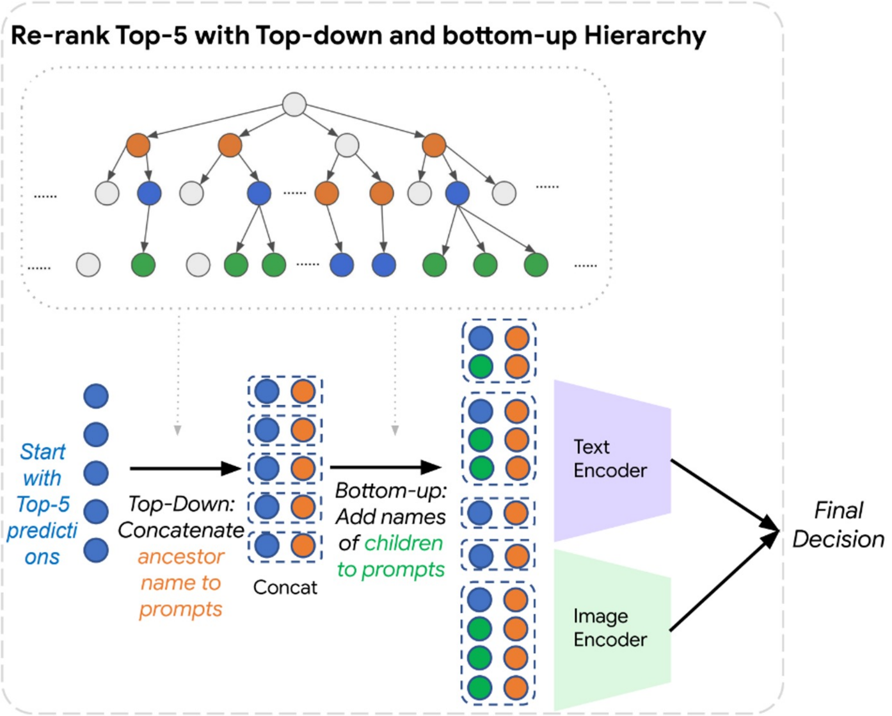
To improve accuracy on low confident set. We re-rank the top-5, and augment the original class name with top-down and bottom-up label augmentation to borrow the wordnet hierarchy knowledge during zero-shot inference.

Step 2: Top-down and bottom-up label augmentation using WordNet hierarchy



Hierarchy-CLIP

Rerank top-5



Given top-5 prediction, we first use top-down WordNet hierarchy to concatenate the ancestor names to prompts: for instance: “husky, dog” “tusker elephant”, which provide context information. Then we use a bottom-up hierarchy to add the children's classes to the candidate class. In the decision, we re-rank the top-5 class based on the highest cosine similarity in each class group after augmentation.

Result 2: Using hierarchy to help improve zero-shot accuracy on low confidence subset

Table 1. CLIP (ViT-B/16) and LiT (ViT-B/32) zero-shot top-1 accuracy comparison between baseline and ours (w/ hierarchy).

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

We conduct zero-shot classification on ImageNet and its variant with both CLIP and LiT models. We find our method significantly improves the accuracy on the low confidence set (over 17 percent point improvement), and overall also improves the whole ImageNet performance (3.6 percent point improvement).

Result 2: Using hierarchy to help improve zero-shot accuracy on low confidence subset

Table 2. Generalizability to non-ImageNet datasets (CLIP (ViT-B/16) zero-shot top-1 accuracy).

Dataset	orig (low)	ours (low)	orig (full)	ours (full)
Caltech-101 [15]	10.6 %	27.2% (+16.6%)	74.1%	77.1% (+3.0%)
Flower102 [17]	20.0%	29.4% (+9.4%)	63.7%	65.3% (+1.6%)
Food-101 [2]	28.2%	49.0% (+20.8%)	84.7%	86.8% (+2.1%)
Cifar-100 [13]	9.4%	17.5% (+8.1%)	31.8%	35.2% (+3.4%)

Our method also show consistent improvement on other datasets: Caltech-101, Flower-102, Food-101 and Cifar-10

Results 3: Our hierarchy-based label augmentation is complementary to prompt ensembling

Table 2. CLIP (ViT-B-16) zero-shot top-1 accuracy comparison with prompt ensemble.

		Ensemble only	Hierarchy and Ensemble
ImageNet	Low conf. set	41.05%	**42.09%**
	Full set	68.48%	**68.86%**
ImageNet-v2	Low conf. set	**36.39%**	36.34%
	Full set	**62.02%**	62.00%
ImageNet-R	Low conf. set	35.13%	**36.12%**
	Full set	60.21%	**60.62%**
ImageNet-Adversarial	Low conf. set	21.13%	**22.00%**
	Full set	30.59%	**31.07%**
ImageNet-Sketch	Low conf. set	**27.13%**	26.56%
	Full set	**48.52%**	48.26%

Our hierarchy-based label augmentation is complementary to prompt ensembling.

Results 4: Ablation Study

Generalizability to other backbones

Table 3. Generalizability to different backbones with CLIP.

backbone	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16	ViT-l/14
ACC (low)	+14.25%	+12.97%	+15.12%	+ 17.13%	+18.89%
ACC (full)	+3.73%	+3.71%	+3.65%	+ 3.60%	+3.23%

Effect of threshold of confidence score on zero-shot accuracy.

Table 5. Effect of threshold of confidence score on zero-shot accuracy.

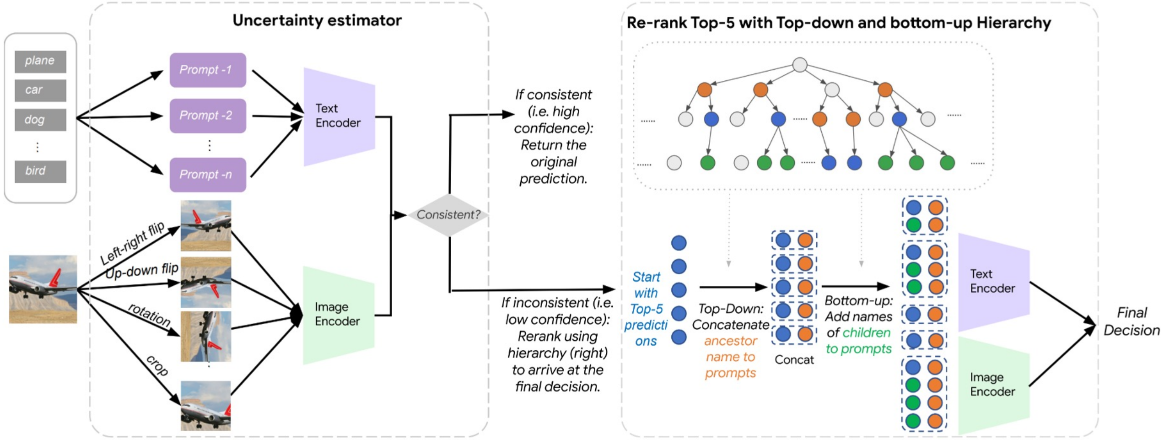
Threshold	Low conf. set size	Acc on low conf. set	Acc on full set
0.47	10000	19.40%	68.72%
0.52	11000	20.82%	68.78%
0.57	12000	22.06%	68.82%
0.62	13000	23.58%	68.85%
0.66	14000	25.01%	**68.88%**
0.70	15000	26.51%	68.86%

We also show the generalization to other backbones and ablation studies.

Conclusion

Paper Code Project page

session and poster ID
WED-AM-272



- **[Confidence Estimation]** We propose a simple yet efficient zero-shot confidence score that is better suited for multi-modal models, based on **predictions' self-consistency under different text prompts and image perturbations.**
- **[Failure Case Analysis]** We identified several failure modes for zero-shot ImageNet classification using multi-modal models.
- **[Improve Top-1 accuracy with Hierarchy]** We develop a label augmentation technique that uses both ancestor and children labels from WordNet. By applying the label augmentation to the previously identified low confidence subset of images, we significantly improve their prediction accuracy
- Our method is hyperparameter-free, requires no additional model training and can be easily scaled to other models.