A. Dataset Statistics

Table 7 depicts detailed statistics for all datasets. For each dataset, we provide in parentheses a one-word description of the type of classes it contains, which we refer to as *super class* of a dataset. We use the same train/dev/test splits of Food-101, Aircraft, Flower-102, UCF-101, and DTD provided by CoOp [74]. For CUB, we randomly sample 10 training images for each category as the development set. For CIFAR-10 and CIFAR-100, we randomly split 10% of the training data as the dev set. For HAM10000, we adopt 80/10/10 splits on the images of each class. For ImageNet, we only evaluate the dev set.

Name	**n. of class**	**n. of Images**		
		Train	Dev	Test
Food-101 (food)	101	50,500	20,200	30,300
FGVC-Aircraft (aircraft)	102	3,334	3,333	3,333
Flower-102 (flower)	102	4,093	1,633	2,463
CUB-200-2011 (bird)	200	3,994	2,000	5,794
UCF-101 (action)	101	7,639	1,898	3,783
DTD (texture)	47	2,820	1,128	1,692
HAM10000 (lesion)	7	8,010	1,000	1,005
RESISC45 (scene)	45	3,150	3,150	25,200
CIFAR-10 (object)	10	45,000	5,000	10,000
CIFAR-100 (object)	100	45,000	5,000	10,000
ImageNet (object)	1,000	1,281,167	50,000	-

Table 7. Detailed statistics of the 11 datasets. The text in parentheses that follows the dataset name corresponds to the super class name, which is used to remove class names in concepts.

B. Implementation Details

B.1. Linear Probe

Following CLIP's implementation of Linear Probe, we use the encoded images, before their projection to the vision-text embedding space, as input to the classifier. We use sklearn's L-BFGS implementation of logistic regression with 1,000 maximum iterations. To determine the best performing values for the L2 regularization strength C, we perform binary search on the validation set initialized with $[1e^6, 1e^4, 1e^2, 1, 1e^{-2}, 1e^{-4}, 1e^{-6}]$. After determining the left and right bounds of C, we iteratively halve the interval with 8 steps to get the final hyperparameter value. We compare our Linear Probe results on ImageNet with CoOp. To perform a fair comparison, we select CLIP-RN50 as the vision encoder and perform 3 random runs to select the few shot images. As shown in Table 8, we marginally outperform CoOp in all data settings.

B.2. Prompt

Table 9 presents the prompts used to query GPT-3. We design 5 general prompts and 5 additional prompts for UCF-101. The general prompts are used for all datasets, with a slight modification: we add the super-class name that describes the type of data present in more fine-grained datasets. For example, when prompting for Flower-102, we add the super class name *flower* after each class name. In this way we reduce ambiguity problems: e.g., for the class *bishop of llandaff*, without the super class name, GPT-3 returns results for *bishop* instead of the *flower*. While this approach reduces ambiguities, it does not completely eliminate them. For example, we found that GPT-3 generates sentences about the *mouse* (device), but in fact, the class *mouse* on ImageNet refers to the animal. Future work can explore better prompting methods, such as providing a detailed definition for each class or designing customized prompts for each dataset.

# of shots	**1**	**2**	**4**	**8**	**16**
CoOp	22.07	31.95	41.29	49.55	55.87
Ours	**22.26**	**32.28**	**41.57**	**49.80**	**55.92**

Table 8. Compare linear probe performance on ImageNet with CoOp. All experiments are based on CLIP-RN50, and we report the average score of 3 random runs.

General Prompt Template
1. describe what the `[CLASS NAME]` looks like:
2. describe the appearance of the `[CLASS NAME]`:
3. describe the color of the `[CLASS NAME]`:
4. describe the pattern of the `[CLASS NAME]`:
5. describe the shape of the `[CLASS NAME]`:
UCF-101 Prompt Template
1. describe what the `[CLASS NAME]` looks like:
2. describe the appearance of the `[CLASS NAME]`:
3. describe how to perform the `[CLASS NAME]`:
4. describe a person performing the `[CLASS NAME]`:
5. describe what can you see when a person is performing the `[CLASS NAME]`:

Table 9. The prompt templates used to generate the raw sentences from GPT-3. The UCF-101 has a different set of prompts, while the other datasets share the same set of general templates.

B.3. T5 concept extractor

The raw outputs of language models are long sentences and sometimes contain class names that need to be removed from the bottlenecks for the sake of interpretability. For example, GPT-3 generates a sentence "*The hen is brown and has a white chest.*" for the class *hen*, which could be decomposed to two concepts: "*brown*" and "*white chest*". We annotate a random sample of 100 sentence-concepts pairs from each of the following datasets: Food-101, CIFAR-100, Aircraft, Flower, and ImageNet. In total, we collect 500 sentences. An example annotation is depicted below:

> The 737-400 has a long and slender fuselage with
> tapered wings and a small tail. (737-400)
> long and slender fuselage; tapered wings; small tail

The class name is concatenated with the raw sentence, and

Dataset	Method	Dev						Test					
		1	2	4	8	16	Full	1	2	4	8	16	Full
Food-101	Linear Prob	58.04	75.24	84.16	**87.48**	**89.87**	**93.11**	57.75	75.34	84.21	**87.90**	**90.02**	**93.17**
	LaBo (Ours)	**80.32**	**84.15**	**85.76**	87.07	88.74	92.53	**80.41**	**84.05**	**85.68**	87.39	88.77	92.45
Aircraft	Linear Prob	27.63	34.86	41.40	**49.72**	**57.91**	**62.89**	28.26	35.07	**41.55**	**50.26**	**56.38**	**64.03**
	LaBo (Ours)	**33.12**	**35.97**	**42.90**	49.08	56.41	61.96	**32.73**	**37.71**	41.04	48.81	54.97	61.42
Flower-102	Linear Prob	**89.20**	**94.06**	**97.00**	**98.40**	**98.91**	**99.11**	**88.06**	**93.65**	**97.67**	**98.56**	**99.32**	**99.45**
	LaBo (Ours)	82.24	88.18	94.92	96.20	98.16	98.65	82.05	90.09	95.21	97.08	98.66	99.35
CUB	Linear Prob	48.55	60.40	**72.50**	**78.25**	**83.35**	**83.60**	47.69	61.06	**72.82**	**79.60**	**83.74**	**84.54**
	LaBo (Ours)	**55.20**	**64.80**	72.45	76.55	79.90	81.00	**54.19**	**64.60**	71.21	77.22	80.69	81.90
UCF-101	Linear Prob	65.54	76.34	85.83	90.25	**93.63**	**98.63**	60.56	73.22	80.62	85.70	**87.63**	**90.67**
	LaBo (Ours)	**80.72**	**83.77**	**88.46**	**90.73**	93.05	97.68	**78.75**	**82.05**	**84.56**	**86.39**	87.39	90.11
DTD	Linear Prob	43.62	53.19	60.55	**68.79**	**74.47**	**80.50**	41.67	51.71	60.76	**69.03**	**74.70**	**81.68**
	LaBo (Ours)	**55.59**	**56.47**	**62.15**	68.44	70.92	76.86	**53.61**	**55.26**	**61.17**	66.43	70.21	77.30
HAM10000	Linear Prob	32.30	**55.40**	45.40	50.90	**63.10**	**84.40**	33.13	**55.32**	44.48	48.26	**61.69**	**83.18**
	LaBo (Ours)	**34.90**	46.40	**45.80**	**54.40**	58.20	81.40	**36.62**	45.17	**45.87**	**52.04**	55.72	81.39
RESISC45	Linear Prob	68.62	**79.10**	**86.72**	**89.89**	**92.49**	**95.24**	67.57	**77.75**	**86.50**	**89.27**	**92.17**	**94.98**
	LaBo (Ours)	**73.02**	76.03	81.37	85.05	88.86	91.65	**73.66**	76.11	81.40	85.71	88.63	91.22
CIFAR-10	Linear Prob	62.36	80.32	92.94	**95.36**	**96.06**	**98.16**	62.44	80.27	92.54	**95.14**	**95.90**	**98.10**
	LaBo (Ours)	**91.24**	**91.04**	**92.98**	94.40	95.06	97.90	**91.06**	**90.79**	**93.03**	94.11	94.93	97.75
CIFAR-100	Linear Prob	39.66	57.84	70.06	**76.52**	**80.34**	**87.70**	39.26	57.35	69.73	**76.22**	**80.16**	**87.48**
	LaBo (Ours)	**62.84**	**66.56**	**71.78**	75.30	78.08	86.82	**62.73**	**65.80**	**70.82**	74.49	77.67	86.04
ImageNet	Linear Prob	42.25	55.71	**64.80**	**71.23**	**75.08**	83.90	-	-	-	-	-	-
	LaBo (Ours)	**51.09**	**57.43**	62.94	68.45	72.60	**83.97**	-	-	-	-	-	-
Average	Linear Prob	52.53	65.68	72.85	**77.89**	**82.29**	**87.93**	51.69	65.13	**72.33**	**77.38**	**81.53**	**87.38**
	LaBo (Ours)	**63.66**	**68.25**	**72.86**	76.88	80.00	86.40	**63.35**	**68.10**	72.08	76.19	79.11	85.72

Table 10. Full results of Linear Prob and LaBo on the development and test sets of 11 datasets.

the concepts are separated by semicolons. We train a T5-large model [45] using the Huggingface API. We add a task prefix - "*extract concepts from sentence:* " for each example. We train the model with Adam optimizer for 5 epochs, setting the batch size to 8 and learning rate to $1e^{-5}$.

B.4. Remove Class Name

After extracting the short concepts using T5, some still contain class names. To ensure there are no class names in the bottleneck, we design two heuristics: (1) If we find the class name in the concept using string match, we replace it with the super class name[10], e.g., the concept "*leaves of the **orange dahlia** are long and narrow*" for the class *orange dahlia* in Flower-102 is modified as "*leaves of the **flower** are long and narrow*". (2) For class names with multiple tokens, the tokens are not always in the same order as the class name. In this case, if a concept with all tokens for the class name is present, we remove it. For instance, the concept "*a **cake** made of **carrot***" for the class *carrot cake* will be deleted. The two heuristics are applied to each concept by considering all class names in the dataset.

B.5. Hyperparameters

We apply grid search with 5 runs to find the best weights for the submodular function for different datasets and shots. We determine the learning rate and batch size by monitoring the validation accuracy with wandb. Table 16 lists all the hyperparameters of our best-performing models.

B.6. Other Details

GPT-3 Generation. Generating 500 sentences for one class takes around 5 minutes by calling the OpenAI APIs. The price of GPT-3-Davinci is $ 0.02 / 1k tokens, and it costs about $ 0.2 for each class.

Running Time. Because we use CLIP with frozen weights, we only need to extract the image features once and reuse them in the rest experiments. Since we only fit a single linear layer, our training time is low. For example, training the full ImageNet for one epoch on an NVIDIA RTX A6000 takes less than 1 minute.

Full Results. The full numerical results are shown in Table 10. Both validation and test accuracy are provided.

C. Additional Analysis

C.1. Activation Function

We ablate the impact of the softmax activation by removing it or replacing it with other activation functions such as ReLU and sigmoid. As shown in Table 11, not using an activation function significantly hurts performance, while

[10] The super class name depends on the datasets. For example, the super class name for the Flower-102 dataset is *flower* (see Table 7).

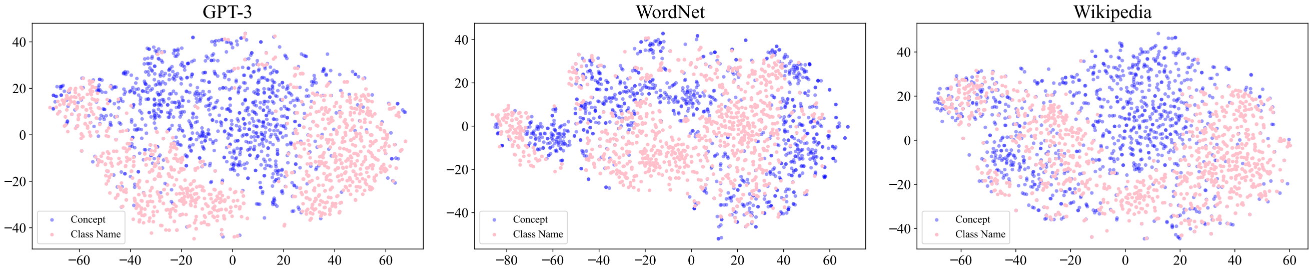


Figure 8. t-SNE visualization of the embeddings of concepts (blue) and class names (pink) on ImageNet. For the three bottlenecks constructed from GPT-3, WordNet, and Wikipedia, we visualize the top-1 concept of each class ranked by the weights of the linear function.

Activation	1	2	4	8	16	Full
-	52.66	58.01	63.02	68.93	73.52	81.32
relu	50.40	53.53	56.61	59.82	61.75	68.01
sigmoid	52.15	57.86	62.59	69.08	73.43	81.42
softmax	**63.03**	**67.79**	**71.88**	**76.08**	**79.10**	**85.71**

Table 11. Compare different activation functions. We report the mean accuracy across the 11 datasets.

GPT-3 type	1	2	4	8	16	Full
Davinci (175B)	**51.09**	**57.43**	**62.94**	**68.45**	**72.60**	83.97
Curie (13B)	45.75	53.89	60.36	66.96	71.65	**84.00**
Babbage (6.7B)	44.61	52.91	60.22	67.06	71.66	83.86
Ada (2.7B)	43.12	53.26	60.99	67.90	72.42	83.96

Table 12. The performance of LaBo on ImageNet using different sizes of GPT-3 to generate concepts. The number in the parenthesis is the number of parameters of the corresponding language model.

using other activation functions performs poorly compared to softmax.

C.2. Language Model Size vs. Performace

We experiment with different sizes of GPT-3: Curie, Babbage, and Ada (sorted from larger to smaller). Figure 12 compares the different GPT-3 variants on ImageNet, showing that larger language models result in better performance, especially in the few show settings. However, there is only a marginal difference in performance when enough data is available.

C.3. Performance of Human-Written Text

Table 13 compares the performance of LaBo between using GPT-3 generated concepts and human-designed concepts sourced from WordNet and Wikipedia. We observe that GPT-3 generated concepts outperform human-written ones in 1-shot experiments, while there is less than 1% drop in performance on average in larger data settings. In addition, our human evaluation on Imagenet (see Figure 5 and 6 in Section 5.3) shows that humans judge the quality of GPT-3 generated concepts to be better than that of human-designed.

We visualize the embeddings of concepts and class names using t-SNE [65] to identify the reason behind the perceived higher quality of GPT-3 concepts. We encode the 1,000 class names of ImageNet using the CLIP text encoder along with the top-1 concept of each class (1,000 concepts in total) from each bottleneck (LaBo, WordNet, and Wikipedia). Figure 8 reflects that, compared to GPT-3, the embeddings of WordNet and Wikipedia concepts have a higher overlap with the embeddings of class names. In other words, Wikipedia and WordNet concepts are more likely to replicate the text features of class names rather than describe the class. This explains why human-written text has higher accuracy but is less interpretable.

Concept Source	1	2	4	8	16	Full
GPT-3	**51.09**	57.43	62.94	68.45	72.60	83.97
Wikipedia	48.76	56.73	63.00	68.96	73.07	**84.07**
WordNet	49.37	**57.84**	**64.10**	**69.92**	**73.35**	83.93

Table 13. The performance of LaBo on ImageNet using different sources of concepts to construct the bottlenecks.

Method	w/ cls	Aircraft	Food	Flower	DTD	UCF
LP	-	39.42	76.99	95.89	68.74	80.04
LaBo	✗	37.29	76.04	92.37	64.78	80.07
CoOp [74]	✓	33.22	**78.45**	**94.97**	**65.37**	78.66
LaBo†	✓	**37.53**	77.83	93.18	**65.37**	**80.10**

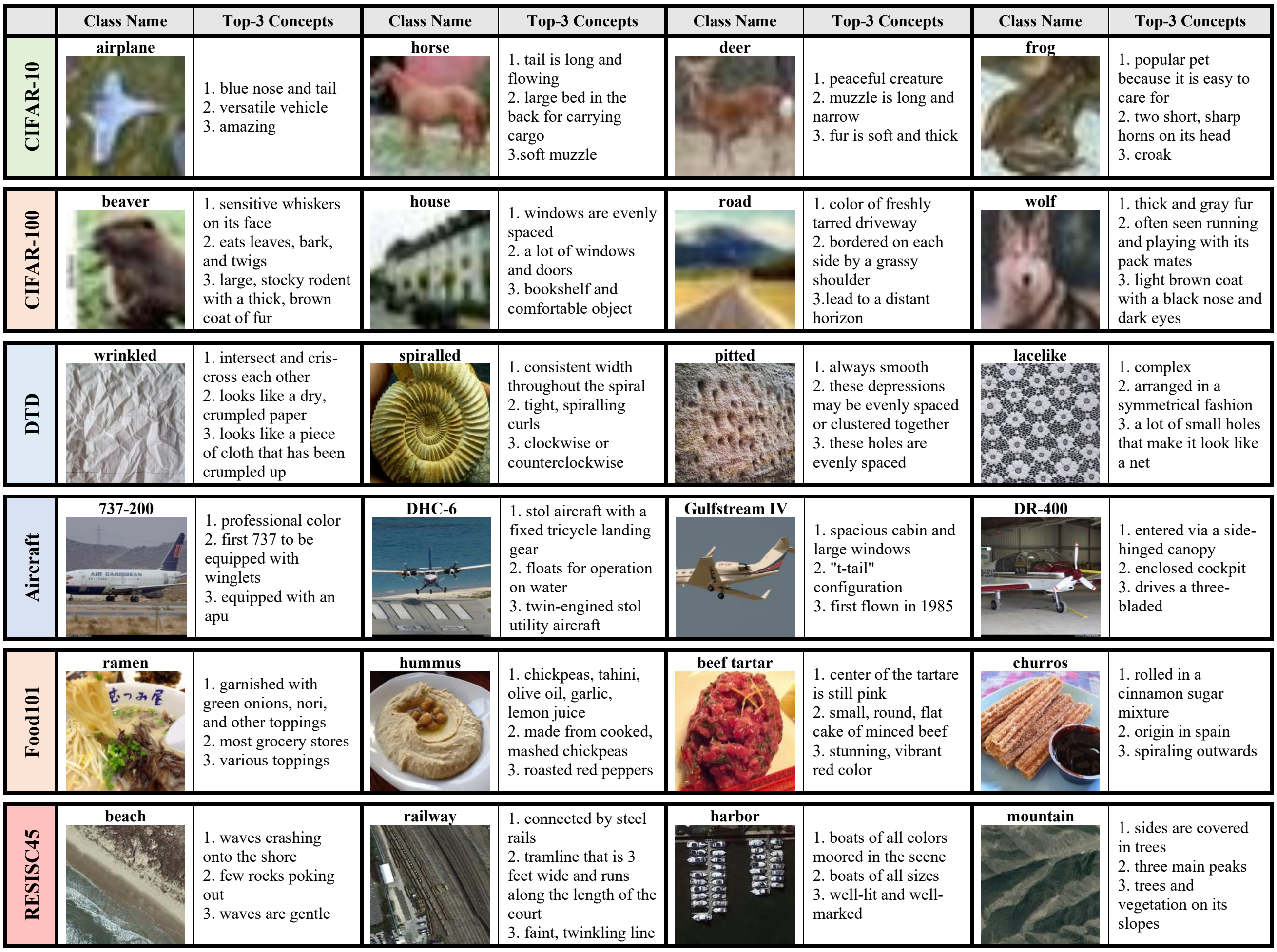
Table 14. Compare LaBo with prompt tuning methods on 5 datasets (16 shots). w/ cls stands for using class names in the context. LaBo† is our method without removing the class names in the concepts. All methods use CLIP-ViT-B/32 as the vision backbone.

C.4. Comparison with the Prompt Tuning Method

Table 14 compares the performance between LaBo and CoOp [74], which employs a soft prompt tuning method (not interpretable) on five datasets. Even though LaBo does not use class names, its performance is similar to that of CoOp. Adding class names to LaBo leads to performance gains, such that it outperforms CoOp on Aircraft and UCF-101.

D. Human Evaluation

We introduce two qualitative metrics to evaluate the automatically generated concept bottlenecks to highlight areas of possible improvement. We introduce two metrics that evalu-



	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
CIFAR-10	airplane	1. blue nose and tail 2. versatile vehicle 3. amazing	horse	1. tail is long and flowing 2. large bed in the back for carrying cargo 3.soft muzzle	deer	1. peaceful creature 2. muzzle is long and narrow 3. fur is soft and thick	frog	1. popular pet because it is easy to care for 2. two short, sharp horns on its head 3. croak
CIFAR-100	beaver	1. sensitive whiskers on its face 2. eats leaves, bark, and twigs 3. large, stocky rodent with a thick, brown coat of fur	house	1. windows are evenly spaced 2. a lot of windows and doors 3. bookshelf and comfortable object	road	1. color of freshly tarred driveway 2. bordered on each side by a grassy shoulder 3.lead to a distant horizon	wolf	1. thick and gray fur 2. often seen running and playing with its pack mates 3. light brown coat with a black nose and dark eyes
DTD	wrinkled	1. intersect and cris-cross each other 2. looks like a dry, crumpled paper 3. looks like a piece of cloth that has been crumpled up	spiralled	1. consistent width throughout the spiral 2. tight, spiralling curls 3. clockwise or counterclockwise	pitted	1. always smooth 2. these depressions may be evenly spaced or clustered together 3. these holes are evenly spaced	lacelike	1. complex 2. arranged in a symmetrical fashion 3. a lot of small holes that make it look like a net
Aircraft	737-200	1. professional color 2. first 737 to be equipped with winglets 3. equipped with an apu	DHC-6	1. stol aircraft with a fixed tricycle landing gear 2. floats for operation on water 3. twin-engined stol utility aircraft	Gulfstream IV	1. spacious cabin and large windows 2. "t-tail" configuration 3. first flown in 1985	DR-400	1. entered via a side-hinged canopy 2. enclosed cockpit 3. drives a three-bladed
Food101	ramen	1. garnished with green onions, nori, and other toppings 2. most grocery stores 3. various toppings	hummus	1. chickpeas, tahini, olive oil, garlic, lemon juice 2. made from cooked, mashed chickpeas 3. roasted red peppers	beef tartar	1. center of the tartare is still pink 2. small, round, flat cake of minced beef 3. stunning, vibrant red color	churros	1. rolled in a cinnamon sugar mixture 2. origin in spain 3. spiraling outwards
RESISC45	beach	1. waves crashing onto the shore 2. few rocks poking out 3. waves are gentle	railway	1. connected by steel rails 2. tramline that is 3 feet wide and runs along the length of the court 3. faint, twinkling line	harbor	1. boats of all colors moored in the scene 2. boats of all sizes 3. well-lit and well-marked	mountain	1. sides are covered in trees 2. three main peaks 3. trees and vegetation on its slopes

Figure 9. Additional qualitative examples for CIFAR-10, CIFAR-100, DTD, Aircraft, Food101 and RESISC45.

	Food	Aircraft	HAM10K	RESISC	Flower	CUB	UCF	DTD	CIFAR10	CIFAR100
Factuality ↑	P@10	P@10	P@10	P@10	P@8	P@10	P@10	P@10	P@10	P@10
LaBo	**33.07**	**11.57**	15.05	14.80	11.48	**27.97**	**37.78**	23.90	14.70	22.48
w/o submod	27.08	8.10	9.57	**16.40**	**18.58**	23.12	37.22	**25.27**	**20.70**	**22.72**
w/o LM	21.63	8.97	**19.71**	12.15	9.98	12.17	20.43	14.83	6.87	14.97
Groundability↑	P@10	P@10	P@10	P@10	P@8	P@10	P@10	P@10	P@10	P@10
LaBo	10.98	8.48	18.83	13.87	9.53	15.63	8.08	8.90	5.70	19.83
w/o submod	**21.52**	**13.67**	17.22	**17.90**	**21.52**	23.07	**29.93**	20.02	**23.10**	21.78
w/o LM	20.58	12.00	**20.00**	14.38	17.93	**25.02**	27.96	**20.31**	7.15	**27.04**

Table 15. Analytic Factuality and Groundability for all datasets except Imagenet (see Figure 5)

ate the bottleneck items along two dimensions: *Factuality* and *Groundability* (see Section 5.3).

Annotator Statistics. Both metrics rely on human annotations, which we collect on Amazon Mechanical Turk. To ensure confidence in the results, we collect 3 annotations per concept. Annotators are paid on average $14.5 per hour, and the total cost of the annotation was $2,100. Our rate was computed by estimating the time it takes to complete the task by 4 different control annotators.[11] In total, our task was completed by a diverse set of 477 annotators. The average pairwise annotator agreement for all annotated data without any pre-processing is 69.83%.

Interface. Figure 11 displays the annotation interface. Given a concept phrase, annotators are prompted to select from 12

[11] Our focus group was graduate students. Since this is not representative of the average population, we doubled the time estimate.

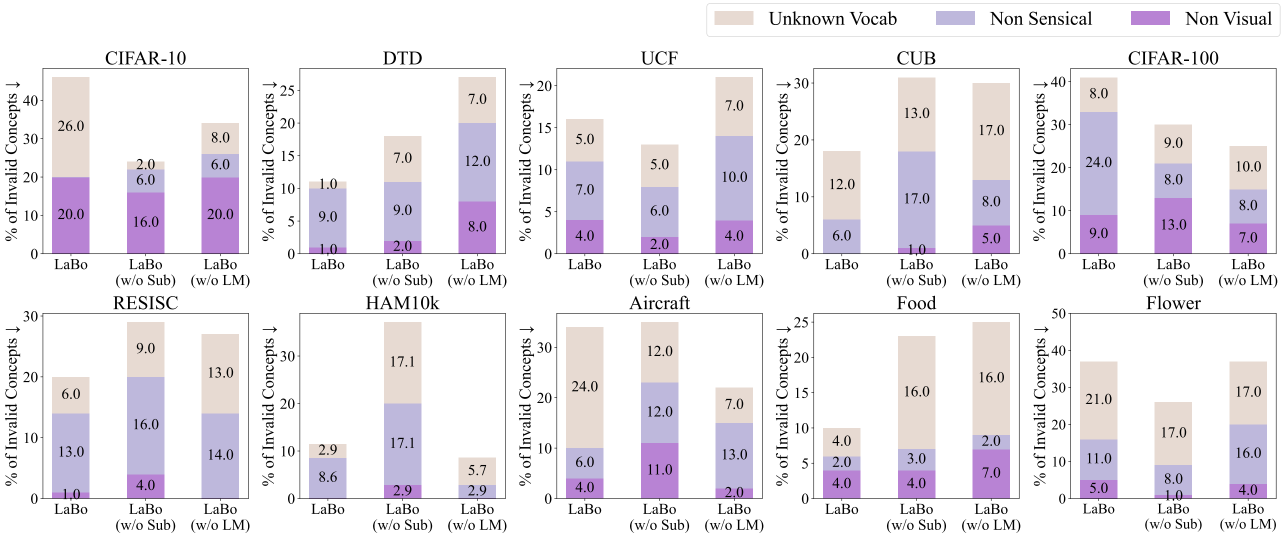
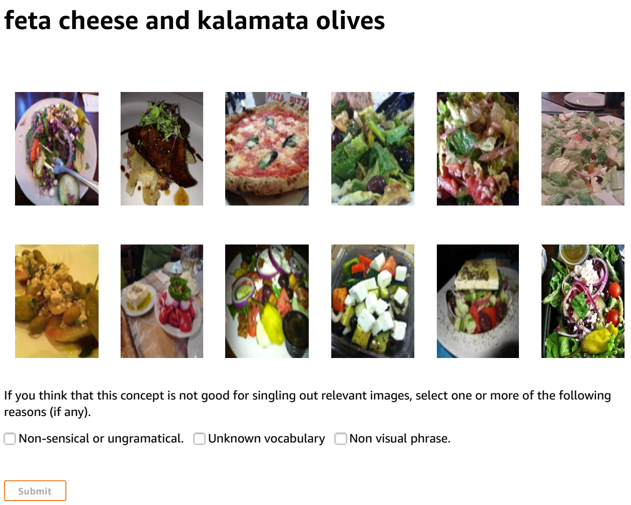


Figure 10. Percentage of invalid concepts identified by humans for different bottlenecks for all 10 datasets except ImageNet (see Figure 6). **Lower** percentage is better.



feta cheese and kalamata olives

If you think that this concept is not good for singling out relevant images, select one or more of the following reasons (if any).

□ Non-sensical or ungramatical. □ Unknown vocabulary □ Non visual phrase.

Figure 11. Sample user interface for measuring *Factuality*. We provide 10 ground truth images with 2 control images randomly positioned. Annotators are required to select the images that can be described by the phrase. The user interface for *Groundability* is identical, but the images presented are the top-10 images in the dataset sorted by CLIP [44] similarity score.

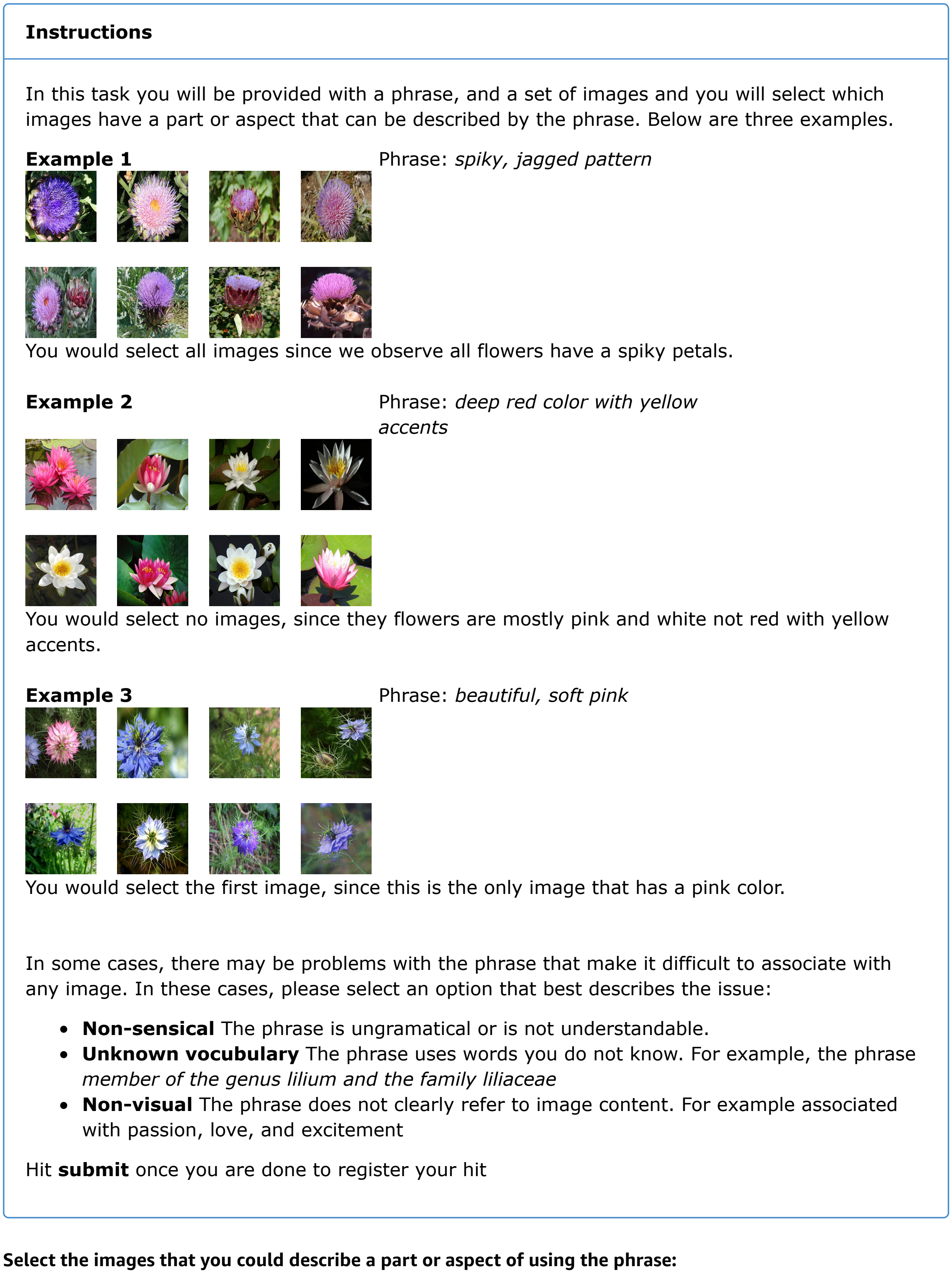
images, 10 of which correspond to the ground truth target corresponding to the concept, and 2 control images randomly sampled from other classes. The user interface was accompanied by a set of instructions presented in Figure 12.

Invalid Annotations. In reporting *Factuality* and *Groundability*, we disregard annotations that select any of the control images unless all annotators failed the control for a particular concept. In total, we disregard 18% of annotations for this reason. In reporting invalid concepts (non-visual, non-sensical, or unknown vocabulary), we consider all annotations but consider a bottleneck invalid if at least 2 out of 3 annotators agree.

Analytic Results. Table 15 displays analytic results of *Factuality* and *Groundability* for all datasets. Figure 10 presents the invalid concept distribution for all datasets separately. It is worth noting the high percentage of non-visual concepts in CIFAR-10 and CIFAR-100 compared to other datasets. We hypothesize that this reflects the annotators' inability to see the images clearly due to the low resolution (see Figure 9) rather than the lack of visual content in the concept. For example, the concepts "small and black" and "blue nose and tail" were annotated as non-visual for CIFAR-10, and the concepts "color of trees and grass" and "two large pincers on its front legs" for CIFAR-100.

E. Qualitative Examples

Figure 9 shows the additional qualitative examples for the rest 6 datasets (CIFAR-10, CIFAR-100, DTD, Aircraft, Food101, and RESISC45).



Instructions

In this task you will be provided with a phrase, and a set of images and you will select which images have a part or aspect that can be described by the phrase. Below are three examples.

Example 1 Phrase: *spiky, jagged pattern*

You would select all images since we observe all flowers have a spiky petals.

Example 2 Phrase: *deep red color with yellow accents*

You would select no images, since they flowers are mostly pink and white not red with yellow accents.

Example 3 Phrase: *beautiful, soft pink*

You would select the first image, since this is the only image that has a pink color.

In some cases, there may be problems with the phrase that make it difficult to associate with any image. In these cases, please select an option that best describes the issue:

- **Non-sensical** The phrase is ungramatical or is not understandable.
- **Unknown vocubulary** The phrase uses words you do not know. For example, the phrase *member of the genus lilium and the family liliaceae*
- **Non-visual** The phrase does not clearly refer to image content. For example associated with passion, love, and excitement

Hit **submit** once you are done to register your hit

Select the images that you could describe a part or aspect of using the phrase:

Figure 12. Instructions provided to annotators to compute *Factuality* and *Groundability*.

	n. of shots	Bottleneck Size	Discriminability (α)	Coverage (β)	Learning Rate	Batch Size
Food-101	1	5,050	$1e^7$	0.5	$1e^{-5}$	16
	2	5,050	$1e^7$	1	$1e^{-4}$	32
	4	5,050	$1e^7$	1	$1e^{-4}$	64
	8	5,050	$1e^7$	1	$1e^{-4}$	128
	16	5,050	$1e^7$	1	$1e^{-4}$	256
	Full	5,050	$1e^7$	5	$1e^{-5}$	1024
Aircraft	1	5,100	$1e^7$	0.5	$5e^{-5}$	16
	2	5,100	$1e^7$	1	$5e^{-5}$	32
	4	5,100	$1e^7$	0.1	$5e^{-5}$	64
	8	5,100	$1e^7$	0	$5e^{-5}$	128
	16	5,100	$1e^7$	1	$5e^{-5}$	256
	Full	5,100	$1e^7$	0.5	$5e^{-5}$	256
Flower-102	1	2,050	$1e^7$	10	$1e^{-5}$	16
	2	2,050	$1e^7$	100	$1e^{-5}$	32
	4	2,050	$1e^7$	10	$1e^{-5}$	64
	8	2,050	$1e^7$	10	$1e^{-5}$	128
	16	2,050	$1e^7$	1	$1e^{-5}$	256
	Full	2,050	$1e^7$	1	$1e^{-5}$	256
CUB	1	2,000	$1e^7$	0	$5e^{-5}$	32
	2	2,000	$1e^7$	0	$5e^{-5}$	64
	4	2,000	$1e^7$	0.1	$5e^{-5}$	128
	8	2,000	$1e^7$	0	$5e^{-5}$	256
	16	2,000	$1e^7$	1	$5e^{-5}$	512
	Full	2,000	$1e^7$	0.1	$5e^{-5}$	512
UCF-101	1	5,050	$1e^7$	1	$1e^{-5}$	8
	2	5,050	$1e^7$	1	$1e^{-5}$	16
	4	5,050	$1e^7$	100	$1e^{-5}$	32
	8	5,050	$1e^7$	10	$1e^{-5}$	64
	16	5,050	$1e^7$	100	$1e^{-5}$	128
	Full	5,050	$1e^7$	10	$1e^{-5}$	256
DTD	1	2,350	$1e^7$	10	$1e^{-5}$	8
	2	2,350	$1e^7$	10	$1e^{-5}$	16
	4	2,350	$1e^7$	5	$1e^{-5}$	32
	8	2,350	$1e^7$	1	$1e^{-5}$	64
	16	2,350	$1e^7$	2.5	$5e^{-5}$	256
	Full	2,350	$1e^7$	7.5	$1e^{-4}$	512
HAM10000	1	350	$1e^7$	0.1	$1e^{-3}$	4
	2	350	$1e^7$	0.1	$1e^{-3}$	4
	4	350	$1e^7$	1	$1e^{-4}$	8
	8	350	$1e^7$	10	$1e^{-3}$	8
	16	350	$1e^7$	15	$1e^{-3}$	16
	Full	350	$1e^7$	0.1	$5e^{-4}$	256
RESISC45	1	2,250	$1e^7$	5	$5e^{-5}$	8
	2	2,250	$1e^7$	5	$5e^{-5}$	16
	4	2,250	$1e^7$	10	$5e^{-5}$	32
	8	2,250	$1e^7$	15	$5e^{-5}$	64
	16	2,250	$1e^7$	15	$5e^{-5}$	128
	Full	2,250	$1e^7$	15	$5e^{-5}$	256
CIFAR-10	1	500	$1e^7$	1	$1e^{-4}$	2
	2	500	$1e^7$	5	$5e^{-4}$	4
	4	500	$1e^7$	5	$1e^{-4}$	8
	8	500	$1e^7$	1	$1e^{-4}$	16
	16	500	$1e^7$	10	$1e^{-4}$	32
	Full	500	$1e^7$	5	$1e^{-4}$	512
CIFAR-100	1	5,000	$1e^7$	7.5	$1e^{-5}$	16
	2	5,000	$1e^7$	2.5	$1e^{-5}$	32
	4	5,000	$1e^7$	7.5	$1e^{-5}$	64
	8	5,000	$1e^7$	7.5	$1e^{-5}$	128
	16	5,000	$1e^7$	5	$1e^{-5}$	256
	Full	5,000	$1e^7$	0	$1e^{-5}$	512
ImageNet	1	50,000	$1e^8$	0	$1e^{-5}$	128
	2	50,000	$1e^8$	0	$1e^{-5}$	256
	4	50,000	$1e^8$	0	$1e^{-5}$	256
	8	50,000	$1e^8$	0	$1e^{-5}$	512
	16	50,000	$1e^8$	0	$1e^{-5}$	1024
	Full	50,000	$1e^8$	0	$1e^{-5}$	2048

Table 16. All hyperparameters used for the main experiments which are tuned on the development set.